Paradox and resolution in electrostatics

James S. Ball* and Richard H. Price†
Department of Physics, University of Utah, Salt Lake City, Utah 84112
(Received 5 October 1976; accepted 14 March 1977)

A closed conducting surface possessing both conduction electrons and muons is hypothesized. A simple and persuasive proof is given, entirely in the framework of classical electrostatics and Newtonian gravity, to show that the region interior to such a surface is shielded from gravity. The resolution of the paradox is explained and a solution is given for the gravitational and electrostatic forces inside a spherical surface of the type hypothesized.

I. A PARADOX IN ELECTROSTATICS

Paradoxes in physics help to sharpen our understanding of physics. At the same time they can be amusing (at least to those who can resolve them); in general, the simpler and more basic the physics is, the more amusing the paradox. We present here a paradox that uses only very basic physics, electrostatics, and Newtonian gravity, and one that we have found to be especially amusing. In spite of its apparent simplicity it has baffled, at least for a few minutes, all those who have been presented with it.

As a starting point in describing this paradox, we use the fact that (with certain idealizing assumptions) a conducting spherical shell, or any other conducting closed container, in a gravitational field will rearrange its electron distribution so that an electron introduced into its interior will not fall.[1] The physical picture is this: The electrons in the conductor are free to move and therefore will settle slightly under the influence of gravity (see Fig. 1). The resulting excess of electrons in the bottom will result in an electrical field in the interior. This electrical field turns out to have precisely the correct value, so that for an electron introduced into the interior, gravitational and electrical forces cancel.

The proof is simple. Let Φ_e be the potential energy of an electron, then

$$\begin{aligned} \Phi_e &= -e\Phi_E + m_e\Phi_G, \\ e &= \text{magnitude of electron charge}, \\ m_e &= \text{electron mass}, \\ \Phi_E &= \text{electrostatic potential}, \\ \Phi_G &= \text{gravitational potential}. \end{aligned} \tag{1}$$

(If the gravitational field is the uniform field at the Earth's surface, then $\Phi_G = gz$, where z is height.) The force on an electron is $\nabla\Phi_e$. At the surface this force must be normal. If $\nabla\Phi_e$ had a component tangent to the surface, then the electrons would respond to this force and move. But by hypothesis the electrons in the surface are in equilibrium; thus, $\nabla\Phi_e$ is normal to the surface, and hence, Φ_e must be constant on the surface.

In the interior we have $\nabla^2\Phi_E = \nabla^2\Phi_G = 0$ since there is neither charge nor mass in the interior. We can then describe Φ_e by

$$\begin{aligned} \Phi_e &= \text{const on surface}, \\ \nabla^2\Phi_e &= 0 \text{ in interior}. \end{aligned}$$

By the uniqueness theorem[2] for Laplace's equation, then, the solution must be

$$\Phi_e = \text{const}$$

in the interior. If an electron is now introduced into the interior, it feels a force $\nabla\Phi_e = 0$. That is, the electron neither falls nor rises.

(There is a tacit assumption in this derivation: We have first of all ignored image forces.[3,4] To deal with image forces we could say that we wait for the electrons in the surface to come to equilibrium. We then glue down the electrons to fix Φ_e. Only then do we introduce a test electron into the interior.)

We are now ready to state the paradox. Suppose there existed a conductor which had not only conduction electrons but also negative muons which, like the electrons, were free to move under the influence of gravity, electrical fields, etc. Let us write down the potential energy for an electron Φ_e and for a muon Φ_μ:

$$\Phi_e = -e\Phi_E + m_e\Phi_G, \tag{2a}$$

$$\Phi_\mu = -e\Phi_E + m_\mu\Phi_G. \tag{2b}$$

Here m_μ is the mass of the muon $\approx 207m_e$. Other symbols have the same meaning as in Eq. (1).

We can repeat the argument we used before for electrons. Φ_e must be constant on the surface, or the electrons, in equilibrium by hypothesis, would move. There is no charge or mass in the interior; hence, $\nabla^2\Phi_e = 0$ and by the uniqueness theorem we have

$$\Phi_e = C_1$$

in the interior, where C_1 is some constant. We can now apply precisely the same argument to Φ_μ: If Φ_μ were not constant on the surface, the muons would feel a force, but like the electrons they are in equilibrium. Again $\nabla^2\Phi_E = 0$ and $\nabla^2\Phi_G = 0$ imply $\nabla^2\Phi_\mu = 0$ in the interior, and the uniqueness theorem forces us to conclude

$$\Phi_\mu = C_2$$

in the interior, where C_2 is some constant.

This all seems reasonable until we realize that it leads to

$$\Phi_E = (m_eC_2 - m_\mu C_1)/e(m_\mu - m_e) = \text{const},$$

$$\Phi_G = (C_2 - C_1)/(m_\mu - m_e) = \text{const},$$

and hence $\nabla\Phi_E = \nabla\Phi_G = 0$ in the interior. The interior has therefore been shielded not only against electrostatic forces, but also against gravitational forces! The interior, our calculation tells us, is a "free-fall" chamber. An astronaut—or any other object—put inside would be weightless. It should be clear that this conclusion is nonsense.

In considering resolutions of this paradox, the reader should realize that neither quantum mechanics nor the atomic nature of matter is relevant. Quantum mechanics

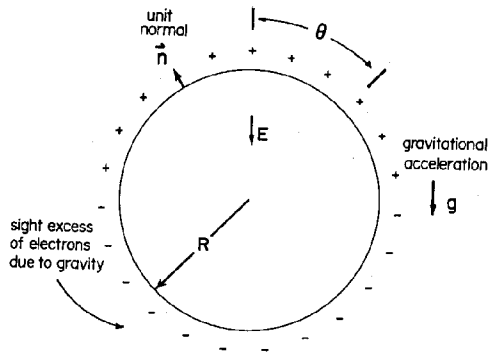

Fig. 1. Charge distribution in a spherical conducting surface in a uniform gravitational field.

must be left out of the picture because we are dealing here with a paradox in *classical* physics. If classical physics were inconsistent at the level of macroscopic electrostatics, it would be very surprising that it was not until the 1930's that classical physics was found to be inadequate. As for the ultimate particle nature of matter, any suspicions that this is the key to the problem can be dispelled by replacing the electrons and muons by two miscible fluids with different charge density to mass density ratios.

II. RESOLUTION OF THE PARADOX

To understand the resolution it is best to start by looking at a weakness in the simple "pure electron" case described by Eq. (1). For simplicity and definiteness we restrict ourselves to the case of a thin spherical shell. The solution for Φ_E is found by using Φ_e = const in Eq. (1). If the gravitational field is that on the Earth's surface, then $\Phi_G = gz$, and *inside* the sphere

$$\Phi_E = m_e gz/e = (m_e g/e)\, r\cos\theta \quad \text{(inside)} \tag{3}$$

(see Fig. 1). By expanding the axisymmetric potential outside in a series of Legendre polynomials, it is clear that we find

$$\Phi_E = (m_e gR^3/r^2 e)\cos\theta \quad \text{(outside)}. \tag{4}$$

The surface charge density can be found simply:

$$\begin{aligned}\sigma &= (1/4\pi)\,(\mathbf{E}_{\text{outside}} - \mathbf{E}_{\text{inside}})\cdot\mathbf{n}\\ &= -(1/4\pi)(\partial/\partial r)(\Phi^{\text{outside}} - \Phi^{\text{inside}})|_{r=R}\\ &= (3/4\pi)(m_e g/e)\cos\theta.\end{aligned} \tag{5}$$

Of greatest interest to us here is the total charge on the bottom half of the sphere:

$$Q_{\text{bottom}} = 2\pi\int_{-1}^{0}\sigma R^2\, d(\cos\theta) = -\frac{3}{4}\frac{m_e g}{e}R^2. \tag{6}$$

The number of *excess* electrons in the bottom half is then

$$|Q_{\text{bottom}}|/e = (3/4)(m_e g/e^2)R^2. \tag{7}$$

We now ask: What if this number is greater than the number of conduction electrons available in the top half?[5] In this case our solution must be incorrect. What happens is that *all* the conduction electrons fall to the lower half, but that is not enough to produce the electrostatic potential of Eqs. (3) and (4) and, hence, to produce Φ_e = const. What

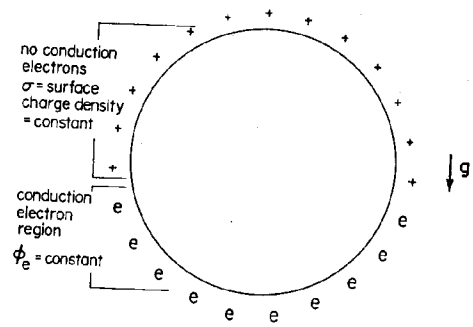

Fig. 2. Pure electron conductor with "not enough" electrons.

then is wrong with the derivation of Φ_e = const given following Eq. (1)? The flaw is that we have taken Φ_e = const on the surface to avoid forces, tangential to the surface, on the conduction electrons. But *if a region of the surface is devoid of electrons, the argument that Φ_e = const on the surface has no basis.* That region of surface need not be, and in general will not be, a Φ_e = const region. For such regions the boundary condition is not Φ_e = const, but rather σ = const (see Fig. 2).

Precisely the same physical principle resolves the muon plus electron paradox. In the surface a muon and an electron clearly cannot be in equilibrium at the same point. Rather, the conduction muons must all fall to positions lower than the conduction electrons. A region of positive surface charge will always develop in a strip of the surface between the muon region and the electron region. (For a spherical surface this is pictured in Fig. 3.) The location of boundaries is determined by the number of conduction electrons and muons. In the case that there are "not enough" conduction electrons and muons, another complication arises. An additional region with no conduction particles will develop at the top of the surface. (For a spherical surface see Fig. 4.)

The derivation that gives rise to the paradox is now seen to be fallacious. Since there are regions devoid of electrons, we cannot claim Φ_e = const on the surface.

This is the conceptual resolution of the paradox. Section III discusses, for the still interested reader, the mathematical problem that must be dealt with to find what the correct potentials are for a surface with two types of conduction particles. Although that problem must be solved numerically in general, we give an approximate solution for the most "realistic" case, that for which the density of conduction particles is large.

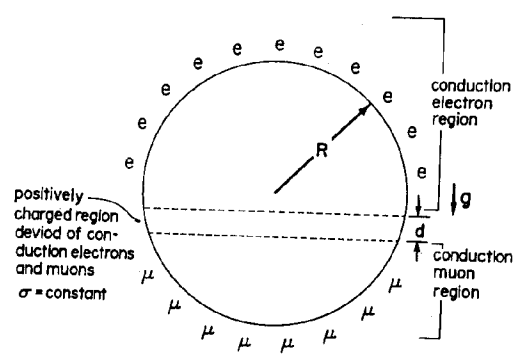

Fig. 3. Spherical surface with conduction electrons and muons.

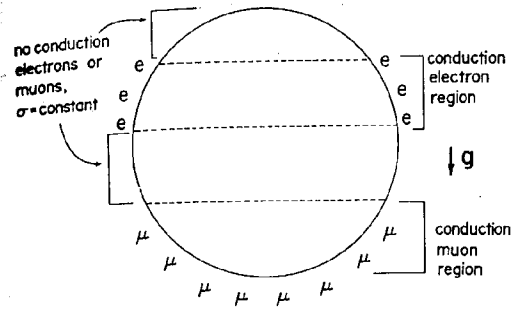

Fig. 4. Spherical surface with "not enough" conduction electrons or muons.

III. NATURE OF THE CORRECT SOLUTION

The simple problem of a conductor with an arbitrarily large number of conduction electrons (and no muons) is one in which we can specify the potentials on a closed surface. The Laplace equation with such boundary conditions (the Dirichlet-type problem) is mathematically the easiest type of potential problem to deal with.

Now let us consider the muon plus electron problem or the "not-enough-electron" problem (for the sphere see Figs. 2–4). In these cases the potential is specified only in the region where conduction particles are located. The other regions are empty of electrons or muons, so the surface charge density is

$$\sigma = \sigma_0 = \text{const}, \tag{8}$$

where σ_0 is the surface charge density of the lattice after conduction particles have been removed. In these regions then, rather than know a potential on the surface, we know the discontinuity in the normal derivative of Φ_E according to

$$\mathbf{n} \cdot (\mathbf{E}_{\text{outside}} - \mathbf{E}_{\text{inside}}) = 4\pi\sigma. \tag{9}$$

Even in the simplest case of spherical geometry it is very difficult to handle the mathematics of such a problem with mixed boundary conditions. (For the "not-enough-electron" situation a systematic analytic approach does exist; it is outlined in the Appendix.)

We will confine ourselves here to the approximate solution of the spherical electron plus muon problem of Fig. 3 with the "realistic" assumption that the width d of the constant density strip is very narrow: $d \ll R$. This is "realistic" because electrical forces on elementary particles are enormously stronger than gravitational forces. It corresponds, furthermore, to the usual tacit assumption about "conductors," that the density of free charge carriers is unbounded. A semiquantitative justification for the $d \ll R$ assumption is given in the Appendix. With this assumption we can idealize the constant density buffer zone as a line. The electrostatic potential is then found by requiring that (i) above the line $\Phi_e = -e\Phi_E + m_e gR\cos\theta$ is constant, (ii) below the line $\Phi_\mu = -e\Phi_E + m_\mu gR\cos\theta$ is constant, and (iii) Φ_E is continuous across the line. Specifically, if the buffer zone is at $\theta = \alpha$, the potential on the surface is

$$\Phi_E = (m_e g/e)R(\cos\theta - \cos\alpha), \quad \theta < \alpha,$$
$$\Phi_E = (m_\mu g/e)R(\cos\theta - \cos\alpha), \quad \theta > \alpha. \tag{10}$$

The potential in the interior can then be found by a number of techniques. The potential on the symmetry axis ($|\cos\theta| = 1$) is particularly easy to find as an integral using the spherical Green's function (details in the Appendix). For the case $\alpha = \pi/2$ the potential on the axis is

$$\Phi_E = (z/2e)g(m_e + m_\mu) + (g/2ez^2)(m_\mu - m_e) \times [(R^2 - z^2)(R^2 + z^2)^{1/2} - R^3], \tag{11}$$

where z is the height above the center of the sphere.

Off the symmetry axis the potential is most easily computed by an expansion in Legendre polynomials (details in the Appendix). For the simplest case, $\alpha = \pi/2$, the results for the total potentials Φ_e and Φ_μ are plotted in Fig. 5. (For clarity this figure depicts the potentials for a mass ratio $m_\mu/m_e = 4$. When the actual physical ratio $m_\mu/m_e \approx 207$ is used, the effect of the electron distribution is obscured.) It should be noted that Φ_e is constant on the top hemisphere but *not* on the bottom, and the opposite is true for Φ_μ.

APPENDIX

In this Appendix we supply some of the mathematical details omitted from the body of the paper.

First, we outline an analytic approach to a mixed boundary value problem, the "not-enough-electron" situation of Fig. 2. This can be transformed by inversion[6] about a point at the north pole of the sphere to a mixed boundary value problem on a plane. In this plane a central disk has a known (nonconstant!) potential Φ_{disk}, and the rest of the plane has a known surface charge density. Such a problem can be solved by removing the central disk and finding the potential everywhere, as a Poisson integral, due to the known surface charge distribution on the remainder of the plane. In particular, we will find some potential Φ' in the region of the disk. We then add a disk with a potential $\Phi'' = \Phi_{\text{disk}} - \Phi'$ so that the superposed potentials have the correct value. The potential due to a disk with a known potential Φ'' can be solved by expansion in oblate spheroidal coordinates. The planar problem is now solved, and we can get the solution to the original problem by reversing the inversion about the north pole.

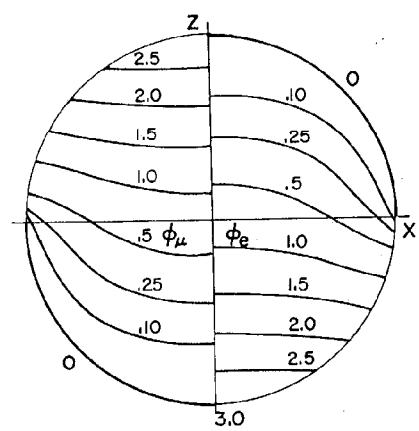

Fig. 5. Equipotential surfaces of Φ_e, the potential energy of an electron, and of Φ_μ, the potential energy of a muon, for the case of a thin "buffer zone" at $\theta = \pi/2$. Equipotentials (wide lines) of Φ_μ are plotted on the left half of the sphere; equipotentials of Φ_e, on the right. For purposes of illustration the mass ratio m_μ/m_e is taken to be 4. The Φ are given in units of gm_eR. Note that, for $z \to -z$, $\Phi_\mu \to \Phi_e$.

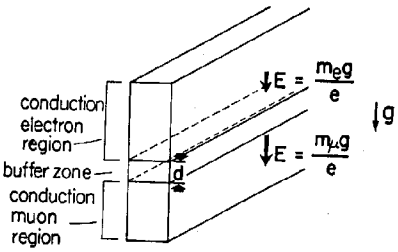

Fig. 6. A piece of the conducting shell in the region of a narrow "buffer zone" of width d.

We next justify the approximation $d \ll R$ used in Sec. III. Figure 6 gives a simplified picture of the buffer zone. From the discontinuity in the E field across this zone we know that the charge density must be

$$\rho = (m_\mu - m_e)g/4\pi e d. \tag{A1}$$

This charge, of course, resides in the excess protons, i.e., those protons not neutralized by bound electrons or muons. The density of these excess protons is

$$n = \rho/e = (1/4\pi)(m_\mu - m_e)g/e^2 d \approx 10^{-4}/d, \tag{A2}$$

where d is given in cm and n in cm^{-3}. Thus, even if $n \approx 1$ cm^{-3}, d is extremely small. For normal electron densities in metallic conductors, d becomes much smaller than atomic dimensions.

Another viewpoint on this question could also be adopted: The usual concept of a conductor tacitly includes the idea of an infinite density sea of charge carriers (and, hence, excess protons). For such a mathematical conductor the buffer zone must have zero width.

With the idealization that the width of the buffer zone is negligible, the problem is to find the potential inside a sphere with boundary values for Φ given by Eq. (10). One form of the solution is given by an integral using the Green's function[7]:

$$\Phi_E(r,\theta) = (1/4\pi) \int \Phi_E(R,\theta',\phi')[R(R^2 - r^2)/(R^2 + r^2 - 2rR\cos\gamma)^{2/3}]\, d(\cos\theta')\, d\phi', \tag{A3}$$

where

$$\cos\gamma \equiv \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi').$$

This integral is particularly useful on the symmetry axis, where $\theta = 0,\pi$ and $r\cos\gamma = z\cos\theta'$, so that

$$\Phi_E(z) = \frac{2\pi}{4\pi e} R^2(R^2 - z^2)g \times \left[m_e \int_{\cos\alpha}^{1} \frac{(\cos\theta' - \cos\alpha)\, d(\cos\theta')}{(R^2 + z^2 - 2Rz\cos\theta')^{3/2}} + m_\mu \int_{-1}^{\cos\alpha} \frac{(\cos\theta' - \cos\alpha)\, d(\cos\theta')}{(R^2 + z^2 - 2Rz\cos\theta')^{3/2}} \right]. \tag{A4}$$

The integrals are straightforward to evaluate and, in the case $\alpha = \pi/2$, give Eq. (11).

For the potential off the symmetry axis the Green's function integrals become unwieldy, and the potential is more easily calculated by expanding its angular dependence in Legendre polynomials:

$$\Phi_E = \sum_{l=0}^{\infty} A_l (r/R)^l P_l(\cos\theta). \tag{A5}$$

For the simplest case, $\alpha = \pi/2$, the coefficients are easily evaluated:

$$A_1 = (gR/2e)(m_e + m_\mu),$$
$$A_l = 0 \quad \text{if } l \text{ odd and } l \neq 1,$$
$$A_l = \frac{gR(m_\mu - m_e)}{e} \frac{(2l+1)(-1)^{l/2}\Gamma(l/2 - \frac{1}{2})}{8\pi^{1/2}\Gamma(2 + l/2)}, \; l \text{ even}. \tag{A6}$$

*Supported in part by National Science Foundation Grant PHY76-14907.

†Supported in part by National Science Foundation Grant PHY74-1611-A03.

[1]See, e.g., A. P. Lightman, W. H. Press, R. H. Price, and S. A. Teukolsky, *Problem Book in Relativity and Gravitation* (Princeton U. P., Princeton, 1975), problem 12.6.

[2]J. D. Jackson, *Classical Electrodynamics*, 2nd ed. (Wiley, New York, 1962), Sec. 1.9.

[3]Reference 2, Sec. 2.2.

[4]There is also another, more subtle assumption. In a real container the lattice of the container's crystal structure will be compressed at the bottom, or stretched at the top, if the container is suspended in a gravitational field. This raises the energy levels for electrons at the bottom relative to those at the top. The equilibrium distribution, then, will have fewer electrons at the bottom than are necessary to cancel gravitational forces. To rid the problem of this complication, we assume our container to be made of a material with an infinitely rigid lattice.

[5]In fact, this is always miniscule for real conductors, so the problem never arises except as a point of principle.

[6]W. K. H. Panofsky and M. Phillips, *Classical Electricity and Magnetism* (Addison-Wesley, Reading, MA 1955), Sec. 3.4.

[7]Reference 2, Sec. 2.6.

Editor's note

This review article, the second in a series, is made possible by a grant from the National Science Foundation.

General relativity primer

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 14 July 1980; accepted for publication 2 December 1981)

In this tutorial article the physical ideas underlying general relativity theory are discussed and the basic mathematical techniques (tensor calculus, Riemann curvature) needed to describe them are developed. The general relativity field equations are presented and are used in several applications including a discussion of black holes.

I. INTRODUCTION

A. Purpose and outline

Special relativity theory (SRT) is a part of the intellectual toolbox of all physicists and a feature of the physicist's education even at the undergraduate level. The novel concepts of SRT, so shocking in 1905, hold no special terror now. The same, regrettably, cannot be said for the general relativity theory (GRT), Einstein's relativistic theory of gravity. The imagery of space-time curvature, and such exotica as black holes, give GRT such a recondite aura that it is too often regarded as hopelessly mystical, even by students and teachers who accept quantum mechanics as a perfectly reasonable description of the world. It is my goal in this article to show that this viewpoint on GRT is unjustified, that relativistic gravity is intuitively accessible and that space-time curvature is a natural conceptual basis for it. More specifically this article presents the mathematical and physical structure of GRT for a student or a teacher of physics, or a physicist in another field, in such a way that these readers can understand how calculations are done in GRT and what they mean. This article then is intended to present in a fairly small number of pages a subject usually dealt with in full length textbooks.[1] This is not, however, a "popular" introduction limited to metaphors, analogies, and word pictures. Such introductions (and in fact Sec. II below could stand alone as one) are of value but they do not teach the theory. They give some answers but not the general scheme for finding answers. This article presents GRT as a physically motivated *mathematical* theory of gravity. The distinction between such a presentation and a popular one is particularly sharp for GRT since the necessary mathematics of tensors is not part of the background of most physicists. To avoid tensor calculus would be to avoid a meaningful expression of the ideas of GRT. To include the mathematics, unfortunately, engenders a great danger, that of the mathematical trees obscuring the physical forest. It is much too easy to forget that all the formulas and transformations, and all the mathematical symbols dripping with subscripts, are part of a description of the physical world. The necessary mathematics of tensors and of curvature are introduced in this article in the most painless and the most physically motivated way I could manage, but I still feel it necessary to urge the reader in the strongest possible terms never to lose sight of the simple underlying physical and geometrical principles.

Clearly in a small article covering a large subject, sacrifices must be made. The most regrettable sacrifice will be the omission of all but a cursory discussion of the stress-energy tensor, the "source" of the gravitational field. Also omitted will be many mathematical details, some of them formal and elegant, some of them tricky and technical, some of them useful for reducing very difficult calculations to merely difficult ones. Missing too will be most of the applications of GRT to problems of current interest. A useful discussion is given, however, of that aspect of GRT that stimulates the most interest and confusion: black holes.

I assume that the reader comes to this article with two prerequisites: First, a familiarity is required with partial differential equations and their application in physics, as would certainly result from, say, a junior- or senior-level course in electrodynamics. Experience with partial differential equations will be necessary for an appreciation of the meaning of the GRT field equations; specific techniques for solving such equations will not be of importance. The second requirement is a comfortable familiarity with SRT, with the Minkowski space-time description of SRT, and with the usual SRT jargon (worldlines, proper time, four-vectors, etc.). The conventions used in this article are outlined in Sec. I B.

This article is organized as follows: Section II presents a heuristic overview of gravity and space-time geometry. The physical ideas that are the basis of the theory are presented without their full mathematical realizations. This section serves as the motivation for the mathematical development in the subsequent two sections. It can also stand alone as a nonmathematical description of the structure of GRT. For some readers the phenomenological picture given in Sec. II may be enough. Others I hope will be inspired to read further in order to understand more quantitatively the ideas of Sec. II. The mathematical tools necessary for a quantitative understanding are developed in Secs. III and IV. In Sec. V the field equations of GRT are presented and discussed, bringing to fruition the seeds planted in Sec. II. Section VI contains a brief discussion of the weak field limit of GRT in which the connection is made with Newtonian gravity and some features of gravitational waves are discussed. The use of the field equations and the interpretation of the theory is exhibited in Sec. VI for the simplest but astrophysically most important case of a spherically symmetric source. In this section the Schwarzschild space-time geometry is derived from the basic starting point of symmetry prin-

The lightning-rod fallacy

Richard H. Price and Ronald J. Crowley[a)]
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 24 April 1984; accepted for publication 24 August 1984)

It is accepted generally that the electric field strength on the surface of an isolated charged conductor is greatest where the surface curvature is greatest. We show here that there is, in fact, no relationship between these two maxima and that, in general, they are located at different points on the surface. Two classes of analytic examples are offered: one using conformal mapping techniques and the other involving small perturbations of a conducting spherical surface.

It is amusing how pervasive a misconception can be, even in as cut and dried a subject as electrostatics. In this paper we confront the "common knowledge" that the electric field at the surface of an isolated conductor is greatest where the curvature is greatest. It is in fact *true* that when the curvature is singular the **E** field is also singular. The coronal discharge near sharp points is exploited in lightning rods and familiar in electrostatic demonstrations. Proofs that $|\mathbf{E}|$ becomes infinite at sharp outer edges and conical apices can be found in standard textbooks.[1]

"Proofs" of the general relationship of $|\mathbf{E}|$ and curvature are rather more vague when curvature is finite. We have surveyed widely used sophomore-level physics texts and have found that the typical approach to the relationship is based on a simple example: Two spheres of different radii are connected by a long fine wire. It is demonstrated that the E-field strength on each of the spheres is inversely proportional to the radius of that sphere. The generalization that

$$|\mathbf{E}| \propto 1/R, \tag{1}$$

where R is the "radius of curvature" of the surface, is then either implied or stated explicitly. In one (older) text this method is pushed further; a sphere of radius R is matched to a point on the conducting surface and it is "proved" that in terms of the potential V (relative to infinity) of the conductor and the "radius of curvature" R at the point

$$|\mathbf{E}| = V/R. \tag{2}$$

A more satisfactory, although heuristic, approach also appears in the texts. Equipotentials very near a nonspherical conducting surface are sketched; it is noted that the equipotentials tend to become spherical at large distances. As a consequence the equipotentials are most closely spaced, and hence the **E** strength is greatest, where the curvature is greatest (see Fig. 1).

In addition to surveying the texts we have harassed colleagues and students at several universities with the question as to where the **E** strength is greatest. The response showed that the lesson in the texts has been well learned.

All this is more than sufficient motivation for us to demonstrate in this article that *on the surface of a conductor there is no general relationship between the location of the maximum of curvature and the maximum of* $|\mathbf{E}|$. Our discussions with our colleagues has forewarned us that there is a strong tendency to try to save some relationship by modifying and specializing it. We will therefore show more specifically that $|\mathbf{E}|$ is not even a *local* maximum where curvature is a *local* maximum. (We will, in fact, give an example in which $|\mathbf{E}|$ is a maximum where curvature is a local minimum.) We will also show that no relationship between the locations of the maxima can be salvaged by requiring that the conductor be convex.

The nonexistence of a relationship between maximum $|\mathbf{E}|$ and the maximum curvature is rooted in the fact that they depend in entirely different ways on the shape of the surface. Curvature depends only locally on the shape of the surface. At any point of the surface the curvature is determined by the first two derivatives, at that point, of the function specifying the surface. The curvature at that point is independent of what the surface does at points a finite distance away. The **E**-field strength, on the other hand, is determined by a solution of Laplace's equation for the electrostatic potential Φ. The solution depends on the specification of boundary conditions, in this instance the value of the potential *everywhere* on a closed surface. Changing the shape of the surface in one region influences the value of the potential at distant locations.

We start our set of examples with a heuristic one, which emphasizes this local dependence of curvature and nonlocal dependence of $|\mathbf{E}|$ on surface shape. The **E** field inside a closed conducting container is zero; the **E** field inside an almost-closed conducting container is almost zero. The almost-closed container shown in Fig. 2 therefore has very weak fields in the hollow that would be enclosed by the container were it not for the narrow gap at the top. At the bottom of this hollow is a small hemispherical pimple. By making the radius of the pimple sufficiently small we can guarantee that the maximum surface curvature is on the

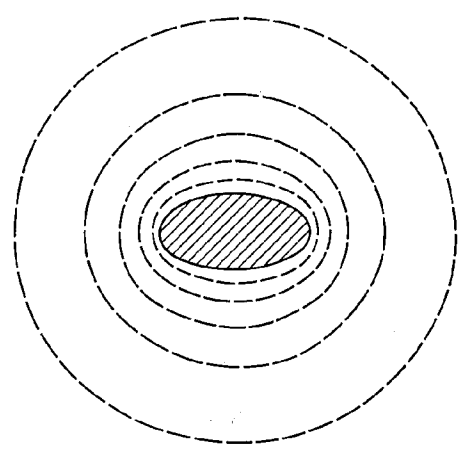

Fig. 1. Proof that $|\mathbf{E}|$ is always greatest where the curvature is greatest. The equipotentials (dashed lines) are most closely spaced, and hence the **E** strength is greatest, where the curvature is greatest.

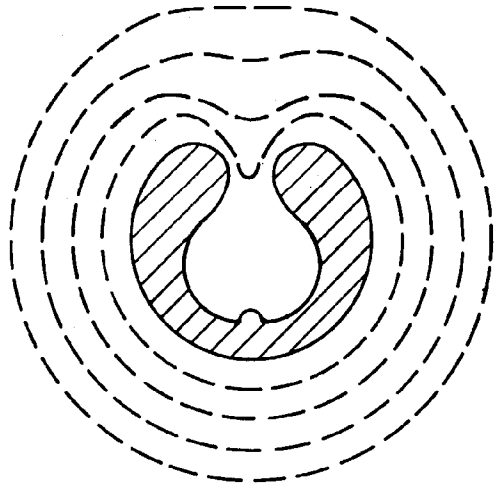

Fig. 2. Proof that $|\mathbf{E}|$ is not always greatest where the curvature is greatest. (Figure 2, like Fig. 1, represents a cross section of a solid conductor formed by rotating the figure about the vertical symmetry axis.) The curvature is greatest at the bottom of the hollow, on the small hemispherical pimple, but the $\mathbf{E}$ field there can be made arbitrarily small by narrowing the gap at the top.

pimple. But the $\mathbf{E}$ field will not or need not be a maximum on the pimple. No matter how sharp the pimple is (i.e., no matter how small its radius) we can make the $|\mathbf{E}|$-field strength on the pimple as weak as we like by making the almost-closed container more nearly closed (i.e., by narrowing the gap at the top).

If we are to exhibit quantitative examples we can no longer avoid a quantitative way of dealing with curvature. In many texts this is done with the "radius of curvature" at a point. This is presumably the radius of the sphere which approximates locally the surface at that point. Such an approximation-by-sphere is, in fact, completely explicit in one of the texts. Despite the texts, *in general a surface near a point cannot be approximated as a sphere and there is no unambiguous meaning to "radius of curvature."*

The correct quantitative description of curvature at a point P of the surface starts with a curve C, in the surface, through that point, as shown in Fig. 3. The curve is parametrized by its arc length s (from any starting point) and at any point we can compute the curve's unit tangent $\mathbf{t} = d\mathbf{r}/ds$. The outer unit normal $\mathbf{n}$ to the surface is defined along the curve, so its derivative $d\mathbf{n}/ds$ can be computed. It is fairly evident that $d\mathbf{n}/ds$ at P does not depend on any of the details of curve C, except its direction at P; any other curve with the same tanget $\mathbf{t}$ at P would lead to the same $d\mathbf{n}/ds$. At point P the scalar $\mathbf{t} \cdot d\mathbf{n}/ds$ then says something about the "bending" of the surface in direction $\mathbf{t}$.

The next truth is not self-evident, but is proved in any book on differential geometry. There are two orthogonal directions in the surface $\mathbf{t}_1$, $\mathbf{t}_2$, for which $\mathbf{t} \cdot d\mathbf{n}/ds$ is a maximum and minimum. These directions are called the principal directions; the associated values of $\mathbf{t} \cdot d\mathbf{n}/ds$ are called the principal curvatures κ_1 and κ_2:

$$\kappa_i = \mathbf{t} \cdot \frac{d\mathbf{n}}{ds}\bigg|_{\text{for } \mathbf{t}_i}, \quad i = 1,\ 2. \tag{3}$$

The reciprocals of the principal curvatures are called the principal radii of curvature[2] R_1, R_2:

$$R_i \equiv 1/\kappa_i. \tag{4}$$

For a nonprincipal direction $\mathbf{t}$, the value of $\mathbf{t} \cdot d\mathbf{n}/ds$ is calculated easily from the angle between $\mathbf{t}$ and the principal directions, and from the values of the principal curvatures or radii. The complete characterization of the bending of a surface at a point then requires the specification of the principal directions and the associated principal curvatures, or principal radii. If the orientation of the bending is not of interest then only the principal curvatures, or radii, need be specified. The orientation-independent information about curvature is often not given as κ_1, κ_2 or R_1, R_2 but is packaged into two types of "average curvature": "Gaussian curvature,"

$$\kappa_G = \kappa_1 \kappa_2 \tag{5}$$

and "mean curvature,"

$$\kappa_M = \tfrac{1}{2}(\kappa_1 + \kappa_2). \tag{6}$$

It is of more than a little interest that the mean curvature is related to a property of the $|\mathbf{E}|$-field strength near a conducting surface. The fractional rate at which $|\mathbf{E}|$ decreases with distance away from the surface of a conductor is given by the mean curvature of the conductor[3]:

$$(1/|\mathbf{E}|)\mathbf{n} \cdot \nabla |\mathbf{E}| = -2\kappa_M.$$

The crucial point of the above discussion is that two numbers characterize the size of curvature, so that the assertion that "$|\mathbf{E}|$ is maximum where curvature is maximum" is not only wrong, it is also ill defined. It is always harder to disprove an ill-stated claim and we will need to show, with our examples, that $|\mathbf{E}|$ is neither an absolute, nor a local, maximum where κ_G *or* κ_M is maximum.

Our first class of detailed examples involves a long conductor of uniform cross section. If we idealize the conductor to be infinitely long, say in the z direction, then the electrostatic potential Φ is independent of z and Laplace's equation $\nabla^2\Phi = 0$ becomes a 2-dimensional problem, in the xy plane. We have exploited complex-variable techniques and conformal transformations to find closed-form solutions of Laplace's equation for a 3-parameter family of conductor cross sections, which more or less resemble Fig. 2 without the pimple. The details of this conformal transformation are tedious enough to be distracting here and are relegated to Appendix A. Here we concentrate on results, presented graphically, of examples for particular geometries.

The ambiguity of "radius of curvature" does not exist in this case. By symmetry the z direction is a principal direction. The corresponding principal radius of curvature is formally infinite. All the information about curvature is therefore contained in one number, the principal curvature κ in the transverse principal direction. (We could of course just as well use the principal radius of curvature, or the mean curvature.) The transverse principal curvature is determined from the curve giving the conductor cross section; it is simply the reciprocal of the radius of the osculat-

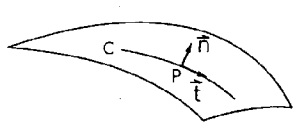

Fig. 3. Curve, unit tangent, and unit surface normal used in quantifying curvature.

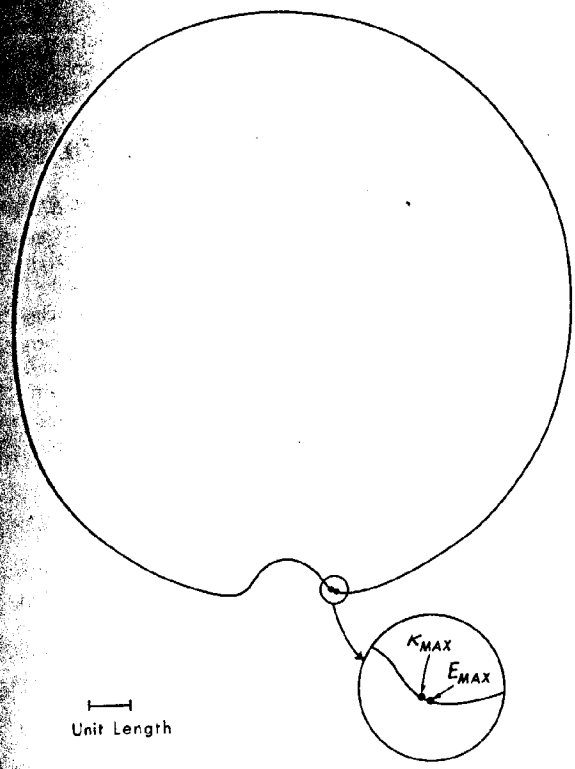

Fig. 4. Cross section of a long conductor. The enlarged detail shows the displacement of the maximum of the surface **E** strength from the maximum curvature.

ing circle at any point of the curve and its calculation is straightforward. (See Appendix A for details.)

Figure 4 shows a conductor cross section for a particular choice of parameters. The greatest curvature occurs near the opening of the gap, at the point labeled $\kappa_{\max}$. The arguments earlier detailed in this paper suggest that the magnitude of **E** is small in the "hollow" and the maximum of $|\mathbf{E}|$ should tend to shy away from the gap. The results presented in Fig. 4 show that this does in fact happen. The maximum of $|\mathbf{E}|$ is further from the gap than the maximum of κ.

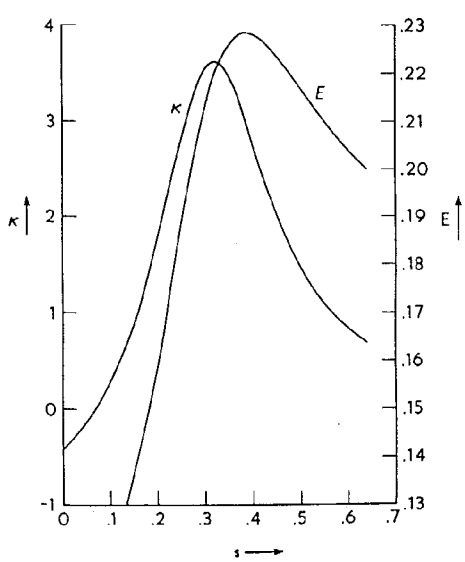

Fig. 5. The variations of curvature κ and **E**-field strength as functions of arc length s, along the cross section illustrated in Fig. 4.

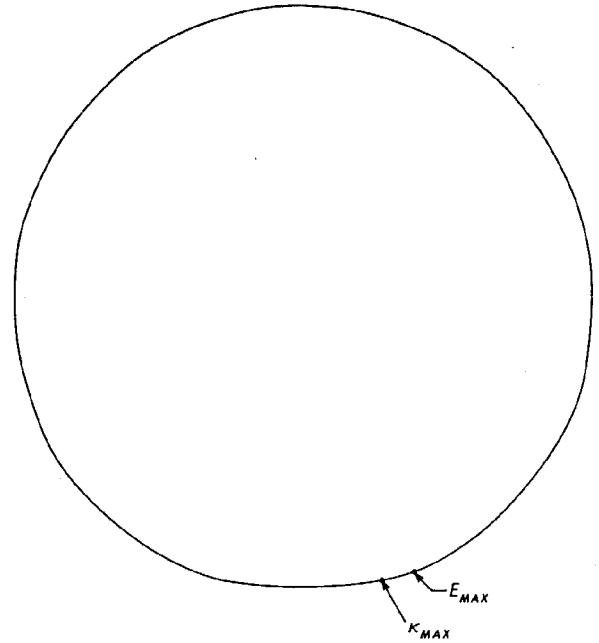

Fig. 6. Convex cross section showing the displacement of the maxima of κ and $|\mathbf{E}|$.

The results are presented in a different way in Fig. 5; the values of $|\mathbf{E}|$ and κ are plotted as functions of s, the arc length along the cross section. The curvature has the physical dimension of inverse length so that numerical values for κ, as well as arc length, require a length scale for the conductor cross section. Values in Fig. 5 correspond to the "unit length" depicted in Fig. 4. The numerical values of $|\mathbf{E}|$ are in arbitrary units. (The $|\mathbf{E}|$ values *could* be related to the length scale and the charge per unit length on the conductor.)

In our discussions of this topic with others we have been accused of cheating by using a conductor which has a concave region. Rather than debate the fairness and relevance of a concave region we present a cross section which is everywhere convex. The cross section in Fig. 6 results from another choice of parameters in our conformal transformation. The convexity of this cross section is evident in Fig. 6 and is supported by the fact that the computed κ is everywhere positive. This example, like that of Figs. 4 and 5, show a displacement of the maxima of κ and $|\mathbf{E}|$.

Our second class of examples is axially symmetric first-order perturbations of a spherical surface of radius R. We take ϵ to be a very small dimensionless number and consider a surface defined by radius r as a function of polar angle θ, according to

$$r = R\left(1 + \epsilon \sum_{n=2}^{\infty} \beta_n P_n(\mu)\right), \quad \mu \equiv \cos\theta, \tag{7}$$

where P_n is the nth Legendre polynomial and the set of coefficients β_n determines the distortion of the sphere. The $n = 0$ and $n = 1$ terms in the series expansion for $r(\theta)$ correspond, respectively, to a change in the size and the origin of

the perturbed sphere, and are of no interest here. It should be noted that for small ϵ all surfaces given by Eq. (7) are everywhere convex. It is very simple to solve, to first order in ϵ, for the surface **E** strength in terms of Q, the total charge on the perturbed sphere. The result (detailed in Appendix B) is

$$|\mathbf{E}| = QR^{-2}\left(1 + \epsilon \sum_{n=2}^{\infty} \beta_n (n-1) P_n(\mu)\right). \tag{8}$$

By symmetry it is clear that the principal directions on the sphere are the ϕ direction (tangents to circles of constant r,θ) which we shall call $\mathbf{t}_1$, and the meridional direction (the "θ direction") $\mathbf{t}_2$. The corresponding principal curvatures to first order in ϵ are shown in Appendix B to be

$$\kappa_1 = R^{-1}\left[1 - \epsilon \sum_{n=2}^{\infty} \beta_n \left(P_n - \mu \frac{d}{d\mu} P_n\right)\right], \tag{9a}$$

$$\kappa_2 = R^{-1}\left[1 + \epsilon \sum_{n=2}^{\infty} \beta_n \times \left([n(n+1) - 1]P_n - \mu \frac{d}{d\mu} P_n\right)\right]. \tag{9b}$$

Since we are working only to first order in ϵ the arithmetic mean of the principal curvatures (κ_M) and the geometric mean (corresponding to the square root of the Gaussian curvature) are the same:

$$\kappa = R^{-1}\left(1 + \frac{1}{2}\epsilon \sum_{n=2}^{\infty} (n-1)(n+2)\beta_n P_n(\mu)\right). \tag{10}$$

Local extrema of κ and $|\mathbf{E}|$ will occur at the poles ($\theta = 0$, π or $\mu = \pm 1$) and at the latitudes at which the derivatives, with respect to μ, of Eqs. (8) and (10) vanish. If $\beta_n = 0$ for all but a single value of n, the local extrema of κ and $|\mathbf{E}|$ will coincide; in general, however, Eqs. (8) and (10) define different functions of μ and, contrary to conventional wi[illegible] only a fortuitous choice of the β_n will lead to coin[illegible] local maxima of $|\mathbf{E}|$ and κ away from the poles.[4] If th[illegible] are chosen such that the maximum of either $|\mathbf{E}|$ or κ [illegible] not occur at a pole then the absolute maximum of $|\mathbf{E}|$ [illegible] will not coincide.

As an interesting specific example we choose β_3 [illegible] $\beta_5 = -1$, and $\beta_n = 0$ for other values of n. For this ch[illegible] depicted in Fig. 7, the sphere is perturbed with a fairly [illegible] topped convex hump around the $\theta = 0$ axis. Because [illegible] hump is flat topped, $\theta = 0$ is a local minimum of the cu[illegible] ture; curvature increases with θ out to $\theta \approx 14.53°$, wh[illegible] is a local maximum; the absolute maximum of curvatu[illegible] at $\theta \approx 111.81°$. The **E**-field strength is greatest at $\theta = 0$ [illegible] that in this example not only is the absolute maximum [illegible] far removed from the absolute maximum of κ, but it co[illegible] cides with a local *minimum* of curvature!

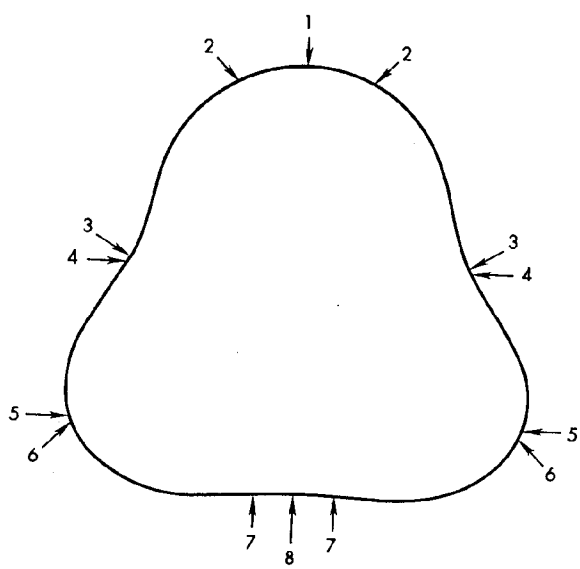

Fig. 7. A perturbed sphere with $\beta_3 = 6$, $\beta_5 = -1$. Figure 7 is drawn with a large value ($\epsilon = 0.05$) of the perturbation parameter to show clearly the nature of the distortion of the sphere. The numbered points correspond to the following extrema: (1) local minimum of κ, absolute maximum of **E**; (2) local maximum of κ; (3) local minimum of **E**; (4) absolute minimum of κ; (5) absolute maximum of κ; (6) local maximum of **E**; (7) local minimum of κ; and (8) local maximum of κ, absolute minimum of **E**.

ACKNOWLEDGMENTS

We would like to thank Dr. Karel Kuchar for an int[illegible] esting argument, which he won, that led to the present ar[illegible] cle. We also thank Dr. Edward Kreusser for amusing su[illegible] gestions in connection with Fig. 2. One of us (RH[illegible] acknowledges gratefully the National Science Foundati[illegible] for support under grant no. PHY81-06909.

APPENDIX A

Our first class of examples is based on a conformal tran[illegible] formation of a crescent in the complex $z = x + iy$ plan[illegible] The crescent is formed from two circular arcs, as pictur[illegible] in Fig. A1, with both going through the points $y = 0$ [illegible] $x = \pm L$. (Here L serves as the unit length, as shown i[illegible] Fig. 4.) The two parameters defining the crescent are mo[illegible] simply taken as l_1 and l_2, the y values of the centers of th[illegible] smaller and larger circles. In terms of these we can write γ, [illegible] the interior angle at the sharp corners of the crescent, as

$$\gamma = \tan^{-1}[L(l_2 - l_1)/(L^2 + l_1 l_2)]. \tag{A1}$$

It is useful also to define the exterior angle at the corner[illegible]

$$\alpha \equiv 2\pi - \gamma \tag{A2}$$

and an auxiliary angle

$$\phi_0 \equiv (\pi/\alpha)\tan^{-1}(L/l_2). \tag{A3}$$

In the formulas for γ and ϕ_0 the $\tan^{-1}$ function is to be taken in the range $(0, \pi/2)$.

For the conformal transformation we will need the com[illegible] plex expression

$$\xi \equiv (z + L)/(z - L). \tag{A4}$$

We define its phase to be zero on the real z axis for $x > L$ and [illegible] to be continuous everywhere outside the crescent. We next [illegible] define a conformal transformation from the complex z [illegible] plane to the complex $W = U + iV$ plane by

$$W = (\xi^{\pi/\alpha} - e^{-2i\phi_0})/(\xi^{\pi/\alpha} - 1). \tag{A5}$$

This transformation maps the exterior of the crescent in the z plane to the exterior of the unit circle in the W plane. Furthermore, it maps ∞ in the z plane to ∞ in the W plane. Finally we define the complex potential

$$F = \Phi + i\psi = \ln W. \tag{A6}$$

Since F, outside the crescent, is an analytic function of z, its real part $\Phi(x,y)$ is harmonic (i.e., $\nabla^2\Phi = 0$). It is also seen easily that $\Phi = 0$ on the crescent (since $|W| = 1$ on the

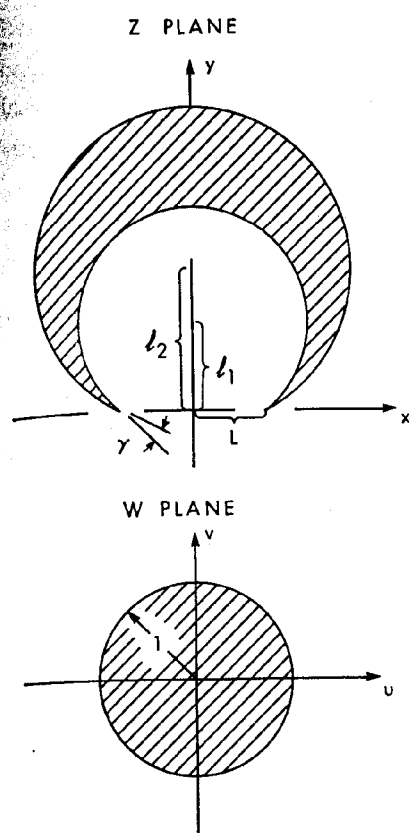

Fig. A1. Geometric parameters used in the conformal transformation, which maps the exterior of the crescent in the z plane to the exterior of the unit circle in the W plane.

mapping of the crescent) and $\Phi \propto \ln(x^2+y^2)$ at large distances from the crescent. Clearly, $\Phi(x,y)$ then represents the electrostatic potential outside a charged 2-dimensional conducting crescent. The explicit function $\Phi(x,y)$, without reference to complex variables, is given by

$$\Phi(x,y) = \frac{1}{2}\ln\left(\frac{R^2+1-2R\cos(\phi+2\phi_0)}{R^2+1-2R\cos\phi}\right), \tag{A7a}$$

where

$$R \equiv \{[(x+L)^2+y^2]/[(x-L)^2+y^2]\}^{\pi/2\alpha} \tag{A7b}$$

and

$$\phi \equiv (\pi/\alpha)\{\delta - \tan^{-1}[2yL/(x^2+y^2-L^2)]\}, \tag{A7c}$$

in which the principal branch $-\pi/2 < \tan^{-1} < \pi/2$ is to be taken and

$$\begin{aligned} \phi &\equiv 0, && \text{if } x^2+(y-l_2)^2 > l_2^2+L^2 \text{ and } x^2+y^2 > L^2,\\ \phi &\equiv \pi, && \text{if } x^2+y^2 < L^2,\\ \phi &\equiv 2\pi && \text{if } x^2+(y-l_1)^2 < l_1^2+L^2 \text{ and } x^2+y^2 > L^2. \end{aligned} \tag{A7d}$$

The crescent is, of course, an equipotential of Φ (in fact the $\Phi = 0$ equipotential) but due to its sharp corners it is not a useful choice as the conductor cross section. Any $\Phi > 0$ equipotential, however, is a smooth curve and may be chosen to represent the surface of a nonsingular conductor. The choice of the equipotential, along with the two geometric parameters l_1, l_2 defines a conducting surface we

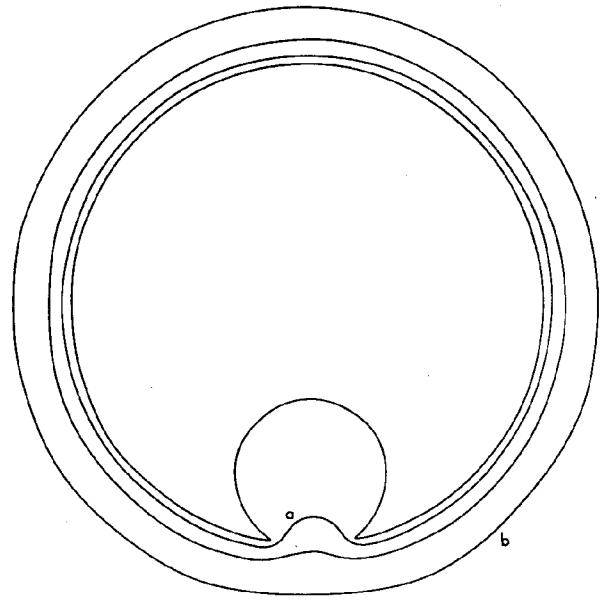

Fig. A2. Equipotentials outside a crescent, as illustrated in Fig. A1, with $l_2 = 6$, $l_1 = 2.645$, so that $\gamma = \pi/16$. Any of these equipotentials may be chosen to represent the surface of a conductor with nonsingular curvature. The equipotentials labeled a and b correspond, respectively, to $\Phi = 0.04$ and 0.22. Equipotential a is the conducting surface used in the example depicted in Figs. 4 and 5. Equipotential b is the basis for the example of Fig. 6.

may use as an example for the relation of $|\mathbf{E}|$ and κ. Typical equipotentials are shown in Fig. A2.

The value of $|\mathbf{E}|$ on any of these surface is calculated in a straightforward manner, in principle, from

$$|\mathbf{E}| = \left[\left(\frac{\partial\Phi}{\partial x}\right)^2 + \left(\frac{\partial\Phi}{\partial y}\right)^2\right]^{1/2} = \left|\frac{dF}{dz}\right|, \tag{A8}$$

where the second equality follows from the Cauchy–Riemann relations. The formula for the radius of curvature of a plane curve can be found in elementary calculus books. Equally well we can start with the expression for the unit normal $\mathbf{n} = \nabla\Phi/|\nabla\Phi|$ and the unit tangent vector $\mathbf{t} = \mathbf{e}_z \times \mathbf{n}$ (where $\mathbf{e}_z$ is the unit vector in the z direction). Differentiation with respect to arc length can be written as

$$\frac{d}{ds} = t_x\frac{\partial}{\partial x} + t_y\frac{\partial}{\partial y}, \tag{A9}$$

so that we have

$$\begin{aligned}\kappa = \mathbf{t}\cdot\frac{d\mathbf{n}}{ds} &= \left[\left(\frac{\partial\Phi}{\partial x}\right)^2 + \left(\frac{\partial\Phi}{\partial y}\right)^2\right]^{-3/2}\\ &\quad\times\left[\left(\frac{\partial\Phi}{\partial x}\right)^2\frac{\partial^2\Phi}{\partial y^2} + \left(\frac{\partial\Phi}{\partial y}\right)^2\frac{\partial^2\Phi}{\partial x^2}\right.\\ &\quad\left. - 2\frac{\partial\Phi}{\partial x}\frac{\partial\Phi}{\partial y}\frac{\partial^2\Phi}{\partial x\partial y}\right].\end{aligned} \tag{A10}$$

The solutions presented in the text require solving Eq. (A7) numerically for x,y pairs corresponding to the value of Φ chosen to define the conducting surface. For the x,y pairs, Eqs. (A8) and (A10) are then used to find $|\mathbf{E}|$ and κ at every point on the surface.

APPENDIX B

For our second class of examples we need to evaluate the curvature expression $\mathbf{t}\cdot d\mathbf{n}/ds$ of Eq. (3) in terms of components with respect to the orthonormal basis vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$, $\mathbf{e}_\phi$ for spherical coordinates. With these basis vectors a unit tangent vector $\mathbf{t}$ and the unit normal $\mathbf{n}$ are written as

$$\mathbf{t} = t^r\mathbf{e}_r + t^\theta\mathbf{e}_\theta + t^\phi\mathbf{e}_\phi,$$
$$\mathbf{n} = n^r\mathbf{e}_r + n^\theta\mathbf{e}_\theta + n^\phi\mathbf{e}_\phi. \tag{B1}$$

The derivative with respect to arc length ds in the direction $\mathbf{t}$ is

$$\frac{d}{ds} = t^r\frac{\partial}{\partial r} + \frac{t^\theta}{r}\frac{\partial}{\partial\theta} + \frac{t^\phi}{r\sin\theta}\frac{\partial}{\partial\phi}. \tag{B2}$$

In computing the derivative of $\mathbf{n}$ we will need the coordinate derivatives of the basis vectors:

$$\frac{\partial\mathbf{e}_r}{\partial\phi} = \sin\theta\mathbf{e}_\phi,\quad \frac{\partial\mathbf{e}_r}{\partial\theta} = \mathbf{e}_\theta,\quad \frac{\partial\mathbf{e}_r}{\partial r} = 0,$$
$$\frac{\partial\mathbf{e}_\theta}{\partial\phi} = \cos\theta\mathbf{e}_\phi,\quad \frac{\partial\mathbf{e}_\theta}{\partial\theta} = -\mathbf{e}_r,\quad \frac{\partial\mathbf{e}_\theta}{\partial r} = 0. \tag{B3}$$

For an axially symmetric surface specified by $r = r(\theta)$ the outward unit normal is given by

$$n^r = r/(r^2 + r'^2)^{1/2},\quad n^\theta = -r'/(r^2 + r'^2)^{1/2}, \tag{B4}$$

where we use prime to denote $d/d\theta$. Symmetry dictates that one of the principal directions is $\mathbf{t}_1 = \mathbf{e}_\phi$. For this direction

$$\frac{d}{ds} = \frac{1}{r\sin\theta}\frac{\partial}{\partial\phi}. \tag{B5}$$

Since n^r and n^θ are independent of ϕ we have

$$\frac{d\mathbf{n}}{ds} = n^r\frac{d}{ds}\mathbf{e}_r + n^\theta\frac{d}{ds}\mathbf{e}_r = \frac{1}{r\sin\phi}\left(n^r\frac{\partial\mathbf{e}_r}{\partial\phi} + n^\theta\frac{\partial\mathbf{e}_\theta}{\partial\phi}\right)$$
$$= \frac{1}{r}(n^r + \cot\theta n^\theta)\mathbf{e}_\phi,$$

so that, for this principal direction,

$$\kappa_1 = \mathbf{t}\cdot\frac{d\mathbf{n}}{ds} = \frac{1}{r}(n^r + \cot\theta n^\theta)$$
$$= \frac{1 - (r'/r)\cot\theta}{(r^2 + r'^2)^{1/2}}. \tag{B6}$$

The second principal direction is given by

$$\mathbf{t}_2 = \frac{1}{(r^2 + r'^2)^{1/2}}(r'\mathbf{e}_r + r\mathbf{e}_\theta), \tag{B7}$$

for which

$$\frac{d}{ds} = \frac{1}{(r^2 + r'^2)^{1/2}}\left(r'\frac{\partial}{\partial r} + \frac{\partial}{\partial\theta}\right), \tag{B8}$$

so that

$$\frac{d}{ds}\mathbf{e}_r = \frac{1}{(r^2 + r'^2)^{1/2}}\mathbf{e}_\theta,$$
$$\frac{d}{ds}\mathbf{e}_\theta = -\frac{1}{(r^2 + r'^2)^{1/2}}\mathbf{e}_r. \tag{B9}$$

We evaluate the derivatives of the components n^r, n^θ with

$$\frac{d}{ds} = \frac{1}{(r^2 + r'^2)^{1/2}}\frac{d}{d\theta}, \tag{B10}$$

where $d/d\theta$ is the total derivative $(r'\,\partial/\partial r + \partial/\partial\theta)$ along the curve $r = r(\theta)$. The result is

$$\kappa_2 = \mathbf{t}\cdot\frac{d\mathbf{n}}{ds} = \mathbf{t}_2\cdot\frac{1}{(r^2 + r'^2)^{1/2}}$$
$$\times\left(\frac{dn^r}{d\theta}\mathbf{e}_r + \frac{dn^\theta}{d\theta}\mathbf{e}_\theta + n^r\mathbf{e}_\theta - n^\theta\mathbf{e}_r\right)$$
$$= (2r'^2 + r^2 - rr'')/[(r^2 + r'^2)^{3/2}].$$

If the surface $r = r(\theta)$ specified by Eq. (7) is and terms of order ϵ^2 and smaller are ignored, Eq. leads to the result for κ_1 given in Eq. (9a). Simi Eq. (B11) leads to Eq. (9b) if the equa $(1-\mu^2)\,d^2P_n/d\mu^2 - 2\mu\,dP_n/d\mu + n(n+1)\,P_n = 0$ the Legendre polynomials is used.

The electrostatic potential everywhere outside an ax symmetric body with total charge Q can be written as

$$\Phi = \frac{Q}{r} + \sum_{n=1}^{\infty}\alpha_n\left(\frac{R}{r}\right)^{n+1}P_n(\mu).$$

At the surface $r = r(\theta)$ specified by Eq. (7) the potenti first order in ϵ is

$$\Phi\Big|_{\text{surface}} = \frac{Q}{R} - \epsilon\frac{Q}{R}\sum_{n=2}^{\infty}\beta_nP_n(\mu) + \sum_{n=1}^{\infty}\alpha_nP_n(\mu).$$

If Φ is to be constant on the surface the right-hand must be independent of μ so that $\alpha_1 = 0$ and

$$\alpha_n = \epsilon(Q/R)\beta_n,\quad n\geqslant 2,$$

and outside the surface, to first order in ϵ,

$$\Phi = \frac{Q}{r} + \epsilon\frac{Q}{R}\sum_{n=2}^{\infty}\beta_n\left(\frac{R}{r}\right)^{n+1}P_n(\mu).$$

The electric field intensity is given by

$$|\mathbf{E}| = \left[\left(\frac{\partial\Phi}{\partial r}\right)^2 + \frac{1}{r^2}\left(\frac{\partial\Phi}{\partial\theta}\right)^2\right]^{1/2}.$$

The $\partial\Phi/\partial\theta$ term contributes to $|\mathbf{E}|$ only to order ϵ^2 higher, so that to first order in ϵ

$$|\mathbf{E}| = \left|\frac{\partial\Phi}{\partial r}\right|$$
$$= \frac{Q}{r^2} + \epsilon\frac{Q}{R}\sum_{n=2}^{\infty}(n+1)\beta_n\frac{R^{n+1}}{r^{n+2}}P_n(\mu).$$

The E-field strength just outside the surface is foun using Eq. (7) in Eq. (B17). The result, to first order in given as Eq. (8).

[a] Permanent address: Department of Physics, California State Uni Fullerton, CA 92634.

[1] See, for example, Secs. 2.11 and 3.4 in J. D. Jackson, *Classical El namics* (Wiley, New York, 1975), 2nd ed.

[2] If the two principal radii are equal at a point then, at that point, is independent of $\mathbf{t}$. At such a point there is only one value of κ and the surface *can* be approximated near the point by a sphere. The po the axis of any surface of revolution have this property of poss unique radius of curvature.

[3] This is given, for example, as Prob. 1.11 in Ref. 1. This shows that the local nature of curvature, and the nonlocal nature of solu Laplace's equation, there can be *some* relationships between the

[4] The same argument can be applied to Eqs. (8) and (9). The fu form for $|\mathbf{E}|$ is different from that of either Eq. (9a) or (9b), so th can be no general relation between the maximum of $|\mathbf{E}|$ and th mum of either principal curvature.

The impossibility of a simple derivation of the Schwarzschild metric

Ronald P. Gruber
3318 Elm Street, Oakland, California 94609

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

Stephen M. Matthews
Lawrence Livermore National Labratories, Livermore, California 94550

William R. Cordwell
Sandia National Laboratories, Albuquerque, New Mexico 87185

Lawrence F. Wagner
VLSI Technology Inc., 1109 McKay Drive, San Jose, California 05131

(Received 2 February 1987; accepted for publication 28 April 1987)

The Schwarzschild metric, the general relativistic description of the space-time outside a spherical mass, has an extremely simple appearance. Because of this many attempts have been made to derive it by combining special relativity with concepts of Newtonian gravitation. It is shown here that such a derivation is impossible. A general discussion is given of the relationship of relativistic gravitation and its Newtonian limit with special emphasis on a particular non-Newtonian effect: spatial curvature.

I. INTRODUCTION

Many of the physical phenomena of interest in connection with relativistically strong gravitational fields can be understood—without any knowledge of general relativity—from the Schwarzschild metric, the "distance formula" for the space-time surrounding a static spherically symmetric gravitating body. For this reason there have been many attempts to "derive" the Schwarzschild metric, without the explicit use of general relativity, by gedanken experiments, plausibility arguments, etc. Such an attempt was published, for example, by Schiff.[1] The flaws in that attempt were subsequently pointed out by Rindler.[2] As Rindler noted, it is the great simplicity of the Schwarzschild metric that allows simple, but incorrect, arguments to give the appearance of a derivation. The appeal of appearances, however, is strong and one finds simple "apparent derivations" still being published.[3]

We wish to point out here that no simple derivation based on the ideas of special relativity and of Newtonian gravitation exists, nor can such a derivation exist. According to relativistic theories, gravity enters the Schwarzschild formula in two guises: (i) It causes inward "pull," e.g., a body in free motion has larger velocity at smaller radius, and (ii) gravity distorts the geometry of three-dimensional space. Simple, pedagogically useful and technically correct arguments can be made for dealing with the "pull" aspect of gravity. (See Appendix.) It is the spatial-distortion aspect of gravity that ensures that too simple a derivation of the Schwarzschild metric must fail.

Although this obstacle is unavoidable, the need for a pedagogical approach to the Schwarzschild metric is real. In this article we try to fill this need by emphasizing that the metric requires that gravity be invoked by the specification of *two* gravitational effects. We then provide an approach to the metric that requires of the reader a familiarity only with special relativity. We believe that this approach has some of the appealing simplicity of earlier pedagogical "derivations," but is technically correct. We present this approach in Sec. II. In Sec. III we discuss the meaning of spatial curvature and some of its counterintuitive consequences. It has been our experience that this aspect of relativistic gravitation is a particular source of confusion.

The Appendix deals with these issues *with* the mathematics of general relativity. In particular we show how the two functional degrees of freedom in the metric are related to relativistic particle mechanics. We also discuss various ways, in addition to the approach presented in Sec. II, in which two pieces of gravitational information can be specified.

II. THE SCHWARZSCHILD METRIC

To describe the geometry of space-time most simply we must make appropriate choices of coordinates of events in space-time. Since the space-time associated with a spherical unchanging gravitational source is static, it is natural to choose a time coordinate t such that the metric formula is independent of t. Because the space-time is spherical we can, at any t, fill space with two-dimensional spherical surfaces. We label such spherical surfaces with a coordinate label r ("Schwarzschild radial coordinate") chosen according to the geometric criterion that the area of the surface at r is $4\pi r^2$. If standard polar coordinates θ and ϕ are introduced on each surface it follows that distances on the surfaces of constant r and constant t are given by $ds^2 = r^2(d\theta + \sin\theta\, d\phi^2)$. The full, four-dimensional, space-time metric must account also for distances involving displacements in time ($dt \neq 0$) and in radius ($dr \neq 0$). By spherical symmetry the extended formula cannot depend on the angular coordinates θ and ϕ, and by the choice of the time coordinate it must also be independent of t. The resulting extended formula, the space-time metric for a spherically symmetric static gravitating source, must then have the form[4]

$$ds^2 = -[F(r)]^2 c^2\, dt^2 + [H(r)]^2\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\,. \tag{1}$$

It is worth emphasizing that we have arrived at Eq. (1) by

consideration of symmetry and coordinate choice only; gravitation, whether that of general relativity or otherwise, has not yet been introduced. The occurrence of two *a priori* unspecified functions $F(r)$ and $H(r)$ is the mathematical reason that two separate gravitational effects must be specified to fix the metric.

Let us consider a coordinate stationary observer, hereafter abbreviated CSO, i.e., we consider an observer with unchanging values of spatial coordinates r, θ, and ϕ. We shall denote the proper time of the CSO (the time measured by the clock carried by the CSO) as $\tau_{\rm CSO}$. From Eq. (1) it is immediately clear that the coordinate time differential dt between two CSO events and the proper time (clock) differential is

$$d\tau_{\rm CSO} = F(r)dt\,. \tag{2}$$

We now invoke the first property of gravity. We imagine a particle that starts from rest at infinity and falls radially inward. Intuitively we know that the inward velocity $v(r)$ of the freely falling particle, *as measured by CSOs*, must increase from its initial zero value at $r=\infty$. Our first specification will be this function:

$$\begin{aligned} v(r) &\equiv \text{velocity of freely falling} \\ &\quad \text{particle as measured by CSO}\,, \\ &= 0, \quad \text{at } r=\infty\,. \end{aligned} \tag{3}$$

Since this piece of information tells us the total mass–energy the CSO will measure for the freely falling particle we call it the "gravitational energy law." It turns out that when this energy law is known the function $F(r)$ can be directly inferred. The pedagogically simplest treatment involves a gedanken experiment with photons; since this treatment is rather standard we relegate it to the Appendix.

The second piece of gravitational information may be specified in several ways. For our purposes here it is simplest to specify what is, in a sense, the "gravitational force law." We again consider a freely falling particle, hereafter FFP, and denote by $d\tau_{\rm FFP}$ the differential of proper time between events experienced by the FFP (i.e., $d\tau_{\rm FFP}$ is the time differential measured by the clock carried with the FFP). A plausible analog of Newtonian radial acceleration that is at the same time relativistically palatable (being constructed of the geometrically defined radial coordinate r and the particle proper time $\tau_{\rm FFP}$) is $d^2r/d\tau_{\rm FFP}^2$. We specify the "gravitational force law" by specifying

$$f(r) \equiv -\frac{d^2r}{d\tau_{\rm FFP}^2}\,. \tag{4}$$

For a gravitating body of mass M, in Newtonian gravitation theory, the energy and force laws are

$$\left.\begin{aligned} v(r) &= -(2GM/r)^{1/2} \\ f(r) &= GM/r^2 \end{aligned}\right\} \begin{array}{l}\text{Newtonian}\\ \text{theory}\end{array}\,. \tag{5}$$

In Newtonian theory, of course, these two functions are not independent, but are necessarily related by $dv^2/dr = -2f$ via Newton's second law. We emphasize here that for relativistic gravitation *there is no a priori logically necessary relation of* $v(r)$ *and* $f(r)$.

We now proceed to show that *if* we knew the energy law and the force law we could find F and H in Eq. (1) and could therefore fully determine the nature of space-time. We will then show that we cannot infer these laws from simple considerations (Newtonian gravitation, special relativity, the equivalence principle, etc.). Rather, a specific theory of gravity, such as Einstein's, is needed. Since the F function follows from the energy law above, as explained in the Appendix, we concentrate here on the more elusive function H that governs spatial distortion. We start by considering two nearby events experienced by a FFP with proper time difference $d\tau_{\rm FFP}$ and coordinate time difference dt. Since the relative velocity of the FFP and the CSOs at radius r is $v(r)$ the CSOs will see a time dilation of the FFP clock and will measure for the events

$$d\tau_{\rm CSO} = d\tau_{\rm FFP}/\sqrt{1-v^2(r)/c^2}\,. \tag{6}$$

But we have seen in Eq. (2) that $d\tau_{\rm CSO} = F(r)dt$ so that

$$F(r)dt = d\tau_{\rm FFP}/\sqrt{1-v^2(r)/c^2}\,. \tag{7}$$

We next relate $v(r)$ to the coordinate velocity dr/dt. By the definition of $v(r)$, as a FFP moves by a CSO

$$\begin{aligned} v(r) &= \frac{d\ (\text{distance measured by CSO})}{d\ (\text{time measured by CSO})} \\ &= \frac{H\,dr}{F\,dt} = \frac{H\,dr}{d\tau_{\rm FFP}}\sqrt{1-\frac{v^2(r)}{c^2}}\,, \end{aligned} \tag{8}$$

or

$$\frac{dr}{d\tau_{\rm FFP}} = \frac{v}{H\sqrt{1-v^2/c^2}}\,. \tag{9}$$

We now differentiate this result and combine it with the force law [Eq. (4)] to arrive at

$$\frac{d^2r}{d\tau_{\rm FFP}^2} = \frac{1}{2}\frac{d}{dr}\left(\frac{v}{H\sqrt{1-v^2/c^2}}\right)^2 = -f(r)\,. \tag{10}$$

(Any question of the validity of the first equality is dispelled in the Appendix, where it is derived from the geodesic equation.) From the second equality and the fact that $v=0$ and $H=1$ at $r=\infty$ we have

$$\frac{v^2}{H^2(1-v^2/c^2)} = -2\int_\infty^r f(r')dr'\,. \tag{11}$$

This equation clearly shows that from the energy law and the force law [i.e., from $v(r)$ and $f(r)$] we may determine $H(r)$.

In order to apply Eq. (11) we must know the form of the functions $v(r)$ and $f(r)$. From some basic considerations, and without a specific reference to a relativistic theory of gravity, we can in fact say quite a bit about $v(r)$ and $H(r)$. We start by listing the only physical quantities relevant to gravitational effects at some point in a spherically symmetric field. These are the radius r, the mass M of the central gravitating object, the gravitational constant G, and the speed of light c. If we limit ourselves to *nonrelativistic* gravity then the speed of light has no relevance and we are limited to r, M, and G. From these quantities there is only one possible way (aside from multiplicative constants) of constructing an expression $(GM/r)^{1/2}$, which is dimensionally a velocity, and only one possible way of constructing an expression GM/r^2, with dimensions of acceleration. This is of course in accord with the Newtonian theory of gravity, in which the energy and force laws have the form in Eq. (5).

When we extend our considerations to relativistic gravity the speed of light becomes relevant and we may construct the dimensionless quantity GM/rc^2. From dimensional considerations it then follows that the energy and

force laws *must* be expressible as

$$v(r) = -(2GM/r)^{1/2}U(GM/rc^2)$$
$$f(r) = (GM/r^2)W(GM/rc^2)\,. \tag{12}$$

Here U and W are some (dimensionless) functions that depend on the details of our relativistic gravity theory. The dimensionless combination GM/rc^2 in these functions tells us "how relativistic" the gravitational field is. For example, for circular planetary motion with speed v_p, at radius r, we have $GM/rc^2 \sim (v_p/c)^2$; for a spherical star of mass M and radius r, the gravitational binding energy is of order GM^2/r and therefore is equivalent to a change in mass GM^2/rc^2 of the star, or a fractional change of GM/rc^2. In "ordinary" astronomy GM/rc^2 is extremely small (e.g., 2×10^{-6} at the Sun's surface) and for most purposes can be approximated as zero. This is, in fact, precisely what is meant by the weak field limit of a gravity theory; i.e., setting $GM/rc^2 = 0$ is the condition that the field is weak enough so that all relativistic effects may be ignored. If we demand that the nonrelativistic theory of gravity (i.e., the weak field limit of our gravity theory) be Newton's theory, then we must have that

$$\begin{matrix} U(x)\to 1 \\ V(x)\to 1 \end{matrix} \quad \text{as } x\to 0\,. \tag{13}$$

[Compare Eqs. (5) and (12).]

This viewpoint may be pushed further. For fields in which relativistic effects are weak (GM/rc^2 is small) but nonnegligible, we may write U and V as the expansions[5]

$$v(r) = -\left(\frac{2GM}{r}\right)^{1/2}\left[1 + A_1\frac{GM}{rc^2} + A_2\left(\frac{GM}{rc^2}\right)^2 + \cdots\right],$$
$$f(r) = \left(\frac{GM}{r^2}\right)\left[1 + B_1\frac{GM}{rc^2} + B_2\left(\frac{GM}{rc^2}\right)^2 + \cdots\right]. \tag{14}$$

The particular numerical values of the coefficients A_i and B_i are determined by the particular theory of gravity we are considering.

We wish now to emphasize very strongly a crucial and often misunderstood point: In arriving at Eq. (12) and the constraint of Eq. (13)] or at Eq. (14), we have satisfied everything that must be satisfied on the basis of elementary conditions; from considerations of compatibility with relativity and compatibility (in the weak field limit) with Newtonian gravity theory we can go no further. To go further, to fix the functions U and V in Eq. (12), or the coefficients A_i, B_i in Eq. (14), we need a detailed theory, such as Einstein's.

Without such (decidedly nonelementary) further information we cannot fully apply Eq. (11). We could, however, get a partial answer by keeping only the lowest order relativistic corrections in Eq. (14) (i.e., the A_1 and B_1 terms) and by computing the lowest order relativistic influence on H. When Eq. (14) is used in Eq. (11) and terms quadratic and higher in GM/rc^2 are neglected we find

$$H^2 \simeq 1 + (2A_1 - \tfrac{1}{2}B_1 + 2)(GM/rc^2)\,. \tag{15}$$

[A similar approximation for F can be found with Eq. (A4) of the Appendix.]

We may take the point of view that the coefficients A_1 and B_1 must be experimentally measured (as they can be with present-day technology). Unless these measurements happen to reveal that $2A_1 - \frac{1}{2}B_1 + 2 = 0$ we must conclude that $H^2 \neq 1$ and space is curved. The condition $2A_1 - \frac{1}{2}B_1 + 2 = 0$ is not logically required, nor does it turn out to be compatible with high-precision measurements which show that $2A_1 - \frac{1}{2}B_1 + 2 \simeq 2$.

The opposite point of view is that A_i and B_i or the functions $v(r)$ and $f(r)$ are to be determined by a theory of gravity. Einstein's gravity theory, general relativity, which has withstood all experimental tests to date turns out to give the simplest form of these functions

$$\left.\begin{matrix} v(r) = -(2GM/r)^{1/2} \\ f(r) = GM/r^2 \end{matrix}\right\} \begin{matrix}\text{general}\\ \text{relativity}\end{matrix}\,. \tag{16}$$

(Equivalently all the A_i and B_i vanish.) In the form of these functions (but certainly not in all of its manifestations!) general relativity turns out to agree with Newtonian theory.

It is now straightforward to use Eq. (16) in Eq. (11) and to find that in general relativity

$$H^2 = (1 - 2GM/rc^2)^{-1} \tag{17}$$

and that, unless the field strength is weak enough so that we approximate $GM/rc^2 = 0$, space is indeed curved.[6]

III. SPATIAL CURVATURE AND INTUITION

The special theory of relativity contradicts our everyday intuition about the absolute nature of space and of time, and about the distinction between them. Perhaps the best known parable for introducing the failure of intuition in this context is the "twin paradox."[7] Here we would like to point out that spatial curvature, as in the Schwarzschild geometry, contradicts another facet of everyday intuition: that the relationship of distances in three-dimensional space is that of Euclidean geometry. In analogy with the twin paradox we shall emphasize that counterintuitive nature of spatial curvature by an example with a snappy title: the "heavy banana paradox."

We consider an ideal massless monkey sitting on a circular ring of circumference C_1. Concentric with the ring is a massless circular plate of circumference C_2. The monkey's arm, of length $(C_1 - C_2)/2\pi$, is just long enough to reach the empty plate. Two massive bananas are now brought in from very large distances along an axis perpendicular to the plane of the ring and plate, until the bananas reach the center and form a spherical mass. During this motion we take the circumference of both the ring and the plate to remain constant. (Such constraints could be imagined to be inherent in the equation of state of the materials making up the ring and the plate. Or they could be built into the program of a computer reading sensors and activating motors to hold constant the distances between particles along the circumferences.) Since the circumferences are constant it would seem clear that neither the ring nor the edge of the plate moves during the arrival of the bananas. However, to his amazement and frustration, the monkey finds that the edge of the plate is now further away and if his arm (like the ring) does not change in length the bananas are out of his reach.

This is easily verified with the mathematics of Sec. II. Combining Eq. (1) with Eq. (17) gives the Schwarzschild spatial metric appropriate either before or after the ban-

anas arrive

$$ds^2 = (1 - 2GM/rc^2)^{-1}\, dr^2 + r^2(d\theta^2 + \sin^2\theta\, d\phi^2)\,. \tag{18}$$

(Here we have set $dt = 0$ to reduce the space-time metric to the spatial metric, i.e., the description of spatial geometry.) Initially $M = 0$ and after the bananas arrive $M > 0$ is the total banana mass. By the geometric nature of the Schwarzschild coordinate, the coordinate radius r_1 of the ring is related to the ring circumference by $C_1 = 2\pi r_1$ and the coordinate r_2 of the edge of the plate is given by $C_2 = 2\pi r_2$. Since, by assumption, neither C_1 nor C_2 changes, it follows that both coordinate locations r_1 and r_2 are unchanged. (This constancy of coordinate position is consistent with the claim that neither the ring nor the plate edge has moved.)

By the distance from the ring to the plate we mean of course a radial distance, one for which we can set $d\theta = d\phi = 0$ along a radial line. Elements of distance along this line are given by $ds = (1 - 2GM/rc^2)^{-1/2}\, dr$ so that the total ring to plate-edge distance is

$$L = \int_{r_2}^{r_1} \left(1 - \frac{2GM}{rc^2}\right)^{-1/2} dr\,. \tag{19}$$

Initially, with $M = 0$, this formula gives $L = r_1 - r_2$, of course, but after the bananas are in place, L is clearly larger than $r_1 - r_2$. For reasonable size massless monkeys and plates, and reasonable mass bananas the effect, of course, is rather small (one part in 10^{27}!) but the different aging of twins in the twin paradox is also small for reasonable choices of velocity.

The point here is one of principle: Two concentric circles have been taken to have constant circumference. Neither one, it can be argued, has therefore moved, and yet they have moved apart in the sense that the radial distance between them has increased. The question whether the ring or plate has moved is at best a vague one and in any case not central to the issue. The explanation lies rather in the fact that the structure of space has been changed by the introduction of the massive bananas, and the relationship of radial and circumferential distances has been altered from that which intuition demands.

APPENDIX

In this Appendix we present first a sketch of the simple standard derivation[8] of the relationship of $F(r)$ and the energy law [i.e., $v(r)$]. We imagine a continuous beam of radiation aimed radially inward from a larger radial coordinate r_1 to a smaller radial coordinate r_0. Since the space-time does not depend on t, the coordinate time required for one wave trough to travel from r_1 to r_0 is the same as that for the subsequent trough. The *coordinate* period Δt (the coordinate time between the arrival of successive wave troughs at a given position in space) is therefore the same at r_1 as at r_0. The *proper* period observed by a CSO at radius r is $\Delta\tau = F(r)\Delta t$ so that

$$\frac{\text{period observed at } r_0}{\text{period observed at } r_1} = \frac{F(r_0)}{F(r_1)}\,. \tag{A1}$$

We now consider the beam to consist of photons of energy $h\nu$ as measured by CSOs, so that

$$\frac{h\nu \text{ at } r_0}{h\nu \text{ at } r_1} = \frac{F(r_1)}{F(r_0)}\,. \tag{A2}$$

This equation gives the blue shift of a photon traveling inward. If we take r_1 to be at spatial infinity (where $F = 1$ if the geometry is to be flat at large radii) and replace r_0 by a general r we find that the inward "falling" photon increases in energy according to

$$\frac{\text{photon energy at } r}{\text{photon energy at } \infty} = \frac{1}{F(r)}\,. \tag{A3}$$

The energy $mc^2(1 - v^2/c^2)^{-1/2}$ of a particle (mass m) must increase in precisely the same way.[9] Thus for a particle starting with $v = 0$ at $r = \infty$ we have

$$\sqrt{1 - v^2(r)/c^2} = F(r)\,. \tag{A4}$$

In the remainder of this Appendix we start with the metric of Eq. (1) and, with the mathematical tools of relativity, give brief derivations of the functions $F(r)$ [or $v(r)$] and $H(r)$ on the one hand, and on the other hand some gravitational effects that, in principle or in practice, can be observed. We use here the notation of the textbook by Misner *et al.* (Ref. 4) including the setting of G and c to unity.

We start with a derivation of Eq. (10) with the mathematics of curved space-time. For radial free-fall motion, the r component of the geodesic equation, in terms of particle four velocity, U^μ gives us

$$0 = U^r{}_{;\nu} U^\nu = \frac{d^2 r}{d\tau^2} + (U^0)^2 \Gamma^r_{00} + (U^r)^2 \Gamma^r_{rr}\,. \tag{A5}$$

If the particle falls from rest at infinity then $U_0 = g_{00} U^0 = -1$. This together with $\mathbf{U}\cdot\mathbf{U} = -1 = g_{00}(U^0)^2 + g_{rr}(U^r)^2$ allows us, after some manipulation, to write Eq. (A5) as

$$\frac{d^2 r}{d\tau^2} = \frac{1}{2}\frac{d}{dr}\left[g^{rr}(U^0 - 1)\right]\,. \tag{A6}$$

In terms of the particle velocity $v(r)$ as measured by a CSO $U^0 = |g^{00}|^{1/2} U^{\hat{0}} = |g^{00}|^{1/2}(1 - v^2)^{-1/2}$. In the notation of Eq. (1) with $U^0 = F^{-1}(1 - v^2)^{-1/2}$ and $g^{rr} = H^{-2}$ we then have

$$\frac{d^2 r}{d\tau^2} = \frac{1}{2}\frac{d}{dr}\left(\frac{F^{-1}(1 - v^2)^{-1/2} - 1}{H^2}\right). \tag{A7}$$

When Eq. (A4) is used to replace F, Eq. (10) follows.

We now turn to the deflection of starlight in Eq. (1). We choose the plane of the motion to be $\theta = \pi/2$ so that the path is given by

$$\begin{aligned} 0 &= -F^2\left(\frac{dt}{d\lambda}\right)^2 + H^2\left(\frac{dr}{d\lambda}\right)^2 + r^2\left(\frac{d\phi}{d\lambda}\right)^2 \\ &= \left[-\frac{F^4(dt/d\lambda)^2}{r^4(d\phi/d\lambda)^2} + \frac{F^2H^2}{r^4}\left(\frac{dr}{d\phi}\right)^2 + \frac{F^2}{r^2}\right] \\ &\quad \times \frac{r^4}{F^2}\left(\frac{d\phi}{d\lambda}\right)^2, \end{aligned} \tag{A8}$$

where λ is an affine parameter. It is straightforward to show, with the geodesic equation, that $r^2\, d\phi/d\lambda$ and $F^2(dt/d\lambda)$ are constants of the motion. Equation (A8) can then be written

$$r^{-4}F^2H^2\left(\frac{dr}{d\phi}\right)^2 + r^{-2}F^2 = \text{constant} = \left(\frac{F(b)}{b}\right)^2, \tag{A9}$$

where b represents the minimum radius of the photon path, at which $dr/d\phi = 0$. We can next use Eq. (A9) to solve for the change in ϕ as r increases from b to ∞. If there were no deflection the answer would be $\pi/2$; the excess change in ϕ

gives the deflection for half the photon path. The total deflection is, therefore,

$$\text{deflection} = 2\left(\int_b^\infty \frac{d\phi}{dr}\,dr - \frac{\pi}{2}\right)$$
$$= 2\left\{\int_b^\infty r^{-1}H(r)\left[\left(\frac{rF(b)}{bF(r)}\right)^2 - 1\right]^{-1/2} \times dr - \frac{\pi}{2}\right\}, \tag{A10}$$

which shows that the deflection involves both the H and the F metric coefficients.

Kepler's law for the coordinate angular velocity $\omega = d\phi/dt$ of a circular orbit is easily found from the geodesic equation. We assume the circular orbit is in the $\theta = \pi/2$ plane and write out the r component of the geodesic equation for particle four velocity $\mathbf{U}$:

$$0 = U^r{}_{;\nu}U^\nu = (U^0)^2\Gamma^r_{00} + (U^\phi)^2\Gamma^r_{\phi\phi}\,. \tag{A11}$$

From this we immediately find

$$\omega^2 = \left(\frac{U^\phi}{U^0}\right)^2 = -\frac{\Gamma^r_{00}}{\Gamma^r_{\phi\phi}} = (2r)^{-1}\frac{d(F^2)}{dr}\,. \tag{A12}$$

It is easily shown that the proper-time angular velocity $\Omega = d\phi/d\tau$ is related to ω by

$$\Omega^2 = \omega^2(F^2 - r^2\omega^2)^{-1}\,. \tag{A13}$$

The fact that ω and Ω are unaffected by the spatial-geometry function $H(r)$ is consistent with the fact that circular orbits do not "explore" regions of different radius.

For a nearly circular planetary orbit in the $\theta = \pi/2$ plane we have

$$1 = F^2\left(\frac{dt}{d\tau}\right)^2 - H^2\left(\frac{dr}{d\tau}\right)^2 - r^2\left(\frac{d\phi}{d\tau}\right)^2. \tag{A14}$$

Since $E = F^2(dt/d\tau)$ and $L = r^2(d\phi/d\tau)$ are constant this result can be rewritten as

$$1 = E^2F^{-2} - L^2\left[r^{-4}H^2\left(\frac{dr}{d\phi}\right)^2 + r^{-2}\right], \tag{A15}$$

which, with $u \equiv M/r$ and $\tilde{L} \equiv L/M$, becomes

$$\tilde{L}^2\left[\left(\frac{du}{d\phi}\right)^2 + H^{-2}u^2\right] = H^{-2}(E^2F^{-2} - 1)\,. \tag{A16}$$

At this point it is useful to expand metric functions F and H as

$$1/H^2(r) = 1 + a_1u + a_2u^2 + \mathcal{O}(u^3)$$
$$1/F^2(r) = 1 + b_1u + b_2u^2 + \mathcal{O}(u^3)\,, \tag{A17}$$

and for orbits near inverse radius u we note that $E^2 = 1 + \mathcal{O}(u)$ and $\tilde{L}^2 = \mathcal{O}(u^{-1})$. We can now expand Eq. (A16) in orders of u, keeping only nonrelativistic terms $[\mathcal{O}(u)]$ and first-order relativistic corrections $[\mathcal{O}(u^2)]$:

$$\tilde{L}^2\left[\left(\frac{du}{d\phi}\right)^2 + u^2 + a_1u^3\right]$$
$$= E^2 - 1 + [(E^2-1)a_1 + E^2b_1]u + E^2(b_2 + a_1b_1)u^2 + \mathcal{O}(u^3)\,. \tag{A18}$$

When this is differentiated with respect to ϕ the result is

$$\tilde{L}^2\left(\frac{2d^2u}{d\phi^2} + 2u + 3a_1u^2\right)$$
$$= (E^2-1)a_1 + E^2b_1 + 2uE^2(b_2 + a_1b_1) + \mathcal{O}(u^2)\,. \tag{A19}$$

For a nearly circular orbit we can substitute $u = u_0 + \psi$ in Eq. (A19) to find

$$\frac{d^2\psi}{d\phi^2} + \kappa^2\psi + \left(\frac{3}{2}\right)a_1\psi^2 + \mathcal{O}(u_0^3) = 0\,,$$
$$\kappa^2 = 1 + 3a_1u_0 - E^2\tilde{L}^{-2}(b_2 + a_1b_1)\,. \tag{A20}$$

Aside from secular effects generated by the ψ^2 term in Eq. (A20) the orbit will be periodic in the variable $\kappa\phi$, with period 2π. The advance of the periastron per orbit will therefore be $2\pi(\kappa^{-1} - 1)$. Since $E^2 = 1 + \mathcal{O}(u_0)$ and, if $\tilde{L}^2$ has the correct Newtonian limit $\tilde{L}^2 = u_0^{-1} + \mathcal{O}(1)$, the advance of the periastron per orbit is[10]

$$2\pi(\kappa^{-1} - 1) \simeq \pi(-3a_1 + b_2 + a_1b_1)u_0\,. \tag{A21}$$

The Newtonian limit [see Eqs. (14), (A4), (A17)] constrains b_1 to be 2, so that the periastron advance per orbit is $\pi(b_2 - a_1)$ and therefore depends both on the relativistic correction to $F(r)$ and on the lowest order spatial curvature.

ACKNOWLEDGMENTS

This letter grew out of conversations, correspondences, and arguments among the authors. During these interactions, which lasted more than a year, useful discussions were carried out with a number of others interested in the subject. We would like to acknowledge the helpful suggestions of Kip Thorne, Christopher Cunningham, Russell Hamilton, and Rob Zako. One of us (RHP) gratefully acknowledges support from the National Science Foundation under grant PHY 8503653.

[1]L. I. Schiff, Am. J. Phys. **28**, 340 (1960).
[2]W. Rindler, Am. J. Phys. **36**, 540 (1968).
[3]See, e.g., M. Harwit, *Astrophysical Concepts* (Wiley, New York, 1973), Sec. 5:13.
[4]The above justification for Eq. (1) is sketchy and incomplete. More detail can be found in R. H. Price, Am. J. Phys. **50**, 300 (1982), Sec. VI A. A still more detailed and rigorous derivation is given in C. W. Misner, K. S. Thorne, and J. A. Wheeler, *Gravitation* (Freeman, San Francisco, 1973), Box 23.3.
[5]Strictly speaking these expansions require the additional reasonable assumption that $U(x)$ and $V(x)$ are analytic at $x \to \infty$.
[6]For pedagogical and conceptual reasons we present a picture in which the relativistic theory fixes $v(r)$ and $f(r)$ and they in turn fix F and H. It is more standard, in fact, to consider the theory to fix F and H and through them to determine $v(r)$ and $f(r)$.
[7]See, e.g., E. F. Taylor and J. A. Wheeler, *Spacetime Physics* (Freeman, San Francisco, 1963), p.71 (where it is called "the clock paradox"), and references given there.
[8]This argument is also presented by Rindler (Ref. 2).
[9]For gedanken experiments which justify this see, e.g., Misner, Thorne, and Wheeler (Ref. 4), p. 187.
[10]For more details on this type of computation of the advance of the periastron see A. S. Eddington, *The Mathematical Theory of Relativity* (Cambridge U. P., Cambridge, 1922), Sec. 40.

211. "The Gravitational Acceleration of Antiprotons," T. Goldman, M. V. Hynes, and M. M. Nieto, Gen. Relativ. Gravit. **18**, 67–70 (1986). The authors address the question of the nature of the gravitational acceleration experienced by antimatter and discuss a way to measure it. (E)

212. "Experimental Evidence for Quantum Gravity?" T. Goldman, R. J. Hughes, and M. M. Nieto, Phys. Lett. B **171**, 217–222 (1986). Experiments are proposed that might reveal classical effects predicted by quantum gravity theories. (I)

213. "On the Gravitational Interactions of Ultra-Relativistic Particles," M. Gasperini, Phys. Lett. B **177**, 51–54 (1986). Ultrarelativistic particles might exhibit a velocity-dependent breakdown of Newton's law and this, in turn, would lead to insights into the various supersymmetric theories of gravity. (A)

214. "Proposed Measurement of the Gravitational Acceleration of the Antiproton," R. E. Brown, in *Intersections Between Particle and Nuclear Physics, American Institute of Physics Conference Proceedings*, Vol. 150, edited by D. F. Geesaman (American Institute of Physics, New York, 1986), pp. 436–444. The fundamental elements of an experimental test of the gravitational acceleration of antimatter are discussed. The experiment involves a large collaboration headed by a team from Los Alamos National Laboratory. (I)

215. "Experimental Particle Physics Without Accelerators," J. Rich, D. Lloyd Owen, and M. Spiro, Phys. Rep. **151**, 239–364 (1987). The authors present a thorough description of the interface between particle physics and gravitation, with in-depth discussion of medium-range forces, and of the time dependence of G and the other fundamental constants. (A)

*216. "Gravity and Antimatter," T. Goldman, R. J. Hughes, and M. M. Nieto, Sci. Am. **256**, 48–56 (1988). The relationship between theories of gravity, quantum mechanics, and the possible behavior of antimatter in a gravitational field is presented to a public audience. (E)

217. "Gravity Measurements on Antimatter and Supergravity," N. Beverini, V. Lagomarsino, G. Manuzio, R. Poggiani, F. Scuri, and G. Torelli, Phys. Lett. B **206**, 533–534 (1988). The forces at work in matter–matter and matter–antimatter experiments are tabulated and discussed. (I)

218. "A New Limit on the Arion Interaction Constant," P. V. Vorobyov and Ya. I. Gitarts, Phys. Lett. B **208**, 146–148 (1988). A SQUID-based search for the weak, gravitylike coupling that would arise from a quasi-magnetic electron–electron interaction yields null results and sets stringent limits on the arionic interaction constant. (A)

219. "The Principle of Equivalence, Quantum Gravity and New Gravitational Forces," M. M. Nieto, T. Goldman, and R. J. Hughes, Aust. Phys. **25**, 259–262 (1988). The nature of the connection between gravity and the other forces of nature (all of which are quantized) can be explored by proper tests of the Principle of Equivalence and Newton's Law of Gravitation. (E)

220. "Antimatter: Its History and Properties," M. M. Nieto and R. J. Hughes, in *Antimatter Science and Technology*," edited by B. W. Augenstein, B. E. Bonner, F. E. Mills, and M. M. Nieto (World Scientific, Singapore, 1988), pp. 228–248. The authors present a very useful survey of our state of knowledge of antimatter. (I)

221. "Search for Anomalous Spin Interaction between Electrons," R. C. Ritter, W.-T. Ni, C. E. Goldblum, and G. T. Gillies, in *Experimental Gravitational Physics, Proceedings of the International Symposium on Experimental Gravitational Physics*, edited by P. F. Michelson, H. En-Ke, and G. Pizzella (World Scientific, Singapore, 1988), pp. 456–460. The preliminary result of a Cavendish experiment using spin-polarized test masses is reported. (I)

222. "Solid Antimatter," M. M. Michaelis and R. Bingham, Laser Part. Beams **6**, 83–91 (1988). Following an earlier suggestion by Robert L. Forward, the authors consider the difficulties involved in producing and dealing with bulk quantities of antimatter. (E)

223. "Test of Equivalence Principle for Neutrinos and Antineutrinos," S. Pakvasa, W. A. Simmons, and T. J. Weiler, Phys. Rev. Lett. **39**, 1761–1763 (1989). Electron neutrinos and their antiparticles emitted by supernova SN 1987A are shown to satisfy the equivalence principle and to set limits on the proposed new long-range forces. (A)

ACKNOWLEDGMENTS

The author is grateful to the Director and staff of the International Bureau of Weights and Measures for their assistance with the original bibliographic work on G that served as the precursor for this Resource Letter. He also thanks Professor R. H. Stuewer, the referees, and the many gravitational physics colleagues who served as sounding boards for ideas during the preparation of this Resource Letter.

The force between two charged wires

Richard H. Price[a] and Richard P. Phillips[b]
University of Utah, Salt Lake City, Utah 84112

(Received 8 August 1988; accepted for publication 4 September 1989)

For two circular cross-section "infinitely long" wires charged to opposite electric potential, it is a standard problem to calculate the potential field, the attractive force per unit wire length, etc. If, however, the potentials are not exactly opposite, the problem is not even well defined. As shown here, the problem becomes well defined when the physical environment of the wires is considered. An effective electrical ground is imposed on the problem either by the presence of nearby conductors or due to the finite length of the wires.

I. INTRODUCTION AND OVERVIEW

A standard configuration in electrostatics involves two long parallel conductors charged to opposite electrical potentials $+V$ and $-V$. If the length of the conductors is much larger than their separation (labeled d in Fig. 1), the standard approach is to take the conductors to be infinitely long and thereby to reduce the problem to one in two-dimensional electrostatics. If the conductor cross sections are circular, the fields can be found in closed form[1] with the use of complex variable techniques, image line charges, or bipolar coordinates. In SI units, λ, the charge per unit

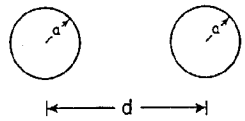

Fig. 1. Two infinitely long parallel conducting cylinders with circular cross sections. If the cylinders have opposite potentials imposed on them, the induced charge per unit length and the force of attraction per unit length can be found in terms of elementary functions.

length, and F, the force of attraction per unit length, are

$$\lambda = 4\pi\epsilon_0 V/\cosh^{-1}[(d^2/2a^2) - 1]$$
$$\approx 2\pi\epsilon_0 V/\ln(d/a),$$
$$F = (8\pi\epsilon_0/d\sqrt{1-4a^2/d^2})$$
$$\times\{V/\cosh^{-1}[(d^2/2a^2) - 1]\}^2$$
$$\approx (2\pi\epsilon_0/d)[V/\ln(a/d)]^2. \quad (1)$$

The approximate expressions given above are the limiting forms that apply when the wires are far apart compared to their radii, that is, when $d \gg a$.

It is easy to lose sight of the crucial role played in this configuration by the choice that the potentials on the two conductors have opposite polarity.[2] Consider, for example, the case in which the same potential $+V$ is applied to both wires. Two different viewpoints lead to very different predictions about the consequence of the applied potential. A "practical" scientist, especially one familiar with electroscopes, would argue that charging two closely spaced wires to the same potential would cause the wires to repel each other.[3] A "mathematical" scientist would argue, quite differently, that the problem is described by Laplace's equation, $\nabla^2\Phi = 0$, with the boundary conditions that the electrostatic potential Φ is $+V$ on the boundaries. Therefore, the unique solution is that $\Phi = +V$ everywhere. But if that were correct, there would be no electric fields (since $\nabla\Phi$ would vanish), no electrical charges on the wires (from Gauss' law), and no force between the wires.

Clearly, some information is missing; the physical problem is incompletely specified. It might seem that what is missing is the charge per unit length on the wires, but to specify that would be to sidestep the real issue: Experimentally, we impose "voltage" on the wires, not charge. Nature figures out what the charge must be. The charge per length should then be a result contained in the solution of a correctly posed mathematical problem, not an input into such a problem.

The difficulty has to do with boundary conditions,[4] in particular with the spatial location of ground, i.e., of zero potential. For the case of opposite potentials on the wires, it is clear that ground is located at the median plane between the wires. If, on the other hand, both cylinders are at $+V$, there is no natural location for ground. In particular, ground cannot in any simple way be at spatial infinity since the mathematics of two-dimensional electrostatics demands that the potential difference between spatial infinity and the wires is infinite, except in the single case that the wires have opposite potentials or, more generally (that is, if the wires have different cross sections), except in the case that the sum of the linear charge densities on the wires is zero.[2]

Where, then, is ground? Let us suppose that the wires are $L = 1$ m long and are separated by $d = 1$ mm. We must

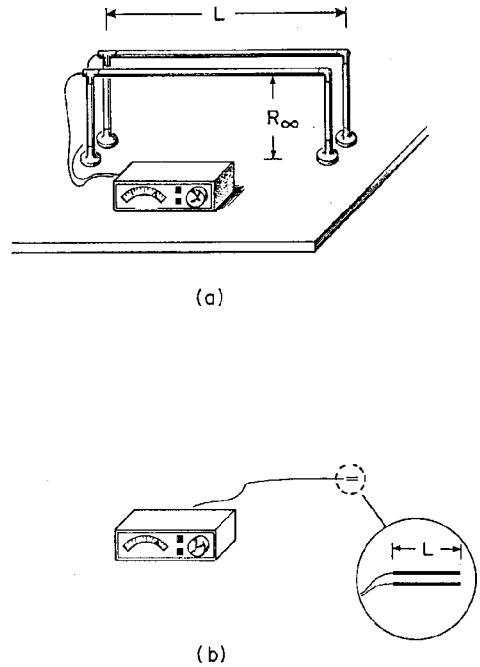

Fig. 2. The two electrostatic environments of two "infinite" parallel wires. In (a) a large metal laboratory table is at a distance R_∞ that is small compared to the wire length L. In (b) the wires are isolated; there are no other electrical elements within distances from the wires many times L.

distinguish two different kinds of experimental environments for these wires. In the first environment [Fig. 2(a)], there are large conductive elements (tables, laboratory apparatus, or whatever) at a distance R_∞, say 1 cm, from the wires. These external conducting elements may be far away in comparison with the separation between the wires, but they are close compared to the wire length. More generally this case is characterized by $L \gg R_\infty$. When this condition applies, it is justified to continue to use a two-dimensional ("infinitely long") viewpoint for the wires,[5] but the large external element must be taken into consideration. This element introduces a large equipotential surface into the neighborhood of the wires, a surface that we can define as "ground." (That is, the voltage on the wires should be taken with reference to the external conductive element. In determining the charge on the wires or the force between the wires, the voltage relative to some nominal circuit ground is irrelevant. It is the difference between the potential of the nearby conductive element and the potential of the wires that has physical consequences.) What is interesting and usually unappreciated about this sort of configuration is that the charges induced on the wires, and the forces between the wires, depend crucially on the location of the external conductive element. Without the location and shape of that element, the problem is incompletely specified.

The second type of electrostatic environment [Fig. 2(b)] is that in which the wires are isolated; there are no relevant external conductive elements. More realistically this means that $L \ll R_\infty$, as would be the case for our 1-m wires if they were many meters from any other conductive element. In this case, we must recognize the fact that the wires exist in a three-dimensional world. At points much closer to the wires than 1 m, and not too near the ends of the

wires, the field structure is approximately two-dimensional, but from many meters away the wires look like a "point" source, not a "line" source. For an isolated point source, of course, the appropriate ground is at spatial infinity. Thus three-dimensional reality can be inserted into the two-dimensional mathematics by imposing an effective ground surface at a distance from the wires of several times L.

The rest of this article will elaborate, with details and model problems, on the central ideas above. Section II deals with the influence of an external conductive element and presents a model problem for which a quantitative description can be given. Section III deals in an approximate quantitative manner with the case of isolated parallel wires.

II. THE INFLUENCE OF GROUND

We consider here the way in which a nearby external conductive element affects the electrical charge induced on wires and the forces on them. An electrostatic configuration with a realistically irregular nearby conducting element does not typically lead to a simple solution, and the need for a numerical solution might obscure the insights that are our goal here. We will therefore use a model problem that has enough flexibility to show the effect of changing the distance to ground and other interesting effects, but which allows a reasonably simple solution. To describe this model problem, in its simplest form, we start with the cylinders in Fig. 1 set to the same potential V. We then choose one of the equipotentials of that solution to be defined as ground, and we characterize the distance to ground as shown in Fig. 3. Once this equipotential is specified as zero potential, the relationship between the potential on the wires and the charge per unit length of the wires is fixed, and the force on each wire can be found.

In the limiting case $a \ll d \ll R_\infty$, it is not difficult to find an approximate solution. We can, in this limit, treat the fields as if they were due to a uniform surface charge distribution, on each cylinder, with charge per unit length λ. Equivalently, we can view the fields as those due to infinitesimal line charges λ at the cylinder axis. (In reality, the charge distributions will not be uniform, but the nonuniformity can be ignored for $d \gg a$ and $d \ll R_\infty$.) The contribution to the potential from each "line charge" is then $\Phi = (\lambda/2\pi\epsilon_0)\ln(R/r)$, where r is the distance from the line charge, and R is an arbitrary constant. The potential V on the surface of one of the cylinders has contributions from both line charges:

$$V \approx \frac{\lambda}{2\pi\epsilon_0}\ln\frac{R}{a} + \frac{\lambda}{2\pi\epsilon_0}\ln\frac{R}{d}.$$

The potential on the grounding sphere at (large) distance R_∞ is

$$0 \approx \frac{\lambda}{2\pi\epsilon_0}\ln\frac{R}{R_\infty} + \frac{\lambda}{2\pi\epsilon_0}\ln\frac{R}{R_\infty}.$$

From these, we find that $R \approx R_\infty$, and that the induced linear charge density is

$$\lambda \approx \{2\pi\epsilon_0 V/[2\ln(R_\infty/d) + \ln(d/a)]\}, \tag{2a}$$

and that the repulsive force per unit length is[6]

$$F \approx \frac{\lambda^2}{2\pi\epsilon_0 d} \approx \frac{2\pi\epsilon_0}{d}\left(\frac{V}{2\ln(R_\infty/d) + \ln(d/a)}\right)^2. \tag{2b}$$

Though the conditions $a \ll d \ll R_\infty$ allow a usefully simple approximation, they turn out to be unnecessary constraints; the electrostatics problem can be solved in closed form for arbitrary values of a, d, and R_∞. The details of the solution are given in the Appendix. Some numerical results for λ and F are presented in Fig. 4 with dashed lines representing the limiting approximations of Eqs. (2). These results show that the approximations are quite accurate except for the case $d/a = 2.1$, in which case the wires are very nearly touching each other, or the case that R_∞/d is near the limiting value ($R_\infty/d = \frac{1}{2} + a/d$) at which the ground surface is touching the wires. Equations (2), as well as the curves in Fig. 4, show that the closer the grounding surface is, the stronger the force between the wires. This agrees with the simple intuitive picture that a close grounding surface implies large potential gradients and therefore large electrical fields, with the consequence of large charges and forces.

We are now in a position to compare the (repulsive) force between the two wires of Fig. 1 symmetrically charged to $+V$ and $-V$. The mathematics in the Appendix covers the general case, but is rather complicated. The essence of the comparison can be seen in the limiting case of widely separated wires ($d \gg a$) with ground a large distance away ($R_\infty \gg d$). From Eqs. (1) and (2), we see that the forces are approximately equal if $(R_\infty/d)^2 \ll d/a$. But for $(R_\infty/d)^2 \gg d/a$, the force of repulsion in the symmetric case is much less than the force of attraction in the antisymmetric case. There is a competition of two geometric effects: the weakening of the force due to a wide separation and the weakening due to a large distance to ground. If the former one dominates, then the forces in the symmetric and antisymmetric cases have the same magnitude.

If the potentials V_1 and V_2 on the two wires in Fig. 1 are neither equal nor opposite, it will be convenient to describe their potentials by

$$V_{\rm av} \equiv \tfrac{1}{2}(V_1 + V_2), \quad V_{\rm diff} \equiv \tfrac{1}{2}(V_1 - V_2),$$

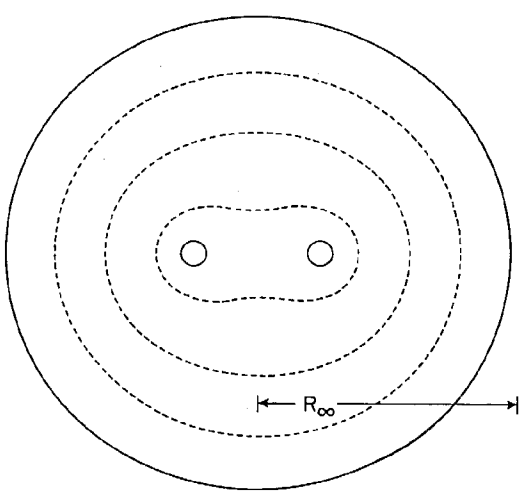

Fig. 3. A model problem for studying the influence of the location of ground. A closed-form solution exists for the potential field of two symmetrically charged circular cross-section conductors. Ground is chosen to be one of the equipotentials and is characterized by the horizontal distance R_∞ from the center of the figure to the equipotential. In the figure, $d/a = 10$, and for the solid curve, $R_\infty/d = 2$. The dashed curves represent other equipotentials, any of which could also be chosen as the zero-potential surface.

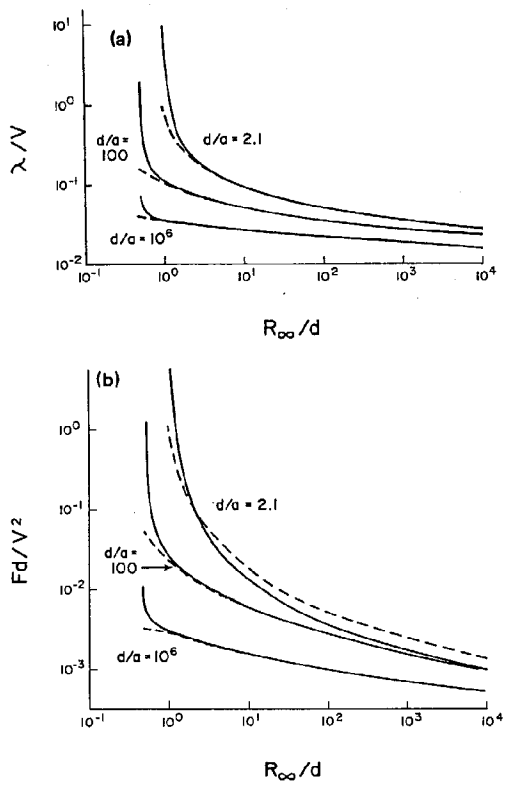

Fig. 4. (a) Induced charge per unit length λ, and (b) force per unit length F, on two wires, both at potential $+V$. Values are given as a function of R_∞/d, the distance to the grounding surface in units of the wire separation. Results are presented for three values of d/a, the separation of the wires in units of the wire radius. Dashed curves show the values given by the limiting form appropriate to large R_∞/d and d/a.

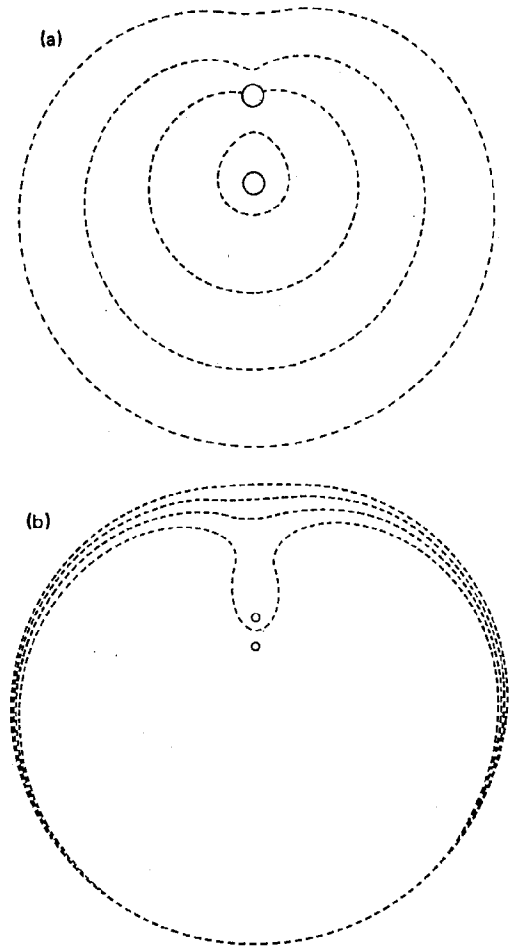

Fig. 5. Equipotentials around two circular cross-section wires with $d/a = 8$, and with $V_{\rm diff} = V_{\rm av}$ in (a) and $V_{\rm diff} = 2V_{\rm av}$ in (b).

so that symmetric and antisymmetric effects will be easily identifiable. If neither $V_{\rm av}$ nor $V_{\rm diff}$ vanishes, several factors complicate the analysis. Foremost is the fact that the equipotentials are now of complex shape. Figure 5 shows examples of this for $V_{\rm diff} = V_{\rm av}$ and $2V_{\rm av}$, both with $d = 8a$. The equipotentials are asymmetric and of complex shape, but in both figures, and in every case except that of $V_{\rm av} = 0$, at sufficient distances the field of the two wires looks like the field of a single wire of potential $V_{\rm av}$. It follows that at sufficiently large values of R_∞/d, equipotentials will always (except in the single case $V_{\rm av} = 0$) be approximately circular and that we can (as we did above for the case $V_{\rm diff} = 0$) use the sufficiently distant equipotentials to represent the location of ground.

The force between the wires, for arbitrary values of $R_\infty/d, d/a$, and $V_{\rm av}/V_{\rm diff}$, is described in the Appendix. As above, the nature of the answer is most easily seen in the limiting case. For widely separated ($d \gg a$) wires with a distant ground ($R_\infty \gg d$), the force of repulsion is

$$F = \frac{2\pi\epsilon_0}{d}\left[\left(\frac{V_{\rm av}}{2\ln(R_\infty/d) + \ln(d/a)}\right)^2 - \left(\frac{V_{\rm diff}}{\ln(d/a)}\right)^2\right]. \tag{3}$$

Note that Eq. (3) agrees with Eq. (1) in the limit $V_{\rm av} = 0$, and with Eq. (2) in the limit $V_{\rm diff} = 0$.

Some interesting patterns can be seen in Eq. (3). In particular, for very thin wires, specifically when $d/a \gg (R_\infty/d)^2$, the attractive and repulsive forces are on an equal footing and approximately cancel if $|V_{\rm av}| \approx |V_{\rm diff}|$. If, on the other hand, $d/a \ll (R_\infty/d)^2$, the attractive part of the force tends to be stronger than the repulsive part, and the overall force can be attractive, even if $V_{\rm av} \gg V_{\rm diff}$.

III. ISOLATED WIRES

If two wires are isolated in space with no nearby grounding electrode, we can consider that there is an effective grounding surface at a value of R_∞ on the order of L, the length of the wires. To see why this is so, we first note that at distances r such that $r \gg d$, but $r \ll L$, the electric potential has the form

$$\Phi = -(1/2\pi\epsilon_0)\lambda_{\rm tot}\ln r + \text{const.} \tag{4}$$

Here, $\lambda_{\rm tot}$ is the sum of the charge per unit length of both wires. If there were a grounding surface at some large $r = R_\infty$, the constant could be expressed in terms of R_∞ to give

$$\Phi = (1/2\pi\epsilon_0)\lambda_{\rm tot}\ln(R_\infty/r). \tag{5}$$

At distances $r \gg L$, the wires can be considered to be a point source with total charge $\lambda_{\rm tot}L$ and therefore with a poten-

tial

$$\Phi = \lambda_{\text{tot}} L / 4\pi\epsilon_0 r. \tag{6}$$

The actual potential in the median plane of the wires (the plane that bisects the wires and is orthogonal to the wires) can be reasonably well approximated by Eq. (5) for $r \ll L$. That is, Eq. (5) should be correct to order of magnitude until r is considerably larger than L. Similarly, Eq. (6) should be approximately true for $r \gg L$. At $r = L$, both equations must be approximately true and this can only be the case if $\ln(R_\infty / L) \approx 1$, i.e., for

$$R_\infty \approx L. \tag{7}$$

The reasoning here is based on the matching of a two-dimensional "line charge" source to its large-distance, three-dimensional "point source" form, and this reasoning can be checked with a relatively simple closed-form solution. We consider the case of a single isolated thin wire of length L. We model this wire as a conducting prolate spheroid of extreme eccentricity (see Appendix). The model has the disadvantage that the cross-sectional radius a is not constant, but varies along the wire approximately as $a \sim \sqrt{(L/2)^2 - z^2}$, where z, the coordinate along the wire, varies from $-L/2$ to $L/2$. This disadvantage is outweighed by two advantages. First, though the "wire" is not constant in cross section, it *is* constant in charge per unit length λ. This would seem to be just as natural a requirement for the three-dimensional extension of a wire that we have previously considered "infinitely long." And we cannot have it both ways. The wire can be constant in cross section *or* in λ, but not in both. The second advantage of this choice of wire, of course, is that the external field can be expressed in closed form.

In the median ($z = 0$) plane of the wire, the external field is

$$\Phi = (\lambda / 4\pi\epsilon_0) \times \ln\{[\sqrt{(2r/L)^2 + 1} + 1][\sqrt{(2r/L)^2 + 1} - 1]\}. \tag{8}$$

The $r \gg L$ limit gives $\Phi = \lambda L / 4\pi\epsilon_0 r$, which agrees, as it must, with Eq. (6). In the opposite limit, $r \ll L$, Eq. (8) becomes $\Phi = (\lambda / 2\pi\epsilon_0)\ln(L/r)$, and we can infer [by comparison with Eq. (5)] that the effective grounding surface for the wire is at $R_\infty = L$. That is, if we are interested in the fields near the wire (at $r \ll L$) where we consider the wire to be two-dimensional, we must locate a cylindrical grounding surface at $r \equiv R_\infty = L$. This is a specific example of the more general conclusion in Eq. (7).

APPENDIX

The detailed calculations justifying the results reported in the text are best done with bipolar coordinates[7] u and v, which are related to Cartesians by

$$x = c \sinh u / (\cosh u - \cos v),$$
$$y = c \sin v / (\cosh u - \cos v). \tag{A1}$$

The coordinate lines for both u and v are circles, as shown in Fig. 6. The cross sections of the wires are represented by the curves $u = +u_0$ and $-u_0$, where the bipolar quantities u_0 and c are related to the quantities a and d of Fig. 1 by

$$u_0 = \cosh^{-1}(d/2a), \quad c = \sqrt{(d/2)^2 - a^2}. \tag{A2}$$

The potential outside the cylinders at $u = \pm u_0$ must satisfy Laplace's equation $\nabla^2\Phi = 0$, which in bipolar coordinates has the form

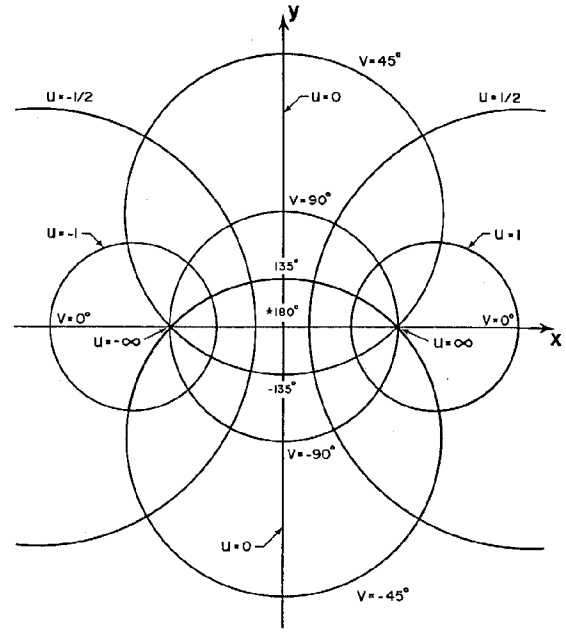

Fig. 6. Bipolar coordinates.

$$\frac{\partial^2\Phi}{\partial u^2} + \frac{\partial^2\Phi}{\partial v^2} = 0. \tag{A3}$$

When the cylinders have opposite potentials $\Phi = \pm V_{\text{diff}}$, the appropriate solution to Eq. (A3) is simply

$$\Phi = K_1 u. \tag{A4}$$

Since $\Phi = V_{\text{diff}}$ on the $u = u_0$ circle, the constant K_1 is fixed, and the potential becomes

$$\Phi = V_{\text{diff}}(u/u_0). \tag{A5}$$

The charge per unit length on the cylinders can be found by computing the charge density $\sigma = -\epsilon_0 \mathbf{n}\cdot\nabla\Phi$, where $\mathbf{n}$ is the unit outward normal. The result of integrating σ around the circumference of the $u = u_0$ circular conductor is a charge per unit length λ_{diff} as given (with V in place of V_{diff}) by Eq. (1). The value of λ_{diff} can be inferred more immediately by noticing that the potential in Eq. (A4) can be viewed as that arising from image line charges $\pm\lambda_{\text{diff}}$ at $x = \pm c, y = 0$. The electrostatic force between the wires can be found by taking the derivative with respect to the separation d, of the electrostatic energy per unit length $V\lambda_{\text{diff}} = (\lambda_{\text{diff}}^2 / 4\pi\epsilon_0)\cosh^{-1}[(d^2/2a^2) - 1]$. Alternatively, and more simply, the force is that between the two image line charges, $(\lambda_{\text{diff}})^2 / 4\pi\epsilon_0 c$, with c given by Eq. (A2).

For two cylinders at $u = \pm u_0$, both charged to the same potential V_{av}, a solution to Eq. (A3) must be found that is even in u, which gives $\Phi = V_{\text{av}}$ at $u = \pm u_0$, and for which the equipotentials at large distances approach circles about the origin. This solution[8] turns out to be

$$\Phi = V_{\text{av}} + K_2\left(|u| - u_0 - 2\sum_1^\infty \frac{e^{-n|u|}}{n}\cos nv + 2\sum_1^\infty \frac{e^{-nu_0}}{n}\frac{\cosh nu}{\cosh nu_0}\cos nv\right). \tag{A6}$$

By computing $-\mathbf{n}\cdot\nabla\Phi$, we find that $K_2 = \lambda_{av}/2\pi\epsilon_0$.

The general potential case, that of the cylinders at two potentials V_1 and V_2, is solved by the superposition of Eqs. (A5) and (A6):

$$\Phi = V_{diff}\frac{u}{u_0} + V_{av} + \left(\frac{\lambda_{av}}{2\pi\epsilon_0}\right) \times\left(|u| - u_0 - 2\sum_1^\infty \frac{e^{-n|u|}}{n}\cos nv + 2\sum_1^\infty \frac{e^{-nu_0}}{n}\frac{\cosh nu}{\cosh nu_0}\cos nv\right). \quad \text{(A7)}$$

In this general solution, however, the constraint λ_{av} has not been resolved. To do this we must specify the location of ground. In the prescription of the text, this means choosing a point on the x axis of Fig. 6 to be at $x = R_\infty$, the point that defines the equipotential at $\Phi = 0$. The $x = R_\infty, y = 0$ point has bipolar coordinate values

$$v = 0, \quad u = u_\infty \equiv 2\tanh^{-1}(c/R_\infty), \quad \text{(A8)}$$

where c is given in terms of a and d by Eq. (A2). By setting Φ to zero at these values of u and v, we find the relationship that determines λ_{av} in terms of the potentials:

$$0 = V_{diff}\frac{u_\infty}{u_0} + V_{av} + \left(\frac{\lambda_{av}}{2\pi\epsilon_0}\right)\left(u_\infty - u_0 - 2\sum_1^\infty \frac{e^{-nu_\infty}}{n} + 2\sum_1^\infty \frac{e^{-nu_0}}{n}\frac{\cosh nu_\infty}{\cosh nu_0}\right). \quad \text{(A9)}$$

In the case of distant ground ($u_\infty \ll 1$) and thin wires ($u_0 \gg 1$), this reduces to

$$\lambda_{av} \approx 2\pi\epsilon_0\{[V_{diff}(u_\infty/u_0) + V_{av}]/(u_0 - 2\ln u_\infty)\} \approx 2\pi\epsilon_0\{[V_{diff}(u_\infty/u_0) + V_{av}]/[\ln(d/a) + 2\ln(R_\infty/d)]\}. \quad \text{(A10)}$$

If V_{diff} (which makes little difference in any case) vanishes, we get the first relation in Eq. (2).

To find the force between the wires, the electrostatic pressure $\epsilon_0 E^2/2$ is computed for the wire on the right. The force per unit area on the wire is then projected in the x direction and integrated over the wire's surface. The force of repulsion is found to be

$$F = \frac{1}{32\pi^2\epsilon_0 c}\int_{-\pi}^{\pi} dv(\cosh u_0 \cos v - 1) \times\left(2\lambda_{av} + \frac{4\pi\epsilon_0 V_{diff}}{u_0} + 4\lambda_{av}\sum_1^\infty e^{-nu_0}\cos nv(1 + \tanh nu_0)\right)^2. \quad \text{(A11)}$$

In the limit of distant ground ($u_\infty \ll 1$) and thin wires ($u_0 \gg 1$), the force becomes

$$F = \frac{2\pi\epsilon_0}{d}\left[\left(\frac{\lambda_{av}}{2\pi\epsilon_0}\right)^2 - \left(\frac{V_{diff}}{u_0}\right)^2\right]. \quad \text{(A12)}$$

Equation (3) follows from Eq. (A12) if λ_{av} is replaced with its value from Eq. (A10) and if the negligible term $V_{diff}(u_\infty/u_0)$ in that equation is ignored.

In Sec. III the solution is used for the field outside a charged conducting prolate spheroid. A family of confocal spheroids, with focii at $z = \pm L/2$, can be parametrized with μ and defined in terms of Cartesian coordinates by

$$\frac{x^2 + y^2}{\mu^2 - 1} + \frac{z^2}{\mu^2} = \left(\frac{L}{2}\right)^2. \quad \text{(A13)}$$

The degenerate case $\mu = 1$ corresponds to the segment of the z axis from $z = -L/2$ to $+L/2$. The field outside a prolate spheroid is given by

$$\Phi = (Q/4\pi\epsilon_0 L)\ln[(u+1)/(\mu - 1)], \quad \text{(A14)}$$

where Q is the total charge on the spheroid. For the extreme $\mu - 1 \ll 1$ case, Q/L can be viewed as the charge per unit length of the wirelike limiting spheroid. In the median ($z = 0$) plane, $\mu = \sqrt{(2r/L)^2 + 1}$ and the potential has the form in Eq. (8).

ACKNOWLEDGMENTS

The insights presented here grew out of work at the Center for Engineering Design on electrical actuation. We gratefully acknowledge the support of the Systems Development Foundation, and DARPA for their support of this work under Contract F33615-86-C-5267. In addition, one of us (R.H.P.) gratefully acknowledges the support of the National Science Foundation under Grant PHY-8503653. We also thank an anonymous referee for helpful suggestions in the presentation of several points.

[a] Department of Physics.
[b] Center for Engineering Design.

[1] This problem is presented or solved, with varying levels of detail, in many texts, e.g., J. D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1962), 2nd ed., Prob. 2.4; L. Page and N. I. Adams, *Principles of Electricity* (Van Nostrand, Princeton, 1969), 4th ed., Sec. 32; E. M. Pugh and E. W. Pugh, *Electricity and Magnetism* (Addison-Wesley, Reading, MA, 1960), Sec. 4-11; W. R. Smythe, *Static and Dynamic Electricity* (McGraw-Hill, New York, 1939), Sec. 4.13.

[2] More generally, the mathematics in the solutions listed in Ref. 1 allows the wires to be at different potentials if they also have different radii. What is crucial is that the radii and potentials be related such that the charge per unit length on the two wires is equal and opposite.

[3] The practical scientist is correct. If the + lead of a dc voltage supply is attached to an electroscope (the moral equivalent of two long wires) and if the − or "ground" lead is ignored, the electroscope *will* deflect. With 1000 V even a fairly crude electroscope shows a strong deflection.

[4] The importance of boundary conditions, and how easily their importance is overlooked, is evident also in the question of the electric field outside a current-carrying wire. See T. N. Sarachman, Am. J. Phys. **37**, 748 (1969); R. Stoeckly, Am. J. Phys. **38**, 934 (1970); D. Marcuse, Am. J. Phys. **38**, 935 (1970); W. T. Scott, Am. J. Phys. **38**, 936 (1970).

[5] Actually, for the "infinitely long" viewpoint to apply, an additional constraint must be imposed: R_∞ must not vary significantly along the wire or along that length of the wire being considered.

[6] There is an additional contribution to the force on each cylinder due to the charge distribution on the grounding surface, but this is negligible in the limit $R_\infty \gg d$.

[7] H. Margenau and G. M. Murphy, *The Mathematics of Physics and Chemistry* (Van Nostrand, Princeton, 1956), 2nd ed., Sec. 5.14. See also Smythe, Ref. 1.

[8] M. P. Sarma and W. Janischewskyj, "Electrostatic field of parallel cylindrical conductors," IEEE Trans. Power Appar. Syst. **88**, 1069 (1969).

[24]D. W. E. Green, "1991 TB_1," IAU Circular 5368 (1991).
[25]In retrospect, it probably would have been sufficient to track only at the sidereal (stellar) rate in this case, because of the brightness of 1991 TB_1. For targets which are significantly fainter than the field stars which would be used as positional references, however, this technique would be valuable as it would increase the signal-to-noise ratio in the image of the asteroid. In fact, for faint asteroids, it might be preferable for the actual simultaneous exposures to be trailed in this fashion for the same reason, especially if the parallax measurement is to be made by the overlay method rather than by using computer software to derive the pixel coordinates of the asteroidal and stellar images.
[26]D. W. E. Green, "1992 AC," IAU Circular 5426 (1992), "1992 AC," IAU Circular 5442 (1992), "1992 AC," IAU Circular 5474 (1992).
[27]D. King-Hele, *Observing Earth Satellites* (Van Nostrand Reinhold, New York, 1983).
[28]D. Tattersfield, *Orbits for Amateurs with a Microcomputer, Volume II* (Wiley, New York, 1987), pp. 118–154.
[29]W. Livingston and D. Talent, "Stalking geosats with a camera," Sky & Telescope **80**, 319 (1990).

Negative mass can be positively amusing

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 30 July 1992; accepted 28 August 1992)

Some insights into gravitation and mechanics, as well as some surprises, result from considering the dynamics of negative mass particles.

Negative masses may be unphysical, but several examples are given of how they can be instructive. In these examples, the dynamics and gravitational interactions involving negative mass particles produce surprising (and amusing) predictions about the motions of the particles. That these apparently impossible motions are, in fact, compatible with physical laws leads to some insights about dynamics and gravitation.

At the outset it must be made clear that the "mass" of a body has several different roles in mechanics: (i) The "inertial mass" of a body refers to its resistance to acceleration; it is the m in $\mathbf{F}=m\mathbf{a}$. (ii) The "gravitational," or "passive gravitational mass," governs how strongly gravity pulls on the body; it is the m in $\mathbf{F}=-m\nabla\Phi$, where Φ is the gravitational potential. (iii) There is also "active gravitational mass," which determines the strength of the gravitational field generated by a body.[1] In standard physical theory all three "masses" are identical.

The equivalence of inertial and passive gravitational mass, in particular, has been verified to high accuracy, and is known by the name "the principle of equivalence." It is this principle that allows us to cancel the masses in Newton's second law, and in the force law for gravity, so that the gravitational acceleration is given by $\mathbf{a}=-\nabla\Phi$. Since no reference to the mass appears, this equation predicts that (in a vacuum) feathers and rocks fall in the same way. We will assume that the equivalence principle holds, so that our negative mass particles have both negative inertial mass, and negative passive gravitational mass.

The first situation to be considered is the fall of a negative mass particle when it is released from rest. The particle, like rocks and feathers, must of course fall downward.[2] For the negative mass particle, of course, the gravitational force on the particle is upward. The particle accelerates (downward) in the direction opposite to the force acting on it due to the particle's negative inertial mass. More interesting than the free-fall of the negative mass particle is how it must be constrained, e.g., before it is released, to prevent it from falling. Since the gravitational force on it is upward, the support force to prevent the particle from falling must be downward. For example, the particle could be tethered by a string and the other end of the string could be pulled downward. This would give us a child-with-a-balloon configuration, with the negative mass particle suspended above the child, who feels an upward pull on the string. There is, of course, a crucial difference between the negative mass particle and an actual balloon. If the string breaks, a balloon—to the child's chagrin—would accelerate upward, whereas the negative mass particle would fall downward.

This becomes even stranger if we replace the child as the agent of downward force by a positive mass particle. Suppose that we have a negative mass particle of -1 kg at the top of the string in the earth's uniform field. Let us put a $+1$-kg mass at the bottom of the string. The $+1$-kg mass pulls downward on the string with a force of 9.81 N, the -1-kg pulls upward on the string with the same force. The string remains under a tension of 9.81 N and the whole configuration—positive mass, negative mass, and string—remains fixed in position falling neither downward or upward. We have created an "antigravity glider." This is acceptable, if not sensible, since the total mass of the configuration is zero.[3] But the consequence of cutting the string may be less acceptable. If the string is cut both particles fall!

We can further offend our intuition with the antigravity glider. Suppose we release from rest the ± 1-kg mass configuration with 10-N tension in the string. At the bottom of the string the 10-N upward tension force will not quite cancel the downward 9.81-N gravitational force. There will be a net upward force of 0.19 N which will result in an upward acceleration of the $+1$-kg mass by 0.19 m s^{-2}. At the top of the string, the 10-N downward tension and the upward 9.81-N gravitational force leave a net 0.19-N

downward force which, acting on a -1-kg mass, results in a 19 m s^{-2} *upward* acceleration. The entire configuration will therefore initially accelerate upward. Since the bottom and the top particles will maintain their separation, the tension in the string will remain 10 N, and the configuration will continue to accelerate upward at the same rate. The rate of acceleration, of course, can be adjusted, or its direction reversed, by adjustments to the tension in the string. Despite initial appearances, the "free acceleration" would not violate physical law. Since the total (inertial) mass of the system is zero, the system can accelerate without the action of any external force. And there is no problem with energy conservation. As the system accelerates upward the increase of gravitational potential energy of the bottom (positive mass) particle is offset by the decrease in gravitational potential of the top (negative mass) particle. Since the particles always have the same velocity the positive kinetic energy of the bottom mass and the negative kinetic energy of the top mass will always add to zero.

The potential of this configuration seem limitless. But harnessing its advantages would require solving technological problems other than simply finding negative masses. One such problem emerges when we consider moving a negative mass particle from place to place. If we have the particle tethered to an elastic string and we tug on the string, the particle responds to the tug by perversely accelerating in the opposite direction, the direction away from the tug. This motion of the particle will stretch the string increasing the tugging force, and increasing the acceleration of the particle in the "wrong" direction. That there should be this awkward instability is inherent in the mathematics of simple harmonic motion. If we characterize the elasticity of the string by spring constant k, the natural frequency for the motions of a particle of mass $-m$, where m is positive, would be $\sqrt{k/(-m)} = \pm i\sqrt{k/m}$. Velocities and accelerations of the particle would in general grow with time t as $\exp \sqrt{k/m}\, t$. In a realistic situation (although the meaning of "realistic" is not quite clear in this context) the string would quickly reach its failure strength and break, and the particle would fly off at high speed.

Clearly, a simple tug on a string is an unsuccessful strategy for bringing a negative mass particle nearer. Pushing the particle away with, say, a pole *would* get the particle going in the right direction, but quick reflexes (that is, monitoring and feedback loops) would be needed to get the pole out of the way so that the particle's acceleration did not grow exponentially in time. Great vigilance would be needed, in fact, at all times with negative mass particles. Ordinary containers which ultimately rely on elastic forces could not easily be used. If the negative mass particle ever impinged on the elastic walls it would, in short order, burst through. Any equilibrium due to elastic forces would be unstable. Convenient confinement (without the action of monitoring and feedback loops) would require a position of stable equilibrium in a force field. One such position for a negative mass particle would be at a local minimum of the gravitational potential since, by the equivalence principle, the particle's motions due to gravity are the same as for a positive-mass particle.

The support of NSF Grant PHY92-07225 is gratefully acknowledged.

[1]Here, we are primarily concerned with inertial and passive gravitational mass, but further amusement can be had from a particle which also has negative "active" gravitational mass. Such a particle will be gravitationally attracted to, but will gravitationally repel, a positive mass particle, and will chase it forever. See Alan P. Lightman, William H. Press, Richard H. Price, and Saul A. Teukolsky, *Problem Book in Relativity and Gravitation* (Princeton U. P., Princeton, 1975), Prob. 13.20.

[2]The equivalence principle has rather the status of a coincidence in Newtonian gravitation theory, but is crucial and central to metric theories of gravitation like general relativity. The fact that a negative mass particle would fall downward is thus required in general relativity. In this connection it is worth mentioning that all the gravitational phenomena described in this article can be described equally well in general relativity. The apparent paradoxes are in no way caused by the shortcomings of Newtonian gravitational theory.

[3]There is nothing fundamentally physically wrong with a weightless string. Such strings can be described in general relativity and are, in quite different contexts, of some interest as "cosmic strings." It is well known that cosmic strings do not induce curvature in the surrounding space-time. [See, for example, T. M. Helliwell and D. A. Konkowski, "Cosmic strings: Gravitation without local curvature," Am. J. Phys. 55, 401–407 (1987).] The string therefore has no gravitational pull, so it should not be surprising (and it is straightforward to verify) that it has no weight.

THE JOY OF INSIGHT

Indeed, the joy of insight is something very important. I myself must say, if I look back at my life as a scientist and a teacher, I think the most important and beautiful moments were when I say, "ah-hah, now I see it a little better," and it is not necessarily when I myself have done something. When I hear a seminar or when I hear a good speaker, then I say, "ah, now I see," this is this joy of insight which pays for all the trouble one has had in this career.

Victor F. Weisskopf, in *Quarks, Quasars, and Quandaries*, edited by Gordon J. Aubrecht (AAPT, College Park, Maryland, 1987), p. 13.

Paradoxical twins and their special relatives

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

Ronald P. Gruber
3318 Elm Street, Oakland, California 94609

(Received 7 August 1995; accepted 2 February 1996)

We present a variation and extension of the twin paradox recently put forth by Boughn. These additional considerations make a strong case that there is no meaning to the question of "where" the aging difference occurs in the twin paradox. © *1996 American Association of Physics Teachers.*

I. INTRODUCTION

A recent article by Boughn,[1] in this journal, presented a relativity parable with great pedagogical power. It is often the case that an important work inspires inferior sequels. In that tradition, we would like to present here additional considerations based on Boughn's twin paradox. Our main point here will be to make the case that the question "where does the asymmetry in aging *really* occur?" is meaningless.

The Boughn paradox involves twins Dick and Jane who start from rest at home (where Mom and Dad remain). They enter their resting rockets which are separated by a distance L along the x axis, and follow identical instructions as they accelerate in the positive x direction. After a while they exhaust rocket fuel, so that they are coasting (not accelerating). Since they have followed identical instructions they must, when they are both coasting, be moving at the same speed relative to Mom and Dad. They therefore have no velocity with respect to each other; they are in the same inertial frame. From the symmetry of the motion, as observed by Mom and Dad, the twins must be the same age when they are coasting. But the twins are now in an inertial frame, call it S', which is moving, say at velocity v, with respect to the frame of Mom and Dad. What is simultaneous to Mom and Dad is not simultaneous in frame S'. As Boughn points out, the twins' birthday—which is simultaneous to Mom and Dad—is observed in frame S' to occur with a time difference $\Delta t' = \gamma v L \equiv vL/\sqrt{1-v^2}$ (where we are using units in which $c=1$). If Jane's position is at larger x then it is she who will be older. The relativity of simultaneity is evident in Fig. 1, in which the worldlines of Dick and Jane are illustrated in the frame of Mom and Dad. Events 1 and $\bar{2}$ are events, like the twins' birthday, which are simultaneous to Mom and Dad. Events 1 and 2 are events which are simultaneous to the twins. It is clear that in S', the final frame of the twins, Jane's birthday is earlier than Dick's.

All this is a straightforward application of the formalism of special relativity. What is not straightforward, and what often confuses students, is the meaning of the result. First, what does the "age" or "aging" of one of the twins mean? The answer, of course, is that each twin can be considered to be a clock. Biological aging is no different in principle from the ticking of the counter in an atomic clock. The age of Jane at event 2, then, can equally well be considered to be the age inferred from her biological aging, or the reading of the atomic clock that was strapped to her wrist and started from zero at the moment Jane was born. It is a well-defined unambiguous, observer-independent value. When we say that the age of Jane at event 2 is greater than the age of Dick at event 1, we are comparing measurements of time about which there is no disagreement.

The issue that is much more slippery is the inference from the ages at events 1 and 2 that Jane has somehow gotten older than Dick. The skeptical student can argue that the age difference between events 1 and 2 is a purely formal, abstract, idea with no physical immediacy. Since events 1 and 2 are in different locations, no direct comparison of the ages of Dick and Jane is possible. For this student it is important to point out that Dick and Jane can simply *visit* one another. Dick can walk over to Jane's position; Jane can do the walking; they can meet somewhere in the middle. It does not matter *as long as they walk slowly*. By comparing their atomic clocks (or biological ages) they will find that indeed Jane in every sense of the word really *is* older by $\gamma v L$. It follows that the age difference has absolute meaning. After Dick and Jane have (*slowly!*) reunited and are at the same location, any observer will see Jane as older by $\gamma v L$.

This point will be important to the argument below, so we give here a demonstration that certain relativistic time differences can be arbitrarily small. To separate this demonstration from the context of Fig. 1, we will consider motion in a frame with coordinates X,T, as depicted in Fig. 2. Two observers, $\mathcal{O}_1$ and $\mathcal{O}_2$, initially are separated by D. For definiteness we will let the observers each move at speed v_s, one moving to the right, one to the left. The time T for them to meet (the time measured by an observer stationary in the frame) is $D/2v_s$. The "proper time" measured by the observers themselves is $D/2v_s\gamma_s = \sqrt{1-v_s^2}\ D/2v_s$. The difference between this proper time and the time measured by the stationary observers is

$$\frac{D}{2v_s}(1-\sqrt{1-v_s^2}). \tag{1}$$

If v_s is much less than the speed of light, this can be approximated as

$$\frac{D}{2v_s}(1-\sqrt{1-v_s^2}) \approx \frac{Dv_s}{4c^2}, \tag{2}$$

where, in the last expression, we have explicitly exhibited the factors of c. For any D, this time difference can be made arbitrarily small, by choosing v_s/c sufficiently small. The time duration observed by moving and stationary observers can therefore be in arbitrarily good agreement. This argument is easily extended to nonsymmetric cases, in which the two observers move through different distances. As long as

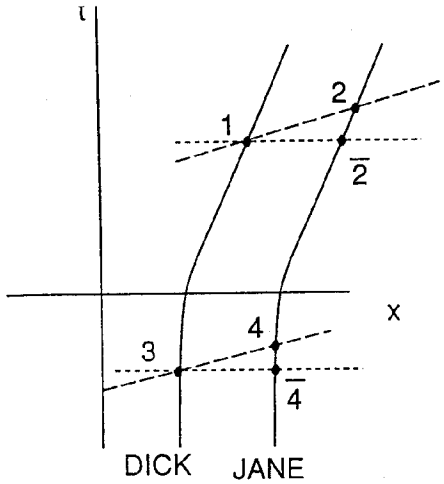

Fig. 1. Worldlines of Dick and Jane in the coordinate system of Mom and Dad. Events 1 and 2 are simultaneous to Dick and Jane; events 1 and $\bar{2}$ are simultaneous to Mom and Dad.

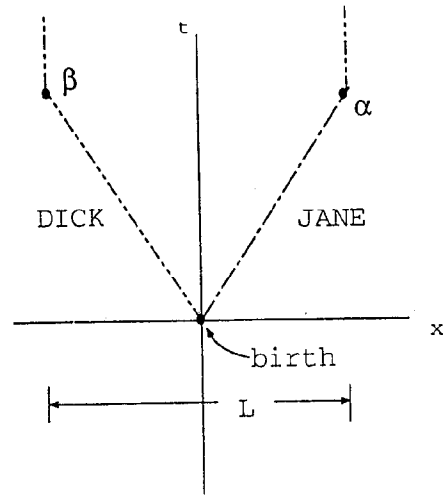

Fig. 3. Dick and Jane, the early years. The twins are born at the same point, then move apart slowly to a separation L.

the observer speeds are arbitrarily slow, all time differences will be arbitrarily small.

Without inducing any new differential aging, Dick and Jane can indeed simply "stroll" to some common meeting place and discover that Jane is now older than Dick. Slow motion also can answer student questions about early family history. Dick and Jane started out spatially very close at birth. How did they get separated by distance L without destroying the simultaneity of their aging? A simple answer is that they slowly drifted apart.

The stage is now set for our extension of the Boughn fable.

II. THE PARABLE OF THE UNCLE

Imagine that a family disagreement with Mom and Dad led to an uncle setting out, before the birth of the twins, and—after some acceleration—coming to rest in a reference frame that happens to be S', the frame in which the twins will eventually come to rest. There is a disturbing paradox in how the uncle will view the voyage of the twins. The uncle *will* see the twins born at the same place, then drift very slowly apart, and then begin their rocket voyage. He will not see them ignite their rocket engines simultaneously (their actions are simultaneous in the frame of Mom and Dad, not that of the uncle) so it is not surprising that the uncle will see the twins, when they finally come to rest in his frame, to have different ages.

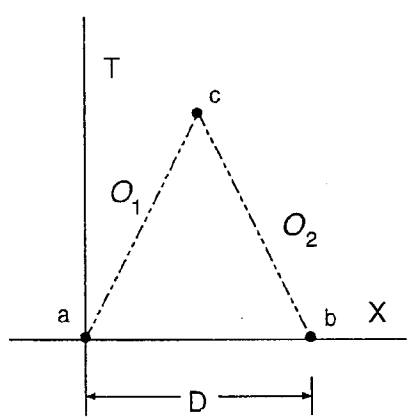

Fig. 2. Worldlines of two observers $\mathcal{O}_1$ and $\mathcal{O}_2$, who meet at event c to compare clock readings.

But there is something quite strange here. To see this most clearly, we can consider the view of things in the frame of the uncle before either of the twins started off in a rocket. In Fig. 1, events 3 and $\bar{4}$, are simultaneous to Mom and Dad and hence cannot be simultaneous in the frame of the uncle. (For the uncle, events 3 and 4 are simultaneous.) If events 3 and $\bar{4}$ represent the twins' birthdays (an event that is simultaneous to the twins and to Mom and Dad) then in the frame of the uncle, Jane's birthday (event $\bar{4}$) came at an earlier time; Jane must be older. With a simple application of the Lorentz transformations we can show that in the frame of the uncle the time difference $\Delta t'$ between events 3 and $\bar{4}$ is $\Delta t' = \gamma v L$. Although this is an indication of an age difference between Dick and Jane, it is not the best measure of that age difference. We carefully defined the "age" of Jane, at any point on her world line, as the time she herself measures since birth. According to this definition the uncle-frame difference between the ages of Dick and Jane should be Jane's age at event 4 minus Dick's age at event 3. It is straightforward to show that this age difference is vL.

By either measure the actuaries in the uncle frame would conclude that aging has taken place *before* the twins enter their rockets. This leads to a disturbing question: How could Dick and Jane possibly have gotten "out of synchronization" to the uncle? He observed them to be born at the same place at the same time, and in subsequently separating they need not have made any fast motions.

The resolution is somewhat surprising: Dick and Jane lose simultaneity during the *arbitrarily slow* motion of their separation! To see this we look back in the family album to find Fig. 3, a depiction, in the Mom and Dad frame, of the early motions of the twins. The early, slow, motions from the ori-

gin (the maternity ward) to events α and β separated the twins by L; each twin was moving at speed v_s which we shall, in a moment, assume to be very small. The twins subsequent stationary "motion" is represented by the vertical segments, which—far in the future—will lead them to rocket ships.

The times t'_α and t'_β, of events α and β, as observed by the uncle, are easily found from the Lorentz transformations, and from the known coordinates of the events in the Mom and Dad frame:

$$t'_\alpha = \gamma[t_\alpha - v x_\alpha] = \gamma[(L/2v_s) - vL/2]. \tag{3}$$

For event β the x coordinate is $x_\beta = -L/2$ so the t' coordinate for the event is

$$t'_\beta = \gamma[(L/2v_s) + vL/2]. \tag{4}$$

The time difference, as measured by the uncle, will therefore be

$$\Delta t' \equiv t'_\beta - t'_\alpha = \gamma v L. \tag{5}$$

This result is independent of v_s; the time difference observed by the uncle remains, no matter how slowly Dick and Jane move. The result is, of course, the result we have already noted as the uncle-measured time difference between the birthdays of Dick and Jane.

We have given different answers to the question "where does the differential aging occur?": (i) It all occurs during the twins' rocket trip. (ii) Some occurs in the twins early (prerocket) years. The lack of a unique answer shows the lack of meaning of the question. Age differences in relativity have a well-defined meaning, but the origin of age differences cannot be assigned to any specific part of a worldline.

[1]S. P. Boughn, "The case of the identically accelerated twins," Am. J. Phys. **57**, 791–793 (1989). Other references to the twin paradox are listed in this paper.

[2]Here and elsewhere we use words like "see" and "view" somewhat inappropriately. We do not mean that the uncle makes observations of distant events by collecting light signals from those events. Rather, we mean that observations are made according to the usual procedure in relativity. Here, that would mean that an observer who is part of the uncle's special relativity reference system, but is located *right at* an event being observed (such as the igniting of a rocket engine), notes the position and time of that event.

Shielding of an oscillating electric field by a hollow conductor

J. M. Aguirregabiria, A. Hernández, and M. Rivas
Fisika Teorikoa, Euskal Herriko Unibertsitatea, P.K. 644, 48080 Bilbao, Spain

(Received 14 December 1995; accepted 4 March 1996)

The electric and magnetic fields for a hollow conducting sphere located in a slowly varying uniform electric field background are computed to first-order in a power series expansion in the field frequency. These results are used to define an equivalent RC circuit and to test the circuit approach which is often used in electromagnetic compatibility (EMC). The case of an infinite cylindrical conducting tube under the influence of the same external field is also analyzed. © *1996 American Association of Physics Teachers.*

I. INTRODUCTION

The knowledge of the penetration of the electric and magnetic fields in electronic equipment is important to properly protect these ever increasingly sensitive devices from external influences. In fact, the shielding of a receptor set from a source of electrical disturbance is an interesting subject of research in electromagnetic compatibility (EMC), which is defined by IEEE as "the ability of a device, equipment or system to function satisfactorily in its electromagnetic environment without introducing intolerable electromagnetic disturbances to anything in that environment."[1]

The physicist's approach to evaluating the electromagnetic shielding is based upon the solution of Maxwell's equations with appropriate boundary conditions on the shielding surfaces, but the mathematical machinery is so complex that, even when the calculations can be carried out, the physical insight is often missed.[2–4] As a consequence, from an engineering point of view, to estimate in practice the electromagnetic field inside the shielding enclosure, it is always necessary to use a simplified theory of electromagnetic shielding.

Among the techniques developed so far in EMC to deal with this kind of calculation we will consider here the so-called "circuit approach" in which the actual physical system is replaced by an equivalent RC circuit. This approach is based upon the fact that the external electromagnetic field will induce on the shielding enclosure a charge distribution which will vary in time because the external field is oscillating. This will produce a current flow in the conductor and it seems rather natural to substitute for the conductor an equivalent electric circuit whose characteristics are defined on heuristic grounds because in general they cannot be computed accurately, not even by numerical simulation.[4]

The main goal of this paper is to analyze a couple of simple but interesting examples in which explicit (although approximate) expressions for these phenomena may be easily computed. In this way we can illustrate and compare the

states for two particles," Phys. Rev. Lett. **72**, 1951 (1994).
[7]H. P. Stapp, *Mind, Matter and Quantum Mechanics* (Springer-Verlag, New York, 1993), pp. 5–9.
[8]T. F. Jordan, "Testing Einstein–Podolsky–Rosen assumptions without inequalities with two photons or particles with spin-$\frac{1}{2}$," Phys. Rev. A **50**, 62–66 (1994).
[9]T. F. Jordan, "Quantum mysteries explored," Am. J. Phys. **62**, 874–880 (1994).
[10]N. D. Mermin, "Quantum mysteries refined," Am. J. Phys. **62**, 880–887 (1994).
[11]U. Fano, "Pairs of two-level systems," Rev. Mod. Phys. **55**, 855–874 (1983).
[12]See, for example, E. Merzbacher, *Quantum Mechanics* (Wiley, New York, 1970), p. 279.
[13]We use the Einstein summation convention for repeated indices throughout the text.
[14]The idea of using the relation $\rho^2=\rho$ to simplify the description of a pure state has been exploited in another context by J. Bohn, "Observable characteristics of pure quantum states," Phys. Rev. Lett. **66**, 1547–1550 (1991).
[15]The precise correspondence is as follows:

$$|\Psi\rangle=1/\sqrt{2}[(|\beta_1\rangle|\beta_2\rangle+|\alpha_1\rangle|\alpha_2\rangle)(C_{zz'}=1,C_{xx'}=1,C_{yy'}=-1);$$

$$|\Psi\rangle=1/\sqrt{2}[(|\beta_1\rangle|\beta_2\rangle-|\alpha_1\rangle|\alpha_2\rangle)(C_{zz'}=1,C_{xx'}=-1,C_{yy'}=1);$$

$$|\Psi\rangle=1/\sqrt{2}[(|\beta_1\rangle|\alpha_2\rangle+|\alpha_1\rangle|\beta_2\rangle)(C_{zz'}=-1,C_{xx'}=1,C_{yy'}=1);$$

$$|\Psi\rangle=1/\sqrt{2}[(|\beta_1\rangle|\alpha_2\rangle-|\alpha_1\rangle|\beta_2\rangle)](C_{zz'}=-1,C_{xx'}=-1,C_{yy'}=-1).$$

The above states, which form an orthonormal and complete set, are often referred to as the "Bell basis;" see, for example, S. L. Braunstein, A. Mann, and M. Revzen, "Maximal violation of Bell inequalities for mixed states," Phys. Rev. Lett. **68**, 3259–3261 (1994).
[16]A. Einstein, B. Podolsky, and N. Rosen, "Can quantum mechanical description of physical reality be considered complete?," Phys. Rev. **47**, 777–780 (1935).
[17]Well, almost any angles. The choice $\theta_a=0$ must be excluded because it does not lead to a distinct set of directions for **a**, **a**′, **b**, and **b**′.
[18]We can even identify a suitable Hardy state if all four directions are specified arbitrarily, as noted by Jordan in Ref. 8. To do this we take $\phi_a=\phi_a'=0$ and $\phi_b=\phi_b'=\pi$ and are left with the task of fixing the four polar angles and Ω. Since $\theta_a-\theta_a'$ and $\theta_b-\theta_b'$ are now fixed we have just three unknowns, which we may take to be θ_a, θ_b, and Ω. The three equations (24)–(26) suffice to determine these unknowns, but the equations are complicated enough that they can be solved only numerically.
[19]J. Torgerson, D. Branning, and L. Mandel, "A method for demonstrating violation of local realism with a two-photon downconverter without use of Bell inequalities," Appl. Phys. B **60**, 267–269 (1995).

The conical resistor conundrum: A potential solution

Joseph D. Romano and Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 23 June 1995; accepted 14 December 1995)

A truncated cone, made of material of uniform resistivity, is given in many introductory physics texts as a nontrivial problem in the computation of resistance. The intended method and answer are incorrect and the problem cannot be solved by elementary means. In this paper, we (i) discuss the physics of current flow in a nonconstant cross-section conductor, (ii) examine the flaws in the "standard" solution for the truncated cone, (iii) present a computed resistance found from a numerically generated solution for the electrical potential in the truncated cone, and (iv) consider whether any problem exists to which the standard solution applies. © *1996 American Association of Physics Teachers.*

In introductory courses students are taught that a solid of electrical resistivity ρ, length L, and constant cross-sectional area A, has a resistance

$$R=\rho L/A. \tag{1}$$

Many (most?) texts[1] for calculus-based courses try to introduce an application of calculus by including a problem in which a resistor has a nonconstant cross section. In these books (usually in Chap. 26±2) the student is asked to find the resistance of a truncated cone, as in Fig. 1. The reader is meant to assume that the resistor is connected via "wires" that end with perfectly conducting disks attached to the end faces of the truncated cone.

The method that is intended can be inferred from the answer. (In the text by Wolfson and Pasachoff[1] the student is explicitly instructed to use this method.) The student is meant to break the cone into isolated differential slabs of resistance (each could be called a pièce de résistance). The slab at location x would have differential thickness dx, and would have a resistance $dR=\rho\, dx/(\pi r^2)$, where r is the radius of the cone's cross section at x. Since $dx/dr=L/(b-a)$ we have $dR=\rho[L/(b-a)](dr/\pi r^2)$. The cone is then

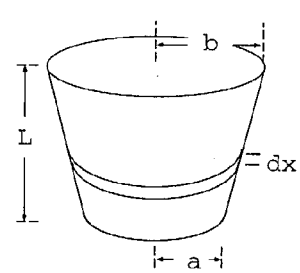

Fig. 1. The standard truncated cone problem.

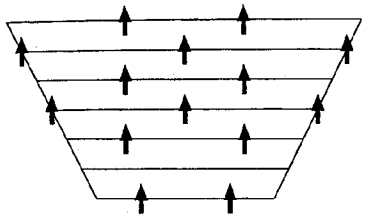

Fig. 2. Planar equipotentials and the implied electric field.

treated as if it is nothing more than a stack of these slabs, and the total resistance is found by summing the resistance of all the slabs:

$$R=\int dR=\rho\,\frac{L}{b-a}\int_a^b \frac{dr}{\pi r^2}=\frac{\rho L}{\pi a b}. \tag{2}$$

This seems plausible at first—the role of A in (1) is played in (2) by the geometric mean of the area of the disks bounding the cone—but the plausibility does not stand up to a close second look. When the slabs are added "in series," it is required that their planar faces be equipotentials. The stack of slabs is electrically equivalent to the cone only if the equipotentials in the current-carrying cone consist of planes perpendicular to the axis, as is shown in Fig. 2. This is impossible. If these *were* the equipotentials then the electrical field (orthogonal to the equipotentials) would be parallel to the axis, as shown in the figure. The current (parallel to the electric field in an Ohmic conductor) would then also be parallel to the cone axis, and therefore not parallel to the sides of the resistor. This would imply that current is flowing in through the sides! The actual equipotentials, therefore, must be curved in order to be perpendicular to the sides of the resistor. A few of the texts suggest that some sort of approximation is being made (about which more below); only Wolfson and Pasachoff[1] point out that the assumption of the planar equipotentials is being made, and that the assumption is wrong.

To discuss what is right, let us first note that there is no charge density in the resistor. If there were charge density, $\nabla\cdot\mathbf{E}$ would be nonzero and, through Ohm's law, this would imply that the divergence of the current flow $\nabla\cdot\mathbf{J}$ is nonzero, and that charge is building up. The mathematical problem of finding the field inside the resistor therefore amounts to finding the solution of Laplace's equation $\nabla^2\Phi=0$ for the electrostatic potential Φ. This equation is to be solved with the following boundary conditions: (1) Φ must be constant, say at values Φ_a and Φ_b, on the disks at the end of the resistor; (2) the gradient $\nabla\Phi$ must be parallel to the resistor sides. Once the solution is found, the current flow is known to be $\mathbf{J}=-\nabla\Phi/\rho$, and integrating the normal component of $\mathbf{J}$ over the area of either of the bounding disks gives the total current I. The resistance is then $|(\Phi_a-\Phi_b)/I|$.

Casting this as a numerical problem is reasonably straightforward, and the solution of that problem, while not trivial, is not daunting with modern computational tools. Our approach has been to formulate the problem in a coordinate system convenient for the boundary conditions, and in which numerical errors are expected to be minimized. The equations based on this coordinate system were solved using the public-domain package *Sparse*.[2] For those who are interested in this sort of thing, details are provided in the Appendix. Here, we want to emphasize the results.

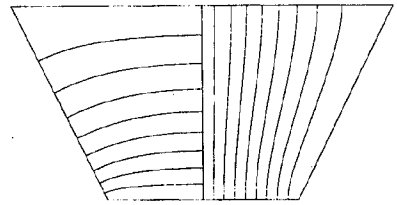

Fig. 3. Equipotentials and flow lines for computed resistance.

We present results first for a resistor with $a/L=1/2$, $b/L=1$. A picture of the current flow lines and the equipotentials are shown in Fig. 3. (Half the lateral cross section is shown for each.) All expected features are evident in these figures. In particular, current does not flow through the resistor's sides and the equipotentials are perpendicular to the sides. At the corners the $\mathbf{E}$ field is required to be perpendicular to the end faces and to be parallel to the sides. To satisfy these incompatible constraints, $\mathbf{E}$ vanishes.

How wrong is the "stack of slabs" solution? Intuition suggests (correctly) that the correct solution will correspond to higher resistance since the correct flows are more constrained than those of the slab solution. (The corner regions, for example, must have reduced current flow.) From the numerical results we find, for our reference case ($a/L=1/2$, $b/L=1$) in Fig. 3, that the resistance is $R=0.692\rho/L$, larger by 9% than the textbook answer given in (2).

Several of the textbooks[3] in the list in Ref. 1 tell the student to assume that the current is uniformly spread over the cross section. We can see no strong case for this assumption being ineluctably linked with the intended method. The current could, for example, be a uniform flow on straight lines that can be traced back to the apex of the cone, as in Fig. 4. This would give a flow pattern that satisfied the correct boundary conditions on the sides of the resistor, but not on the disks that truncate it. This flow pattern would be precisely correct if the resistor were formed by the intersection of a cone with a spherical shell. The resistor would then be truncated not by disks, but by the spherical caps shown dashed in the figure. This replacement might in principle (though not plausibly) be taken by the student to be the natural interpretation of the instructions to assume uniform flow. This same implausibly perverse student would compute a resistance of

$$R=\frac{\rho}{2\pi a b}\,\frac{(b-a)^2}{\sqrt{L^2+(b-a)^2}-L}. \tag{3}$$

This calculation is a straightforward application of geometry and does not involve calculus. It gives, furthermore, an answer $R=0.674\rho/L$ for our reference problem (that of Fig. 3)

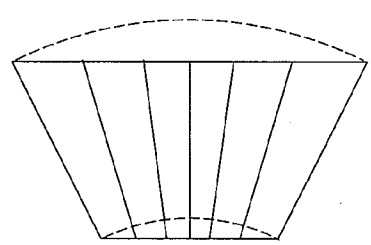

Fig. 4. Another type of "uniform" flow.

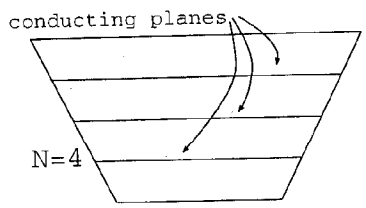

Fig. 5. Resistor segmented by conducting planes.

which is less than 3% off from the correct (numerical) answer, while the expected textbook method is off by almost 9%.

One can make other approximations using a portion of a spherical shell. In particular one could use a spherical shell which is just barely completely enclosed within the truncated cone, or one could use a shell which just barely encloses the truncated cone. The former approximation, which provides a lower bound on the resistance, gives $R=0.595\rho/L$ for our reference case, while the latter approximation, which provides an upper bound, gives $R=0.833\rho/L$.

The textbook by Keller, Gettys, and Skove[1] suggests, in its presentation of the problem, that the answer expected from the student will apply only if the resistor does not taper very much, i.e., for $(b-a)\ll L$. Intuition suggests (and numerical models confirm) that the textbook formula becomes a good approximation in this limit. This does provide some justification for the stack of slabs approach as an easy way of deriving an approximation, but it should not be taken seriously as unique or efficient. An example of an alternative approach which is equally justified is that of (3). In the limit $(b-a)\ll L$ this becomes $\rho L/\pi ab$ and, as we have seen, gives a better approximation, at least for some range of parameters.

We have asked ourselves what might be the problem for which the textbook solution is the correct answer. The result of this inverse problem solving is to imagine thin conducting planes used to break the resistor into N segments, as in Fig. 5. (These segments are different from the slabs of the textbook method. For one thing, the segments are tapered; the slabs are not.) We can compute the resistance of a segmented cone by using our numerical program to compute the resistance of each segment. Table I shows the results of segmentation for a resistor with the geometry ($a/L=1/2$, $b/L=1$) of that in Fig. 3. As the number of segments increases from $N=1$ (the unsegmented resistor) to $N=8$, the excess of the computed resistance over the textbook answer decreases. In the limit of an infinite number of segments, the textbook answer would be reached.

The transverse conducting planes do not give us a very satisfying physical problem, but they do point us in the direction of an interesting physical problem: a truncated conical resistor made of material with *anisotropic* resistivity. The resistivity parallel to the cone axis should be ρ, while the resistivity in the transverse directions vanishes. For such a resistor the method and the answer of the textbooks is correct. This, of course, is a very inappropriate problem for an introductory course, but the same, or worse, must be said for the problem as it presently appears in textbooks.

Table I. Resistance of a cone with N segments.

N	Computed R/Textbook R
1	1.087
2	1.064
4	1.040
8	1.022

ACKNOWLEDGMENTS

We wish to thank Christopher Johnson for directing us to an appropriate package for numerically solving our equations and Carlton DeTar for helping with its use. We thank anonymous referees for useful suggestions, in particular the idea of upper and lower bounds on the resistance using spherical shell approximations. This work was partially supported by the National Science Foundation under Grant No. PHY9207225.

APPENDIX

We choose to have the apex defining the conical resistor to be at the origin of a system of cylindrical oordinates r,ϕ,z, in terms of which we define new coordinates η, ζ by

$$\eta\equiv\frac{1}{2}\frac{r^2}{z^2}\quad \zeta\equiv\ln\left(\frac{z}{L}\right). \tag{4}$$

The surfaces of constant η are conical surfaces with the same apex (like the flow lines shown in Fig. 4). The squared form of r/z in (4) was chosen to eliminate the usual r^{-1} factor that enters the axi-symmetric (ϕ-independent) Laplace equation. The logarithmic form of z in (4) was chosen to eliminate the z-dependent factors that arise from the definition of η.

The boundaries of the resistor are constant coordinate surfaces. The equipotential disks truncating the resistor are at $\zeta=\zeta_a\equiv\ln[a/(b-a)]$ and $\zeta=\zeta_b\equiv\ln[b/(b-a)]$. The sides of the resistor are at a $\eta=\eta_0\equiv\frac{1}{2}[(b-a)/L]^2$. In terms of these coordinates the axisymmetric Laplace equation takes the form

$$\frac{\partial^2\Phi}{\partial\zeta^2}-\frac{\partial\Phi}{\partial\zeta}+2\eta(1+2\eta)\frac{\partial^2\Phi}{\partial\eta^2}+(2+6\eta)\frac{\partial\Phi}{\partial\eta}-4\eta\frac{\partial^2\Phi}{\partial\zeta\,\partial\eta}=0. \tag{5}$$

The boundary conditions on the disks are straightforward: At ζ_a and ζ_b the potentials are taken to be $\Phi=0$ and $\Phi=1$. At the resistor sides, $\eta=\eta_0$, the condition is that $\nabla\Phi$ is parallel to the surface $\eta=\eta_0$. Since the η, ζ coordinates are not orthogonal, some care must be used in computing the gradient of Φ and in evaluating the condition of parallelism at the side. The result is that at $\eta=\eta_0$ we must have

$$(1+2\eta)\frac{\partial\Phi}{\partial\eta}-\frac{\partial\Phi}{\partial\zeta}=0. \tag{6}$$

Although the axis is not a physical boundary, $\eta=0$ is a boundary of our coordinate region. The appropriate condition at the axis is simply the Laplace equation (5) with $\eta=0$.

The range of η and ζ was discretized into an $n\times n$ square grid, and the Laplace equation, and boundary conditions at $\zeta=\zeta_a,\zeta_b$ and $\eta=0$, η_0, were written as difference equations on this grid. These difference equations form an $n\times(n-2)$ set of linear equations in which the unknowns are the values of Φ at the grid points. This set of equations was solved numerically with the *Sparse* package[2] which solves explicitly

(rather than iteratively). From the computed solution for Φ near the bounding disks, the current was computed as described in the text, and a value for the resistance was found.

To estimate and improve the accuracy of the result for the computed resistance we used a version of Richardson extrapolation. The resistance computed for different grid size n, with n ranging from 20 to 100, was found to follow the pattern $R(n)=A+B/n$. The resistor value $R(\infty)$ was taken to be A. The uncertainty in A depended on the scatter in the values of B found for different values of n (due to round-off error, $1/n^2$ truncation terms, etc.). From this we can say with some confidence that the computed resistances were accurate to around 0.1%. This uncertainty is much smaller than the differences (several percent) between our numerically computed resistances and the textbook values given by (2).

[1]We do not claim to have made an exhaustive search of the texts. Of 13 calculus-based introductory texts we looked into, the truncated cone problem was found in the following nine: W. P. Crummett and A. B. Western, *University Physics* (Brown, Dubuque, 1994), Chap. 26, Prob. 49; P. M. Fishbane, S. Gasiorowicz, and S. T. Thornton, *Physics for Scientists and Engineers* (Prentice-Hall, Englewood Cliffs, NJ, 1993), Chap. 27, Prob. 55; D. Halliday and R. Resnick, *Fundamentals of Physics* (Wiley, New York, 1981), Chap. 28, Prob. 25; A. Hudson and R. Nelson, *University Physics* (Harcourt Brace Jovanovich, New York, 1982), Chap. 24, Prob. 24C-1; F. J. Keller, W. E. Gettys, and M. J. Skove, *Physics* (McGraw-Hill, New York, 1993), Chap. 24, Prob. 5; H. C. Ohanian, *Physics* (Norton, New York, 1985), Chap. 28, Prob. 25; R. A. Serway, *Physics for Scientists and Engineers* (Saunders, Philadelphia, 1992), 3rd ed., Chap. 27, Prob. 68; P. A. Tipler, *Physics for Scientists and Engineers* (Worth, New York, 1991), Chap. 22, Prob. 69; R. Wolfson and J. A. Pasachoff, *Physics* (Harper Collins, New York, 1995), Chap. 27, Prob. 69.

[2]K. S. Kundert and A. Sangiovanni-Vincentelli, *Sparse* (University of California, Berkeley, 1988).

[3]Of those listed in Ref. 1, texts which instruct the student to assume uniform current flow are those by: Crummett and Western; Halliday and Resnick; Hudson and Nelson; Serway. In the Keller, Gettys and Skove text, the student is told that uniform current flow follows from the assumption of a small taper angle for the cone.

Lagrangian field theories and energy-momentum tensors

Gerardo Muñoz
Department of Physics, California State University, Fresno, Fresno, California 93740-0037

(Received 17 July 1995; accepted 14 November 1995)

We discuss the implications of Poincaré invariance within the context of Lagrangian field theories. It is shown that the correct implementation of this invariance leads in a straightforward manner to a conserved energy-momentum tensor which is both symmetric and gauge invariant. © *1996 American Association of Physics Teachers.*

I. INTRODUCTION

Few things are more frustrating to students than to be led through a long, formal argument only to be told at the end that the result obtained is incorrect and must somehow be fixed by an auxiliary procedure. This is particularly harmful if the formal argument involved turns out to be one of the mathematical cornerstones of modern physics. Unless the discussion includes a re-examination of the analysis to find out exactly what went wrong, the students will be left with the paradoxical feeling that a supposedly very general theorem produces unacceptable answers when applied to certain specific situations. Quite understandably, later on they will be reluctant to think about any physical problem in terms of the tools provided by such a theorem.

The case we have in mind is the typical derivation of the energy-momentum tensor for relativistic field theories. Most textbooks begin the discussion by writing down a Lagrangian $\mathscr{L}$, and arguing that this Lagrangian must be invariant under a spacetime translation. The details of the derivation vary from text to text, but the end result invariably yields the energy-momentum tensor (our conventions are the same as Jackson's;[1] see the Appendix):

$$T^{\mu}_{\nu}=\frac{\partial\mathscr{L}}{\partial(\partial_{\mu}A_{\lambda})}\,\partial_{\nu}A_{\lambda}-\delta^{\mu}_{\nu}\mathscr{L}$$
$$=-\frac{1}{4\pi}F^{\mu\lambda}\partial_{\nu}A_{\lambda}+\frac{1}{16\pi}\,\delta^{\mu}_{\nu}F^{\rho\lambda}F_{\rho\lambda}\,, \tag{1}$$

when applied to the free electromagnetic field, $\mathscr{L}=-(1/16\pi)F^{\rho\lambda}F_{\rho\lambda}$.[1–4] It is then pointed out that this so-called *canonical* energy-momentum (or stress) tensor is unacceptable for a number of reasons. First, it is clearly not a gauge-invariant quantity. Second, it is not symmetric, thereby ruining conservation of angular momentum. Third, its components do not reproduce the standard definitions of energy density and momentum density. Fourth, it is not traceless, which contradicts the fact that our starting Lagrangian is conformally invariant (i.e., photons are massless). Some technical details of the derivations add to the confusion: In some instances, it is apparently crucial to assume a Lagrangian with an explicit dependence on the spacetime position[2]—a dependence which is clearly absent in $\mathscr{L}=-(1/16\pi)F^{\rho\lambda}F_{\rho\lambda}$—while in other cases a *local* translation $x^{\mu}\rightarrow x'^{\mu}=x^{\mu}+a^{\mu}(x)$ seems necessary,[3] even though in special relativity we deal exclusively with *global* translations.

Zero time dilation in an accelerating rocket

Ronald P. Gruber
3318 Elm Street, Oakland, California 94609

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 31 January 1997; accepted 9 April 1997)

We give a variation of the twin paradox of special relativity in which the relationship of acceleration of the rocket twin and time dilation is clarified. © *1997 American Association of Physics Teachers.*

Numerous articles have struggled with the student misconception that in the classical relativistic twin paradox, the asymmetric aging is caused by acceleration. In most introductory physics texts it is correctly pointed out that the viewpoints of the stay-at-home twin and the rocket twin are not equivalent; the acceleration of the rocket twin distinguishes her from that of the unaccelerated stay-at-home twin. There is a danger that students can infer that the acceleration is in some sense the direct cause of the aging. Certain texts are worse than others in this regard. In a popular calculus-based textbook,[1] for example, the traveler twin has aged less than the stay-at-home twin "because her bodily processes slowed down during her travels in space." Later, it is pointed out that there is an asymmetry in the motion of the twins. It is the rocket twin's "experience of forces when her spaceship turned around, with [the stay-at-home] twin not subject to such forces." To some students this would certainly be an invitation to understand that the force necessary to cause the acceleration is directly responsible for slowing the aging process.

The relationship of acceleration and differential aging is discussed at some length in the recent article by Debs and Redhead,[2] in which a history of the problem is also presented. The solution offered in that article involves doing away with a definitive meaning for simultaneity in an inertial frame. While this is quite interesting in connection with the logical structure of special relativity, it is of little value to the student encountering the twin paradox for the first time.

For students just starting their study of relativity, a more appropriate article is the recent variation on the twin paradox given by Boughn.[3] In Boughn's version, twins experience differential aging, although their history of accelerations is identical. This certainly helps to dispel the idea of any direct connection between acceleration and differential aging, but even here the acceleration is needed to cause the differential aging, and there is the danger that it could be seen as a direct cause of the differential aging. It would seem, therefore, that the best way to make the case that acceleration per se is not the root of asymmetric aging is to give an example of one without the other. Without acceleration[4] the twins cannot meet a second time. This precludes a twin paradox without acceleration. Here, we give the opposite: a simple twin paradox with acceleration, but (in a limit) no asymmetric aging.

To illustrate our point we consider a rocket undergoing periodic motion as illustrated in Fig. 1, motion that—aside from relativistic considerations—would be simple harmonic motion:

$$x=\frac{V_{\max}}{\omega}\sin\omega t, \tag{1}$$

where x and t are the coordinates of a fixed frame on earth. It is clear that the maximum speed achieved by the rocket twin, relative to the earth, is $V_{\max}$. The acceleration of the rocket, i.e., the magnitude of the rocket's 4-acceleration, has a maximum magnitude of $V_{\max}\omega$, which occurs at times $\omega t=\pm\pi/2,\pm 3\pi/2,\ldots$. These relativistic results agree perfectly with the Newtonian answers. This is no surprise; the maximum acceleration occurs when the particle has zero velocity relative to the fixed frame on earth, and the relativistic and nonrelativistic results are therefore the same.

The time ticked by clocks on the rocket, that is, the "proper time" τ of the rocket, is related to the earth time t, by $d\tau=dt\sqrt{1-(V_{\max}/c)^2\cos^2\omega t}$. A rocket trip starting and ending at the earth will take an integer number, of half cycles of the oscillatory motion, starting, say at $t=0$ and lasting until $\Delta t=n\pi/\omega$. For such a trip the proper time (i.e., the time measured by the rocket's own clocks) will be

$$\Delta\tau=n\int_0^{\pi/\omega}\sqrt{1-(V_{\max}/c)^2\cos^2\omega t}\,dt, \tag{2}$$

so that the ratio of elapsed rocket time to elapsed earth time is given by

$$\frac{\Delta\tau}{\Delta t}=\frac{2}{\pi}\int_0^{\pi/2}\sqrt{1-(V_{\max}/c)^2\cos^2\theta}\,d\theta. \tag{3}$$

This integral can be evaluated numerically [either by direct numerical integration, or by the fact that the integral is $E(V_{\max}/c)$, where E is the complete Legendre elliptic function of the second kind]. The ratio of $\Delta\tau$ to Δt is shown in

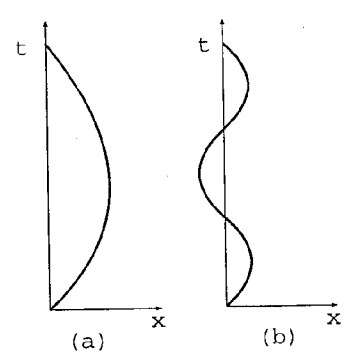

Fig. 1. Two possible rocket-twin worldlines with the same starting and stopping times and the same maximum velocity $V_{\max}$, and hence the same time dilation. The world line in (b) has maximum acceleration three times that for the world line in (a).

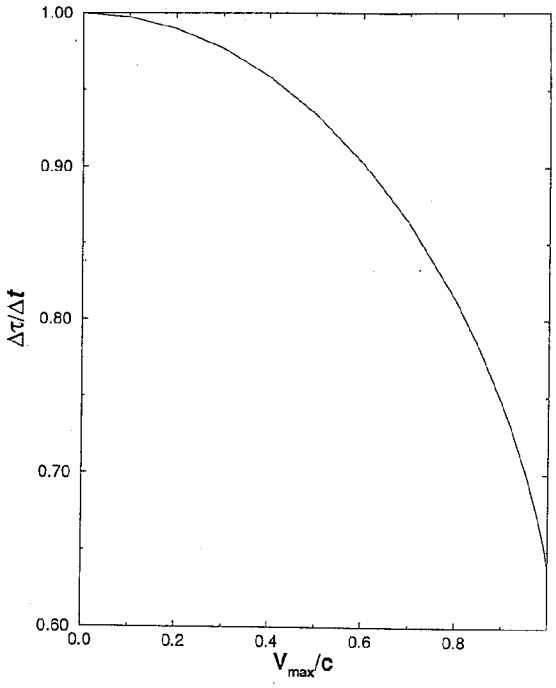

Fig. 2. The time dilation (aging of rocket twin/aging of earth twin) is shown as a function of the maximum velocity V_{max}, for the harmonic motion described in the text.

Fig. 2. It varies from unity (for $V_{max}=0$) to $2/\pi=0.6366...$ (for V_{max} approaching c). What is crucial to note is that this time-dilation effect is independent of ω! The maximum acceleration is $V_{max}\omega$ (and, from dimensional considerations, *any* measure of acceleration will be proportional to $V_{max}\omega$). It follows that time dilation and acceleration can be chosen independently.

The relationship, or lack of one, between the acceleration and the time dilation may be best seen with a numerical example. Let us suppose that for medical reasons the rocket twin chooses a maximum acceleration of g, the familiar free-fall acceleration at the earth surface. If she makes a simple one-dimensional one year trip (six months out, six months back) we calculate:

$$\omega=\pi/1\ \text{yr}=9.94\times10^{-8}\ \text{s}^{-1},\quad V_{max}/c=g/\omega c=0.329. \tag{4}$$

From Eq. (3) we find that $\Delta\tau/\Delta t=0.97$, so that 3% of the rocket twin's year, or about 11 days, will be "lost" as seen by the earthbound twin. If, on the other hand, the rocket twin went back and forth (as in Fig. 1) many times, the answer would be very different. Had she made 100 "legs" to the trip (that is, 100 half cycles rather than just one) her ω would be larger by a factor of 100, and her V_{max}/c, therefore, smaller by a factor of 100. A computation with Eq. (2) shows that in this case her time dilation would be $\Delta\tau/\Delta t=0.999\,997$. She would rejoin her earthbound sister younger than her twin only by a little more than a minute.

One of us (RHP) gratefully acknowledges support from the National Science Foundation under Grant No. PHY-9507719.

[1]R. A. Serway, *Principles of Physics* (Saunders, Philadelphia, 1994), p. 266.
[2]T. A. Debs and M. L. G. Redhead, "The 'twin paradox' and the conventionality of simultaneity," Am. J. Phys. **64**, 384–392 (1996).
[3]S. P. Boughn, "The case of the identically accelerated twins," Am. J. Phys. **57**, 791–793.
[4]This is only true in the absence of gravitational fields. In the curved space–time of general relativity, unaccelerated worldlines (space–time geodesics) can have multiple crossins.

THE ROLE OF THE EXPERIMENTAL PHYSICIST

My role has been that of an experimental physicist who, by observation and measurement of the properties and operation of the physical world, supplies the data that may lead to the formulation of conceptual structures. The consistency of the consequences of a conceptual structure with the data of physical experiment determines the validity of that structure as a description of the physical universe. Our early predecessors observed Nature as she displayed herself to them. As knowledge of the world increased, however, it was not sufficient to observe only the most apparent aspects of Nature to discover her more subtle properties; rather, it was necessary to interrogate Nature and often to compel Nature, by various devices, to yield an answer as to her functioning. It is precisely the role of the experimental physicist to arrange devices and procedures that will compel Nature to make a quantitative statement of her properties and behavior.

Polykarp Kusch, "The magnetic moment of the electron" (Nobel Lecture, December 12, 1955, reprinted in *Nobel Lectures, Physics*, Vol. 3, 1942–1962, Elsevier Amsterdam, 1964).

Aim high and go far—Optimal projectile launch angles greater than 45°

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

Joseph D. Romano
Department of Physics, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin 53201

(Received 27 February 1997; accepted 21 October 1997)

For ideal projectile motion, which starts and ends at the same height, maximum range is achieved when the firing angle is 45°. If air resistance is taken into account, the optimal angle is somewhat less than 45° and this is often considered obvious. We show here that it is not obvious by considering drag forces with different dependence on projectile speed. In some cases maximum range is achieved for launch angles greater than 45°. Simple physical arguments are given which help explain results that were found by computing trajectories and ranges.

If mechanics is food for thought, then projectile motion provides much of the starch at the beginning of introductory physics courses. One of the more interesting results served is the relationship of the range (horizontal distance traveled) of a projectile and the angle θ of inclination at its launch. For a fixed initial speed, and for negligible air resistance, the range is a function of θ that is symmetric about 45°. The range for 30°, for example, is the same as the range for 60°. The optimal angle, i.e., the angle for maximum range, is 45°.

A more difficult question is what the optimal angle is if air resistance is taken into account.[1] It is semicommon knowledge that maximum range, when the resistance of still air is taken into account, occurs for angles less than 45°.[2] But is this answer obvious? An informal survey of our colleagues revealed that the immediate reaction is "yes, it's obvious." This seems to be related to the fact that a lower trajectory reduces the time and distance of flight, therefore minimizing the time and distance over which the drag force due to the air is acting. This point, in fact, is explicitly made in one of the texts.[3] Another introductory textbook[4] clearly treats the issue as obvious, since the question is assigned to the students at the end of the chapter on two-dimensional motion. At this point in the text, Newton's laws have not even been introduced, let alone a description of air resistance. The implication is that the decrease in the optimal launching angle is a robust result, independent of the details of the resistive force. This, in fact, is precisely what seems to be proved in a recent publication in this Journal,[5] i.e., that for any force directed opposite to the velocity of the projectile, the optimal angle cannot be greater than 45°. The proof, however, suffers from a flaw in the formulation of the problem,[6] so the question remains open and interesting.

One can get a first insight into the nature of the problem by considering 30° and 60° launches, cases that have equal range in the absence of air resistance. The 60° trajectory does indeed give air drag more time and distance to act. But for the 30° trajectory the projectile spends a larger fraction of its time at speeds near the maximum speed. This suggests that the more strongly we make the resistive force increase with velocity, the more it will act to shorten the 30° trajectory relative to the 60° trajectory.

Figure 1 shows the results of a numerical computation to test this idea. The figure shows 30° and 60° trajectories for two types of resistance: (i) "standard" air drag proportional to v^2, where v is the speed of the projectile; and (ii) drag proportional to v^8. In both cases, the resistive force acts antiparallel to the velocity vector. The results show—as predicted by the above argument—that for drag $\sim v^8$, the 30° trajectory is shortened more drastically by resistance than the 60° trajectory; for drag $\sim v^2$, the 60° trajectory is more strongly affected.

The above results encourage further investigation into the way trajectories are affected. To do this, we have computed trajectories based on the following equation of motion for drag forces proportional to the nth power of projectile speed:

$$\frac{d\mathbf{v}}{dt} = -\frac{\lambda}{m}\mathbf{v}v^{n-1} - g\hat{\mathbf{j}}, \tag{1}$$

where $\mathbf{v}$ is the velocity of a particle of mass m, which is affected by a gravitational acceleration g, taken to act in the negative y direction. We limit our attention to drag forces that vary as a power of the projectile speed.

It is useful to note at the outset some scaling simplifications. If we denote the initial speed of the projectile by v_0, then all length scales can be scaled by the natural length scale v_0^2/g of the problem. For the dynamics, we scale time by the natural time scale v_0/g, replacing t by $T \equiv tg/v_0$. Thus, if we introduce the dimensionless velocity $\mathbf{u} \equiv \mathbf{v}/v_0$, Eq. (1) reduces to

$$\frac{d\mathbf{u}}{dT} = -k\mathbf{u}u^{n-1} - \hat{\mathbf{j}}, \tag{2}$$

where $k \equiv \lambda v_0^n/mg$ is the ratio of the initial drag force to the weight of the projectile. Since the initial condition in Eq. (2) is $u_0 = 1$, the problem is completely specified by the drag parameters k and n, and by the launch angle θ.

Numerical solution of the above equation of motion confirms that aiming high is often the way to go far. In addition to debunking the "obvious" nature of the "always aim low" philosophy, the numerical results, and other considerations, lead to the following observations about optimal angles.

(i) The optimal angle of launch is *greater than* 45° when n, the exponent of the velocity dependence, is larger than some critical value n_{crit}, which is around 3.5.

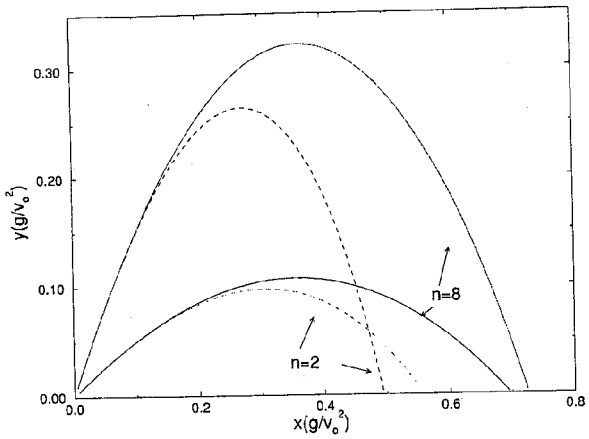

Fig. 1. Trajectories are shown for 30° and 60° launch angles, in the case of two types of drag; one for which $n=2$, and one for which $n=8$. The $n=2$ drag has a stronger shortening effect on the 60° trajectory, whereas the $n=8$ drag has a stronger shortening effect on the 30° trajectory. The x,y coordinates are normalized by the natural length scale v_0^2/g of the problem, where g is the acceleration of gravity and v_0 is the initial speed of the projectile. For both the $n=2$ and $n=8$ drag laws, the dimensionless indicator of damping strength, $\lambda v_0^n/mg$, has been taken to be unity. Note also that $x=1$ corresponds to maximum range for the drag-free case.

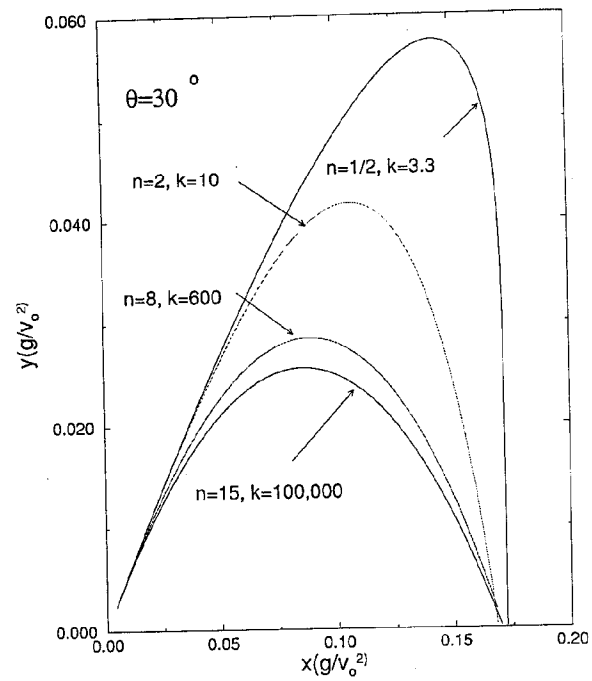

Fig. 2. The shape of 30° trajectories for various drag laws. The low n laws produce a very skewed orbit with a descent much steeper than ascent. For large n, the trajectories are very nearly parabolic. The x,y coordinates are normalized by the natural length scale v_0^2/g of the problem, where g is the acceleration of gravity and v_0 is the initial speed of the projectile.

The precise value of n_{crit} depends on the strength of the drag (i.e., on k), increasing as drag strength increases.

(ii) For n much less than n_{crit} and strong drag, the trajectories are very skewed; the descent from the maximum height is much steeper than the ascent to that height. There is no such distortion of the trajectories when n is much greater than n_{crit}. This point is illustrated in Fig. 2.

(iii) For $n=8$ and $k=100$, the optimal angle is found to be 47.0°. For $n>n_{\text{crit}}$ parameters could not be found for which the optimal angle is much larger than 47°. This is quite different from the $n<n_{\text{crit}}$ case in which strong drag leads to very shallow optimal angles.

(iv) For *extremely* strong drag and large n, the optimal angle is, as in the case of small n, less than 45°.

Most of these results can be understood as more than just computer output. The most basic issue, the n dependence of the optimal angle, can be understood with a calculation based on the two ways in which drag affects the range: First, it reduces the time the projectile is in the air, and second, it reduces the horizontal velocity. This can be quantified with a calculation in the limit of weak drag, i.e., a calculation to first order in the drag parameter k. This weak-limit calculation is useful in that it gives a definitive proof that θ_{opt} can be more than 45°, a proof that is independent of the difficulties that can cloud numerical results. This calculation unfortunately is not light reading, and has been relegated to the Appendix. It is recommended only to readers with the requisite skepticism and tolerance for details. The results of that calculation are rather more interesting than the calculation itself, and are shown in Fig. 3. Since the deviation of the optimal angle from 45° is proportional to k, we plot $\delta\theta/k \equiv (\theta_{\text{opt}}-45°)/k$, where θ_{opt} is the angle for maximum range. A key result is that $n_{\text{crit}}=3.4148...$ in the limit of weak drag.

The calculation in the Appendix, then, can be taken as an "explanation" of observation (i) above. Explanations of observations (ii)–(iv) above lie in a rather simple picture of the effect of air drag for large n. In the case of large n, the drag—if it is of any importance at all—is ferociously strong at the beginning of the launch, immediately slows down the projectile to a speed at which the drag is a small force compared to the weight of the projectile, and thereafter is unimportant. Large n drag, therefore, is confined to a very small portion of the beginning of the trajectory. This by itself gives an immediate explanation of observation (ii) and of the results presented in Fig. 2. A calculation based on this picture explains the remaining observations.

To do the calculation, let us imagine that the drag is effective only during the very small initial portion of the trajectory pictured in Fig. 4, and that the subsequent motion is a drag-free parabola. Let us denote by v_{trans} the speed of the projectile at which the drag/no-drag transition occurs, and let us denote by H the height at which the transition occurs. Further, let us suppose that during the strong drag phase the velocity vector is rotated downward by gravity an amount $\delta\theta_{\text{grav}}$. If a projectile is launched from height H, the angle for maximum range to a target at zero height is less than 45°. It is straightforward to show that for small H (i.e., for $H \ll v_{\text{trans}}^2/g$) the optimal angle is $45° - \delta\theta_H$, where

$$\delta\theta_H = \tfrac{1}{2} g H / v_{\text{trans}}^2. \tag{3}$$

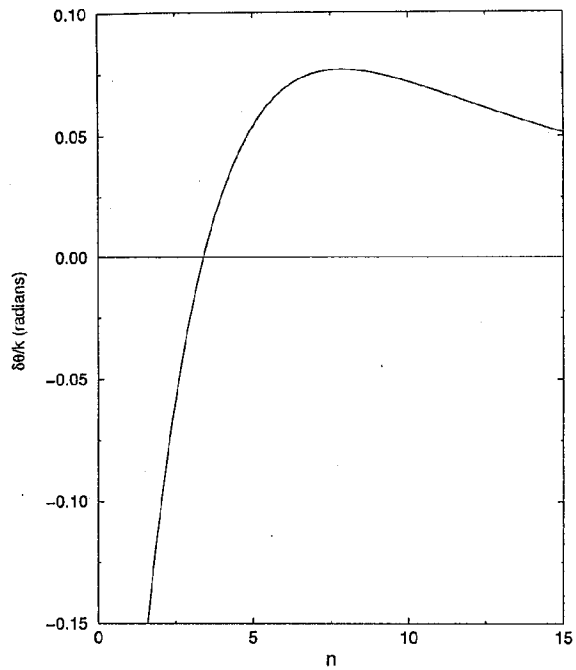

Fig. 3. The deviation of the optimal angle from 45° as a function of the drag exponent n, in the limit of weak drag. The deviation $\delta\theta$ is proportional to k, the dimensionless constant expressing the ratio of the initial drag force to the weight of the projectile. Note that the optimal angle changes from below 45° to above 45° at around $n=3.41$.

If the time for the projectile to ascend to the transition is τ, then, due to the influence of gravity, the projectile will have a velocity vector at the transition that is rotated downward from its original direction by an amount

$$\delta\theta_{\text{grav}} = g\tau/\sqrt{2}v_{\text{trans}}, \tag{4}$$

where we have assumed that the angle during the drag phase was approximately 45°.

The important question is whether $\delta\theta_H$ or $\delta\theta_{\text{grav}}$ is larger. If $\delta\theta_H$ is larger, then one needs to adjust the initial firing angle to somewhat below 45°, so that the projectile will be optimally aimed when it starts its drag-free motion. If, on the other hand, $\delta\theta_{\text{grav}}$ is larger, gravity rotates the velocity vector too far down, and one must fire a bit above 45° for the projectile to be optimally aimed when it starts its drag-free motion.

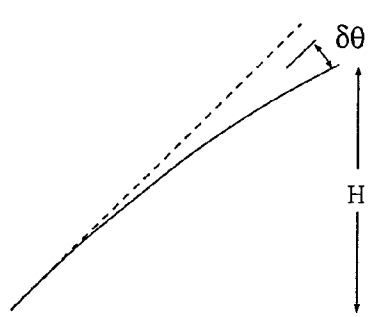

Fig. 4. The transition from drag-dominated to drag-free motion in the case of very large n.

To get insight into the relative size of $\delta\theta_H$ and $\delta\theta_{\text{grav}}$, let us write $H=\bar{a}\tau^2/2\sqrt{2}$, where $\bar{a}$ is the appropriate average acceleration during the drag phase. In terms of $\bar{a}$, we have

$$\frac{\delta\theta_H}{\delta\theta_{\text{grav}}} = \frac{\bar{a}\tau}{4v_{\text{trans}}}. \tag{5}$$

Roughly speaking, $\bar{a}\tau$ is the amount by which the speed of the projectile was reduced during the drag phase. For weak drag, this reduction will be a small fraction of the initial speed v_0, and we will have $\bar{a}\tau \ll v_0 \approx v_{\text{trans}}$. This argument then predicts that for weak drag and large n, the ratio in Eq. (5) will be much smaller than unity, and hence one must aim higher than 45° for maximum range. This conclusion is in agreement with the weak drag analysis of the Appendix.

On the other hand, for very strong drag—i.e., drag for which $v_{\text{trans}} \ll v_0$—the ratio in Eq. (5) will be large. In this case, the need for a reduction in the angle due to the height is larger than the rotation of the velocity by gravity. This argument, then, predicts that in the case of large n and very strong drag, the optimal angle is less than 45°, just as it is for small n drag of any strength. This prediction has been confirmed with numerical integration of the equations of motion. The numerical problem is delicate since very large n requires a very small step size in time. To find numerical solutions, we were forced to use an adaptive step-size routine and a very small value of large n. More specifically, we used $n=4$, which is a value of n just large enough so that for weak drag the optimal angle is above 45°. We computed trajectories for $k=0.1$ and for $k=10^{15}$, and found an optimal angle above 45° for the first case, and approximately 42° for the second.

This strong drag reversal explains why, for large n, it is impossible to have an optimal angle much above 45°. For small k (i.e., for weak drag) the increase of the optimal angle above 45° is proportional to k. If we try to increase the optimal angle by increasing k, we leave the weak drag regime and find that we are in fact decreasing the optimal angle. It almost seems that there is a moral lesson here about trying to aim too high, but a consideration of that hypothesis goes beyond the scope of this paper.

ACKNOWLEDGMENTS

We thank Apostol Gramada for suggesting that the optimal angle might depend on the details of drag, and George Williams for discussions and for useful suggestions. We also thank Keith Lockitch for supplying us with the adaptive step-size code, which we used to verify the large n, strong drag prediction. We gratefully acknowledge the support of the National Science Foundation under Grant Nos. PHY-9507719 at Utah and PHY-9507742 at Wisconsin.

APPENDIX

Here we analyze the solutions of Eq. (2) to first order in k for a projectile starting with velocity components $(u_{x0}, u_{y0}) = (\cos\theta, \sin\theta)$. We start with the y (vertical) component of the equation, initially limited to the time $T=0$ to $T=T_1$, during which the projectile is rising:

$$\frac{du_y}{dT}=-1-ku_y[u_x^2+u_y^2]^p. \tag{6}$$

Here we have introduced $p\equiv(n-1)/2$. Actually, for analyzing the ascending phase of the motion, it is more convenient to work with the time-reversed equation:

$$\frac{d\tilde{u}_y}{d\tilde{T}}=-1+k\tilde{u}_y[\tilde{u}_x^2+\tilde{u}_y^2]^p, \tag{7}$$

where $\tilde{u}_x\equiv dx/d\tilde{T}$ and $\tilde{u}_y\equiv dy/d\tilde{T}$. Equation (7) is obtained from Eq. (6) by introducing a new time variable $\tilde{T}\equiv T_1-T$. [A tilde ($\tilde{}$) is used to distinguish a time-reversed variable from its "forward-in-time" counterpart.] In this description, the projectile descends from time $\tilde{T}=0$ to $\tilde{T}=T_1$, starting with vertical velocity $\tilde{u}_y=0$ and ending with $\tilde{u}_y=-u_{y0}$. We chose to solve the time-reversed equation because of its simpler initial condition.

We can find $\tilde{u}_y$ as a function of $\tilde{T}$ by approximating $\tilde{u}_x\approx-u_{x0}$ (correct to lowest order in k) and integrating Eq. (7):

$$\tilde{T}=\int_0^{\tilde{T}}d\tilde{T}=\int_0^{\tilde{u}_y}\frac{d\tilde{u}_y}{-1+k\tilde{u}_y[u_{x0}^2+\tilde{u}_y^2]^p}. \tag{8}$$

To first order in k, this gives

$$\tilde{T}=-\tilde{u}_y-k\int_0^{\tilde{u}_y}\tilde{u}_y[u_{x0}^2+\tilde{u}_y^2]^p d\tilde{u}_y \tag{9}$$

$$=-\tilde{u}_y+\frac{k}{2(p+1)}[u_{x0}^{2(p+1)}-(u_{x0}^2+\tilde{u}_y^2)^{(p+1)}] \tag{10}$$

$$=-\tilde{u}_y+\frac{k}{2(p+1)}[u_{x0}^{2(p+1)}-(u_{x0}^2+\tilde{T}^2)^{(p+1)}], \tag{11}$$

where the approximation $\tilde{u}_y\approx-\tilde{T}$ (correct to lowest order in k) has been used on the right-hand side. We can now find the maximum height H of the trajectory by integrating $\tilde{u}_y$ from $\tilde{T}=T_1$ (when the projectile is at ground level) to $\tilde{T}=0$ (when the projectile is at its peak):

$$H=\int_{T_1}^0\tilde{u}_y d\tilde{T} \tag{12}$$

$$=\frac{1}{2}T_1^2-\frac{k}{2(p+1)}\left[u_{x0}^{2(p+1)}T_1-\int_0^{T_1\approx u_{y0}}(u_{x0}^2+\tilde{T}^2)^{p+1}d\tilde{T}\right] \tag{13}$$

$$=\frac{1}{2}T_1^2-\frac{k}{2(p+1)}\left[u_{x0}^{2(p+1)}T_1-u_{y0}^{2p+3}\times\int_0^1(\cot^2\theta+\xi^2)^{p+1}d\xi\right]. \tag{14}$$

Here, we have replaced u_{x0}/u_{y0} by $\cot\theta$, where θ is the angle at which the projectile is launched.

We now consider the descent of the projectile, and reset the (forward-in-time) clock so that it starts at $T=0$. We let T_2 stand for the time at which the projectile reaches ground level (i.e., the starting height) again. For this case, we choose to solve the forward-in-time Eq. (6). This equation is identical to Eq. (7) except for the sign of the drag term. Also, the initial condition ($u_y=0$ at $T=0$) in terms of the forward-in-time variables is identical to that for the ascent of the projectile in terms of the time-reversed variables. Thus, by simply reversing the sign of the drag term and replacing T_1 everywhere by T_2, it follows immediately from Eq. (14) that H is also given by

$$H=\frac{1}{2}T_2^2+\frac{k}{2(p+1)}\left[u_{x0}^{2(p+1)}T_2-u_{y0}^{2p+3}\times\int_0^1(\cot^2\theta+\xi^2)^{p+1}d\xi\right]. \tag{15}$$

By subtracting (14) from (15), and using $\frac{1}{2}(T_2^2-T_1^2)\approx u_{y0}(T_2-T_1)$ (which is correct to first order in k), it follows that

$$T_2-T_1=\frac{k}{(p+1)}\left[u_{y0}^{2(p+1)}\int_0^1(\cot^2\theta+\xi^2)^{p+1}d\xi-u_{x0}^{2(p+1)}\right]. \tag{16}$$

Also, by setting $\tilde{u}_y=-u_{y0}$ at $\tilde{T}=T_1$ in Eq. (10), we get

$$2T_1=2u_{y0}-\frac{k}{(p+1)}[1-u_{x0}^{2(p+1)}]. \tag{17}$$

Finally, we add these last two results to obtain T_{tot}, the total time of flight of the projectile:

$$T_{\text{tot}}\equiv T_1+T_2=2u_{y0}+\frac{k}{(p+1)}\times\left[u_{y0}^{2(p+1)}\int_0^1(\cot^2\theta+\xi^2)^{p+1}d\xi-1\right]. \tag{18}$$

Next we consider the equation

$$\frac{du_x}{dT}=-ku_x[u_x^2+u_y^2]^p \tag{19}$$

for the horizontal motion, and in the drag term make the lowest order approximation $u_x\approx u_{x0}$ and $u_y\approx(u_{y0}-T)$:

$$\frac{du_x}{dT}=-ku_{x0}[u_{x0}^2+(u_{y0}-T)^2]^p. \tag{20}$$

We now integrate this equation to arrive at

$$u_x=u_{x0}-ku_{x0}\int_0^T[u_{x0}^2+(u_{y0}-\bar{T})^2]^p d\bar{T}. \tag{21}$$

To find the range R, we integrate u_x from $T=0$ to $T=T_{\text{tot}}$, using the approximation $T_{\text{tot}}\approx 2u_{y0}$ in terms that are already first order in k:

$$R=u_{x0}T_{\text{tot}}-ku_{x0}\int_0^{2u_{y0}}dT\int_0^T[u_{x0}^2+(u_{y0}-\bar{T})^2]^p d\bar{T}. \tag{22}$$

The double integral in the expression can be simplified by changing the order of integration:

$$\int_0^{2u_{y0}} dT \int_0^T [u_{x0}^2+(u_{y0}-\bar{T})^2]^p d\bar{T}$$

$$= \int_0^{2u_{y0}} d\bar{T} \int_{\bar{T}}^{2u_{y0}} [u_{x0}^2+(u_{y0}-\bar{T})^2]^p dT \tag{23}$$

$$= \int_0^{2u_{y0}} (2u_{y0}-\bar{T})[u_{x0}^2+(u_{y0}-\bar{t})^2]^p d\bar{T} \tag{24}$$

$$= \int_0^{2u_{y0}} u_{y0}[u_{x0}^2+(u_{y0}-\bar{T})^2]^p d\bar{T} + \int_0^{2u_{y0}} (u_{y0}-\bar{T}) \times [u_{x0}^2+(u_{y0}-\bar{T})^2]^p d\bar{T}. \tag{25}$$

The second integral above vanishes by symmetry, and the first can be written as

$$2u_{y0}^{2(p+1)} \int_0^1 (\cot^2\theta+\xi^2)^p d\xi. \tag{26}$$

We can now use this result, and Eq. (18), in Eq. (22) to arrive at an expression for the range that is correct to first order in k. We express the result in terms of the launch angle θ, replacing u_{x0}, u_{y0} by $\cos\theta$, $\sin\theta$. The result is

$$R = 2\sin\theta\cos\theta + k\mathscr{F}(\theta), \tag{27}$$

where

$$\mathscr{F}(\theta) \equiv \frac{\cos\theta}{(p+1)}\left[(\sin\theta)^{2(p+1)}\int_0^1 (\cot^2\theta+\xi^2)^{p+1} d\xi - 1\right] - 2\cos\theta\,(\sin\theta)^{2(p+1)}\int_0^1 (\cot^2\theta+\xi^2)^p d\xi. \tag{28}$$

To find the critical value of n, we now differentiate $\mathscr{F}(\theta)$ with respect to θ and evaluate it at 45°:

$$\mathscr{F}'|_{45°} = \frac{1}{\sqrt{2}}\left[\frac{1}{(p+1)}\left(1+\frac{2p+1}{2^{p+1}}\int_0^1 (1+\xi^2)^{p+1} d\xi\right) - \frac{2p+3}{2^p}\int_0^1 (1+\xi^2)^p d\xi + \frac{2p}{2^{(p-1)}}\int_0^1 (1+\xi^2)^{p-1} d\xi\right]. \tag{29}$$

This expression is easily evaluated numerically and is found to be negative for small p, changing to positive at $p = 1.2074\ldots$. This means that in the limit of weak drag, the optimal angle changes from less than 45° to greater at $n = 2p+1 = 3.4148\ldots$.

For weak drag the optimal angle can be written as $45° + \delta\theta$, where $\delta\theta$ is first order in k. From the definition of $\mathscr{F}$ it follows that $\delta\theta/k = \frac{1}{4}\mathscr{F}'|_{45°}$. It is this quantity that is plotted in Fig. 3. As a check, we computed $\delta\theta/k$ directly without simplifying the double integrals occurring in Eq. (22). The results were in excellent agreement with those given by Eq. (29) and in Fig. 3.

[1]We do not wish to wander too deeply into real world effects. For projectiles traveling very large distances, the decrease in atmospheric density favors high angles of launch. [See D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics* (Wiley, New York, 1981), 5th ed., Question 16, p. 73.] Here we consider a flat earth and homogeneous atmosphere.

[2]P. J. Brancazio, "Trajectory of a fly ball," Phys. Teach. **23** (1), 20–23 (1985).

[3]L. S. Lerner, *Physics for Scientists and Engineers* (Jones and Bartlett, Sudbury, MA, 1996), p. 120.

[4]R. A. Serway, *Physics for Scientists and Engineers* (Saunders, Philadelphia, 1992), 4th ed., Question 19, p. 95.

[5]C. W. Groetsch, "On the optimal angle of projection in general media," Am. J. Phys. **65** (8), 797–799 (1997).

[5]Richard H. Price and Joseph D. Romano, "Comment on 'On the optimal angle of projection in general media,' by C. W. Groetsch," Am. J. Phys. **66** (2), 114 (1998.

SCORNING THE BASE DEGREES

We need notice at the moment only that the choice of the simplest law that fits the facts is an essential part of procedure in applied mathematics, and cannot be justified by the methods of deductive logic. It is, however, rarely stated, and when it is stated it is usually in a manner suggesting that it is something to be ashamed of. We may recall the words to Brutus.

But 'tis a common proof
That lowliness is young ambition's ladder,
Whereto the climber upward turns his face;
But when he once attains the upmost round,
He then unto the ladder turns his back,
Looks in the clouds, scorning the base degrees
By which he did ascend.

Harold Jeffreys, *Theory of Probability* (Oxford University Press, 1939), p. 4.

Comment on "On the optimal angle of projection in general media," by C. W. Groetsch [Am. J. Phys. 65 (8), 797–799 (1997)]

Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

Joseph D. Romano
Department of Physics, University of Wisconsin—Milwaukee, Milwaukee, Wisconsin 53201

(Received 20 October 1997; accepted 21 October 1997)

In an accompanying paper[1] we give analytic and numerical demonstrations that to achieve maximum range of a projectile moving in a dissipative medium, the launch angle must, in some cases, be greater than 45°. In his recent article, C. W. Groetsch[2] appears to have proven that this is impossible. The flaw in the proof lies in the manner in which the physical problem is formulated as a mathematical problem. We have discussed this question of formulation with Professor Groetsch, who agrees that it is flawed, and that his analysis does not show that optimal angles greater than 45° are impossible. We describe here the problem with the formulation.

At the beginning of that paper, a function of time $f(t)$ is defined to be the resistive force per unit velocity. The motion of a projectile, moving under this force, is then analyzed for arbitrary $f(t)$. The launch angle θ enters the analysis only as an initial condition for the differential equations of motion. In the analysis, the range of the particle R is then found in terms of this initial parameter θ, and it is shown that if $f(t)$ is non-negative (if the resistance always opposes the motion), then the angle θ that maximizes R cannot be larger than 45°.

The flaw lies in the fact that f, the resistive force per unit velocity, at any moment of time t, depends on the speed v, and angle θ, at which the projectile was launched. In other words, f should really be written as $f(t,v,\theta)$. In finding the optimal angle, the angle for maximum range, the dependence of f on θ must be taken into account. This dependence was omitted.

Note that Groetsch's analysis is valid if the resistive force is proportional to the velocity. In this case f is a constant, and hence has no dependence on θ.

[1]Richard H. Price and Joseph D. Romano, "Aim high and go far—Optimal projectile launch angles greater than 45°," Am. J. Phys. **66** (2), 109–113 (1998).

[2]C. W. Groetsch, "On the optimal angles of projection in general media," Am. J. Phys. **65** (8), 797–799 (1997).

Applying relativistic electrodynamics to a rotating material medium

Charles T. Ridgely
Department of Physics and Astronomy, California State University, Long Beach, California 90840

(Received 22 April 1996; accepted 3 September 1997)

We apply relativistic electrodynamics to a rotating linear medium. Covariant field equations are used to derive general field equations in a rotating coordinate system. We argue that the relation between fields in the presence of matter and those in a vacuum is necessarily dependent upon the coordinate system used. Constitutive equations are then derived in the rotating and laboratory reference frames. We find that our constitutive equations in the laboratory frame agree with Minkowski's constitutive equations, derived on the basis of special relativity in 1908. Thus we conclude that special relativity can be used in the analysis of experiments involving rotational motion. To exemplify the use of special relativity, we derive an experimentally observed result of a 1913 experiment performed by Wilson and Wilson in which a polarizable, permeable cylinder was rotated in a uniform, axially directed magnetic field.

I. INTRODUCTION

For more than 80 years, the compatibility of special relativity and classical electrodynamics has been generally accepted; however, there still exists one area of the subject which apparently leads to some confusion. How does one apply relativistic electrodynamics to a material medium? More particularly, how does one deal with rotating material media? There is no ambiguity when the relative motion is uniform; however, as pointed out in a recent article by Pellegrini and Swift,[1] rotational motion of a material medium

A circular twin paradox

Maria B. Cranor, Elizabeth M. Heider, and Richard H. Price
Department of Physics, University of Utah, Salt Lake City, Utah 84112

(Received 28 October 1999; accepted 28 March 2000)

In the special relativistic twin paradox presented here, each twin lives on one ring of a counterrotating pair of infinitesimally separated rings, so that the twins travel on the same circular path but in opposite directions. The observers on the ring of one twin should see the clock of the other twin slowed by time dilation, but at each meeting of the twins symmetry demands that they agree on the amount of time that has passed since their previous meeting. The resolution of the paradox focuses attention on the relation of time dilation to clock synchronization.

I. INTRODUCTION

Twin paradoxes have played an important role in the pedagogical history of special relativity theory; generations of physics students have been challenged by their professors to apply their newly minted understanding of relativistic principles to the resolution of these famously counterintuitive problems. Continuing in this venerable tradition, we present a twin paradox here, but one that differs significantly from the familiar rocket parable. Although this paradox appears in Lightman *et al.*,[1] the solution provided there is brief, mathematical, and formal, and is not accessible to the full range of students for whom the paradox would be of interest. Some elements of the resolution of the paradox are related to issues that have been discussed by several authors[2–4] but the connection to our paradox is not straightforward. Here we present an analysis that should be both mathematically and physically intelligible to beginning relativity students. In order to optimize the accessibility of the basic ideas, we relegate all calculations to the Appendix.

In our parable, Lisa, the protagonist of the paradox, lives on a ring of radius R with a team of observers stationed at every point on her ring. Bart lives on an identical ring. The rings are negligibly separated, rotating at equal and opposite angular velocity ω about a common axis. This common axis is at rest in a "Lab" frame, through which Bart, Lisa, and all of Lisa's observers move at speed $v=\omega R$. Bart moves through the Lab at this speed in the counterclockwise direction while the observers on Lisa's ring move through the Lab at this speed in the clockwise direction. The twins will pass each other periodically, their negligible separation allowing each to read the other's clock.

At a certain moment, Bart and Lisa happen to be at the same place, and they notice that their clocks both read $t=0$. To the observers on Lisa's ring (see Fig. 1) Bart's clock flies by at speed $v_{\rm rel}=2v/(1+v^2/c^2)$. (Recall that relative speeds do not add simply in special relativity.) Due to time dilation, Lisa's team observes Bart's clock ticking more slowly than their own clocks by a relative Lorentz factor $\gamma_{\rm rel}$ that works out to be $\gamma_{\rm rel}=(1+v^2/c^2)/(1-v^2/c^2)$. Since Bart's clock agrees with Lisa's as he passes her at $t=0$, time dilation means that his clock will lag behind the clock of the next of Lisa's observers that he passes. As he passes successive members of Lisa's observing team, his clock should be seen to lag further and further behind. One half-rotation later, the observer he passes will be Lisa. We conclude that when Bart passes her their clocks will disagree; Bart's clock must lag behind Lisa's.

This is nonsense, of course. Very convincing arguments support our intuition that Bart's and Lisa's clocks will agree at each meeting if they agree at the first meeting. One of these arguments follows from the point of view of the laboratory observers. They see Bart and Lisa tracing out identical motions, except one is going clockwise and the other counterclockwise. This "handedness" of the motion can have no effect on the rate of ticking of clocks, so the clock of Bart and the clock of Lisa must tick off the same number of seconds between meetings (though neither ticks off the same number of seconds as the Lab clocks). Another obvious "twin" argument further convinces us that Bart's clock cannot lag Lisa's at their subsequent meetings: In the very same way we argued that Bart's clock must lag behind Lisa's we could have argued (based on a corps of special relativity observers on Bart's ring) that Lisa's clock will lag behind Bart's at future meetings.

"Clearly" there is something wrong with our argument about time dilation and Bart's clock, but what precisely is wrong?

II. CLOCK SYNCHRONIZATION ON A ROTATING RING

Special relativistic time dilation must be considered in the context of clock synchronization. Suppose that Lisa's closest neighbor in the counterclockwise direction is Milhouse. As Bart passes first Lisa and then Milhouse, time dilation is inferred from the comparison of Bart's clock with Lisa's and then with Milhouse's. We have set up our parable so that the clocks of Bart and Lisa both read $t=0$ at the first event. The comparison at the second event (to determine whether Bart's clock lags, leads, or neither) depends on the setting of Milhouse's clock, which in turn depends on how *his* clock was originally related to Lisa's.

Time dilation will only be observed from a reference frame in which the clocks are appropriately synchronized. Clarification of our paradox thus requires careful deconstruction of the clock synchronization issues implicit in the situation we have described.[5] One prescription for synchronizing two clocks is that given by Einstein and called "Einstein synchronization" in the special relativity literature. In this prescription for clocks A and B,[6] "...a common time for A and B...cannot be defined at all unless we establish by definition that the 'time' required by light to travel from A to B

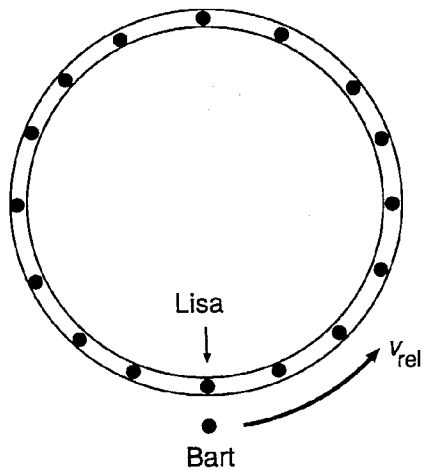

Fig. 1. As viewed in a rotating frame in which Lisa's ring is unmoving, Bart moves in the counterclockwise direction, past the observers on Lisa's ring, at a speed $v_{rel}=2v/(1+v^2/c^2)$.

equals the 'time' it requires to travel from B to A. Let a ray of light start [as shown in Fig. 2] at the A time t_A from A towards B. At the 'B time' t_B let it be reflected from B in the direction of A, and arrive again at A in the 'A time' t'_A. The two clocks are Einstein synchronized if

$$t_B - t_A = t'_A - t_B.'' \tag{1}$$

The synchronization of clocks on *rotating* rings presents a special challenge. Here we describe four ways one might choose to synchronize such clocks. In the first two methods, the rings are imagined to be initially nonrotating, that is, initially at rest in the Lab. The clocks are synchronized while at rest, and then the rings are set into rotation. In the last two methods, synchronization is done while the rings are already rotating.

A. Method 1

Before the rings are set into motion, the ring observers, sitting at rest in the Lab, may decide to synchronize their clocks according to the principles of Einstein synchronization, i.e., by exchanging light signals. Lisa, ensconced at point A, noting that her clock registers t_A, fires a laser beam at Milhouse, her "next door neighbor" in the counterclockwise direction, who is stationed at point B. At t_B, he receives and reflects the beam back to her; she receives the signal at t'_A. Lisa sends Milhouse a slip of paper upon which is written the value of $(t'_A+t_A)/2$, with instructions that his clock should have had that reading at t_B. Milhouse adjusts his clock accordingly. This procedure is followed from observer to observer around the ring, and we imagine the limit of an infinite number of observers with infinitesimal separation. Einstein synchronization on the not-yet spinning rings thus

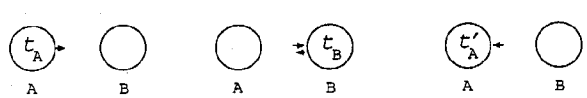

Fig. 2. Einstein synchronization of clock B with clock A. A light signal is sent from A to B and back to A. Clock B is adjusted so that $t_B=(t_A+t'_A)/2$.

allows Lisa's team to constitute themselves correctly as a special relativity reference frame; that is, Lisa's observers will measure the rate of a clock moving by them at speed v to be too slow by $1/\sqrt{1-v^2/c^2}$.

With clocks synchronized in what appears to be an incontrovertibly correct way, the rings are now uniformly set into rotation. (Here "uniformly" means that all points on the ring are treated identically.)

B. Method 2

Before the rings are set into uniform motion, that is, when they are at rest in the Lab frame, the clocks on Lisa's ring can be synchronized by an even easier method. Upon receiving a flash from a "Big Lab Clock"[2] stationed at the center of rotation, the ring observers can all set their clocks to read $t=0$. It seems obvious (and is true!) that this method produces the same results as the classical Einstein synchronization used in the first stage of Method 1. As in Method 1, the rings are gradually and uniformly put into motion after the synchronization process is completed.

C. Method 3

In this method, Lisa's crew chooses to synchronize their clocks with their ring already in motion. A rather obvious way to do this is to have a light flash at the center of rotation, all observers having been instructed to set their clocks to $t=0$ at the moment they see the flash. This method is the same as Method 2 except that the rings are in motion when the procedure is performed.

D. Method 4

In Method 4, Lisa and her team of observers attempt to employ Einstein synchronization on the already-rotating ring. Lisa sends a light signal to Milhouse, the first observer in the counterclockwise direction, and an infinitesimal distance from her. That observer sends back a light signal. The times of arrival of the light signals are used in the same manner as in Method 1 to synchronize the clocks. This synchronization is then continued, proceeding around the ring in the counterclockwise direction.

E. Comparison

In Method 1, before the rings are put into rotation, the clocks will be correctly synchronized for Lisa's crew to make standard special relativity observations. If the Lab observers also have clocks that are correctly synchronized, then simultaneity will mean the same thing to Lisa's observers and to the Lab observers. When the clocks of each of Lisa's observers strike midnight, the clock of each nearby Lab observer will have the same reading—say 2:23 a.m.

It is clear that this is also true for Method 2. Since the "flash at the center" process favors no particular location on the ring, a moment of simultaneity on the ring (the same clock reading for all of Lisa's observers) is also a moment of simultaneity in the Lab (the same reading for all Lab observer clocks).

This very same argument applies equally well to the synchronization by Method 3. Though the ring is now moving during the synchronization process, the "flash at the center" again favors no particular observer. It follows that a moment

of simultaneity (all clocks have the same reading) on a ring will also be a moment of simultaneity in the Lab.

From the above arguments we conclude that Methods 1, 2, and 3 all provide the same sort of synchronization. As we will show in the following, Method 4 is different.

III. THE PARADOX RESOLVED

If Lisa and her nearest neighbor Milhouse were properly synchronized to be part of a special relativity reference frame moving through the Lab, then events (like the striking of midnight) that are simultaneous to Lisa and Milhouse cannot be simultaneous in the Lab frame. If Methods 1, 2, or 3 are used for synchronization of ring clocks, then events that are simultaneous to Lisa and Milhouse will also be simultaneous to the Lab observers. It follows that Lisa and Milhouse, and more generally the entire set of observers on Lisa's ring, are not correctly synchronized to constitute special relativity reference frames. This explains what we already know must be true: There will be no lagging of Bart's clock observed as it passes each of Lisa's observers. *For the relativistically inappropriately synchronized clocks of Lisa's observers, there is no time dilation of Bart's clock.*

This conclusion may seem to some readers to evade, not resolve, the paradox. We have explained away the awkward implications of time dilation by using synchronization that does not produce time dilation. We justify the inclusion of these "inappropriate" methods of synchronization with two arguments: First, this helps to emphasize the connection between clock synchronization and time dilation; second, a student would ask why such obvious methods of synchronization are not used.

In any case, the paradox cannot be evaded if Lisa's clocks are synchronized by Method 4. In this case, the clocks of Lisa and Milhouse are "correctly" synchronized. If Lisa and Milhouse are negligibly separated on the ring, their readings will differ negligibly from readings done in a standard special relativity reference frame that is instantaneously comoving with them. Lisa and Milhouse *will* therefore observe Bart's clock to run slowly and Bart's clock *will* lag Milhouse's when they pass each other. The resolution of the paradox now takes a very different form: When Method 4 is used, there will be a discontinuity in synchronization. Suppose Einstein synchronization is used starting with Lisa, proceeding to Milhouse, and proceeding around the ring until the clock of the last observer, call her Selma, is synchronized. In the case that the angular separation of Lisa's team of observers is negligibly small, Lisa at angle 0 and Selma at angle 2π are at the same place. But their clocks will not agree. Due to the discontinuity of Method 4 synchronization, Selma's clock will lead (i.e., will have a higher reading) than Lisa's by an amount that we show in the Appendix to be $Disc=(2\pi R/c)(v/c)/\sqrt{1-v^2/c^2}$.

When calculated in detail, the time dilation of Bart's clock, moving past Lisa's observers, turns out to show that Bart's clock will lag Selma's by precisely the amount *Disc*, when Bart reaches Selma. At the very same location as Selma is Lisa, whose clock lags Selma's by *Disc*, and hence agrees perfectly with Bart's clock.

The Appendix gives the details of computation of the discontinuity *Disc*, and shows that this discontinuity makes time dilation compatible with the comparison of Lisa's clock to Bart's and to the clock of a Lab observer.

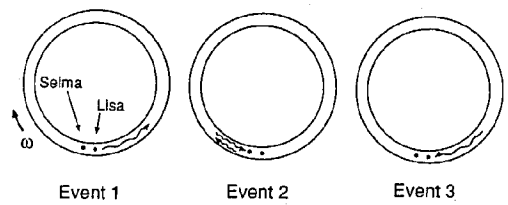

Fig. 3. Method 4 used to synchronize Lisa's clock with that of another observer.

IV. CONCLUSION

The resolution of the paradox of the counterrotating twins depends on the method that is used to synchronize clocks on a ring. If the clocks are synchronized "uniformly" (with no particular position on the ring singled out), then the resolution is that the observers so synchronized will not measure any time dilation. If, on the other hand, the clocks are synchronized by Einstein synchronization, starting with one particular observer, there will be a discontinuity in synchronization at the location of that observer, and this discontinuity permits both time dilation and the agreement of the twins' clocks at every meeting.

APPENDIX: DISCONTINUITY DUE TO METHOD 4 SYNCHRONIZATION

Figure 3 shows the three events needed to synchronize the clock of another observer with Lisa's clock. Let us say that Lisa is at Lab position $\phi=0$, and she will synchronize clocks with one of her observer friends at Lab angle $\phi=\Delta\phi$. (In Fig. 3, the friend is shown as Selma, the last of Lisa's observers, located at $\phi=\Delta\phi=2\pi$.) In the calculations below, the coordinate ϕ is measured relative to the Lab system. The measure $\Delta\phi$ refers to the angular position of the observer friend as measured on Lisa's ring, but it has the same value as observed in the Lab frame. In either frame it simply indicates the angular displacement of the observer as a fraction of a complete circle. If the observer friend were diametrically opposite Lisa, both the observers on Lisa's ring and the Lab observers would describe the angular difference as $\Delta\phi=\pi$.

For definitiveness let us say that Lisa's ring of observers is rotating through the Lab in the clockwise, or negative, direction, at angular velocity ω, as shown in the figure. This means that Bart will be traveling in the counterclockwise, or positive, direction relative to Lisa's observers; if Bart is to be observed by a synchronized team of Lisa's observers, synchronization must proceed in the counterclockwise direction.

Event 1 of the synchronization procedure is for Lisa to send a photon in the positive direction. Let us say that this event occurs at Lab time $t=0$. At event 2, Lisa's friend receives that photon and sends one back to Lisa in the negative direction. Between events 1 and 2 the photon moves through the Lab according to $\phi=ct/R$, and the position of Lisa's friend is given by $\phi=\Delta\phi-\omega t$. By solving these two equations for the intersection of friend and photon, we find that event 2 occurs at Lab angular location and at Lab time

$$\phi_2=\frac{\Delta\phi}{1+v/c},\quad t_2=\frac{R}{c}\frac{\Delta\phi}{1+v/c}, \tag{2}$$

where $v\equiv\omega R$ is the speed of Lisa's observers through the Lab. Between events 2 and 3, Lisa moves through the Lab

according to $\phi=\omega t$, and the photon from her friend moves through the Lab according to $\phi=\phi_2-c(t-t_2)/R$. These two motions intersect at event 3, at Lab time:

$$t_3=2\frac{R}{c}\frac{\Delta\phi}{1-v^2/c^2}. \tag{3}$$

Due to the time dilation of Lisa's clock with respect to the Lab system, Lisa's clock will read $\tau_3=t_3/\gamma$ at event 3, where $\gamma\equiv 1/\sqrt{1-v^2/c^2}$, or $\tau_3=2(R/c)\gamma$. According to the prescription for Einstein synchronization, Lisa's friend, at $\phi=\Delta\phi$, will be given instructions to adjust her clock so that it would have read

$$\text{clock setting at event } 2=(\tau_3+\tau_1)/2=\tau_3/2=(R/c)\gamma\Delta\phi \tag{4}$$

at event 2.

Now let us suppose that Lisa's friend is Selma, at $\phi=2\pi$, so that she is at the same position as Lisa. At event 2, the Lab time, from Eq. (2), is $t_2=(R/c)\Delta\phi/(1+v/c)$, and Lisa's clock reads t_2/γ, while Selma's Einstein synchronized clock reads $\tau_3/2=(R/c)\gamma\Delta\phi$. Thus at event 2 Lisa's clock will lag Selma's clock by $\tau_3/2-t_2/\gamma$, and hence by

$$Disc=\frac{\tau_3}{2}-\frac{t_2}{\gamma}=\frac{2\pi Rv}{c^2}\gamma, \tag{5}$$

where we have used $\Delta\phi=2\pi$.

1. Bart's time dilation lag

Consider Bart moving in the counterclockwise direction past Lisa's observers, with his relative velocity $v_{\rm rel}=2v/(1+v^2/c^2)$ and relative Lorentz factor $\gamma_{\rm rel}=(1+v^2/c^2)/(1-v^2/c^2)$. The rate at which Bart's clock advances, compared to the readings on the clocks he passes, is given by the usual time dilation relationship $\Delta t_{\rm Bart}=\Delta t_{\rm LR}/\gamma_{\rm rel}$. Here the subscript "LR" indicates "Lisa's ring." Bart is comparing clocks not with Lisa, but with other observers on her ring. At the completion of his circumnavigation of Lisa's ring, he will encounter Selma, and his clock will read $\tau_{\rm Selma}/\gamma_{\rm rel}$, and hence will lag behind hers by $\tau_{\rm Selma}(1-1/\gamma_{\rm rel})$. To find Selma's clock reading at that event we note that Lisa has moved through the Lab by π and therefore Lisa's clock will read $\pi/(\omega\gamma)$. Selma's must therefore read

$$\tau_{\rm Selma}=\pi/(\omega\gamma)+Disc=\frac{\pi\gamma}{\omega}\left(1+\frac{v^2}{c^2}\right). \tag{6}$$

Bart's reading will lag by

$$\tau_{\rm Selma}\left(1-\frac{1}{\gamma_{\rm rel}}\right)=\frac{\pi\gamma}{\omega}\left(1+\frac{v^2}{c^2}\right)\left(1-\frac{1}{\gamma_{\rm rel}}\right)=2\pi\gamma Rv/c^2. \tag{7}$$

But this is the same as the amount *Disc*, by which Lisa's clock lags behind Selma's! Thus Bart's clock *does* undergo time dilation. It lags behind Selma's clock. But due to the discontinuity in synchronization, Lisa's clock lags by precisely the same amount, and Bart and Lisa will have clock readings that agree, as of course they must.

2. Time dilation of a Lab observer

Consider now a Lab observer, that is, an observer fixed in position in the Lab frame. Suppose that such an observer happens to be fixed at an infinitesimal distance from Lisa's ring. Such an observer will be measured by Lisa's observers to be moving past them, as shown in Fig. 4, at speed v in the counterclockwise direction. While a study of clock comparisons between this observer and observers on Lisa's ring is not necessary to resolve the Bart–Lisa paradox, it does further illuminate the role of the synchronization discontinuity on Lisa's ring.

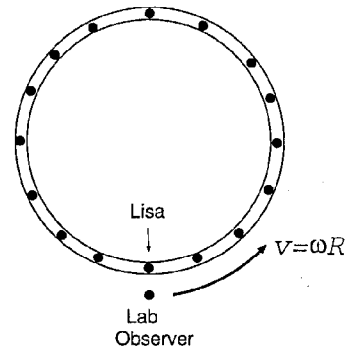

Fig. 4. As viewed in a rotating frame in which Lisa's ring is unmoving, an observer who is fixed in the Lab frame moves in the counterclockwise direction, past the observers on Lisa's ring, at a speed $v=\omega R$.

Let us suppose that at the moment the Lab observer passes Lisa, his clock and Lisa's clock both read $t=0$. After a Lab time $2\pi/\omega$ has passed, this Lab observer will encounter Selma, and his clock will lag hers by $\tau_{\rm Selma}(1-1/\gamma)$ due to time dilation. As above we can argue that $\tau_{\rm Selma}$ is greater than Lisa's clock reading by *Disc*, and hence $\tau_{\rm Selma}=2\pi/(\omega\gamma)+Disc=2\pi\gamma/\omega$. The amount by which the Lab observer's clock lags Selma's is therefore $(2\pi\gamma/\omega)(1-1/\gamma)$. Since Selma's clock leads Lisa's clock by *Disc*, this means that the Lab clock will lead Lisa's clock by

$$Disc-\frac{2\pi\gamma}{\omega}\left(1-\frac{1}{\gamma}\right)=\frac{2\pi}{\omega}\left(1-\frac{1}{\gamma}\right). \tag{8}$$

This lead is precisely what the Lab observer must observe since, due to time dilation, Lisa's clock has been ticking slowly, relative to Lab clocks, by the factor γ, as it has moved through the Lab by an angle of 2π.

3. Synchronization around the ring versus comoving synchronization

We have described Method 4 synchronization as being carried out with light signals propagating on a circular path, yet we have treated the observers on Lisa's ring (aside from the discontinuity) as if they were special relativity observers. More specifically, we have claimed that those observers, from Lisa to Selma, would measure the same time dilation as would observers in a properly synchronized inertial reference frame. Here we give a detailed justification for this. We show that the Method 4 synchronization of two nearby observers on Lisa's ring differs negligibly from the synchronization of a momentarily comoving inertial reference frame.

To do this we consider, just as we did in our discussion of Method 4 synchronization, Lisa and an observer friend at $\Delta\phi$. As in the earlier discussion, event 2 will be the reception of a photon by the friend, and the emission of the return photon. Let us invoke a Cartesian spatial coordinate system $\{x,y\}$ in the Lab, as shown in Fig. 5, with origin at Lisa at the

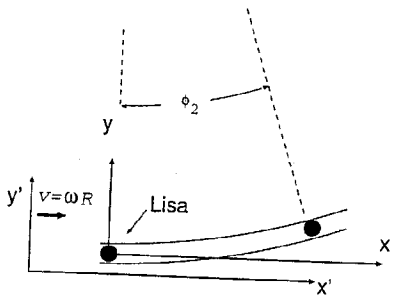

Fig. 5. Method 4 synchronization (as viewed in the Lab frame) and a momentarily comoving inertial reference frame.

moment ($t=0$) of event 1, and with the x axis in the direction of the emitted photon. The Lab coordinates of event 2 are then

$$t=t_2, \quad x=R\sin\phi_2, \quad y=R(1-\cos\phi_2), \tag{9}$$

where t_2 and ϕ_2 are given in Eq. (2). We now invoke an inertial reference frame $\{t',x',y'\}$ that is momentarily comoving with Lisa. That is, at event 1 this inertial frame is moving in the positive x direction with speed $v=\omega R$, so that instantaneously Lisa is at rest in this frame. By a straightforward Lorentz transformation the time coordinates of event 2 in this momentarily comoving frame are

$$t'=\gamma\left(t+\frac{v}{c^2}x\right)=\gamma\left(t_2+\frac{v}{c^2}R\sin\phi_2\right)$$
$$=\frac{R}{c}\gamma\Delta\phi\left(\frac{1}{1+v/c}+\frac{v/c}{\Delta\phi}\sin\left(\frac{\Delta\phi}{1+v/c}\right)\right). \tag{10}$$

The setting on the clock of Lisa's friend at event 2 is given in Eq. (4). It follows that this setting will lead the setting t' in the inertial frame by $\Delta t\equiv(R/c)\gamma\Delta\phi-t'$, or

$$\Delta t=\frac{v/c}{1+v/c}\frac{R}{c}\gamma\Delta\phi\left(1-\frac{1+v/c}{\Delta\phi}\sin\left(\frac{\Delta\phi}{1+v/c}\right)\right). \tag{11}$$

Suppose Lisa's ring is occupied by N equally spaced observers (with the last observer, Selma, at the same position as Lisa), then $\Delta\phi=2\pi/N$. In the limit that N is very large

$$\Delta t=\frac{v/c}{1+v/c}\frac{R}{c}\gamma\frac{2\pi}{N}\left(1-\frac{N(1+v/c)}{2\pi}\right.$$
$$\left.\times\sin\left(\frac{2\pi}{N(1+v/c)}\right)\right)\xrightarrow{N\to\infty}\mathcal{O}(N^{-3}). \tag{12}$$

If the synchronization of Lisa's ring were carried out from Lisa to Selma, by using momentarily comoving inertial frames, the setting of Selma's clock would be different by only $N\Delta t=\mathcal{O}(N^{-2})$ from the setting arrived at with Method 4. In the limit that N is large (i.e., in the limit that $\Delta\phi$ is small), this is negligible.

[1]A. P. Lightman, W. H. Press, R. H. Price, and S. A. Teukolsky, *Problem Book in Relativity and Gravitation* (Princeton U.P., Princeton, 1975), pp. 6 and 7 and 138–140.

[2]T. A. Weber, "Measurements on a rotating frame in relativity, and the Wilson and Wilson experiment," Am. J. Phys. **65** (10), 946–953 (1997).

[3]O. Grøn, "Relativistic description of a rotating disk," Am. J. Phys. **47** (10), 869–876 (1975).

[4]C. Giannoni and O. Grøn, "Rigidly connected accelerated clocks," Am. J. Phys. **47** (5), 431–435 (1979).

[5]It is worth pointing out that many (most?) special relativity paradoxes exploit confusion about simultaneity. Clock synchronization can be considered a specific application of defined simultaneity.

[6]A. Einstein, "On the electrodynamics of moving bodies," in *The Principle of Relativity* (Dover, New York, 1952), pp. 37–55.

grand Ampère
No 484 p 760
mpère vol 2
0 p 773
mpère vol 2
0 p 743
hys. 4 170
lle 21 29
nève: Rambo
981 Einfache
und Gegenwar
3
phy ed C C
1 p 139

Eur. J. Phys. 6 (1985) 287–294. Printed in Northern Ireland

Vector spherical harmonics and their application to magnetostatics

R G Barrera†, G A Estévez‡ and J Giraldo§||

† Institute of Physics, University of Mexico, Mexico 20, DF, Mexico
‡ Department of Mathematics, Chemistry, Physics and Computer Science, Inter American University, San Germán, PR 00753, USA
§ Departamento de Física, Universidad Nacional de Colombia, Bogotá, DE, Colombia

Received 29 June 1984, in final form 27 March 1985

Abstract An alternative and somewhat systematic definition of the vector spherical harmonics, in analogy with the commonly used scalar spherical harmonics, is presented. The new set of vector spherical harmonics satisfies the properties of orthogonality and completeness, and is compared with other existing definitions of vector spherical harmonics. Some applications to problems in magnetostatics are illustrated.

Zusammenfassung Eine alternative und mehr systematische Definition vektorieller Kugelfunktionen, analog zu den gewöhnlich verwendeten skalaren Kugelfunktionen wird vorgestellt. Der neue Satz vektorieller Kugelfunktionen erfüllt die Orthogonalitäts- und Vollständigkeitsbedingungen: er wird verglichen mit vorhandenen anderen Definitionen vektorieller Kugelfunktionen und zur Veranschaulichung auf Probleme der Magnetostatik angewendet.

1. Introduction

Vector spherical harmonics (VSH) have been used in the expansion of plane waves to study the absorption and scattering of light by a sphere (see, for example, Bohren and Huffman 1983). They have also been widely used in nuclear and atomic physics (see, for example, Blatt and Weisskopf 1978).

The definitions of the various existing sets of VSH in different fields of physics are often dictated by convenience. For example, one method of defining such sets makes use of an operator which is proportional to the usual orbital angular momentum operator of quantum mechanics. When this operator acts upon the scalar spherical harmonics (SSH) function, it generates one (out of a triad of) VSH. The purpose of this note is to develop an alternate set of VSH which is particularly useful in classical electrodynamics. An alternative simple magnetostatic multipole moments and related electromagnetic problems.

2. Review of scalar spherical harmonics

This review is intended primarily to define notation and underscore the parallel between the use of SSH and the VSH to be introduced in §3. Familiarity with the properties and uses of the SSH at the level of development presented in the standard electromagnetism text of Jackson (1975) will be assumed. Where possible we will follow the notation of Jackson.

A crucial property of the SSH, $Y_{lm}(\theta, \phi)$, is the completeness or closure relation, i.e. any arbitrary function g of θ, ϕ can be expanded as

where $d\Omega = \sin\theta\, d\theta\, d\phi$ and the integral $\int d\Omega$ is over the whole range of angles θ, ϕ. The coefficients are given by

$$A_{lm} = \int d\Omega\, Y^*_{lm}(\theta, \phi) g(\theta, \phi). \tag{2.3}$$

The evaluation of the coefficients is often simplified using the symmetries of the SSH, namely,

$$\begin{aligned} Y_{lm}(\theta, \phi + \pi) &= (-1)^m Y_{lm}(\theta, \phi) \\ Y_{lm}(\pi - \theta, \phi) &= (-1)^{l+m} Y_{lm}(\theta, \phi) \\ Y_{lm}(\pi - \theta, \phi + \pi) &= (-1)^l Y_{lm}(\theta, \phi). \end{aligned} \tag{2.4}$$

One of the reasons why the SSH are useful in physics is that they behave in an exemplary way when operated upon by the Laplacian ∇^2:

$$\nabla^2 Y_{lm} = -\frac{l(l+1)}{r^2} Y_{lm}. \tag{2.5}$$

The simplifications that can be achieved with spherical harmonic expansions are evident in the Poisson equation $\nabla^2 \Phi_E = -4\pi\rho$. Expanding both Φ_E and ρ,

$$\Phi_E(r, \theta, \phi) = \sum_l \sum_m C_{lm}(r) Y_{lm}(\theta, \phi) \tag{2.6}$$

$$\rho(r, \theta, \phi) = \sum_l \sum_m \rho_{lm}(r) Y_{lm}(\theta, \phi) \tag{2.7}$$

the Poisson equation becomes

$$\begin{aligned} \nabla^2 \Phi_E &= \sum\sum \nabla^2 C_{lm} Y_{lm} = \sum\sum (\nabla^2 C_{lm}) Y_{lm} \\ &\quad + \sum\sum C_{lm} \nabla^2 Y_{lm} \\ &= \sum\sum \left[\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} C_{lm}\right) - \frac{l(l+1)}{r^2}\right] Y_{lm} \\ &= -4\pi\rho = -4\pi \sum\sum \rho_{lm} Y_{lm}. \end{aligned} \tag{2.8}$$

Since the coefficients of the expansion for $\nabla^2\Phi_E$ must match the coefficient for ρ we are left with

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} C_{lm}\right) - \frac{l(l+1)}{r^2} C_{lm} = -4\pi\rho_{lm}. \tag{2.9}$$

The angular dependences have in effect cancelled out and we need deal only with an ordinary differential equation.

To achieve a general solution of the differential equation given by equation (2.9) we start by considering the very special case $\rho_{lm} = \delta(r - r')$ and the differential equation

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} f(r)\right) - \frac{l(l+1)}{r^2} f(r) = -4\pi\delta(r' - r). \tag{2.10}$$

For $r > r'$ or $r < r'$ the right-hand side of equation (2.10) vanishes and the solution is simple:

$$\begin{aligned} f &= Cr^l & r &< r' \\ f &= Dr^{-(l+1)} & r &> r'. \end{aligned} \tag{2.11}$$

Notice that we have chosen only solutions that are well behaved at the origin and at infinity. Continuity of f at $r = r'$ requires that $D = C(r')^{2l+1}$. To obtain the remaining coefficient, C, the differential equation must be integrated over an infinitesimal range of r from $r = r' - \varepsilon$ to $r = r' + \varepsilon$. This yields $C = [4\pi/(2l+1)]r'^{l+1}$. Details of the procedure to obtain C are to be found in Jackson (1975, §3.9). Introducing the usual notation

$$(r_<, r_>) = \begin{cases} (r, r') & \text{for } r > r' \\ (r', r) & \text{for } r < r' \end{cases} \tag{2.12}$$

the solution, i.e. the Green function, can be written

$$f(r, r') = \frac{4\pi}{2l+1} \frac{r_<^l}{r_>^{l+1}} (r')^2. \tag{2.13}$$

To obtain the solution to equation (2.9) we need only recognise that any ρ_{lm} can be written as a superposition of delta functions. As is well known

$$\rho_{lm}(r) = \int \rho_{lm}(r')\delta(r - r')\, dr'. \tag{2.14}$$

Therefore the solution to equation (2.9) must be a similar superposition of fs as given by equation (2.13):

$$C_{lm}(r) = \int \rho_{lm}(r') \frac{4\pi}{2l+1} \frac{r_<^l}{r_>^{l+1}} (r')^2\, dr'. \tag{2.15}$$

If we now require that the charge distribution ρ be bounded in extent, and we ask explicitly for the value of C_{lm} at a value of r outside the source, we have $r = r_>$ in the integrand and

$$C_{lm}(r) = \frac{4\pi}{(2l+1)r^{l+1}} \int \rho_{lm}(r') r'^{l+2}\, dr'. \tag{2.16}$$

The quantities q_{lm},

$$\begin{aligned} q_{lm} &\equiv \int \rho_{lm}(r) r^{l+2}\, dr \\ &= \int \rho(r, \theta, \phi) Y^*_{lm}(\theta, \phi) r^l\, d^3x \end{aligned} \tag{2.17}$$

($d^3x \equiv r^2\, dr\, d\Omega$) which are characteristic of the charge distribution are called its *multipole moments*, and the expression that results from substituting equations (2.16) and (2.17) back in equation (2.6),

$$\Phi_E = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \frac{4\pi}{2l+1} \frac{q_{lm}}{r^{l+1}} Y_{lm}(\theta, \phi) \tag{2.18}$$

VECTOR SPHERICAL HARMONICS AND THEIR APPLICATION TO ELECTROMAGNETISM

Notes by Richard Price and Tracy Steelhammer

I. REVIEW OF (SCALAR) SPHERICAL HARMONICS

The reader is assumed to have studied ordinary spherical harmonics and to have a familiarity with their properties and uses on the level, e.g. of the development given in "Classical Electrodynamics" by J. D. Jackson. (We use Jackson's notational conventions for the ordinary spherical harmonics.) The very brief review here serves only to define notation and underscore the parallels between the use of ordinary spherical harmonics and the vector harmonics to be introduced in the next section.

A crucial property of the spherical harmonics $Y_{\ell m}(\theta,\phi)$ is their completeness, i.e. any function F of θ,ϕ can be expanded as

$$F(\theta,\phi) = \sum F_{\ell m} Y_{\ell m}(\theta,\phi) \tag{1.1}$$

(We adopt the notation here and throughout that an unadorned summation sign $\sum$ is to be read as $\sum_{\ell=0}^{\infty} \sum_{m=-\ell}^{+\ell}$). If F is a function of other variables, e.g. r and t, then the expansion coefficients $F_{\ell m}$ are functions of these additional variables. The computation of the expansion coefficients is made relatively simple by the orthogonality condition

$$\int Y^*_{\ell m}(\theta,\phi) Y_{\ell' m'}(\theta,\phi)\, d\Omega = \delta_{\ell\ell'}\delta_{mm'} \tag{1.2}$$

(Notation: $d\Omega \equiv \sin\theta\, d\theta\, d\phi$ and the integral $\int d\Omega$ is assumed to be over the whole range of angles θ,ϕ) so that the coefficients are given by

$$F_{\ell m} = \int Y^*_{\ell m}(\theta,\phi) F\, d\Omega \tag{1.3}$$

The symmetries of the $Y_{\ell m}$'s :

$$Y_{\ell m}(\theta, \phi+\pi) = (-1)^m Y_{\ell m}(\theta, \phi)$$
$$Y_{\ell m}(\pi-\theta, \phi) = (-1)^{\ell+m} Y_{\ell m}(\theta, \phi)$$
$$Y_{\ell m}(\pi-\theta, \phi+\pi) = (-1)^{\ell} Y_{\ell m}(\theta, \phi) \quad , \qquad (1.4)$$

often simplify the evaluations of the coefficients.

The spherical harmonics are useful in physics because they behave in an exemplary way under the influence of the Laplacian operator ∇^2 :

$$\nabla^2 Y_{\ell m} = -\frac{\ell(\ell+1)}{r^2} Y_{\ell m} \qquad (1.5)$$

The simplifications that can be achieved with spherical harmonic expansions are evident in the Poisson equation $\nabla^2\Phi = -4\pi\rho$. If we expand both Φ and ρ,

$$\Phi(r,\theta,\phi) = \sum C_{\ell m}(r) Y_{\ell m}(\theta,\phi) \qquad (1.6)$$
$$\rho(r,\theta,\phi) = \sum \rho_{\ell m}(r) Y_{\ell m}(\theta,\phi) \qquad (1.7)$$

then the Poisson equation becomes

$$\nabla^2\Phi = \sum \nabla^2 C_{\ell m} Y_{\ell m} = \sum \{\nabla^2 C_{\ell m}\} Y_{\ell m} + \sum C_{\ell m} \nabla^2 Y_{\ell m}$$
$$= \sum \left\{ \frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} C_{\ell m}\right) - \frac{\ell(\ell+1)}{r^2} \right\} Y_{\ell m} = -4\pi\rho = -4\pi \sum \rho_{\ell m} Y_{\ell m} \qquad (1.8)$$

Since the coefficients of the expansion for $\nabla^2\Phi$ must match the coefficient for ρ we are left with

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} C_{\ell m}\right) - \frac{\ell(\ell+1)}{r^2} C_{\ell m} = -4\pi \rho_{\ell m} \quad . \qquad (1.9)$$

The angular dependences have in effect cancelled out and we need deal only with an ordinary differential equation.

The differential equation (1.9) is of a type we will encounter later; it warrants some comment. To achieve a general solution to Eq. (1.9) we start by considering the very special case $\rho_{\ell m} = \delta(r-r')$ and the differential equation

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d}{dr} f(r)\right) - \frac{\ell(\ell+1)}{r^2} f(r) = -4\pi\,\delta(r-r') \, . \qquad (1.10)$$