

# MINKOWSKI SPACETIME AND SPECIAL RELATIVITY

Gregory L Naber

California State University, Chico, USA

## Introduction

Minkowski spacetime is generally regarded as the appropriate mathematical context within which to formulate those laws of physics that do not refer specifically to gravitational phenomena. Here we shall describe this context in rigorous terms, postulate what experience has shown to be its correct physical interpretation, and illustrate by means of examples its appropriateness for the formulation of physical laws.

## Minkowski Spacetime and the Lorentz Group

**Minkowski spacetime**  $\mathcal{M}$  is a 4-dimensional real vector space on which is defined a bilinear form  $\mathbf{g} : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$  that is symmetric ( $\mathbf{g}(v, w) = \mathbf{g}(w, v)$  for all  $v, w \in \mathcal{M}$ ), nondegenerate ( $\mathbf{g}(v, w) = 0$  for all  $w \in \mathcal{M}$  implies  $v = 0$ ) and of index 1 (there exists a basis  $\{e_1, e_2, e_3, e_4\}$  for  $\mathcal{M}$  with

$$\mathbf{g}(e_a, e_b) = \eta_{ab} = \begin{cases} 1 & \text{if } a = b = 1, 2, 3 \\ -1 & \text{if } a = b = 4 \\ 0 & \text{if } a \neq b \end{cases} .$$

$\mathbf{g}$  is called a **Lorentz inner product** for  $\mathcal{M}$  and any basis of the type just described is an **orthonormal basis** for  $\mathcal{M}$ . We shall often write  $v \cdot w$  for the value  $\mathbf{g}(v, w)$  of  $\mathbf{g}$  on  $(v, w) \in \mathcal{M} \times \mathcal{M}$ . A vector  $v \in \mathcal{M}$  is said to be **spacelike**, **timelike**, or **null** if  $v \cdot v$  is positive, negative, or zero, respectively, and the set  $C_N$  of all null vectors is called the **null cone** in  $\mathcal{M}$ .

If  $\{e_1, e_2, e_3, e_4\}$  is an orthonormal basis and if we write  $v = v^1 e_1 + v^2 e_2 + v^3 e_3 + v^4 e_4 = v^a e_a$  (using the Einstein summation convention, according to which a repeated index, one subscript and one superscript, is summed over its possible values) and  $w = w^b e_b$ , then

$$v \cdot w = v^1 w^1 + v^2 w^2 + v^3 w^3 - v^4 w^4 = \eta_{ab} v^a w^b .$$

In particular,  $v$  is null if and only if

$$(v^4)^2 = (v^1)^2 + (v^2)^2 + (v^3)^2$$

(hence the name null “cone” for  $C_N$ ). Timelike vectors are “inside” the null cone and spacelike vectors are “outside”.

⟨ Figure 1 here ⟩

We select some orientation for the vector space  $\mathcal{M}$  and will henceforth consider only oriented, orthonormal bases for  $\mathcal{M}$ . From the Schwartz Inequality for  $\mathbb{R}^3$  one can show (Theorem 1.3.1, [3]) that, if  $v$  is timelike and  $w$  is either timelike or null and nonzero, then  $v \cdot w < 0$  if and only if  $v^4 w^4 > 0$  in any orthonormal basis. In particular, one can define an equivalence relation on the set of all timelike vectors by decreeing that two such,  $v$  and  $w$ , are equivalent if and only if  $v \cdot w < 0$ . For reasons that will emerge shortly we then say that  $v$  and  $w$  have the same **time orientation**. There are precisely two equivalence classes, one of which we select and designate **future directed**. Timelike vectors in the other class are then called **past directed**. One can show (Section

1.3 and Corollary 1.4.5, [3]) that this classification can be extended to nonzero null vectors as well (but *not* to spacelike vectors). We will call an oriented, orthonormal basis **time oriented** if its timelike vector  $e_4$  is future directed and will consider only these in what follows. An oriented, time oriented, orthonormal basis for  $\mathcal{M}$  will be called an **admissible basis**. If  $\{e_1, e_2, e_3, e_4\}$  and  $\{\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4\}$  are two such bases and if we write

$$\begin{aligned} e_b &= \Lambda^1_b \hat{e}_1 + \Lambda^2_b \hat{e}_2 + \Lambda^3_b \hat{e}_3 + \Lambda^4_b \hat{e}_4 \\ &= \Lambda^a_b \hat{e}_a, \quad b = 1, 2, 3, 4, \end{aligned} \quad (1)$$

then the matrix  $\Lambda = (\Lambda^a_b)$  ( $a =$  row index,  $b =$  column index) can be shown to satisfy the following three conditions (Section 1.3, [3]).

1. (Orthogonality)  $\Lambda^T \eta \Lambda = \eta$

where  $T$  means transpose and

$$\eta = (\eta_{ab}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

2. (Orientability)  $\det \Lambda = 1$ .

3. (Time Orientability)  $\Lambda^4_4 \geq 1$ .

We shall refer to any  $4 \times 4$  matrix  $\Lambda = (\Lambda^a_b)$  satisfying these three conditions as a **Lorentz transformation** (although one often sees the adjectives **proper** and **orthochronous** appended to emphasize conditions #2 and #3, respectively). The set  $\mathcal{L}$  of all such matrices forms a group under matrix multiplication that we call simply the **Lorentz group**. It is a simple matter to show (Lemma 1.3.4, [3]) from the orthogonality condition #1 that, if  $\Lambda^4_4 = 1$ , then  $\Lambda$  must be of the form

$$\begin{pmatrix} & & & 0 \\ (R^i_j) & & & 0 \\ & & & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where  $(R^i_j)$  is an element of  $SO(3)$ , i.e., a  $3 \times 3$  orthogonal matrix with determinant 1. The set  $\mathcal{R}$  of

all matrices of this form is a subgroup of  $\mathcal{L}$  called the **rotation subgroup**. Although it will play no role in what we do here, it should be pointed out that in many applications (e.g., in particle physics) it is necessary to consider the larger group of transformations of  $\mathcal{M}$  generated by the Lorentz group and spacetime translations ( $x^a \rightarrow x^a + \Lambda^a$ , for some constants  $\Lambda^a$ ,  $a = 1, 2, 3, 4$ ). This is called the **inhomogeneous Lorentz group**, or **Poincaré group**.

## Physical Interpretation

For the purpose of describing how one is to think of Minkowski spacetime and the Lorentz group physically it will be convenient to distinguish (intuitively and terminologically, if not mathematically) between a “vector” in  $\mathcal{M}$  and a “point” in  $\mathcal{M}$  (the “tip” of a vector). The points in  $\mathcal{M}$  are called **events** and are to be thought of as actual physical occurrences, albeit idealized as “point-events” which have no spatial extension and no duration. One might picture, for example, an instantaneous collision, or explosion, or an “instant” in the history of some point material particle or photon (“particle of light”).

Events are observed and identified by the assignment of coordinates. We will be interested in coordinates assigned in a very particular way by a very particular type of observer. Specifically, our **admissible observers** preside over three-dimensional, right-handed, Cartesian spatial coordinate systems, relative to which photons always move along straight lines in any direction. With a single clock located at the origin, such an observer can determine the speed  $c$  of light *in vacuo* by the so-called **Fizeau procedure** (emit a photon from the origin when the clock there reads  $t_1$ , bounce it back from a mirror located at  $(x^1, x^2, x^3)$ , receive the photon at the origin again when the clock there reads  $t_2$  and set  $c = 2\sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2} / (t_2 - t_1)$ ). Now place an identical clock at each spatial point and synchronize them by emitting from the origin a spherical electromagnetic wave (photons in all directions) and setting the clock whose location is  $(x^1, x^2, x^3)$  to read  $\sqrt{(x^1)^2 + (x^2)^2 + (x^3)^2} / c$  at the instant

the wave arrives. An observer now assigns to an event the three spatial coordinates of the location at which it occurred in his coordinate system as well as the time reading on the clock at that location at the instant the event occurred. We shall assume also that our admissible observers are **inertial** in the sense of Newtonian mechanics (the trajectory of a particle on which no forces act, when described in terms of the coordinates just introduced, is a point or a straight line traversed at constant speed). It is an experimental fact (and quite a remarkable one) that all of these admissible observers (whether or not they are in relative motion) agree on the numerical value of the speed of light *in vacuo* ( $c \approx 3.00 \times 10^{10}$  cm/sec). We shall exploit this fact at the outset to have all of our admissible observers measure time in units of distance by simply multiplying their time coordinates  $t$  by  $c$ . The resulting time coordinate is denoted  $x^4 = ct$ . In these units all speeds are dimensionless and the speed of light *in vacuo* is 1.

In our mathematical model  $\mathcal{M}$  of the world of events this very subtle and complex notion of an admissible observer is fully identified with the conceptually very simple notion of an admissible basis  $\{e_1, e_2, e_3, e_4\}$ . If  $x \in \mathcal{M}$  is an event and if we write  $x = x^a e_a$ , then  $(x^1, x^2, x^3)$  are the spatial and  $x^4$  is the time coordinate supplied  $x$  by the corresponding observer. If  $\{\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4\}$  is another basis/observer related to  $\{e_1, e_2, e_3, e_4\}$  by (1) and if we write  $x = \hat{x}^a \hat{e}_a$ , then

$$\hat{x}^a = \Lambda^a_b x^b, \quad a = 1, 2, 3, 4. \quad (2)$$

Thus, Lorentz transformations relate the space and time coordinates supplied any given event by two admissible observers. If  $(\Lambda^a_b) \in \mathcal{R}$ , then the two observers differ only in the orientation of their spatial coordinate axes. On the other hand, for any real number  $\theta$  one can define an element  $L(\theta)$  of  $\mathcal{L}$  by

$$L(\theta) = \begin{pmatrix} \cosh \theta & 0 & 0 & -\sinh \theta \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sinh \theta & 0 & 0 & \cosh \theta \end{pmatrix} \quad (3)$$

and, if two admissible bases are related by this

Lorentz transformation, then the coordinate transformation (2) becomes

$$\begin{aligned} \hat{x}^1 &= (\cosh \theta) x^1 - (\sinh \theta) x^4 \\ \hat{x}^2 &= x^2 \\ \hat{x}^3 &= x^3 \\ \hat{x}^4 &= -(\sinh \theta) x^1 + (\cosh \theta) x^4 \end{aligned} \quad (4)$$

Letting  $\beta = \tanh \theta$  (so that  $-1 < \beta < 1$ ) and suppressing  $\hat{x}^2 = x^2$  and  $\hat{x}^3 = x^3$  one obtains

$$\begin{aligned} \hat{x}^1 &= \frac{1}{\sqrt{1-\beta^2}} x^1 - \frac{\beta}{\sqrt{1-\beta^2}} x^4 \\ \hat{x}^4 &= -\frac{\beta}{\sqrt{1-\beta^2}} x^1 + \frac{1}{\sqrt{1-\beta^2}} x^4 \end{aligned} \quad (5)$$

This corresponds to two observers whose spatial axes are oriented as shown in Figure 2 with the hatted coordinate system moving along the common  $x^1$ -,  $\hat{x}^1$ -axis with speed  $|\beta|$ , to the right if  $\beta > 0$  and to the left if  $\beta < 0$ .

(Figure 2 here)

We remark that, reverting to traditional time units,  $\beta = \frac{v}{c}$ , where  $|v|$  is the relative speed of the two coordinate systems, and (5) becomes what is generally referred to as a ‘‘Lorentz transformation’’ in elementary expositions of special relativity, i.e.,

$$\begin{aligned} \hat{x}^1 &= \frac{x^1 - vt}{\sqrt{1 - v^2/c^2}} \\ \hat{t} &= \frac{t - \frac{v}{c^2} x^1}{\sqrt{1 - v^2/c^2}} \end{aligned} \quad (6)$$

There is a sense in which, to understand the kinematic effects of special relativity, it is enough to restrict one’s attention to the so-called **special Lorentz transformations**  $L(\theta)$ . Specifically, one can show (Theorem 1.3.5, [3]) that if  $\Lambda \in \mathcal{L}$  is any Lorentz transformation, then there exists a real number  $\theta$  and two rotations  $R_1, R_2 \in \mathcal{R}$  such that  $\Lambda = R_1 L(\theta) R_2$ . Since  $R_1$  and  $R_2$  involve no relative motion, all of the kinematics is contained in  $L(\theta)$ .

We shall explore these kinematic effects in more detail shortly.

Now suppose that  $x$  and  $x_0$  are two distinct events in  $\mathcal{M}$  and consider the displacement vector  $x - x_0$  from  $x_0$  to  $x$ . If  $\{e_1, e_2, e_3, e_4\}$  is an admissible basis and if we write  $x = x^a e_a$  and  $x_0 = x_0^a e_a$ , then  $x - x_0 = (x^a - x_0^a)e_a = \Delta x^a e_a$ . If  $x - x_0$  is null, then

$$(\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2 = (\Delta x^4)^2$$

so the spatial separation of the two events is equal to the distance light would travel during the time lapse between the events. The same must be true in any other admissible basis since Lorentz transformations are the matrices of linear maps that preserve the Lorentz inner product. Consequently, all admissible observers agree that  $x_0$  and  $x$  are “connectible by a photon”. They even agree as to which of the two events is to be regarded as the “emission” of the photon and which is to be regarded as its “reception” since one can show (Theorem 1.3.3, [3]) that, when a vector is either timelike or null and nonzero, the sign of its fourth coordinate is the same in every admissible basis (because  $\Lambda^4_4 \geq 1$ ). Thus,  $x^4 - x_0^4$  is either positive for all admissible observers ( $x_0$  occurred before  $x$ ) or negative for all admissible observers ( $x_0$  occurred after  $x$ ). Since photons move along straight lines in admissible coordinate systems we adopt the following terminology. If  $x_0, x \in \mathcal{M}$  are such that  $x - x_0$  is null, then the straight line in  $\mathcal{M}$  containing  $x_0$  and  $x$  is called the **worldline of a photon** in  $\mathcal{M}$  and is to be thought of as the set of all events in the history of some particle of light that “experiences” both  $x_0$  and  $x$ .

Let us now suppose instead that  $x - x_0$  is timelike. Then, in any admissible basis,

$$(\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2 < (\Delta x^4)^2$$

so the spatial separation of  $x_0$  and  $x$  is less than the distance light would travel during the time lapse between the events. In this case one can prove (Section 1.4, [3]) that there exists an admissible basis  $\{\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4\}$  in which  $\Delta \hat{x}^1 = \Delta \hat{x}^2 = \Delta \hat{x}^3 = 0$ , i.e., there is an admissible observer for whom the

two events occur at the same spatial location, one after the other. Thinking of this location as occupied by some material object (e.g., the observer’s clock situated at that point) we find that the events  $x_0$  and  $x$  are both “experienced” by this material particle and that, moreover,  $\sqrt{|\mathbf{g}(x - x_0, x - x_0)|}$  is just the time lapse between the events recorded by a clock carried along by this material particle. To any other admissible observer this material particle appears “free” (not subject to forces) because it moves on a straight line with constant speed. This leads us to the following definitions. If  $x_0, x \in \mathcal{M}$  are such that  $x - x_0$  is timelike, then the straight line in  $\mathcal{M}$  containing  $x_0$  and  $x$  is called the **worldline of a free material particle** in  $\mathcal{M}$  and  $\sqrt{|\mathbf{g}(x - x_0, x - x_0)|}$ , usually written  $\tau(x - x_0)$ , or simply  $\Delta\tau$ , is the **proper time separation** of  $x_0$  and  $x$ . One can think of  $\tau(x - x_0)$  as a sort of “length” for  $x - x_0$  measured, however, by a clock carried along by a free material particle that experiences both  $x_0$  and  $x$ . It is an odd sort of length, however, since it satisfies not the usual triangle inequality, but the following “reversed” version.

**Reversed Triangle Inequality** (Theorem 1.4.2, [3]): *Let  $x_0, x$  and  $y$  be events in  $\mathcal{M}$  for which  $y - x$  and  $x - x_0$  are timelike with the same time orientation. Then  $y - x_0 = (y - x) + (x - x_0)$  is timelike and*

$$\tau(y - x_0) \geq \tau(y - x) + \tau(x - x_0) \quad (7)$$

*with equality holding if and only if  $y - x$  and  $x - x_0$  are linearly dependent.*

The sense of the inequality in (7) has interesting consequences about which we will have more to say shortly.

Finally, let us suppose that  $x - x_0$  is spacelike. Then, in any admissible basis

$$(\Delta x^1)^2 + (\Delta x^2)^2 + (\Delta x^3)^2 > (\Delta x^4)^2$$

so the spatial separation of  $x_0$  and  $x$  is greater than the distance light could travel during the time lapse

that separates them.. There is clearly no admissible observer for whom the events occur at the same location. No free material particle (or even photon) can experience both  $x_0$  and  $x$ . However, one can show (Section 1.5, [3]) that, given any real number  $T$  (positive, negative, or zero), one can find an admissible basis  $\{\hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_4\}$  in which  $\Delta\hat{x}^4 = T$ . Some admissible observers will judge the events simultaneous, some will assert that  $x_0$  occurred before  $x$ , and others will reverse the order. Temporal order, cause and effect, have no meaning for such pairs of events. For those admissible observers for whom the events are simultaneous ( $\Delta\hat{x}^4 = 0$ ), the quantity  $\sqrt{\mathbf{g}(x - x_0, x - x_0)}$  is the distance between them and for this reason this quantity is called the **proper spatial separation** of  $x_0$  and  $x$  (whenever  $x - x_0$  is spacelike).

For any two events  $x_0, x \in \mathcal{M}$ ,  $\mathbf{g}(x - x_0, x - x_0)$  is given in any admissible basis by  $(\Delta\hat{x}^1)^2 + (\Delta\hat{x}^2)^2 + (\Delta\hat{x}^3)^2 - (\Delta\hat{x}^4)^2$  and is called the **interval** separating  $x_0$  and  $x$ . It is the closest analogue in Minkowskian geometry to the (squared) length in Euclidean geometry. It can, however, assume any real value depending on the physical relationship between the events  $x_0$  and  $x$ . Historically, of course, it was the various physical interpretations of this interval that we have just described which led Minkowski [2] to the introduction of the structure that bears his name.

## Kinematic Effects

All of the well-known kinematic effects of special relativity (the addition of velocities formula, the relativity of simultaneity, time dilation and length contraction) follow easily from what we have done. Because it eases visualization and because, as we mentioned earlier, it suffices to do so, we will limit our discussion to the special Lorentz transformations.

Let  $\theta_1$  and  $\theta_2$  be two real numbers and consider the corresponding elements  $L(\theta_1)$  and  $L(\theta_2)$  of  $\mathcal{L}$  defined by (3). Sum formulas for  $\sinh \theta$  and  $\cosh \theta$  imply that  $L(\theta_1)L(\theta_2) = L(\theta_1 + \theta_2)$ . Defining  $\beta_i = \tanh \theta_i$ ,  $i = 1, 2$ , and  $\beta = \tanh(\theta_1 + \theta_2)$ , the

sum formula for  $\tanh \theta$  then gives

$$\beta = \frac{\beta_1 + \beta_2}{1 + \beta_1 \beta_2} . \quad (8)$$

The physical interpretation is simple. One has three admissible observers whose spatial axes are related in the manner shown in Figure 2. If the speed of the second relative to the first is  $\beta_1$  and the speed of the third relative to the second is  $\beta_2$ , then the speed of the third relative to the first is not  $\beta_1 + \beta_2$  as one's Newtonian predisposition would lead one to expect, but rather  $\beta$ , given by (8). This is the **relativistic addition of velocities formula**.

We have seen already that, when the interval between  $x_0$  and  $x$  is spacelike, the events will be judged simultaneous by some admissible observers, but not by others. Indeed, if  $\Delta x^4 = 0$  and the observers are related by (5), then  $\Delta\hat{x}^4 = -\frac{\beta}{\sqrt{1-\beta^2}} \Delta x^1 = -\beta \Delta\hat{x}^1$  which will not be zero unless  $\beta = 0$  and so there is no relative motion ( $\Delta\hat{x}^1$  cannot be zero since then  $\Delta\hat{x}^a = 0$  for  $a = 1, 2, 3, 4$  and  $x = x_0$ ). This phenomenon is called the **relativity of simultaneity** and we now construct a simple geometrical representation of it.

Select two perpendicular lines in the plane to represent the  $x^1$ - and  $x^4$ -axes (the Euclidean orthogonality of the lines has no physical significance and is unnecessary, but makes the pictures easier to draw). The  $\hat{x}^1$ -axis will be represented by the straight line  $\hat{x}^4 = 0$  which, from (5), is given by  $x^4 = \beta x^1$  (in Figure 3 we have assumed that  $\beta > 0$ ). Similarly, the  $\hat{x}^4$ -axis is identified with the line  $x^4 = \frac{1}{\beta} x^1$ . Since Lorentz transformations leave the Lorentz inner product invariant, the hyperbolas  $(x^1)^2 - (x^4)^2 = k$  coincide with  $(\hat{x}^1)^2 - (\hat{x}^4)^2 = k$  and we calibrate the axes accordingly, e.g., the branch of  $(x^1)^2 - (x^4)^2 = 1$  with  $x^1 > 0$  intersects the  $x^1$ -axis at the point  $(x^1, x^4) = (1, 0)$  and intersects the  $\hat{x}^1$ -axis at the point  $(\hat{x}^1, \hat{x}^4) = (1, 0)$ . This necessitates a different scale on the hatted and unhatted axes, but one can show (Section 1.3, [3]) that, with this calibration, all coordinates can be obtained geometrically by projecting parallel to the opposite axis (e.g., the  $x^4$ - and  $\hat{x}^4$ -coordinates of an event result from pro-

jecting parallel to the  $x^1$  - and  $\hat{x}^1$  - axes, respectively).

⟨ Figure 3 here ⟩

Thus, a **line of simultaneity** in the hatted (respectively, unhatted) coordinates is parallel to the  $\hat{x}^1$  - (respectively,  $x^1$  -) axis so that, in general, a pair of events lying on one will not lie on the other (note, however, that these lines are “really” 3-dimensional hyperplanes so what appears to be a point of intersection is actually a 2-dimensional “plane of agreement”, any two events in which are judged simultaneous by both observers).

For any two events whatsoever the relationship between the time lapse  $\Delta\hat{x}^4$  in the hatted coordinates and the time lapse  $\Delta x^4$  in the unhatted coordinates is, from (5),

$$\Delta\hat{x}^4 = -\frac{\beta}{\sqrt{1-\beta^2}} \Delta x^1 + \frac{1}{\sqrt{1-\beta^2}} \Delta x^4$$

so the two are generally not equal. Consider, in particular, two events on the worldline of a point *at rest* in the unhatted coordinate system, e.g., two readings on the clock at rest at the origin in this system. Then  $\Delta x^1 = 0$  so

$$\Delta\hat{x}^4 = \frac{1}{\sqrt{1-\beta^2}} \Delta x^4 > \Delta x^4 .$$

This effect is entirely symmetrical since, if  $\Delta\hat{x}^1 = 0$ , then (5) implies

$$\Delta x^4 = \frac{1}{\sqrt{1-\beta^2}} \Delta\hat{x}^4 > \Delta\hat{x}^4 .$$

Each observer judges the other’s clocks to be running slow. This phenomenon is called **time dilation** and is clearly visible in the spacetime diagram in Figure 4 (e.g., both observers agree on the time reading “0” for the clock at the origin of the unhatted system, but the line  $\hat{x}^4 = 1$  intersects the worldline of the clock, i.e., the  $x^4$  - axis, at a point *below*  $(x^1, x^4) = (0, 1)$ ).

⟨ Figure 4 here ⟩

We should emphasize that this phenomenon is quite “real” in the physical sense. For example, certain types of elementary particles (mesons) found in cosmic radiation are so short-lived (at rest) that, even if they could travel at the speed of light, the time required to traverse our atmosphere would be some ten times their normal life span. They should not be able to reach the earth, but they do. Time dilation “keeps them young” in the sense that what seems a normal life time to the meson appears much longer to us.

Finally, since admissible observers generally disagree on which events are simultaneous and since the only way to measure the “length” of a moving object (say, a measuring rod) is to locate its endpoints “simultaneously”, it should come as no surprise that length, like simultaneity, and time, depends on the admissible observer measuring it. Specifically, let us consider a measuring rod lying at rest along the  $\hat{x}^1$  - axis of the hatted coordinate system. Its “length” in this coordinate system is  $\Delta\hat{x}^1$ . The worldlines of its endpoints are two straight lines parallel to the  $\hat{x}^4$  - axis. If the unhatted observer locates two events on these worldlines “simultaneously” their coordinates will satisfy  $\Delta x^4 = 0$  and, by (5)  $\Delta\hat{x}^1 = \frac{1}{\sqrt{1-\beta^2}} \Delta x^1$  so

$$\Delta x^1 = \sqrt{1-\beta^2} \Delta\hat{x}^1 < \Delta\hat{x}^1$$

and the moving measuring rod appears contracted in its direction of motion by a factor of  $\sqrt{1-\beta^2}$ . As for time dilation, this phenomenon, known as **length contraction**, is entirely symmetrical, quite real, and clearly visible in a spacetime diagram (Figure 5).

⟨ Figure 5 here ⟩

## The Relativity Principle

We have found that admissible observers can disagree about some rather startling things (whether or not two events are simultaneous, the time lapse between two events even when no one thinks they are simultaneous, and the length of a measuring rod). This would be a matter of no concern at all, of course, if one could determine, in any given

situation, who was really right. Surely two events are either simultaneous or they are not and we need only sort out which admissible observer has the correct view of the situation? Unfortunately (or fortunately, depending on one's point-of-view) this distinction between the judgements made by different admissible observers is precisely what physics forbids.

**The Relativity Principle** ([2]): *All admissible observers are completely equivalent for the formulation of the laws of physics.*

We must be clear that this is not a mathematical statement. It is rather a statement about the physical world around us and how it should be described, gleaned from observations, some of which are complex and subtle and some of which are commonplace (a passenger in a smooth, quiet airplane travelling at constant groundspeed cannot “feel” his motion relative to the earth). It is a powerful guide for constructing the laws of relativistic physics, but even more fundamentally it prohibits us from regarding any particular admissible observer as having a privileged view of the universe. In particular, we are forbidden from attaching any objective significance to such questions as, “were the two supernovae simultaneous?”, “How long did the meson survive?”, and “What is the distance between the Crab Nubula and Alpha Centauri?” This is severe, but one must deal with it.

## Particles and 4-Momentum

If  $I \subseteq \mathbb{R}$  is an interval, then a map  $\alpha : I \rightarrow \mathcal{M}$  is a **curve** in  $\mathcal{M}$ . Relative to any admissible basis we can write

$$\alpha(\xi) = x^a(\xi) e_a$$

for each  $\xi \in I$ . We shall assume that  $\alpha$  is **smooth** in the sense that each  $x^a(\xi)$ ,  $a = 1, 2, 3, 4$ , is infinitely differentiable ( $C^\infty$ ) on  $I$  and the **velocity vector**

$$\alpha'(\xi) = \frac{dx^a}{d\xi} e_a$$

is nonzero for every  $\xi \in I$  (we adopt the usual custom, in a vector space, of identifying the tangent

space at each point with the vector space itself). This definition of smoothness clearly does not depend on the choice of admissible basis for  $\mathcal{M}$ . The curve  $\alpha$  is said to be **spacelike**, **timelike**, or **null** if  $\alpha'(\xi) \cdot \alpha'(\xi) = \eta_{ab} \frac{dx^a}{d\xi} \frac{dx^b}{d\xi}$  is positive, negative, or zero, respectively, for each  $\xi \in I$ . A timelike curve  $\alpha$  for which  $\alpha'(\xi)$  is future directed for each  $\xi \in I$  is called a **timelike worldline** and its image is identified with the set of all events in the history of some (not necessarily free) point material particle. If  $I = [\xi_0, \xi_1]$  and  $\alpha : [\xi_0, \xi_1] \rightarrow \mathcal{M}$  is a timelike worldline, then the **proper time length** of  $\alpha$  is defined by

$$\begin{aligned} L(\alpha) &= \int_{\xi_0}^{\xi_1} \sqrt{|\mathbf{g}(\alpha'(\xi), \alpha'(\xi))|} d\xi \\ &= \int_{\xi_0}^{\xi_1} \sqrt{-\eta_{ab} \frac{dx^a}{d\xi} \frac{dx^b}{d\xi}} d\xi \end{aligned}$$

and interpreted as the time lapse between the events  $\alpha(\xi_0)$  and  $\alpha(\xi_1)$  as recorded by a clock carried along by the particle whose worldline is  $\alpha$ . This interpretation is easily motivated by writing out a Riemann sum approximation to the integral and appealing to our interpretation of the proper time separation  $\Delta\tau = \sqrt{-\eta_{ab} \Delta x^a \Delta x^b}$ . There are subtleties, however, both mathematical and physical (Section 1.4, [3]). The mathematical ones are addressed by the following result (which combines Theorems 1.4.6 and 1.4.8, [3]).

**Theorem:** *Let  $x_0$  and  $x$  be two events in  $\mathcal{M}$ . Then  $x - x_0$  is timelike and future directed if and only if there exists a timelike worldline  $\alpha : [\xi_0, \xi_1] \rightarrow \mathcal{M}$  in  $\mathcal{M}$  with  $\alpha(\xi_0) = x_0$  and  $\alpha(\xi_1) = x$  and, in this case,*

$$L(\alpha) \leq \tau(x - x_0) \quad (9)$$

*with equality holding if and only if  $\alpha$  is a parametrization of a timelike straight line.*

The inequality (9) asserts that if two material particles experience both  $x_0$  and  $x$ , then the one that

is free (and so can be regarded as at rest in some admissible coordinate system) has longer to wait for the occurrence of the second event (moving clocks run slow). For many years this basically obvious fact was christened “The Twin Paradox”.

Just as a smooth curve in Euclidean space has an arc length parametrization, so a timelike worldline has a **proper time parametrization** defined as follows. For each  $\xi$  in  $[\xi_0, \xi_1]$  let

$$\tau = \tau(\xi) = \int_{\xi_0}^{\xi} \sqrt{|\mathbf{g}(\alpha'(\zeta), \alpha'(\zeta))|} d\zeta$$

(the proper time length of  $\alpha$  from  $\alpha(\xi_0)$  to  $\alpha(\xi)$ ). Then  $\tau = \tau(\xi)$  has a smooth inverse  $\xi = \xi(\tau)$  so  $\alpha$  can be reparametrized by  $\tau$ . We will abuse our notation slightly and write

$$\alpha(\tau) = x^a(\tau) e_a.$$

The velocity vector with this parametrization is denoted

$$U = U(\tau) = \frac{dx^a}{d\tau} e_a,$$

called the **4-velocity** of the worldline and is the unit tangent vector field to  $\alpha$ , i.e.,

$$U(\tau) \cdot U(\tau) = -1 \quad (10)$$

for each  $\tau$ . An admissible observer is, of course, more likely to parametrize a worldline by his own time coordinate  $x^4$ . Then

$$\alpha'(x^4) = \frac{dx^1}{dx^4} e_1 + \frac{dx^2}{dx^4} e_2 + \frac{dx^3}{dx^4} e_3 + e_4$$

so

$$|\mathbf{g}(\alpha'(x^4), \alpha'(x^4))| = 1 - \|\vec{V}\|^2$$

where

$$\|\vec{V}\| = \sqrt{\left(\frac{dx^1}{dx^4}\right)^2 + \left(\frac{dx^2}{dx^4}\right)^2 + \left(\frac{dx^3}{dx^4}\right)^2}$$

is the usual magnitude of the particle’s velocity vector

$$\vec{V} = \vec{V}(x^4) = \frac{dx^1}{dx^4} e_1 + \frac{dx^2}{dx^4} e_2 + \frac{dx^3}{dx^4} e_3 = V^i e_i$$

in the given admissible coordinate system. One finds then that

$$U = \left(1 - \|\vec{V}\|^2\right)^{-\frac{1}{2}} \left(\vec{V} + e_4\right). \quad (11)$$

We shall identify a **material particle** in  $\mathcal{M}$  with a pair  $(\alpha, m)$ , where  $\alpha$  is a timelike worldline and  $m$  is a positive constant called the particle’s **proper mass** (or **rest mass**). If each  $\frac{dx^a}{d\xi}$ ,  $a = 1, 2, 3, 4$ , is constant, then  $(\alpha, m)$  is a **free material particle** with proper mass  $m$ . The **4-momentum** of  $(\alpha, m)$  is defined by  $P = mU$ . Thus,

$$P \cdot P = -m^2. \quad (12)$$

In any admissible basis we write

$$\begin{aligned} P &= P^a e_a = m U^a e_a = m \frac{dx^a}{d\tau} e_a \\ &= m \left(1 - \|\vec{V}\|^2\right)^{-\frac{1}{2}} \left(\vec{V} + e_4\right). \end{aligned} \quad (13)$$

The “spatial part” of  $P$  in these coordinates is  $\vec{P} = \frac{m}{\sqrt{1 - \|\vec{V}\|^2}} \vec{V}$  which, for  $\|\vec{V}\| \ll 1$ , is approximately  $m\vec{V}$ . Identifying  $m$  with the inertial mass of Newtonian mechanics (measured by an observer for whom the particle’s speed is small), this is simply the classical momentum of the particle. Somewhat more explicitly, if one expands  $\frac{1}{\sqrt{1 - \|\vec{V}\|^2}}$  by the

Binomial Theorem one finds that

$$\begin{aligned} P^i &= \frac{m}{\sqrt{1 - \|\vec{V}\|^2}} V^i \\ &= m V^i + \frac{1}{2} m V^i \|\vec{V}\|^2 + \dots, \quad i = 1, 2, 3, \end{aligned} \quad (14)$$

which gives the components of the classical momentum plus “relativistic corrections”. In order to preserve a formal similarity with Newtonian mechanics one often sees  $\frac{m}{\sqrt{1 - \|\vec{V}\|^2}}$  referred to as the “relativistic mass” of the particle, but we shall avoid this terminology. The fourth component of  $P$  is given by

$$\begin{aligned} P^4 &= -P \cdot e_4 \\ &= \frac{m}{\sqrt{1 - \|\vec{V}\|^2}} = m + \frac{1}{2} m \|\vec{V}\|^2 + \dots \end{aligned} \quad (15)$$

The appearance of the term  $\frac{1}{2} m \|\vec{V}\|^2$  corresponding to the Newtonian kinetic energy suggests that  $P^4$  be denoted  $E$  and called the **total relativistic energy** measured by the given admissible observer for the particle.

$$E = -P \cdot e_4 . \quad (16)$$

Now, one must understand that the concept of “energy” in physics is a subtle one and simply giving  $-P \cdot e_4$  this name does not ensure that there is any physical content. Whether or not the name is appropriate can only be determined experimentally. In particular, one should ask if the appearance of the term  $m$  in (15) is consistent with the view that  $P^4$  represents the “energy” of the particle. Observe that if  $\|\vec{V}\| = 0$  (i.e., if the particle is at rest relative to the given observer), then (15) gives

$$E = m \left( = m c^2 , \text{ in standard units } \right) , \quad (17)$$

which we interpret as saying that, even when the particle is at rest, it still has energy. If this is really “energy” in the physical sense, then it should be possible to liberate and use it. That this is, indeed, possible has, of course, been rather convincingly demonstrated.

Next we observe that not only material particles, but also photons possess “momentum” and “energy” and therefore should have 4-momentum (witness, for example, the photoelectric effect in which photons collide with and eject electrons from their orbits in an atom). Unlike a material particle, however, a photon’s characteristic feature is not proper mass, but frequency  $\nu$ , or wavelength  $\lambda = \frac{1}{\nu}$ , related to its energy  $\mathcal{E}$  by  $\mathcal{E} = h\nu$  ( $h$  being Planck’s constant) and these are highly observer dependent (Doppler effect). There is, moreover, no “proper frequency” analogous to “proper mass” since there is no admissible observer for whom the photon is at rest. In an attempt to model these features we consider a point  $x_0 \in \mathcal{M}$ , a future directed null vector  $N$  and an interval  $I \subseteq \mathbb{R}$ . The curve  $\alpha : I \rightarrow \mathcal{M}$  defined by

$$\alpha(\xi) = x_0 + \xi N \quad (18)$$

is a parametrization of the worldline of a photon through  $x_0$ . Being null,  $N$  can be written in any admissible basis as

$$N = ( -N \cdot e_4 ) \left( \vec{d} + e_4 \right) , \quad (19)$$

where

$$\begin{aligned} \vec{d} = & \left[ (N \cdot e_1)^2 + (N \cdot e_2)^2 \right. \\ & \left. + (N \cdot e_3)^2 \right]^{-\frac{1}{2}} \left[ (N \cdot e_1) e_1 \right. \\ & \left. + (N \cdot e_2) e_2 + (N \cdot e_3) e_3 \right] \end{aligned} \quad (20)$$

is the direction vector of the worldline in the corresponding spatial coordinate system. Now, by analogy with (16), we define a **photon** in  $\mathcal{M}$  to be a curve in  $\mathcal{M}$  of the form (18), take  $N$  to be its **4-momentum** and define the **energy**  $\mathcal{E}$  of the photon in the admissible basis  $\{e_1, e_2, e_3, e_4\}$  by

$$\mathcal{E} = -N \cdot e_4 . \quad (21)$$

Then, by (19),

$$N = \mathcal{E} \left( \vec{d} + e_4 \right) . \quad (22)$$

The corresponding **frequency**  $\nu$  and **wavelength**  $\lambda$  are then defined by  $\nu = \mathcal{E}/h$  and  $\lambda = 1/\nu$ . In another admissible basis one has  $N = \hat{\mathcal{E}}(\vec{\hat{d}} + \hat{e}_4)$ , where  $\vec{\hat{d}}$  and  $\hat{\mathcal{E}}$  are defined by the hatted versions of (20) and (21). One can then show (Section 1.8, [3]) that

$$\begin{aligned} \frac{\hat{\mathcal{E}}}{\mathcal{E}} = \frac{\hat{\nu}}{\nu} &= \frac{1 - \beta \cos \theta}{\sqrt{1 - \beta^2}} \\ &= (1 - \beta \cos \theta) + \frac{1}{2} \beta^2 (1 - \beta \cos \theta) + \dots , \end{aligned} \quad (23)$$

where  $\beta$  is the relative speed of the two spatial coordinate systems and  $\theta$  is the angle (in the unhatted spatial coordinate system) between the direction  $\vec{d}$  of the photon and the direction of motion of the hatted spatial coordinate system. Equation (23) is the formula for the **relativistic Doppler effect** with the first term in the series being the classical formula.

We conclude this section by examining a few simple interactions between particles of the sort modelled by our definitions, assuming only that 4-momentum is conserved in the interaction. For convenience, we will use the term **free particle** to refer to either a free material particle or a photon. If  $\mathcal{A}$  is a finite set of free particles, then each element of  $\mathcal{A}$  has a unique 4-momentum which is a future directed timelike or null vector. The sum of any such collection of vectors is timelike and future directed except when all of the vectors are null and parallel, in which case the sum is null and future directed (Lemma 1.4.3, [3]). This sum we call the **total 4-momentum of  $\mathcal{A}$** . Now we formulate a definition which is intended to model a finite set of free particles colliding at some event with a (perhaps new) set of free particles emerging from the collision (e.g., an electron and proton collide, with a neutron and neutrino emerging from the collision). A **contact interaction in  $\mathcal{M}$**  is a triple  $(\mathcal{A}, x, \tilde{\mathcal{A}})$ , where  $\mathcal{A}$  and  $\tilde{\mathcal{A}}$  are two finite sets of free particles, neither of which contains a pair of particles with linearly dependent 4-momenta (which would presumably be physically indistinguishable) and  $x \in \mathcal{M}$  is an event such that

1.  $x$  is the terminal point of all of the particles in  $\mathcal{A}$  (i.e., for each worldline  $\alpha : [\xi_0, \xi_1] \rightarrow \mathcal{M}$  of a particle in  $\mathcal{A}$ ,  $\alpha(\xi_1) = x$ ),
2.  $x$  is the initial point of all the particles in  $\tilde{\mathcal{A}}$ , and
3. the total 4-momentum of  $\mathcal{A}$  equals the total 4-momentum of  $\tilde{\mathcal{A}}$ .

Properly #3 is called the **conservation of 4-momentum**. If  $\mathcal{A}$  consists of a single free particle, then  $(\mathcal{A}, x, \tilde{\mathcal{A}})$  is called a **decay** (e.g., a neutron decays into a proton, an electron and an antineutrino).

Consider, for example, an interaction  $(\mathcal{A}, x, \tilde{\mathcal{A}})$  for which  $\tilde{\mathcal{A}}$  consists of a single photon. The total 4-momentum of  $\tilde{\mathcal{A}}$  is null so the same must be true of  $\mathcal{A}$ . Since the 4-momenta of the individual particles in  $\mathcal{A}$  are timelike or null and future directed their sum can be null only if they are, in fact, all null and parallel. Since  $\mathcal{A}$  cannot contain distinct photons with parallel 4-momenta, it must consist of

a single photon which, by #3, must have the same 4-momentum as the photon in  $\tilde{\mathcal{A}}$ . In essence, “nothing happened at  $x$ ”. We conclude that *no nontrivial interaction of the type modelled by our definition can result in a single photon and nothing else*. Reversing the roles of  $\mathcal{A}$  and  $\tilde{\mathcal{A}}$  shows that, if 4-momentum is to be conserved, *a photon cannot decay*.

Next let us consider the decay of a single material particle into two material particles, e.g., the spontaneous disintegration of an atom through  $\alpha$ -emission. Thus, we consider a contact interaction  $(\mathcal{A}, x, \tilde{\mathcal{A}})$  in which  $\mathcal{A}$  consists of a single free material particle of proper mass  $m_0$  and  $\tilde{\mathcal{A}}$  consists of two free material particles with proper masses  $m_1$  and  $m_2$ . Let  $P_0$ ,  $P_1$  and  $P_2$  be the 4-momenta of the particles of proper mass  $m_0$ ,  $m_1$  and  $m_2$ , respectively. Then  $P_0 = P_1 + P_2$ . Appealing to the Reversed Triangle Inequality, the fact that  $P_1$  and  $P_2$  are linearly independent and future directed, and (12) we conclude that

$$m_0 > m_1 + m_2 . \quad (23)$$

The excess mass  $m_0 - (m_1 + m_2)$  of the initial particle is regarded, via (17), as a measure of the amount of energy required to split  $m_0$  into two pieces. Stated somewhat differently, when the two particles in  $\tilde{\mathcal{A}}$  were held together to form the single particle in  $\mathcal{A}$ , the “binding energy” contributed to the mass of this latter particle.

Reversing the roles of  $\mathcal{A}$  and  $\tilde{\mathcal{A}}$  in the last example gives a contact interaction modelling an **inelastic collision** (two free material particles with masses  $m_1$  and  $m_2$  collide and coalesce to form a third of mass  $m_0$ ). The inequality (23) remains true, of course, and a somewhat more detailed analysis (Section 1.8, [3]) yields an approximate formula for  $m_0 - (m_1 + m_2)$  which can be compared (favorably) with the Newtonian formula for the loss in kinetic energy that results from the collision (energy which, classically, is viewed as taking the form of heat in the combined particle). An analysis of the interaction in which both  $\mathcal{A}$  and  $\tilde{\mathcal{A}}$  consist of an electron and a photon yields (Section 1.8, [3]) a formula for the so-called **Compton effect**. Many more such examples

of this sort are treated in great detail in Chapter VI, § 14 of [6].

## Charged Particles and Electromagnetic Fields

A **charged particle** in  $\mathcal{M}$  is a triple  $(\alpha, m, q)$ , where  $(\alpha, m)$  is a material particle and  $q$  is a nonzero real number called the **charge** of the particle. Charged particles do two things of interest to us. By their very presence they create electromagnetic fields and they also respond to the electromagnetic fields created by other charges.

Charged particles “respond” to an electromagnetic field by experiencing changes in 4-momentum. The quantitative nature of this response, i.e., the **equation of motion**, is generally taken to be the so-called **Lorentz 4-Force Law** which expresses the proper time rate of change of the particle’s 4-momentum at each point of the worldline as a *linear* function of the 4-velocity. Thus, at each point  $\alpha(\tau)$  of the worldline

$$\frac{dP(\tau)}{d\tau} = q \tilde{F}_{\alpha(\tau)}(U(\tau)) \quad (24)$$

where  $\tilde{F}_{\alpha(\tau)} : \mathcal{M} \rightarrow \mathcal{M}$  is a linear transformation determined, in each admissible coordinate system, by the classical electric  $\vec{E}$  and magnetic  $\vec{B}$  fields (here we are assuming that the contribution of  $q$  to the ambient electromagnetic field is negligible, i.e.,  $(\alpha, m, q)$  is a “test charge”). Let us write (24) more simply as

$$\tilde{F}(U) = \frac{m}{q} \frac{dU}{d\tau}. \quad (25)$$

Dotting both sides of (25) with  $U$  gives

$$\begin{aligned} \tilde{F}(U) \cdot U &= \frac{m}{q} \frac{dU}{d\tau} \cdot U = \frac{m}{2q} \frac{d}{d\tau} (U \cdot U) \\ &= \frac{m}{2q} \frac{d}{d\tau} (-1) = 0. \end{aligned}$$

Since any future directed timelike unit vector  $u$  is the 4-velocity of some charged particle, we find that  $\tilde{F}(u) \cdot u = 0$  for any such vector. Linearity then

implies  $\tilde{F}(v) \cdot v = 0$  for any timelike vector. Now, if  $u$  and  $v$  are timelike and future directed, then  $u+v$  is timelike so  $0 = \tilde{F}(u+v) \cdot (u+v) = \tilde{F}(u) \cdot v + u \cdot \tilde{F}(v)$  and therefore  $\tilde{F}(u) \cdot v = -u \cdot \tilde{F}(v)$ . But  $\mathcal{M}$  has a basis of future directed timelike vectors so

$$\tilde{F}(x) \cdot y = -x \cdot \tilde{F}(y) \quad (26)$$

for all  $x, y \in \mathcal{M}$ . Thus, at each point, the linear transformation  $\tilde{F}$  must be *skew-symmetric with respect to the Lorentz inner product*. One could therefore model an electromagnetic field on  $\mathcal{M}$  by an assignment to each point of a skew-symmetric linear transformation whose job it is to assign to the 4-velocity of a charged particle whose worldline passes through that point the change in 4-momentum that the particle should expect to experience because of the presence of the field. However, a slightly different perspective has proved more convenient. Notice that a skew-symmetric linear transformation  $\tilde{F} : \mathcal{M} \rightarrow \mathcal{M}$  and the Lorentz inner product together determine a bilinear form  $F : \mathcal{M} \times \mathcal{M} \rightarrow \mathbb{R}$  given by

$$F(x, y) = \tilde{F}(x) \cdot y$$

which is also skew-symmetric ( $F(y, x) = \tilde{F}(y) \cdot x = -F(x, y)$ ) and that, conversely, a skew-symmetric bilinear form uniquely determines a skew-symmetric linear transformation. Now, an assignment of a skew-symmetric bilinear form to each point of  $\mathcal{M}$  is nothing other than a *2-form* on  $\mathcal{M}$  and it is in the language of forms that we choose to phrase classical electromagnetic theory (a concise introduction to this language is available, for example, in Chapter 4 of [5]).

Nature imposes a certain restriction on which 2-forms can reasonably represent an electromagnetic field on  $\mathcal{M}$  (“Maxwell’s equations”). To formulate these we introduce a **source 1-form**  $J$  as follows: If  $x^1, x^2, x^3, x^4$  is any admissible coordinate system on  $\mathcal{M}$ , then

$$J = J_1 dx^1 + J_2 dx^2 + J_3 dx^3 - \rho dx^4 \quad (27)$$

where  $\rho : \mathcal{M} \rightarrow \mathbb{R}$  is a **charge density** function and  $\vec{J} = J_1 e_1 + J_2 e_2 + J_3 e_3$  is a **current density**

vector field (these are to be regarded as the usual “smoothed out”, pointwise versions of “charge per unit volume” and “charge flow per unit area per unit time” as measured by the corresponding admissible observer). Now, our formal definition is as follows: The **electromagnetic field on  $\mathcal{M}$**  determined by the source 1-form  $J$  on  $\mathcal{M}$  is a 2-form  $F$  on  $\mathcal{M}$  that satisfies **Maxwell’s equation**

$$dF = 0, \tag{28}$$

and

$$*d *F = J. \tag{29}$$

A few comments are in order here. We have chosen units in which not only the speed of light, but also various other constants that one often finds in Maxwell’s equations (the dielectric constant  $\epsilon_0$  and magnetic permeability  $\mu_0$ ) are 1 and a factor of  $4\pi$  in (29) is “normalized out”. The  $*$  in (29) is the Hodge star operator determined by the Lorentz inner product and the chosen orientation of  $\mathcal{M}$ . This is a natural isomorphism

$$* : \Omega^p(\mathcal{M}) \longrightarrow \Omega^{4-p}(\mathcal{M}), \quad p = 0, 1, 2, 3, 4$$

of the  $p$ -forms on  $\mathcal{M}$  to the  $(4-p)$ -forms on  $\mathcal{M}$  and is most simply defined as follows: Let  $x^1, x^2, x^3, x^4$  be any admissible coordinate system on  $\mathcal{M}$ . If  $1 \in \Omega^0(\mathcal{M})$  is the constant function (0-form) on  $\mathcal{M}$  whose value is  $1 \in \mathbb{R}$ , then

$$*1 = dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4$$

is the volume form on  $\mathcal{M}$ . If  $1 \leq i_1 < \dots < i_k \leq 4$ , then  $*(dx^{i_1} \wedge \dots \wedge dx^{i_k})$  is uniquely determined by

$$\begin{aligned} & (dx^{i_1} \wedge \dots \wedge dx^{i_k}) \wedge *(dx^{i_1} \wedge \dots \wedge dx^{i_k}) \\ &= -dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4. \end{aligned}$$

Thus, for example,  $*dx^2 = dx^1 \wedge dx^3 \wedge dx^4$ ,  $*(dx^1 \wedge dx^2) = -dx^3 \wedge dx^4$ ,  $*(dx^1 \wedge dx^2 \wedge dx^3 \wedge dx^4) = -1$ , etc. It follows that, if  $\mu$  is a  $p$ -form on  $\mathcal{M}$ , then

$$**\mu = (-1)^{p+1} \mu \tag{30}$$

(a more thorough discussion is available in Chapter V A3 of [1]). In particular, (29) is equivalent to

$$d *F = *J. \tag{31}$$

On regions in which there are no charges, so that  $J = 0$ , (28) and (31) become the **source free Maxwell equations**

$$dF = 0 \tag{32}$$

and

$$d *F = 0, \tag{33}$$

i.e., both  $F$  and  $*F$  are closed 2-forms.

Any 2-form  $F$  on  $\mathcal{M}$  can be written in any admissible coordinate system as  $F = \frac{1}{2} F_{ab} dx^a \wedge dx^b$  (summation convention!), where  $(F_{ab})$  is the skew-symmetric matrix of components of  $F$ . In order to make contact with the notation generally employed in physics we introduce the following names for these components:

$$(F_{ab}) = \begin{pmatrix} 0 & B^3 & -B^2 & E^1 \\ -B^3 & 0 & B^1 & E^2 \\ B^2 & -B^1 & 0 & E^3 \\ -E^1 & -E^2 & -E^3 & 0 \end{pmatrix} \tag{34}$$

Thus,

$$\begin{aligned} F &= E^1 dx^1 \wedge dx^4 + E^2 dx^2 \wedge dx^4 + E^3 dx^3 \wedge dx^4 + \\ & B^3 dx^1 \wedge dx^2 + B^2 dx^3 \wedge dx^1 + B^1 dx^2 \wedge dx^3. \end{aligned} \tag{35}$$

Computing  $*F$ ,  $dF$ ,  $d *F$  and  $*d *F$  and writing  $\vec{E} = E^1 e_1 + E^2 e_2 + E^3 e_3$  and  $\vec{B} = B^1 e_1 + B^2 e_2 + B^3 e_3$  one finds that  $dF = 0$  is equivalent to

$$\operatorname{div} \vec{B} = 0 \tag{36}$$

and

$$\operatorname{curl} \vec{E} + \frac{\partial \vec{B}}{\partial t} = \vec{0}, \tag{37}$$

while  $*d*F = J$  is equivalent to

$$\operatorname{div} \vec{E} = \rho \quad (38)$$

and

$$\operatorname{curl} \vec{B} - \frac{\partial \vec{E}}{\partial t} = \vec{J}. \quad (39)$$

Equations (36)-(39) are the more traditional renderings of Maxwell's equations.

In another admissible coordinate system  $\hat{x}^1, \hat{x}^2, \hat{x}^3, \hat{x}^4$  on  $\mathcal{M}$  (related to the first by (2)) the 2-form  $F$  would be written  $F = \frac{1}{2} \hat{F}_{ab} d\hat{x}^a \wedge d\hat{x}^b$ . Setting  $\hat{x}^a = \Lambda^a_\alpha x^\alpha$  and  $\hat{x}^b = \Lambda^b_\beta x^\beta$  gives  $F = \frac{1}{2} (\Lambda^a_\alpha \Lambda^b_\beta \hat{F}_{ab}) dx^\alpha \wedge dx^\beta$  so

$$F_{\alpha\beta} = \Lambda^a_\alpha \Lambda^b_\beta \hat{F}_{ab}, \quad \alpha, \beta = 1, 2, 3, 4. \quad (40)$$

Now, suppose that we wish to describe the electromagnetic field of a uniformly moving charge. According to the Relativity Principle, it matters not at all whether we view the charge as moving relative to a "fixed" admissible observer, or the observer as moving relative to a "stationary" charge. Thus, we shall write out the field due to a charge fixed at the origin of the hatted coordinate system ("Coulomb's Law") and transform, by (40), to an unhatted coordinate system moving relative to it. Relative to  $\hat{x}^1, \hat{x}^2, \hat{x}^3, \hat{x}^4$ , the familiar inverse square law for a fixed point charge  $q$  located at the spatial origin gives  $\vec{B} = \vec{0}$  and  $\vec{E} = \frac{q}{\hat{r}^3} \vec{\hat{r}}$ , where  $\vec{\hat{r}} = \hat{x}^1 \hat{e}_1 + \hat{x}^2 \hat{e}_2 + \hat{x}^3 \hat{e}_3$  and  $\hat{r} = ((\hat{x}^1)^2 + (\hat{x}^2)^2 + (\hat{x}^3)^2)^{\frac{1}{2}}$  (note that  $\vec{E}$  is defined only on  $\mathcal{M} - \operatorname{Span}\{\hat{e}_4\}$ ). Thus,

$$\left( \hat{F}_{ab} \right) = \frac{q}{\hat{r}^3} \begin{pmatrix} 0 & 0 & 0 & \hat{x}^1 \\ 0 & 0 & 0 & \hat{x}^2 \\ 0 & 0 & 0 & \hat{x}^3 \\ -\hat{x}^1 & -\hat{x}^2 & -\hat{x}^3 & 0 \end{pmatrix}. \quad (41)$$

It is a simple matter to verify that, on its domain,  $(\hat{F}_{ab})$  satisfies the source free Maxwell equations. Taking  $\Lambda$  to be the special Lorentz transformation

corresponding to (5) and writing out (40) with  $(\hat{F}_{ab})$  given by (41) yields

$$\begin{aligned} E^1 &= q \left( \frac{\hat{x}^1}{\hat{r}^3} \right) & B^1 &= 0 \\ E^2 &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{\hat{x}^2}{\hat{r}^3} \right) & B^2 &= \frac{-q\beta}{\sqrt{1-\beta^2}} \left( \frac{\hat{x}^3}{\hat{r}^3} \right) \\ E^3 &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{\hat{x}^3}{\hat{r}^3} \right) & B^3 &= \frac{q\beta}{\sqrt{1-\beta^2}} \left( \frac{\hat{x}^2}{\hat{r}^3} \right). \end{aligned} \quad (42)$$

We wish to express these in terms of measurements made by the unhatted observer at the instant the charge passes through his spatial origin. Setting  $x^4 = 0$  in (5) gives  $\hat{x}^1 = \frac{1}{\sqrt{1-\beta^2}} x^1$ ,  $\hat{x}^2 = x^2$  and  $\hat{x}^3 = x^3$  and so  $\hat{r}^2 = \frac{1}{1-\beta^2} (x^1)^2 + (x^2)^2 + (x^3)^2$  which, for convenience, we write  $r_\beta^2$ . Making these substitutions in (42) gives

$$\begin{aligned} \vec{E} &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{1}{r_\beta^3} \right) (x^1 e_1 + x^2 e_2 + x^3 e_3) \\ &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{1}{r_\beta^3} \right) \vec{r} \end{aligned} \quad (43)$$

and

$$\begin{aligned} \vec{B} &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{1}{r_\beta^3} \right) (0e_1 - \beta x^3 e_2 + \beta x^2 e_3) \\ &= \frac{q}{\sqrt{1-\beta^2}} \left( \frac{1}{r_\beta^3} \right) ((\beta e_1) \times \vec{r}), \end{aligned} \quad (44)$$

for the field of a charge moving uniformly with velocity  $\beta e_1$  at the instant the charge passes through the origin. Observe that when  $\beta \ll 1$ ,  $r_\beta \approx r$  so (43) says that the electric field of a slowly moving charge is approximately the Coulomb field. When  $\beta \ll 1$ , (44) reduces to the **Biot-Savart Law**.

Let us consider one other simple application, i.e., the response of a charged particle  $(\alpha, m, q)$  to an electromagnetic field which, for some admissible observer, is constant and purely magnetic. For simplicity we assume that, for this observer  $\vec{E} = \vec{0}$  and

$\vec{B} = be_3$ , where  $b$  is a nonzero constant. The corresponding 2-form  $F$  has components

$$(F_{ab}) = \begin{pmatrix} 0 & b & 0 & 0 \\ -b & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

(from (34)). The corresponding linear transformation  $\tilde{F}$  has the same matrix relative to this basis so, with  $\alpha(\tau) = x^a(\tau)e_a$  and  $U(\tau) = U^a(\tau)e_a$ , the Lorentz 4-Force Law (25) reduces to the system of linear differential equations

$$\begin{cases} \frac{dU^1}{d\tau} = \frac{bq}{m} U^2 \\ \frac{dU^2}{d\tau} = -\frac{bq}{m} U^1 \end{cases} \quad \begin{cases} \frac{dU^3}{d\tau} = 0 \\ \frac{dU^4}{d\tau} = 0 \end{cases}.$$

The system is easily solved and the results easily integrated to give

$$\begin{aligned} \alpha(\tau) = & x_0 + a \sin\left(\frac{bq\tau}{m} + \phi\right) e_1 \\ & + a \cos\left(\frac{bq\tau}{m} + \phi\right) e_2 \\ & + c\tau e_3 + \left(1 + \frac{a^2 b^2 q^2}{m^2} + c^2\right) \tau e_4 \end{aligned} \quad (45)$$

where  $x_0 = x_0^a e_a \in \mathcal{M}$  is constant and  $a, \phi$  and  $c$  are real constants with  $a > 0$  (we have used  $U \cdot U = -1$  to eliminate one other arbitrary real constant). Note that, at each point on  $\alpha$ ,  $(x^1 - x_0^1)^2 + (x^2 - x_0^2)^2 = a^2$ . Thus, if  $c \neq 0$  the spatial trajectory in this coordinate system is a helix along the  $e_3$ -direction (i.e., along the magnetic field lines). If  $c = 0$  the trajectory is a circle in the  $x^1 x^2$ -plane. This case is of some practical significance since one can introduce constant magnetic fields in a bubble chamber so as to induce a particle of interest to follow a circular path. We show now how to measure the charge-to-mass ratio for such a particle. Taking  $c = 0$  in (45) and computing  $U(\tau)$ , then using (11) to solve for the

coordinate velocity vector  $\vec{V}$  of the particle gives

$$\vec{V} = \frac{abq/m}{\sqrt{1 - \|\vec{V}\|^2}} \left( \cos\left(\frac{bq\tau}{m} + \phi\right) e_1 + \sin\left(\frac{bq\tau}{m} + \phi\right) e_2 \right).$$

From this one computes

$$\|\vec{V}\|^2 = \left(1 + \frac{m^2}{a^2 b^2 q^2}\right)^{-1}$$

(note that this is a constant). Solving this last equation for  $\frac{q}{m}$  (and assuming  $q > 0$  for convenience) one arrives at

$$\frac{q}{m} = \frac{1}{a|b|} \frac{\|\vec{V}\|}{\sqrt{1 - \|\vec{V}\|^2}}.$$

Since  $a, b$  and  $\|\vec{V}\|$  are measurable one obtains the desired charge-to-mass ratio.

To conclude we wish to briefly consider the existence and use of ‘‘potentials’’ for electromagnetic fields. Suppose  $F$  is an electromagnetic field defined on some connected, open region  $X$  in  $\mathcal{M}$ . Then  $F$  is a 2-form on  $X$  which, by (28), is closed. Suppose also that the second deRham cohomology  $H^2(X; \mathbb{R})$  of  $X$  is trivial (since  $\mathcal{M}$  is topologically  $\mathbb{R}^4$  this will be the case, for example, when  $X$  is all of  $\mathcal{M}$ , or an open ball in  $\mathcal{M}$ , or, more generally, an open ‘‘star-shaped’’ region in  $\mathcal{M}$ ). Then, by definition, every closed 2-form on  $X$  is exact so, in particular, there exists a 1-form  $A$  on  $X$  satisfying

$$F = dA. \quad (46)$$

In particular, such a 1-form  $A$  always exists *locally* on a neighborhood of any point in  $X$  for any  $F$ . Such an  $A$  is not uniquely determined, however, because, if  $A$  satisfies (46), then so does  $A + df$  for any smooth real-valued function (0-form)  $f$  on  $X$  ( $d^2 = 0$  implies  $d(A + df) = dA + d^2 f = dA = F$ ). Any 1-form  $A$  satisfying (46) is called a (**gauge**) **potential** for  $F$ . The replacement  $A \rightarrow A + df$  for some  $f$  is called a **gauge transformation** of the potential and the freedom to make such a replacement without altering (46) is called **gauge freedom**.

One can show that, given  $F$ , it is always possible

to locally solve  $dA = F$  for  $A$  subject to an arbitrary specification of the 0-form  $*d*A$ . More precisely, if  $F$  is any 2-form satisfying  $dF = 0$  and  $g$  is an arbitrary 0-form, then locally, on a neighborhood of any point, there exists a 1-form  $A$  satisfying

$$dA = F \quad \text{and} \quad *d*A = g \quad (47)$$

(a more general result is proved in Appendix 2 of [4] and a still more general one in Section 2.9 of this same source). The usefulness of the second condition in (47) can be illustrated as follows. Suppose we are given some (physical) configuration of charges and currents (i.e., some source 1-form  $J$ ) and we wish to find the corresponding electromagnetic field  $F$ . We must solve Maxwell's equations  $dF = 0$  and  $*d*F = J$  (subject to whatever boundary conditions are appropriate). Locally, at least, we may seek instead a corresponding potential  $A$  (so that  $F = dA$ ). Then the first of Maxwell's equations is *automatically* satisfied ( $dF = d(dA) = 0$ ) and we need only solve  $*d*(dA) = J$ . To simplify the notation let us temporarily write  $\delta = *d*$  and consider the operator  $\Delta = d \circ \delta + \delta \circ d$  on forms (variously called the Laplace-Beltrami operator, Laplace-de Rham operator, or Hodge Laplacian on Minkowski spacetime). Then

$$\Delta A = d(\delta A) + \delta(dA) = d(*d*A) + *d*(dA). \quad (48)$$

According to the result quoted above we may narrow down our search by imposing the condition  $*d*A = 0$ , i.e.,

$$\delta A = 0 \quad (49)$$

(this is generally referred to as imposing the **Lorentz gauge**). With this (48) becomes  $\Delta A = *d*(dA)$  and to satisfy the second Maxwell equation we must solve

$$\Delta A = J. \quad (50)$$

Thus, we see that the problem of (locally) solving Maxwell's equations for a given source  $J$  reduces to that of solving (49) and (50) for the potential  $A$ . To understand how this simplifies the problem we note

that a calculation in admissible coordinates shows that the operator  $\Delta$  reduces to the componentwise D'Alembertian  $\square$ , defined on real-valued functions by

$$\square = \frac{\partial^2}{\partial(x^1)^2} + \frac{\partial^2}{\partial(x^2)^2} + \frac{\partial^2}{\partial(x^3)^2} - \frac{\partial^2}{\partial(x^4)^2}.$$

Thus, equation (50) decouples into four scalar equations

$$\square A_a = J_a, \quad a = 1, 2, 3, 4, \quad (51)$$

each of which is the well-studied inhomogeneous wave equation.

See also: **General Relativity: Introduction, Electromagnetism, Gauge Theory: Introduction.**

## References

- [1] Choquet-Bruhat Y, De Witt-Morette C and Dillard-Bleick M (1977) *Analysis, Manifolds and Physics*. Amsterdam: North-Holland.
- [2] Einstein A, et al (1958) *The Principle of Relativity*. New York: Dover.
- [3] Naber GL (1992) *The Geometry of Minkowski Spacetime*. New York, Berlin: Springer-Verlag.
- [4] Parrott S (1987) *Relativistic Electrodynamics and Differential Geometry*. New York, Berlin: Springer-Verlag.
- [5] Spivak M (1965) *Calculus on Manifolds*. New York: W A Benjamin.
- [6] Synge JL (1972) *Relativity: The Special Theory*. Amsterdam: North-Holland.