International Journal of Theoretical Physics, Vol. 27, No. 10, 1988

A Discrete Geometry: Speculations on a New Framework for Classical Electrodynamics

Geoffrey Hemion[1]

Received March 8, 1988

An attempt is made to describe the basic principles of physics in terms of discrete partially ordered sets. Geometric ideas are introduced by means of an action at a distance formulation of classical electrodynamics. The speculations are in two main directions: (i) Gravity, one of the four elementary forces of nature, seems to be fundamentally different from the other three forces. Could it be that gravity can be explained as a natural consequence of the discrete structure? (ii) The problem of the observer in quantum mechanics continues to cause conceptual problems. Can quantum statistics be explained in terms of finite ensembles of possible partially ordered sets? The development is guided at all stages by reference to the simplest, and most well-established principles of physics.

1. INTRODUCTION

The model that physicists use to describe the physical world is based on a differential manifold. It is thought that the curvature of the manifold itself provides an explanation of gravity. Within the manifold, further structures are defined—vector fields, particle paths, and so forth—and these are taken to account for the behavior of physical material. This picture, or *Weltbild*, is so generally accepted, and it is based on such a long history of physical research, that it might seem almost unthinkable to question it. And yet the purpose of this paper is to advance the proposition that other mathematical models may also be chosen.

While such a step may be condemned by many cautious readers, nevertheless it is interesting to recall the "General Remark D" of Einstein

[1]Fakultät für Mathematik, Universität Bielefeld, Bielefeld, West Germany.

0020-7748/88/1000-1145$06.00/0

(1956). He wrote, "One can give good reasons why reality cannot at all be represented by a continuous field. From the quantum phenomena it appears to follow with certainty that a finite system of finite energy can be completely described by a finite set of numbers (quantum numbers). This does not seem to be in accordance with a continuum theory, and must lead to an attempt to find a purely algebraic theory for the description of reality. But nobody knows how to obtain the basis of such a theory."

Unfortunately, both mathematics and physics have recently become increasingly subdivided into narrow areas of specialization. The area of mathematics that claims to have physical relevance is called "mathematical-physics." It is usually considered to be a special subject within the field of mathematical analysis—that is to say, the study of differential structures in general. New and speculative papers in mathematical physics are often concerned with very abstract analytic systems. Also, much recent work in probability theory can be thought of as falling into this category. On the other hand, the areas of mathematics that may be relevant to Einstein's problem—graph theory, lattice theory, and so on—are not usually considered to be a part of "mathematical physics." Thus, practical research in those subjects is concerned, for example, with questions that might have relevance in information theory. As a result there appears to have been virtually no research done on Einstein's problem.

And yet it cannot be said that the model of the physical world in terms of differential manifolds provides a perfectly clear and satisfactory description of all physical phenomena. On the contrary, the idea is often expressed that there is something wrong with our basic understanding of physics, and that therefore "new ideas" are called for. But what new ideas can there be? Surely, too many famous and competent physicists have devoted their energies to this question, but without result. How is it possible to overturn the work of centuries of philosophical thought?

It seems to me that another approach is called for. The goal is, according to Einstein, to find a representation of nature that avoids the use of a continuous field. The alternative to a continuum is a discrete set: a set that is such that the elements do not have arbitrarily near neighbors. But then, logically enough, the laws of physics, dealing with forces, interactions, and in general relationships between the elements of the set, must be transmitted over finite (not infinitesimal) distances. Thus, the path to a successful discrete theory of physics must in some way make use of an "action-at-a-distance" theory.

This conclusion is at once obvious and trivial, but it is still necessary to explain it further. The concept of action at a distance seems to be misunderstood and little known: it is often erroneously asserted that the theory of relativity is incompatible with any action-at-a-distance theory.

But this is far from being the case. On the contrary, the theory of action at a distance that I shall discuss can *only* be sensibly formulated in relativistic terms. This is the historical reason that physicists resorted to the concept of ether, or, in its modern guise, "space-time," in the 19th century.

The theory of action at a distance has been considered and developed by many famous people, including Gauss, Schwarzschild, Tetrode, Fokker, Feynman, and, more recently, Hoyle and Narlikar. A truly simple and all-encompassing relativistic formulation was discovered by Fokker (1929), and Section 2 of the present paper is devoted to a description of the theory. For simplicity I shall call it "Fokker's theory." In my opinion this formulation provides a key to the understanding of classical physics in a differential-free mathematical framework.

Section 3 is concerned with the development of a discrete mathematical structure based on Fokker's theory. Some of the consequences and defects of this structure are investigated there by means of examples. The goal is to show that the discrete framework I formulate is in some sense equivalent to the usual continuous framework provided by $\mathbf{R}^4$. But is it possible sensibly to compare two so different models? This question is also dealt with.

I define and work within a very specific discrete model. This gives the advantage of having a specific framework upon which to base my arguments. But this way of describing my ideas brings with it a very real danger; some readers may gain the impression that I am only prepared to advocate this one model, to the exclusion of all other possible discrete models. Nothing could be further from the truth. In fact, I am aware of many points where this model could be altered or changed to advantage. I can only hope that others will become interested in the idea of a discrete (rather than continuous, or differential) geometry for physics, and that they will also propose possible new solutions to some of the problems that lead to conceptual difficulties in modern physics.

While Section 2 is essentially nothing more than an account of a number of long-established classical results and Section 3, while new, follows conventional ways of thinking, Sections 4 and 5 are of a much more speculative nature. Section 4 considers a number of simple and "classical" results in the general theory of relativity, and attempts to argue that the present discrete framework may produce structures that lead to similar results. Section 5 considers a number of elementary phenomena in the theory of quantum mechanics and the question of whether or not it is possible to interpret these phenomena in terms of discrete structures.

Perhaps the main question is: what is the reason for being dissatisfied with the conventional description of physics in terms of differential manifolds? There are of course technical problems with some recently proposed field theories, which have led people to construct simple "lattice models."

But my principal motivation has been confined to a very much more basic level indeed: every elementary textbook on quantum mechanics begins, sometimes even in the introduction, with a discussion of the "basic mystery" of quantum mechanics. Feynman *et al.* (1965) expresses this particularly well. But certainly any mathematician must be dissatisfied with this state of affairs. Thus, this paper can be considered to begin and to end in this "loose end" in the introduction to any book on quantum mechanics. Section 5 can only be understood in these terms.

Finally, Section 6 discusses some aspects of Hoyle and Narlikar's work on quantum electrodynamics. Section 7 concludes with a number of further speculations and open questions.

It is certainly not the case that the ideas presented here add up to a coherent and definitive alternative to the present geometric foundations of physics. Perhaps such a goal can never be achieved when one considers the diverse—sometimes even divergent—directions of modern physics. But it is at least my thesis that the standard lattice model, which is often considered to be convenient to use, is not particularly appropriate, and therefore the present work can be considered as providing some arguments for the use of more interesting discrete geometric models in theoretical physics.

2. ACTION AT A DISTANCE IN CLASSICAL ELECTRODYNAMICS

2.1. Maxwell's Equations

The theory of classical electrodynamics was developed during the 19th century in parallel with an increasingly refined basis of experimental observation. The goal was to provide a coherent and accurate explanation of the physical world, and yet on both counts the theory has always been deficient. The fact that it is inaccurate has led to its being discarded in favor of the quantum theory. It might well be argued that we will never achieve a final and definitive theory of physics, but on the other hand the quantum theory is itself based on the concepts of classical electrodynamics, and so the classical theory retains a position of importance. The fact that it poses many difficult—and still unsolved—mathematical questions is perhaps a property that it must share with any reasonable theory of the physical world.

Of course Maxwell (1831-1879) played a central and decisive role in the formulation of the theory of classical electrodynamics. Nevertheless it might be interesting to recall that during the 1830s Gauss (1777-1855) also devoted a great deal of time to the problem of electricity and magnetism. The information available to him was hardly less than what Maxwell had.

And in fact Gauss was (together with W. Weber) the editor of a contemporary journal on magnetism. Yet Gauss failed to produce a viable theory.

In a letter to Weber (Gauss, 1845) he wrote, "I would doubtless have published my researches long ago were it not that, at the time I gave them up, I had failed to find what I regarded as the keystone, *Nil actum reputans si quid superesset agendum*: namely the derivation of the additional forces—to be added to the interaction of electrical charges at rest, when they are both in motion—from an action which is propagated not instantaneously but in time, as is the case with light."

As we shall see, this basic strategy of Gauss is correct, but unfortunately in the 1830s he did not have the means of bringing the idea to fruition. It turns out that such interactions, both in space and in time, can only be sensibly formulated within the framework of the theory of relativity. Thus, according to Wheeler and Feynman (1949), "Field theory taught gradually and over seven decades' difficult lessons about constancy of light velocity, about relativity of space and time, about advanced and retarded forces, and in the end made possible by this circuitous route the theory of direct interparticle interaction which Gauss had hoped to achieve in one leap."

Maxwell's equations are

$$\nabla \times E = -\frac{1}{c}\frac{\partial H}{\partial t}, \qquad \nabla \times H = \frac{1}{c}\left(4\pi J + \frac{\partial E}{\partial t}\right)$$
$$\nabla \cdot E = 4\pi\rho_e, \qquad \nabla \cdot H = 0 \tag{1}$$

where E is the electric field, H is the magnetic field, t is time, c is the speed of light, ρ_e is a scalar field representing the charge density, and J is a vector field representing the three-dimensional electrical currents. One may also write $J = \rho_e v$, where v is a vector field representing the velocity of the charge, whose electrical density is ρ_e. The various fields E, H, ρ_e, and J are defined in three-dimensional Euclidean space $\mathbf{R}^3$. The theory describes a one-parameter family of such fields, indexed by the real parameter t. In addition, it is necessary to specify the effect of the electromagnetic fields on charged matter, and this is done by means of the Lorentz equation.

$$f = \rho_e\left(E + \frac{v \times H}{c}\right) \tag{2}$$

where f is the force density (that is, the force per unit volume).

Now it is obvious that all of these equations, when taken together, produce a theory of such great mathematical complexity that only a few simple classes of solutions are known. But this complexity alone is not the main problem. Even given that one is able to produce new solutions, it is still unclear what relevance they might have to the description of the real

physical world. Thus, especially during the early 20th century, the theory was altered and qualified. For example, the idea that matter can be thought of as a continuous fluid gradually lost influence, and instead people came to think of pointlike particles moving through a vacuous space. It became necessary to change (at least) the conception of the quantities ρ_e, J, and f from the original idea of smooth fields. Then the theory of relativity changed the basic framework within which the equations had a meaning.

In fact, the theory of relativity allows a simplification of the equations. One may write

$$F = \begin{pmatrix} 0 & H_z & -H_y & -E_x \\ -H_z & 0 & H_x & -E_y \\ H_y & -H_x & 0 & -E_z \\ E_x & E_y & E_z & 0 \end{pmatrix} \tag{3}$$

and F is considered to represent an antisymmetric tensor field in four-dimensional Euclidean space $\mathbf{R}^4$. The vector J is defined on $\mathbf{R}^4$ to be

$$(J_1, J_2, J_3, J_4) = (J_x, J_y, J_z, \rho_e) \tag{4}$$

Then Maxwell's equations become

$$F_{ij,j} = 4\pi J_i, \qquad F_{ij,k} + F_{jk,i} + F_{ki,j} = 0 \tag{5}$$

where i, j, and k run from 1 to 4 , and the summation convention

$$F_{ij,j} = \sum_{j=1}^{4} \frac{\partial F_{ij}}{\partial x_j} \tag{6}$$

is being used. The partial derivative here is to be understood in the sense of covariant derivatives in the theory of differential manifolds. We have chosen the unit of "time" along the x_4 axis in such a way that the speed of light is 1. The space $\mathbf{R}^4$ is considered to be the usual pseudo-Riemannian manifold of the special theory of relativity. Also, the Lorentz equation can be rewritten within this framework, but I prefer to defer this to the sequel.

This formulation of Maxwell's equations leads to a still simpler formulation when one observes that it is possible to find a so-called "vector potential" A, which is a vector field on $\mathbf{R}^4$ satisfying

$$F_{ij} = \frac{\partial A_j}{\partial x_i} - \frac{\partial A_i}{\partial x_j} \tag{7}$$

It may be assumed that the gauge condition $A_{i,i} = 0$ holds, and thus that the wave equations

$$A_{i,jj} = 4\pi J_i \tag{8}$$

for $i = 1, \ldots, 4$ also hold.

Many textbooks describe these equations, the assumptions and observations that lie behind them, and also a few solutions in simple situations. It is not the purpose of this paper to attempt to provide a comprehensive treatment of such well-known results. But it will be worthwhile and instructive for our later purposes to examine two particular solutions now.

2.2. Some Simple Solutions to Maxwell's Equations

Solution 1. Plane Waves. This first solution to Maxwell's equations is concerned with the situation when space and time are devoid of electrical charge. Thus, J and ρ_e are zero and we have $A_{i,jj}=0$ for $i=1,\ldots,4$. Now there are certainly very many different, but similar, solutions in this case. One class of solutions can be identified by simply taking $A_1=A_3=A_4=0$. We can restrict the situation still more by assuming that the remaining component, namely A_2, depends only on the variables x_1 and x_4, which, to return to the more traditional notation, I will call x and ct. Thus we obtain the equation

$$\frac{\partial^2 A_2}{\partial x^2}-\frac{\partial^2 A_2}{c^2\partial t^2}=0 \tag{9}$$

We may choose any smooth, real function, say $u:\mathbf{R}\to\mathbf{R}$. Then certainly $A_2=u(x-ct)$ provides a solution. In addition, the equation $A_{i,i}=0$ is trivially satisfied. One traditional possibility is to take u as being a trigonometric function, for example $u(x-ct)=\sin(x-ct)$. This leads to $F_{21}=H_z=-\partial A_2/\partial x=-\cos(x-ct)$; similarly $F_{24}=E_y=-\cos(x-ct)$, we have $F_{12}=-F_{21}$, $F_{42}=F_{24}$, and all other F_{ij} are 0. This solution represents a plane polarized wave traveling in the x direction with velocity c.

Now there is certainly nothing unusual in all of this. As far as the *practical* application of physics is concerned, such electromagnetic plane wave solutions are of great importance—from, say, the design of a simple capacitor to the use of very long-baseline interferometry in radio astronomy. But notwithstanding such practical considerations, it must be admitted that if our goal is to be the understanding of the *basic* principles of physics, then all such solutions must be discarded from the outset. For example, it is known that the universe is in some sense expanding, and this fact is in conflict with our plane wave solution. But it is also clear that the universe does, in fact, contain electrically charged material.

The linearity of Maxwell's equations allow one simply to add in any of these vacuum solutions to a given solution, thus obtaining a new solution. However, such a procedure leads to difficulties as soon as the effect of the Lorentz equation is brought into consideration.

Solution 2. A Uniformly Charged Spherical Ball. The second solution I will consider concerns once again a universe without electrical charge,

except for a stationary spherical ball of uniform charge density ρ and radius ε, located at the point $(0, 0, 0) \in \mathbf{R}^3$. One solution for this problem (others may be obtained by adding in additional vacuum solutions as in case 1 above) involves both the electrical field E and the magnetic field H being constant in time. But of course this implies that H vanishes. Outside the ball, we have E being radially symmetric about the origin $(0, 0, 0)$ and proportional in strength to the inverse of the distance to the origin.

Specifically, choose $A_1 = A_2 = A_3 = 0$ and $A_4 = k_1/r$ [where $r = (x^2+y^2+z^2)^{1/2}$] for $r > \varepsilon$ and $A_4 = k_2 r^2$ for $r \leq \varepsilon$, where k_1 and k_2 are appropriate constants. Once again the equation $A_{i,i} = 0$ is trivially satisfied for all r, and the Laplacian, considered in spherical coordinates,

$$\frac{\partial}{r^2 \partial r}\left(r^2 \frac{\partial A_4}{\partial r}\right) \tag{10}$$

vanishes for $r > \varepsilon$. When $r \leq \varepsilon$ we have no electrical currents, so that $J_i = 0$ for $i = 1, 2, 3$. But $J_4 = \rho$, so that we must have $A_{4,jj} = 4\pi\rho$, and thus $k_2 = 4\pi\rho/6$. In order to have A_4 continuous, it is then necessary to have $k_1 = 4\pi\rho\varepsilon^3/6$.

This solution involves a continuous, fluidlike ball of electrical material. Such a fluid may have represented the picture that physicists in the 19th century found to be appropriate, and in fact Lorentz used such a model of the electron in his attempt to show that the mass was of electromagnetic origin. But this is very much removed from the accepted ideas of today. In fact, it seems now to be generally accepted that (as far as quantum mechanics allows the discussion of such concepts) the electron is purely pointlike.

One way of dealing with such pointlike electrons (see, for example, Dirac, 1938) is to consider our uniform ball solution with a constant total charge, but with the radius ε tending to zero, and thus the charge density ρ tending to infinity. Thus we end up with the solution $A_4 = k/r$ for some constant k. Of course this solution is not continuous, or even well defined, at the origin. But still it is possible, following Dirac, to extend the theory to include the possibility of such generalized functions. There are, however, a number of problems with this picture.

For example, there is the problem of the self-interaction of such an electron with its own electrical field. The field is infinite at the particle, and so something other than the simple Lorentz equation may be necessary to describe its motion. (The reader should be wary of dismissing this problem with the thought that it is of little concern in the quantum theory. On the contrary, as already noted, the quantum theory is defined in terms of classical electrodynamics, and in any case, the quantum theory has itself a collection of similar "divergences.") Dirac shows, using a kind of perturbation theory argument, that such a classical electron may, by itself, begin to accelerate,

achieving an exponentially increasing velocity through the force of its own classical electromagnetic field. He shows how to eliminate such physically unrealistic solutions, but the price that must be paid is the phenomenon of "preacceleration." That is, an electron that is about to be disturbed by an external field must "anticipate" the field with a prior movement of its own. This would appear to violate the principle of causality. But even without discussing such esoteric paradoxes, it is easy to see that in every neighborhood of such an electron there would be an infinite total field energy, and that would also appear to conflict with a number of established principles of physics.

2.3. Pointlike Particles

In this section I show how it is possible to deal in a systematic way with the electromagnetic fields generated by collections of pointlike, electrically charged particles.

Definition. Let $\gamma : \mathbf{R} \to \mathbf{R}^4$ be a smooth path. I will say that γ is *timelike* if γ is order-preserving. The order on $\mathbf{R}$ is, of course the usual total ordering. On $\mathbf{R}^4$ the Lorentz ordering will be used. Thus, if X, $Y \in \mathbf{R}^4$ are two points with coordinates $X = (x_1, x_2, x_3, x_4)$ and $Y = (y_1, y_2, y_3, y_4)$, then $X < Y$ if both $x_4 < y_4$ and

$$(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2 < (x_4 - y_4)^2 \tag{11}$$

From now on, *particles* will be considered to be timelike paths that are infinitely long in both directions. Let Γ be some locally finite collection of nonintersecting particles. That is, given any compact region K in $\mathbf{R}^4$, then at most finitely many particles of Γ meet K. In this situation, it is possible to define the electrical currents in terms of generalized functions. For each particle $\gamma \in \Gamma$ we assign a real number e_γ corresponding to the electrical charge of γ. Furthermore, at a given point $\gamma(\tau)$ of γ let $v(\tau) = \gamma'(\tau)$ be the velocity of γ, a 4-vector. Note that the path γ is parameterized here simply with the time coordinate $\tau = x_4$. Another possibility is to take the *proper-time parameterization.* This is such that at all points the particle $\gamma(s)$ has $|\partial\gamma/\partial s| = 1$ with respect to the Lorentz metric on $\mathbf{R}^4$. That is,

$$\left|\frac{d\gamma}{ds}\right| = \left[\left(\frac{d\gamma_4}{ds}\right)^2 - \sum_{i=1}^{3}\left(\frac{d\gamma_4}{ds}\right)^2\right]^{1/2} \tag{12}$$

We can now define the electrical current of the particle γ to be the vector field on $\mathbf{R}^4$ given by $J(x) = e_\gamma v(\tau)\delta(x - \gamma(\tau))$, where $x = (x_1, x_2, x_3, x_4)$ and $\tau = x_4$.

Such electrical currents, expressed in terms of generalized functions, can be inserted into the equation $A_{i,jj} = 4\pi J_i$. Standard solutions are given by the so-called Liénard-Wiechert potentials [consult any textbook on

classical electrodynamics (e.g., Eyges, 1972) for a derivation of these solutions]

$$A_{\pm}(\mathbf{x}, \tau) = e_{\gamma} \int v(\tau) \frac{\delta(\mathbf{x}, \tau \pm |\mathbf{x} - \gamma(\tau)|)}{|\mathbf{x} - \gamma(\tau)|} d(\tau) \tag{13}$$

Here $\mathbf{x}$ is the projection of a point $x \in \mathbf{R}^4$ onto $\mathbf{R}^3$, and thus we will also consider $\gamma(\tau)$ here to denote the projection of the point of γ (namely the image on τ) onto $\mathbf{R}^3$.

In the special case that $\gamma(\tau) = (0, 0, 0)$, $\forall \tau \in \mathbf{R}$, it is clear that both of the solutions $A_{\pm}$ reduce to the solution found in the last section. In the more general case A_+ and A_- are different, and they are called the advanced and retarded potentials, respectively.

As we have seen, it is possible to express the Liénard-Wiechert potentials (and thus the electromagnetic fields) in a simple way if the generating particle is stationary. The case of uniform motion without acceleration can also be deduced from this solution by observing that one need only change the frame of reference (in the sense of relativity) in order to reduce such motion to the stationary case. But more general Liénard-Wiechert solutions, associated with accelerated particles, cannot in general be expressed in terms of simple functions.

It is interesting to observe that the retarded solutions can be thought of as representing electromagnetic radiations traveling *forward* in time, while the advanced solutions represent radiations traveling *backward.* From the point of view of classical electrodynamics, both solutions are valid, or any combination of the two. The idea of forward- and backward-directed radiations stems from the idea of cause and effect. The cause of the radiation is the presence and motion of the particle γ at some point $\gamma(\tau)$. The effect is then felt along the "lightcones" [that is, the points of $\mathbf{R}^4$ with vanishing Lorentz distance from $\gamma(\tau)$] above and below $\gamma(\tau)$. The "normal" situation is that the cause precedes the effect in time. That is, the effect is only felt along the light cone above the point $\gamma(\tau)$, and this is represented by the retarded Liénard-Wiechert potential. Thus, it would seem to be a natural idea to exclude the advanced potentials from further consideration. However, as we shall see, it is by no means clear that the relationship of cause and effect in physical processes implies the vanishing of the advanced potentials. On the contrary, the validity of the action-at-a-distance theory that I will consider depends on the existence of advanced potentials that do not violate the principle of cause and effect.

2.4. Action at a Distance in Physics

The concept of "action at a distance" is usually associated with Newtonian gravitation. The force of gravity within this theory is

transmitted instantaneously between widely separated objects. Now it is easy to show that if the premises of the theory of relativity are accepted, then such an instantaneous action at a distance will lead to a violation of the principle of cause and effect, as discussed in the last section.

But there are other theories of action at a distance that are not in conflict with the theory of relativity. One such theory involves the study of Hamiltonian systems with constraints (see, for example, Llosa, 1981). On the other hand, the theory that I shall discuss (sometimes called the "many-time" theory) is based on the consequences of the Liénard–Wiechert potentials. It is fair to say that this theory represents the solution to the problem that Gauss posed in the 1830s, and thus it is just as much a traditional approach to the problem of electrodynamics as is Maxwell's theory.

What reasons are there for examining such unusual theories? To begin with, I have already noted some of the standard difficulties with the Maxwell theory. One could hope to overcome these difficulties by attempting to find a new framework for electrodynamics. But perhaps the most important reason has to do with Mach's principle.

Consider, for example, a rotating bucket of water (Figure 1). The surface of the water assumes a concave shape, even though the water itself may be stationary with respect to the bucket. How can we explain this curvature?

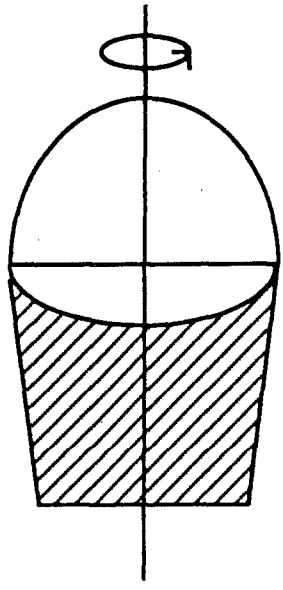

Fig. 1

It is necessary to say that the bucket is not in an inertial frame of reference, and that for an observer who happens to be in such a frame—who is, so to speak, standing "outside" the bucket—it is obvious that the water experiences a centrifugal force. But then what is more special about the inertial frame in comparison with the rotating frame? In the end one is reduced to an appeal to Mach's principle: that the distant matter of the universe somehow influences the seemingly local phenomena of physics.

One might think that this appeal to Mach's principle is a natural one in the framework of Maxwell's theory: it is only necessary to define certain conditions "at infinity" within a given space, and then to apply the uniqueness theorems for the solutions of partial differential equations. But what are these conditions at infinity? And indeed, while it is true that the space $\mathbf{R}^4$ together with the Lorentz metric provides an approximate local model for physics, it is also true that it fails badly as a cosmological model for such global solutions. Certainly Einstein, who took Mach's principle very seriously, was surprised by Gödel's (1949) demonstration of the existence of a rotating universe that conforms with the general theory of relativity. Thus, it is apparent that one should approach with caution any too easy discussion of local field theories in terms of vague boundary conditions.

To be quite specific, the reader should consider some particular physical movement, for example, lifting a book from a table. Certain forces are involved: there is the inertia of the book, the pages might move relative to the binding. The question is, should one consider these phenomena as being purely local, so that eventually some small electromagnetic disturbances might propagate themselves though space and time to the distant stars, and thus Mach's principle will in some way be satisfied? Or should one think of such things as being *directly and immediately* caused by an interaction with all of the matter in the universe, even the most distant?

This question is a very philosophical one, but not without some practical value, as Einstein's experience suggests. The question is also a mathematical one. Can the relativistically invariant classical electrodynamics be formulated in terms of direct interactions between widely separated particles, and if so, are the two theories equivalent?

2.5. Fokker's Action Principle

The formulation of the theory of action at a distance that I shall discuss is due to Fokker (1929) and so I shall refer to it as Fokker's theory. It is based on the idea of pointlike particles moving through space-time. In its initial formulation, the theory is only concerned with these particles, and it ignores completely the question of electromagnetic fields. The theory is expressed as a variational principle. Fokker considers a locally finite collection Γ of particles represented by smooth timelike paths in $\mathbf{R}^4$. Given such

a particle $\gamma \in \Gamma$, two real numbers e_γ and m_γ are associated with γ, representing the electrical charge and the mass of γ, respectively. The next step is to define the quantity

$$J_\Gamma = -\sum_{\gamma\in\Gamma} m_\gamma \int_{-\infty}^{+\infty} \left|\frac{d\gamma}{du}\right| du + \sum_{\gamma<\xi} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (e_\gamma e_\xi)\delta(s^2_{\gamma(u),\xi(v)})\gamma'\xi'\, du\, dv \quad (14)$$

Of course, one immediately notices that J_Γ does not appear to be properly defined! Even if we confine ourselves to the relatively simple first term, it is clear that the integral diverges; then we are asked to sum over a possibly infinite set, indexed by Γ, of such integrals. The second term, if we look so far, only serves to make matters worse! (It should be noted here that Fokker, writing in 1929, did not use the notation of generalized functions. In fact his more traditional notation is in many ways more readable than the modern treatments.)

Fokker's principle is that the collection of paths Γ should have the property that

$$J_\Gamma = \text{extremum} \quad (15)$$

In what sense is J_Γ to be evaluated, and with respect to what is J_Γ an extremum? These questions will gradually be dealt with in the rest of this section. But for the moment one should bear in mind that we are dealing with a variational principle, and the things to be varied are the elements of Γ. Imagine that a compact set $K \subset \mathbf{R}^4$ is given, and a variation of Γ that is confined to the interior of K. Then one can reasonably expect that J_Γ (or something like J_Γ) will only change by a finite amount, and thus the condition that J_Γ should be an extremum can be sensibly interpreted.

The most immediate task is to define the various quantities in J_Γ. To begin with, all of the expressions in the first term—and in particular $|d\gamma/du|$—have already been defined. As for the second term, we will be dealing with the expression

$$S = \int_{t_1}^{t_2}\int_{-\infty}^{+\infty} (e_\gamma e_\xi)\delta(s^2_{\gamma(u),\xi(v)})\gamma'\xi'\, du\, dv \quad (16)$$

for real numbers $t_1 < t_2$ and disjoint $\gamma, \xi \in \Gamma$. Here $\gamma'\xi'$ is defined by

$$\gamma'\xi' = \frac{d\gamma_4}{du}\frac{d\xi_4}{du} - \sum_{i=1}^{3}\frac{d\gamma_i}{du}\frac{d\xi_i}{du} \quad (17)$$

The expression $s^2_{\gamma(u),\xi(v)}$ is defined as follows. Let $p = (p_1, \ldots, p_4)$, $q = (q_1, \ldots, q_4)$ be points in $\mathbf{R}^4$. Then

$$s^2_{p,q} = (p_4 - q_4)^2 - \sum_{i=1}^{3}(p_i - q_i)^2 \quad (18)$$

It is simple consequence of the definition of the Dirac δ-function that

$$\int_{-\infty}^{+\infty} \delta(s^2_{a,\xi(v)})\, dv = \int_{-\infty}^{+\infty} \frac{1}{2r}[\delta(t-r)+\delta(t+r)]\, dv \tag{19}$$

where $t = t_v = a_4 - \gamma_4(v)$ and

$$r = r_v = \left\{ \sum_{i=1}^{3} [a - \xi_i(v)]^2 \right\}^{1/2} \tag{20}$$

We have here the more familiar one-dimensional Dirac generalized function, and so S can be expressed in the following way:

$$S = (e_\gamma e_\gamma) \int_{t_1}^{t_2} \left(\frac{(\gamma'\xi')_{\text{adv}}}{2r_{\text{adv}}} + \frac{(\gamma'\xi')_{\text{ret}}}{2r_{\text{ret}}} \right) dv \tag{21}$$

where, for a given u, $r_{\text{adv}} = r_{\gamma(u),v^*}$, where v^* is the unique real number with the property that $r_{\gamma(u),v^*} = t_{\gamma(u),v^*}$. Similarly, $r_{\text{ret}} = r_{\gamma(u),v'}$, where v' is such that $r_{\gamma(u),v'} = -t_{\gamma(u),v'}$. The product $\gamma'\xi'$ is evaluated by taking γ' at $\gamma(u)$ on γ, and then ξ' at the points on ξ diagonally above and below $\gamma(u)$.

The only thing that remains to be explained is the inequality $\gamma < \xi$ in the expression for J_Γ. The set Γ is certainly countable, and therefore we may find an injection $\Gamma \to \mathbf{Z}$, the integers, thus defining a total ordering of Γ. Any such total ordering of Γ will do, and its only purpose is to ensure that the integral will be evaluated exactly once for each pair of disjoint paths in Γ. The possibility of self-interaction of a particle on itself is thus excluded at the outset, and hence one of the great difficulties of classical field theory simply plays no role in Fokker's theory.

2.6. A Correspondence between Fokker's Theory and the Maxwell Theory

In essence, the first term of J_Γ describes a relativistic version of Newton's first law. The second term is a relativistic version of Coulomb's law. I will now show that the variational principle J_Γ = extremum results in a law of particle motion corresponding with the Lorentz equation. To begin with, it should be remarked that the notion of *fields* (in the sense of mappings from $\mathbf{R}^4$ to some other structures) was not necessary in the description of Fokker's theory. All that was used was the idea of particle paths and distances along the paths. In order to establish a correspondence with the more conventional ideas, it will be necessary to identify some quantity within the Fokker theory to correspond with the idea of an electromagnetic field. To this end, let $x \in \mathbf{R}^4$ be such that x lies on no path in Γ. Let

$$A^{(\xi)}(x) = \int_{-\infty}^{+\infty} \delta(s^2_{a,\xi(v)}) \frac{d\xi}{dv}\, dv \tag{22}$$

With this definition, each point $x \in \mathbf{R}^4$ not on ξ is associated with a vector

$$A^{(\xi)}(x) = (A_1^{(\xi)}(x), A_2^{(\xi)}(x), A_3^{(\xi)}(x), A_4^{(\xi)}(x)) \tag{23}$$

Using (19), one sees that $A^{(\xi)}$ is a smooth vector field on $\mathbf{R}^4 - \{\xi\}$. I will show that this field corresponds to the usual Liénard-Wiechert potentials generated by the particle ξ. The fact that $A^{(\xi)}$ has the same *formal* structure as the sum of one-half the advanced and one-half the retarded Liénard-Wiechert potentials shows that $A^{(\xi)}$ certainly satisfies Maxwell's equations. But more importantly (remembering that Fokker's theory is a theory of *particles* rather than fields), it is necessary to show that other particles $\gamma \in \Gamma - \{\xi\}$ describe paths that obey the Lorentz equation with respect to the sum of the electromagnetic fields $A^{(\xi)}$ generated by all of the particles $\xi \in \Gamma - \{\gamma\}$. Note that the Lorenntz equation for a particle γ in a classical electromagnetic field given by F is

$$m_\gamma \gamma_i'' = e_\gamma F_{ik} \gamma_i' \tag{24}$$

where, for example, $\gamma_i' = d\gamma_i(u)/du$ for $i = 1, \ldots, 4$.

I shall follow the reasoning in Wheeler and Feynman (1949), but observing where necessary at which points additional assumptions are needed. To begin with, it will be convenient to choose the proper time parameterization for γ. Thus, assume that $|d\gamma/du| = 1$ at all points in the image of γ. Consider a variation of (14), where we restrict out attention to just two particles $\gamma, \xi \in \Gamma$. Define

$$J_{\gamma,\xi} = -m_\gamma \int_{t_1}^{t_2} \left| \frac{d\gamma}{du} \right| du + e_\gamma e_\xi \int_{t_1}^{t_2} \int_{-\infty}^{+\infty} \delta(s^2_{\gamma(u),\xi(v)}) \gamma' \xi' \, du \, dv \tag{25}$$

Thus, $J_{\gamma,\xi}$ can be considered as being part of one of the terms in J_Γ. We will consider a variation of the path γ between $\gamma(t_1)$ and $\gamma(t_2)$. Let a new smooth path $\beta : \mathbf{R} \to \mathbf{R}^4$ be given, with the property that $\beta(u) = g(u)$, for all $u \in \mathbf{R} - [t_1, t_2]$. (Of course, it can no longer be assumed that β has the proper time parameterization.) Then, for some sufficiently small interval $[-\varepsilon, +\varepsilon]$ centered on zero, we may define a one-parameter group of smooth, timelike paths γ^r, indexed by $r \in [-\varepsilon, +\varepsilon]$, as

$$\begin{aligned} \gamma^r(u) &= \gamma(u) + r[\beta(u) - \gamma(u)] \\ &= \gamma(u) + r\sigma(u) \qquad \text{where } \sigma = \beta - \gamma \end{aligned} \tag{26}$$

Now let

$$J^r_{\gamma\xi} = J_{\gamma,\xi} - J_{\gamma^r,\xi} \tag{27}$$

Our assumptions imply that $J^r_{\gamma\xi}$ is differentiable at zero. At this stage we define

$$J^r_\gamma = \sum_{\xi \neq \gamma} J^r_{\gamma\xi}$$

and assert that also J^r_γ is differentiable at zero. But of course this is not in general true; it is necessary to *postulate* that J^r_γ both exists and is differentiable at zero. Then the variational hypothesis implies that

$$\left.\frac{dJ^r_\gamma}{dr}\right|_{r=0} = 0 \tag{28}$$

for all $\gamma \in \Gamma$.

One may write

$$\frac{dJ^r_{\gamma\xi}}{dr} = -m_\gamma \int_{t_1}^{t_2} \frac{d}{dr}\left|\frac{d\gamma^r}{du}\right| du + e_\gamma e_\xi \int_{t_1}^{t_2} \frac{d}{dr}\left[\int_{t_1}^{t_2}\int_{-\infty}^{+\infty} \delta(s^2_{\gamma(u),\xi(v)})\gamma'^r \xi' \, dv\right] du \tag{29}$$

In order to evaluate this expression, let us begin with the first integral. We have

$$\frac{d}{dr_{r=0}}\left|\frac{d\gamma^r}{du}\right| = \frac{d}{dr}\left|\frac{d\gamma}{du} + r\frac{d\sigma}{du}\right| = \frac{d\gamma}{du}\frac{d\sigma}{du} \tag{30}$$

(Remember that γ is assumed to have the proper-time parameterization.) Using partial integration, we obtain

$$\int_{t_1}^{t_2} \frac{d}{dr}\left|\frac{d\gamma^r}{du}\right| du = \left(\frac{d\gamma}{du}\sigma\right)_{t_2}^{t_1} + \int_{t_1}^{t_2} \frac{d^2\gamma}{du^2}\sigma \, du \tag{31}$$

As for the second expression, we have

$$\begin{aligned} & e_\xi \int_{t_1}^{t_2} \frac{d}{dr}\left[\int_{-\infty}^{+\infty} \delta(s^2_{\gamma(u),\xi(v)})\gamma'^r \xi' \, dv\right] du \\ &= \int_{t_1}^{t_2} \frac{d}{dr}[A^{(\xi)}(\gamma^r(u))\gamma'^r] \, du \\ &= \int_{t_1}^{t_2} \frac{d}{dr}[A^{(\xi)}(\gamma + r\sigma)(\gamma' + r\sigma')] \, du \\ &= \int_{t_1}^{t_2} \left\{\sigma_m \frac{\partial A^{(\xi)}(\gamma)}{\partial x_m}\gamma' + r\sigma' \frac{d}{dr}[A^{(\xi)}(\gamma + r\sigma)] + \sigma' A^{(\xi)}(\gamma + r\sigma)\right\} du \end{aligned} \tag{32}$$

The second term in this expression is zero at $r = 0$. The third term can be changed by means of partial integration:

$$\begin{aligned} \int_{t_1}^{t_2} \sigma' A^{(\xi)}(\gamma + r\sigma) &= [\sigma A^{(\xi)}]_{t_2}^{t_1} - \int_{t_1}^{t_2} \sigma \frac{d}{du}(A^{(\xi)}) \, du \\ &= -\int_{t_1}^{t_2} \sigma_m \frac{\partial A^{(\xi)}_m}{\partial x_n}\gamma'_n \, du \end{aligned} \tag{33}$$

Then, combining all terms, we obtain

$$\left.\frac{dJ^r_{\gamma\xi}}{dr}\right|_{r=0} = \int_{t_1}^{t_2} \sigma_m \left[-m_\gamma \frac{d^2\gamma_m}{du4} + e_\gamma \gamma'_n \left(\frac{\partial A_n^{(\xi)}}{\partial x_m} - \frac{\partial A_m^{(\xi)}}{\partial x_n} \right) \right] du \tag{34}$$

If we sum over all particles $\xi \neq \gamma$ in Γ (a questionable idea if Γ is infinite), then the derivative should be zero for all such variations σ of γ, and so we conclude that

$$m_\gamma \frac{d^2\gamma_m}{du^2} = e_\gamma \sum_{\xi \neq \gamma} F_{mn}^{(\xi)} \gamma'_n \tag{35}$$

This is the Lorentz equation, but without the terms associated with the self-interaction of the particle γ upon itself.

In this section a number of requirements have been placed on the set of continuous paths Γ. The goal has been to achieve a framework sufficiently restrictive to allow equation (35) to be deduced. The description of the conditions that could define such a framework has been vague. (A property it shares with the existing literature on the subject.) My interest has been to provide a sketch of the methods used in the standard classical theory sufficient to make the present discrete description more comprehensible. But perhaps some readers may be encouraged to go further and pursue a more thorough investigation of these conditions. Such a program would certainly be interesting in its own right and could be expected to have relevance for the discrete theory as well. In lieu of such results, I will retreat to the strategy of simply making a definition: namely, a space consisting of a set of continuous paths Γ that is such that all of the calculations of this section are valid will be called a "Fokker space."

2.7. Action and Reaction in Fokker's Theory

It is interesting and useful to seek further correspondences between Fokker's theory and the usual formulation. In particular, the idea of energy and momentum and the conservation law concerning these quantities play important roles in classical electrodynamics. The fact that particles obey the usual law of motion (35) shows that, *locally at least*, conservation of energy and momentum holds. Can a *global* conservation law also be deduced? Consider a particle γ in a set of particles Γ in $\mathbf{R}^4$. For convenience concentrate on the case $\Gamma = \{\gamma, \xi\}$. For each point $\gamma(t)$ in the image of γ, one may define the energy-momentum vector G^γ to be $m_\gamma \, d\gamma/du$. Note that if we revert to the more conventional units of time, such that the speed of light is given by $c \approx 300{,}000$ km/sec, and if we parameterize γ with this time (rather than using the proper-time parameterization), then we obtain

the familiar expressions

$$\mathbf{G}^\gamma = \frac{m_\gamma c v}{(1 - v^2/c^2)^{1/2}}, \qquad G_4^\gamma = \frac{m_\gamma c^2}{(1 - v^2/c^2)^{1/2}} \tag{36}$$

where $\mathbf{G}^\gamma$ represents the 3-vector of momentum in $\mathbf{R}^3$ and G_4^γ is the energy of the particle.

In general, this vector can be expected to change as we move along γ, and the rate of change is given by $dG^\gamma/du = m_\gamma d^2\gamma/du^2$. But after examining (35), we obtain

$$\begin{aligned}
\frac{dG_m^\gamma}{du} = m_\gamma \frac{d^2\gamma_m}{du^2} &= e_\gamma \sum_{\xi \neq \gamma} F_{mn}^{(\xi)} \gamma_n' \\
&= e_\gamma \gamma_n' \left(\frac{\partial A_n^{(\xi)}}{\partial x_m} - \frac{\partial A_m^{(\xi)}}{\partial x_n} \right) \\
&= e_\gamma e_\xi \gamma_n' \left[\frac{\partial}{\partial x_m} \int_{-\infty}^{+\infty} \delta(s^2_{\xi(u),x}) \frac{d\xi_n}{du}\, du \right. \\
&\quad \left. - \frac{\partial}{\partial x_n} \int_{-\infty}^{+\infty} \delta(s^2_{\xi(u),x}) \frac{d\xi_m}{du}\, du \right] \\
&= 2 e_\gamma e_\xi \int_{-\infty}^{+\infty} \delta'(s^2_{\xi(u),x})(r_m \gamma' \cdot \xi' - \xi_m \gamma' \cdot r)\, du
\end{aligned} \tag{37}$$

We have used the result that

$$\int f(x) \left\{ \frac{\partial}{\partial y} [\delta(g(x, y))] \right\} dx = \int f(x) \left\{ \frac{\partial}{\partial y} [g(x, y)] \right\} \delta'(g(x, y))\, dx$$

for test functions f and smooth functions g of two variables. The vector r is determined by the point x and the points $\xi(u)$ on the image of ξ that have vanishing Lorentz distance to x. One can take r to be the sum of half the retarded and advanced distances. At this stage it is convenient to use the well-known result

$$\int f(x)\, \delta'(x)\, dx = -f'(0)$$

for suitable test functions f, and in particular

$$\int \delta'(x)\, dx = 0$$

This enables us to write

$$2 e_\gamma e_\xi \int_{-\infty}^{+\infty} \delta'(s^2_{\xi(u),x}) \gamma_m' \xi' \cdot r\, du = 0 \tag{38}$$

and thus

$$\frac{dG_m^\gamma}{du} = 2e_\gamma e_\xi \int_{-\infty}^{+\infty} \delta'(s^2_{\xi(u),x})(r_m\gamma' \cdot \xi' - \xi'_m\gamma' \cdot r - \gamma'_m\xi' \cdot r)\, du \qquad (39)$$

Now the interesting thing here is that if we look at G^ξ rather than G^γ, then we obtain a similar expression, but with the sign of r reversed. (If only the retarded distances were to be taken—as in the Maxwell theory—then the expressions would be identical, up to the change in sign.) Wheeler and Feynman next define the expression

$$\begin{aligned} G_m(\gamma, \xi) = {} & m_\xi \xi'_m + m_\gamma \gamma'_m \\ & + 2e_\gamma e_\xi \int_u^{+\infty} \int_{-\infty}^{v} - \int_{-\infty}^{u} \int_v^{+\infty} \delta'(s^2_{\xi(u),\gamma(v)}) \\ & \times (r_m\gamma' \cdot \xi' - \xi'_m\gamma' \cdot r - \gamma'_m\xi' \cdot r)\, du\, dv \end{aligned} \qquad (40)$$

This is the mth component of the total energy-momentum vector G for the system $\Gamma = \{\gamma, \xi\}$, where $m = 1, \ldots, 4$. Clearly G is constant along the paths, since the partial derivatives of G with respect to path lengths vanish. The cases $m = 1, 2, 3$ are the components of three-dimensional momentum, and $m = 4$ is the energy. For the case of two stationary particles, the expression for the energy reduces to the usual Coulomb expression $e_\gamma e_\xi / R$, where R is the three-dimensional distance between them. More generally, the principle of conservation of energy implies that if the signs of the electrical charges of the two particles are similar, then—assuming that the two particles satisfy Fokker's variational priciple—the two particles cannot approach one another more closely than some given distance related to the fixed energy of the system.

2.8. Some Simple Solutions to the Fokker Theory

In this section we examine a number of configurations of particle paths in $\mathbf{R}^4$ that satisfy either Fokker's variational principle that J_Γ = extremum or at least the weaker condition (35). These solutions will illustrate, in particular, the fact that any action-at-a-distance theory can only be considered in global terms.

Solution 1. At Most One Charged Particle. To begin, there are the trivial cases. These are that $\Gamma = \varnothing$, or that Γ contains only uncharged particles ($e_\gamma = 0$, $\gamma \in \Gamma$), or that at most one particle in Γ has a nonvanishing electrical charge.

Solution 2. Two Positively Charged Particles Lying in a Plane. The simplest nontrivial case is represented by two charged particles $\Gamma = \{\gamma, \xi\}$. Assume that $e_\gamma = e_\gamma = m_\gamma = m_\gamma = 1$ and that $\gamma_2 = \xi_2 = \gamma_3 = \xi_3 = 0$, so that the particles lie in the x-t hyperplane in $\mathbf{R}^4$.

Now it might be thought that the class of all Γ satisfying Fokker's variational principle subject to all these restrictions must be very small, or even empty. But this is by no means the case. On the contrary:

Theorem 2.1. Let (u_1, u_2), $u_1 < u_2$, be an open interval in $\mathbf{R}$, and choose $f: (u_1, u_2) \to \mathbf{R}$ to be some differentiable function with the property that $|f'(x)| < 1 - \varepsilon$ for some $\varepsilon > 0$ such that f' is convergent from the left at u_2 and from the right at u_1. Assume further that f is twice differentiable. Then there exists a set $\Gamma = \{\gamma, \xi\}$ satisfying (35) such that $\gamma_1(x) = f(x)$ and $\gamma_4(x) = x$ for all $x \in (u_1, u_2)$.

Proof. We are given the segment $\gamma(x)$, $x \in (u_1, u_2)$. Our first task is to construct ξ, and also $\gamma(x)$, for $x \notin (u_1, u_2)$. The restriction on the absolute value of the derivative of f ensures that γ is timelike between u_1 and u_2. Choose the point $P = (P_1, P_2, P_3, P_4) \in \mathbf{R}^4$ such that $P_1 = P_2 = 0$ and such that the conditions $\gamma_4(u_0) - P_4 = \gamma_1(u_0) - P_1$ and $\gamma_1(u_0) - P_1 = \gamma_1(u_0) - P_1$ hold. Thus P is *determined by the pair* $(\gamma(u_1), \gamma(u_2))$ in an obvious way. Let $v_4 = P_4$ and define $\xi(v_4) = P$. Next choose P^* to be *any* point with $P_4^* - \gamma_4(u_1) = P_1^* - \gamma_1(u_1)$ and $|P_4^* - P_4| < |P_1^* - P_1|$ (Figure 2).

Thus, we shall say that P^* is *diagonally above* $\gamma(u_2)$, and *above* P. We define $v_2 = P_4^*$ and $\xi(v_1) = P$. One sees then that we also have a great deal of freedom in choosing ξ. In fact, subject to a few small restrictions, we can also choose ξ to be almost arbitrary between $\xi(v_1)$ and $\xi(v_2)$. The restrictions are, again, that ξ be twice differentiable between v_1 and v_2 and that $|\xi_1'| < 1$ in this region. Furthermore, we assume that the second derivatives of ξ_1 exist from the right at $\xi(v_1)$ and from the left at $\xi(v_2)$. Equation (35) then gives a condition that $\xi_1''(v_1)$—evaluated from the right—must fulfill. But given this specification at v_1, equation (35) then gives a condition that $\xi_1''(v_2)$ must fulfill.

It may now be assumed that some suitable ξ has thus been defined in $[v_1, v_2]$. The next step in the construction is to extend smoothly the definition of γ above v_2 up to a point that is diagonally above $\xi(v_2)$. This is just a matter of finding the unique solution to (35) in this region. One then extends the definition of ξ above v_2, and so on. The region of definition of γ and ξ can also be extended below u_1 and v_1, respectively, in an analogous manner. The fact that e_γ and e_ξ are of the same sign ensures that at each stage of the construction the paths are disjoint. It is further necessary to show that the paths that have thus been constructed are infinitely long in both directions. This follows from the energy-momentum conservation principle that Γ must satisfy. ■

We now have a large class of sets that satisfy the principle of stationary action with respect to neighboring, sufficiently smooth paths. Of course this does not yet imply (15), even when we restrict our attention to local (but

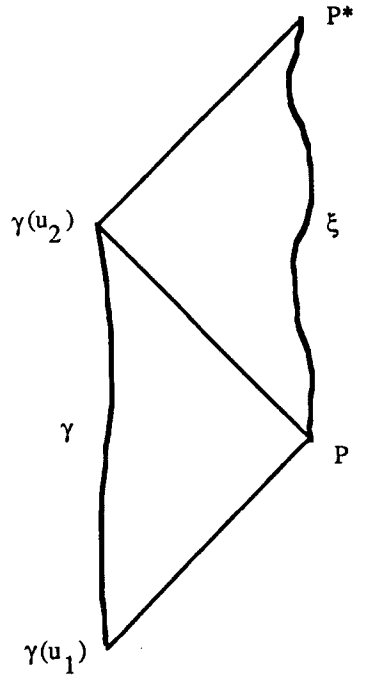

Fig. 2

possibly not particularly smooth) variations. But at least we can prove a number of related results concerning the smoothness of paths as a "local" property.

Definition 1. Let Γ be a locally finite collection of particles in $\mathbf{R}^4$. I call Γ *locally proper at* $\gamma(t)$, where $\gamma \in \Gamma$ and $t \in \mathbf{R}$, if there exists a $\delta > 0$ and $\Delta > 0$ such that for all smooth variations of γ between $\gamma(t-\delta)$ and $\gamma(t+\delta)$, the expression J^r_γ defined in Section 2.6 is well defined and continuous in r for $|r| < \Delta$. I call Γ *proper* if this condition holds for all $\gamma \in \Gamma$ and all $\gamma(t)$, $t \in \mathbf{R}$.

Definition 2. Let Γ be locally proper at $\gamma(t)$. Then Γ is a *local minimum* at $\gamma(t)$ if there exists a $\delta > 0$ such that for all smooth variations of γ between $\gamma(t-\delta)$ and $\gamma(t+\delta)$, the inequality $J^0_\gamma \leq J^r_\gamma$ holds, $\forall r$ with $|r| < \Delta$, for a suitable $\Delta > 0$.

Definition 3. With Γ, γ, t as above, Γ is an *extended local minimum* at $\gamma(t)$ if it is a local minimum at $\gamma(t)$ with respect to variations β of γ that are smooth, except possibly at a single point between $\gamma(t-\delta)$ and $\gamma(t+\delta)$ where β is continuous, but possibly β' is not continuous. At this point,

however, both β'_+ and β'_- (the limits of the derivatives in the approach to the point of singularity) exist. (Note that the expression J_Γ is also meaningful with respect to collections of paths Γ that are only piecewise smooth.)

Theorem 2.2. Let $\Gamma = \{\gamma, \xi\}$ be a set of two smooth, disjoint paths in $\mathbf{R}^4$. Assume that e_γ, e_ξ, m_γ, m_ξ are all positive. Assume further that Γ is a local minimum at $\gamma(t)$ for some $t \in \mathbf{R}$. Then Γ is also an extended local minimum at $\gamma(t)$.

Proof. In order to produce a contradiction, assume that there does exist a piecewise smooth variation β of γ between $\gamma(t-\delta)$ and $\gamma(t+\delta)$ for an appropriate $\delta > 0$, which is such that $J^1_\Gamma < J^0_\Gamma$ with respect to β. Assume that β is smooth, except for a single point $\beta(t_0)$, where it is continuous, but not differentiable. Now for any small ε with the property that $0 < \varepsilon < \delta$, we can alter β between $\beta(t-\varepsilon)$ and $\beta(t+\varepsilon)$ in such a way that we obtain a smooth variation β_* that agrees with β below $\beta(t-\varepsilon)$ and above $\beta(t+\varepsilon)$. Furthermore, $\beta'_*(s)$ is between $\beta'(t_0-\varepsilon)$ and $\beta'(t_0+\varepsilon)$ in value. But if (14) is examined, noting that for small ε the electromagnetic field due to ξ is nearly constant, it is seen that the difference between the value of J^1_Γ due to β and the value of J^1_Γ due to β_* has as upper limit an expression of the form const $\times \varepsilon$. Thus, by choosing ε small enough, we can reduce this difference to below the value $|J^1_\Gamma - J^0_\Gamma|/2$, contradicting the assumption that Γ is a local minimum at $\gamma(t)$. ■

This theorem shows that if one knows that one of the paths in Solution 2 is smooth, then the other path should be smooth as well. (One must be careful here. I have *not* shown that the sets produced according to Solution 2 are local extrema—or even extended local extrema—at all points!) This would appear to lead to the idea that piecewise-smooth paths—with genuine kinks—can never appear in solutions to Fokker's variational principle. But consider the following idea. In the construction of sets Γ according to Solution 2, we required that the paths be smooth, and in particular smooth at the points P and P^*. What if we no longer require the extension of ξ, say, above P^* to be differentiable at P^*? In this case the advanced Liénard-Wiechert potential below P^* is discontinuous at $\gamma(u_2)$. Thus, the construction of γ above $\gamma(u_2)$ must also produce a kink at $\gamma(u_2)$. One may approximate this solution by paths that are smooth, but sharply curved near P and P^*, etc. Now the principle of conservation of momentum shows that they approach a definite solution with a kink. Thus, one can imagine such two-particle solutions to Fokker's principle in which an infinite chain of such discontinuous exchanges of energy-momentum along the light cones between the two particles occur, similar to the photon exchanges of quantum electrodynamics.

Solution 3. Two Spiralling Particles. The last solution I will consider is concerned with the case $\Gamma = \{\gamma, \xi\}$, with both particles having the same mass and a nonzero charge, satisfying $e_\gamma = -e_\xi$. Let Γ describe a double helix,

$$\gamma(t) = (\sin(kt), \cos(kt), 0, t)$$

$$\xi(t) = (-\sin(kt), -\cos(kt), 0, t)$$

where $k < 1$. Then it is clear from the symmetry of this set that, for an appropriate choice of the mass $m_\gamma = m_\xi$, the paths satisfy (35).

This example illustrates an important difference between the Fokker and the Maxwell theories. In the case of the Maxwell theory, two such spiralling particles will transfer energy to an electromagnetic field, and thus the spiral is unstable. The Fokker theory requires an "absorbing universe" (this concept will be developed in the sequel) in order to mimic the Maxwellian electromagnetic fields. If the universe is not a complete absorber, then the "fields" of the Fokker theory do not have the same properties as those of the Maxwell theory. This situation is an obvious consequence of the basic action-at-a-distance philosophy: in the absence of an absorber there can be no radiation!

2.9. The Principle of Cause and Effect

It was mentioned in Section 2.3 that for any reasonable theory of the physical world the effects of a given physical process should only be felt *afterward.* Thus, effect should follow cause in time. Now, if the advanced Liénard–Wiechert potentials are simply excluded, as is usual in the Maxwell theory, then all processes in classical electrodynamics progress forward in time, and thus the effect will always follow the cause. But, as we have seen, solutions to Fokker's variational principle represent a mixture of half advanced and half retarded fields. This appears to violate the principle of cause and effect, and for this reason the Fokker theory was considered for some years to be invalid. Nevertheless, Wheeler and Feynman (1945) showed that it may be possible to have advanced fields without violating this principle. In this section I will examine some of Wheeler and Feynman's arguments.

To begin with, I must point out a number of difficulties. "Does the Fokker theory violate the principle of cause and effect?" In a sense this question can never be satisfactorily answered. The problem is that the question has more to do with philosophy than it does with mathematics. What shall we consider to be a cause, and what an effect? Perhaps one could even restrict one's definition of the idea of a "cause" to be just an action that one can think of as being initiated by a human being. But the reduction of the question to such an obscure level of philosophical debate

could hardly lead to a satisfactory result! It is better to seek a mathematical formulation.

One could, for example, ask whether a given set Γ of particles satisfying Fokker's variational principle could occur in a solution to the Maxwell theory with purely retarded potentials. This would certainly imply that Fokker's theory does not violate the principle of cause and effect. Unfortunately, such an approach must lead to insuperable mathematical difficulties: it is simply impossible to work with exact global solutions to the Maxwell theory: one is forced to adopt the usual assumptions—vague statistical arguments, assumptions concerning the regularity of global solutions, and even unrealistic cosmological models. But still, granted these limitations, it is worthwhile to see whether or not an affirmative answer to our basic question can at least be made to *appear* to be plausible.

Wheeler and Feynman discuss four different derivations of the relevant effect, involving various assumptions. Perhaps the most appealing and instructive derivation for us is the fourth one. This derivation makes use of the results of Dirac, already mentioned in Section 2.2. These results concern the radiative damping force in classical electrodynamics.

To be specific, let us return to the Lorentz equation, as formulated in Section 2.6. It was asserted there that $m_\gamma \gamma_i'' = e_\gamma F_{ik} \gamma_i'$, $i = 1, \ldots, 4$. But which fields F_{ik} should be taken? In the Fokker theory one has no choice—one must take the sum of half the retarded and half the advanced fields, with self-interactions excluded—but what about the Maxwell theory? It would be possible in this setting to choose just the retarded Liénard-Wiechert potentials produced by the other particles $\xi \neq \gamma$ in Γ. But such a choice would be false, since it would ignore the (infinite and "singular") field of the particle γ itself. It would also fail to account for the so-called radiative damping effect, in which any accelerated classical charged particle emits electromagnetic radiation, and thus loses energy (an effect that can be observed in simple experiments). According to the Maxwell theory, this damping effect must occur even in a universe that is empty, except for the one particle. Thus, it is reasonable to try to account for the radiative damping effect in terms of the self-interaction of the particle on itself.

Dirac writes $(2e_\gamma/3)(\gamma_i''' \gamma_j' - \gamma_j''' \gamma_i')$ to describe the additional radiation damping term—to be added onto the retarded fields F_{ij} coming in from the other particles in Γ—which results in an electromagnetic field with the property that the resulting Lorentz equation describes the motion of the particle γ correctly. Using a Taylor expansion to describe the self-interaction of γ, he concludes that this is equal to $\frac{1}{2} e_\gamma (F^\gamma_{ij,\mathrm{ret}} - F^\gamma_{ij,\mathrm{adv}})$, where F^γ_{ret} and F^γ_{adv} are the retarded and advanced fields due to γ, respectively. (Note that the fields are to be taken *directly at the particle* γ, where they are "infinite"!) Granted the validity of Dirac's formula, one must ask how it fits in with

the Fokker theory, where such advanced and retarded fields play an important role.

Wheeler and Feynman consider the following, somewhat unrealistic model. They imagine that everything should be described within the framework of $\mathbf{R}^4$ and consider the motion of some particle γ in a universe of particles Γ. All of these particles produce their own electromagnetic fields; in particular, the field produced by the particle γ is given by

$$F^\gamma = \tfrac{1}{2}(F^\gamma_{\text{ret}} + F^\gamma_{\text{adv}}) \tag{41}$$

Now they assume that this universe of particles is, in some sense, of limited extent, so that it is possible to speak about points "outside" the universe of particles. They assume that the radiation F^γ causes movements of the surrounding particles of Γ, so that eventually the radiation is "absorbed." Thus, they write

$$\tfrac{1}{2}\sum_\Gamma (F^\gamma_{\text{ret}} + F^\gamma_{\text{adv}}) = 0 \qquad \text{(outside the absorber)} \tag{42}$$

where the sum is taken over all particles $\gamma \in \Gamma$. Then then claim that this equation implies the vanishing of both the retarded and the advanced fields separately,

$$\begin{aligned} &\sum_\Gamma F^\gamma_{\text{ret}} = 0 \qquad \text{(outside)} \\ &\sum_\Gamma F^\gamma_{\text{adv}} = 0 \qquad \text{(outside)} \end{aligned} \tag{43}$$

But this in turn implies that

$$\sum_\Gamma (F^\gamma_{\text{ret}} - F^\gamma_{\text{adv}}) = 0 \qquad \text{(outside the absorber)} \tag{44}$$

At this stage it is remarked that the difference of the advanced and retarded fields has no singlarities, so that it must vanish everywhere. (This assumption brings with it at least the requirement that the particle paths be sufficiently smooth! Dirac (1938) showed that for such paths, the difference is free of singularities.) Thus, the electromagnetic field at the particle γ in the Fokker theory, due to all the other particles $\xi \in \Gamma - \{\gamma\}$, is given by

$$\begin{aligned} &\sum_{\xi \neq \gamma} \tfrac{1}{2}(F^\xi_{\text{ret}} + F^\xi_{\text{adv}}) \\ &\quad = \sum_{\xi \neq \gamma} F^\xi_{\text{ret}} + \tfrac{1}{2}(F^\gamma_{\text{ret}} - F^\gamma_{\text{adv}}) - \sum_\Gamma \tfrac{1}{2}(F^\xi_{\text{ret}} - F^\xi_{\text{adv}}) \\ &\quad = \sum_{\xi \neq \gamma} F^\xi_{\text{ret}} + \tfrac{1}{2}(F^\gamma_{\text{ret}} - F^\gamma_{\text{adv}}) \end{aligned} \tag{45}$$

This is precisely the sum of the pure *retarded* fields coming from all the other particles in the universe Γ, plus the usual radiation damping term of the Maxwell theory! This means that despite the presence of advanced

fields, the principle of cause and effect can be made to hold in the Fokker theory.

One sees that the key to this argument is the valdity of the formula

$$\sum_{\Gamma} (F^{\xi}_{\text{ret}} - F^{\xi}_{\text{adv}}) = 0 \tag{46}$$

which was derived from the separate vanishing of the sums of the advanced and retarded fields "outside the absorber."

2.10. Perfect Absorbers of Radiation

Is it possible to prove that the formula (46) is true for all sets of particles Γ that satisfy Fokker's variational principle? The answer is no.

Counterexample 1. Let $\Gamma = \{\gamma, \xi\}$ be the set of Solution 3 in Section 2.8. Let

$$A_{\text{adv}} = A^{\gamma}_{\text{adv}} + A^{\xi}_{\text{adv}}, \qquad A_{\text{ret}} = A^{\gamma}_{\text{ret}} + A^{\xi}_{\text{ret}}$$

Write $A^* = A_{\text{ret}} - A_{\text{adv}}$. From the symmetry of the set it is clear that along the axis $\{x \in \mathbf{R}^4 : x_1 = x_2 = x_3 = 0\}$ we have $A^* = 0$. In particular, we must have $\partial A^*/\partial x_4 = 0$ here. Now let 0 be the point $P = (0, 0, 0, 1) \in \mathbf{R}_4$ on the axis. Clearly P lies diagonally above the points $(1, 0, 0, 0) \in \gamma$ and $(-1, 0, 0, 0) \in \xi$. Furthermore, the distance from P to each of these two points is the same, and since the three-dimensional velocities of the particles γ and ξ are of identical magnitudes and opposite directions, the retarded field A_{ret} must vanish at P. But this is true along the line $L = \{(0, y, 0, t) \in \mathbf{R}^4 : t = (1 + y^2)^{1/2}\}$. On the other hand, depending on the slope of the spiral determined by γ and ξ (the constant k in Solution 3), as we progress along the line L from P we approach more nearly one of the particles and get farther away from the other. Thus, the difference between the advanced distances changes as we progress along L through P, and so $\partial A_{\text{adv}}/\partial x_2 \neq 0$ at P. This implies that $\partial A^*/\partial x_2 \neq 0$ at P, and hence $F_{\text{ret}} - F_{\text{adv}} \neq 0$ at P also, where F_{ret} and F_{adv} are, respectively, the retarded and advanced fields generated by γ and ξ together.

This example shows that the principle of cause and effect, as described in Section 2.9, does not hold for all possible solutions to Fokker's variational principle: it is necessary to look for solutions that are, additionally, "perfect absorbers" of radiation. How can this idea of a perfect absorber be defined?

The intuitive idea is simply to imagine a particle γ undergoing various accelerations and then to consider the retarded radiations being emitted by the particle. One imagines that the universe contains a large collection of other particles that, in the absence of the influence of γ, would behave in a certain way. But then the radiations due to γ can be thought of as perturbing the other particles, thus causing them to emit both additional retarded and advanced radiations. The retarded radiations from γ are,

according to this picture, perfectly absorbed if, when added to all the additional radiations due to the perturbations, the result approaches zero sufficiently rapidly, far from γ. The first three derivations in Wheeler and Feynman (1945) are based on such reasoning. But how are we to understand it? Since the Fokker theory is based on a global variational principle, it is difficult to see what meaning can be attached to the idea of small perturbations to other *global* solutions to the Fokker variational principle. In fact, the argument is really based on the Maxwell theory, and it is tacitly assumed that sensible solutions to the Maxwell theory are, at the same time, also solutions to the Fokker theory. Now, in view of the seeming impossibility of actually obtaining sensible global solutions to either theory, it is perhaps best to adopt this style of argument, and thus simply to assume that all global solutions to the Maxwell theory, which appear to have the property of perfect absorption of radiation in this sense, are also global solutions to the Fokker theory.

However, even if we accept this reasoning, there is still another, seemingly irrefutable objection to Wheeler and Feynman's argument. This objection is that everything is symmetric in time: it is possible to simply exchange the qualifiers "adv" and "ret" throughout all of the formulas in Section 2.9, and after this exchange the validity of the formulas must remain unchanged! Wheeler and Feynman contend with this argument by noting that there is a certain asymmetry in time associated with the "initial conditions." But rather than pursuing such delicate reasoning, it seems best simply to note that every global solution is necessarily asymmetric in time: the universe is expanding!

Thus, an interesting project would be to examine the absorber properties—both in the direction of the past and the future—of various cosmological models, and see whether they are consistent, according to these ideas, with (46). This has been done by Hogarth (1962) and Roe (1969) for various conformally flat models. The idea is that the principle of cause and effect will be best explained by a cosmological model that exhibits perfect absorption in the future and imperfect absorption in the past. It is certainly possible, and perhaps worthwhile, to debate the question of whether or not the observational evidence tends to support or refute various cosmological models. But, as a practical matter, from now on I simply assume that the Fokker theory provides a viable basis for classical electrodynamics.

3. DISCRETE MODEL FOR CLASSICAL ELECTRODYNAMICS

3.1. Relativity and the Ising Model

In this section I will be concerned with an attempt to find a discrete mathematical framework for space-time, whereby the considerations of

classical electrodynamics will play an important role. It seems to be most natural to begin by discussing a simple mathematical model that has often been used and found to be of value. This is the well-known Ising model. It is, for a given positive integer n, the set of points in $\mathbf{R}^n$ with integral coordinates. One can also write $\mathbf{Z}^n$ to describe this set more succinctly.

The Ising model can be most sensibly applied to the analysis of crystals, periodic building structures, and the like. But it is also possible to contemplate using it to describe the structure of four-dimensional space-time itself. Thus, whereas the space $\mathbf{R}^4$ is to be found in many places throughout any physics text, it could be proposed simply to substitute everywhere the space $\mathbf{Z}^4$ for $\mathbf{R}^4$. This procedure could be justified with the thought that if the scale of distances were chosen to be sufficiently large, then the discrete structure would become so fine as to be undetectable by means of any practical physical experiment. This procedure would have the merit of allowing the usual limiting operations of analysis to be applied, at least down to a very fine scale, but on the other hand, ultimately the discrete structure will become evident, so that the convergence problems of modern physics could be avoided.

Such a plan has, however, one very severe drawback. That is that the theory of relativity—which is universally acknowledged as providing the foundation for any discussion of the basic principles of physical space and time—will be violated in an essential and unavoidable way.

The theory of relativity has as its main premise the idea that all possible inertial frames of reference are equally valid. Thus, for example, there can be no physical experiment whose ultimate result is the determination of the speed of the earth through the classical "ether." Now, it must be admitted that the measurements of the cosmological background radiations would appear to provide a class of experiments that do, in fact, violate this basic principle of relativity. But if we add the qualification that the experiments should only test local phenomena, then the principle will continue to hold.

If we were to assume now that the Ising model formed the basis of physical space-time, then it would be possible to consider the following experiment. The experiment would consist of taking finite rectangular boxes B and counting the number of points of space-time within B. It would be found that for certain orientations of typical boxes—particularly those that are very thin—a small movement will suddenly result in the box containing either very many, or very few, points. For other orientations of B this phenomenon will not be observed. By this method the orientation of the underlying Ising model could be determined, and thus the preferred inertial frame of reference would be discovered.

One might think, then, that it is necessary to reject the principle of relativity if we are to continue with an investigation of discrete spaces as

candidates for providing good models of space-time. But this is by no means the case. On the contrary, our simple experiment suggests a way of defining a class of suitable relativistically invariant discrete models.

Consider some distribution of points throughout $\mathbf{R}^4$ with the property that (1) for any finite Borel set B, the number of points in B is finite, (2) the expected number of points in B is proportional to $\mu(B)$, the measure of B, and (3) the number of points in B_1 is independent of the number of points in B_2 for disjoint sets $B_1 \cap B_2 = \varnothing$. Now it is well known that these conditions define a Poisson process on $\mathbf{R}^4$ (see, for example, Cox and Isham, 1980). If we were to take such a process as providing in some way a basis for a discrete theory of space-time, then, by definition, our counting experiment would fail to reveal any preferred orientation of the underlying model. Note that here I am making use of the following result.

Theorem 3.1. Let $\psi: R_4 \to R_4$ be a Lorentz transformation. Then for any Borel set $B \in \mathbf{R}^4$ we have $\mu(B) = \mu(\psi(B))$.

A *Lorentz transformation* is by definition any affine transformation of $\mathbf{R}^4$ that leaves the line element $ds^2 = dx^2 + dy^2 + dz^2 - c\,dt^2$ invariant. (The four coordinate axes of $\mathbf{R}^4$ can be denoted x, y, z, t, whereby the last coordinate is taken to be the time of a typical point.)

Proof of Theorem 3.1. Without loss of generality we may assume that B is the standard four-dimensional cube B_s, the coordinates of all of whose points have values between 0 and 1. By performing translations and rotations where necessary, it is also possible to assume that $\psi(0) = 0$, and that ψ leaves the y and z axes invariant. Let p_1 be the point $(1, 0, 0, 0)$, and let $\psi(p_1) = (x_1, 0, 0, t_1)$. Then we have $t_1 = (1 + x_1^2)^{1/2}$. Similarly, if $p_2 = (1, 0, 0, 1)$, then we must have $\psi(p_2) = (x_2, 0, 0, t_2)$, where $x_2 = t_2$, and we may calculate that $x_2 = x_1 + (1 + x_1^2)^{1/2}$. Elementary geometric arguments show that the Euclidean volume of $\psi(B)$ is

$$\frac{1}{2}\left[\sqrt{2} \times x_2 \times \left(\frac{\sqrt{2}}{x_2}\right)\right] = 1$$

and therefore the measure of B is invariant under ψ. ■

At this stage, then, we have managed to improve on the simple Ising model for four-dimensional space-time: it is only necessary to examine the sample points of a standard Poisson process on $\mathbf{R}^4$. The space obtained is relativistically invariant.

Now it may be argued that my objection to the Ising model is unfair. According to this view, the actual structure of space-time is unimportant, and it should play no further role in the processes of physics; it should do nothing more than provide an unobtrusive backdrop for the physical events occurring within the space-time. Such a view may well reflect the philosophy

of the theory of relativity. Such a view may indeed also be satisfied by the picture of space-time in terms of an abstract probability space.

Nevertheless, the basic geometry of space-time should, in fact, reflect the properties of the observable world. This must lead us to question not only the Ising model, but also the model of space-time as a probability space. After all, it is one of the tenets of modern physics that only that which can be *observed* should be allowed to have a place in physics. The Ising model loses its validity as soon as it can be observed, and thus it fails this test in a most profound way. My objection to the probability space model is only slightly more subtle. It is true that attempts have been made to understand the probabilistic character of quantum mechanics in terms of an underlying probabilistic space-time. [See, for example, Nelson (1967) in this connection.] But the fact is that this way of going about it is based on the Euclidean space $\mathbf{R}^4$, which is, after all, the space with which one is supposed to be dissatisfied! The use of $\mathbf{R}^4$ and the formulation of physics in terms of differential equations implies an infinitely fine and complicated structure that we will never be able to observe in practical experiments.

3.2. Discrete, Partially Ordered Sets

If we decide to reject the idea of Euclidean space as providing a basis for the geometry of physics, then what alternatives can there possibly be? It seems best to begin by discussing things in a very abstract setting. Surely the idea of a partially ordered set must play a central role in the formulation of any physical theory. The ordering reflects the ordering of time, and by taking the definition of a *partially ordered set*, we are doing nothing more than excluding the logically impossible situation of having time running around in circles. Indeed, the space $\mathbf{R}^4$, together with the usual ordering of the theory of special relativity, is itself a partially ordered set. Therefore I will begin by considering as a model some partially ordered set W, with additional properties yet to be determined.

The main property one would like W to have is that it be *discrete.* Certainly $\mathbf{R}^4$ is not discrete, but the definition I have in mind is the following.

Definition 1. A partially ordered set W is *discrete* if, for any two elements $a, b \in W$, the set $W^a \cap W_b = \{w \in W: a \leq w \leq b\}$ is finite.

Obviously there are many possible discrete, partially ordered sets (p.o.s.). Any finite p.o.s. is discrete. Also, $\mathbf{Z}^n$ (the set of points in $\mathbf{R}^n$ with integral-valued coordinates) for any positive n, taken together with the usual Lorentz ordering, is discrete. (The Lorentz ordering is given by

$$a \leq b \Leftrightarrow \{a_n \leq b_n \text{ and } (a_1 - b_1)^2 + \cdots + (a_{n-1} - b_{n-1})^2 \leq (a_n - b_n)^2\}$$

for $a, b \in \mathbf{R}^n$.)

$\mathbf{Z}^n$ is also discrete with respect to the ordering given by

$$a \leq b \Leftrightarrow \{a_i \leq b_i, \forall i = 1, \ldots, n\}$$

One could go on and examine many further sets—but most have little to do with the requirements set by physics—and thus it seems reasonable to look for a somewhat stronger condition.

Definition 2. A partially ordered set W is *strongly discrete* if for any two elements $a, b \in W$, the set $\Lambda_{ab} = \{w \in W: w \leq a \text{ and not } w \leq b\}$ is finite.

Clearly, for all $a, b \in W$ we have $W^a \cap W_b \subset \Lambda_{ba} \cup \{a\}$, so that strongly discrete implies discrete. But, for all $n > 1$, $\mathbf{Z}^n$ combined with either of the orderings defined above is discrete, but not strongly discrete. On the other hand, any finite p.o.s. is also strongly discrete. Are there infinite strongly discrete p.o.s.? A trivial example is $\mathbf{Z}$, the set of all integers, with the usual total ordering. There are other, equally trivial examples, which are also equally inappropriate as models for space-time. But does there exist a strongly discrete p.o.s. that gives a "reasonable" approximation to $\mathbf{R}^n$, $n > 1$, say? Consider the following.

Example 1. Let W be a subset of $\mathbf{Z}^2$, with the Lorentz ordering. In fact, W is a subset of the set $Y = \{(z, -2^n) \in \mathbf{Z}^2: n \in \mathbf{N}, \text{ the positive integers}\}$. W consists of the set of points $(z, u) \in Y$ with the property that $z \equiv 0 \pmod{2u}$.

This example is obviously discrete, but is it strongly discrete? To see that is, one need only note that there can exist no two points (z_1, u_1), $(z_2, u_2) \in W$ such that $z_1 \pm z_2 = u_1 \pm u_2$. For, assume that, say, (z_1, u_1) and (z_2, u_2) are two different points such that $z_1 - z_2 = u_1 - u_2$ and $|u_1| > |u_2|$. Then we would also have $(0, u_1)$ and $(z_2 - z_1, u_2)$ satisfying this condition, where both $(0, u_1)$ and $(z_2 - z_1, u_2)$ are elements of W. But then since $W \subset Y$, we must have $|z_2 - z_1| + |u_2| = 2^n$ and $|u_2| = 2^m$ for some $n, m \in \mathbf{N}$, and at the same time we are required to have $z_2 - z_1 \equiv 0 \pmod{2u_2}$, a contradiction. Therefore any point $(z, u) \in \mathbf{Z}^2$ has at most two elements of W diagonally below it. Given any two points $s, t \in W$, the set $W^s \cap W_t$ is finite (W^s is the set of elements of W above s and W_t is the set of elements of W below t). Hence W is strongly discrete.

The set W can be extended into the upper half-plane of $\mathbf{Z}^2$. For example, let $W^* = W \cup \mathbf{Z}^2_+$, where $\mathbf{Z}^2_+ = \{(z, u) \in \mathbf{Z}^2: u \geq 0\}$. Then W^* with the Lorentz ordering is also strongly discrete. W^* has the property of being evenly distributed throughout "horizontal" slabs in $\mathbf{R}^2$. But, at least in $\mathbf{Z}^2_-$, it is very unevenly distributed in vertical slabs. In fact, W rapidly becomes very thinly spread out as we go downward through W. Also, it is not difficult to see that examples similar to this W are possible in dimensions higher than two.

Is this change in the density of points in the direction of the ordering an undesirable property in a model for space-time? Certainly it is a departure from the usual requirement of homogeneity in time, as exhibited by Minkowski space (the space $\mathbf{R}^4$ together with the Lorentz metric structure). But on the other hand, a metric that changes with time is just what is needed to model the expanding universe.

Example 2. Take $\mathbf{Z}^4$, together with the Lorentz ordering. We have seen that this set is not strongly discrete. However, we can choose any positive integer $N_0 \in \mathbf{N}$ and then define an equivalence relation on $\mathbf{Z}^4$ by means of

$$(x_1, y_1, z_1, t_1) \approx (x_2, y_2, z_2, t_2) \Leftrightarrow x_1 - x_2 \equiv 0 \bmod N_0$$
$$y_1 - y_2 \equiv 0 \bmod N_0$$
$$z_1 - z_2 \equiv 0 \bmod N_0$$

We can denote the set of equivalence classes by W_T, a set embedded in the space consisting of the product of a 3-torus and the real line. Given two points $q_1, q_2 \in W_T$, then $q_1 < q_2 \Leftrightarrow \exists p_1, p_2 \in \mathbf{Z}^4$ such that $p_i \approx q_i$, $i = 1, 2$, and $p_1 < p_2$. Clearly this makes W_T a discrete p.o.s. The fact that it is strongly discrete follows from the fact that the three-dimensional torus is compact.

It might be objected that the set W_T of Example 2 cannot provide a good model of the universe, since it is known that the later is infinite in extent. But is this true? Segal (1976) has given good reasons to support the idea that a space such as $S^3 \times \mathbf{R}$, i.e., the product of the 3-sphere and the real line, could provide a better model for cosmology than the models usually considered by physicists. For example, he shows that the cosmological redshift of light can arise through an interesting effect, more subtle than the simple Doppler explanation that is usually invoked.

His main premise seems to be the idea that when making astronomical observations from the earth, we naturally view things in the tangent space $\mathbf{R}^3 \times \mathbf{R}$. Thus, the interpretation of these observations implies (assuming $S^3 \times \mathbf{R}$ is the proper space for cosmology) some particular mapping $\mathbf{R}^3 \times \mathbf{R} \to S^3 \times \mathbf{R}$, where the null point is mapped onto the point where the astronomical observer happens to be situated. Clearly, this mapping cannot preserve distances. The standard idea is to assume that, in any case, there is no distortion in the units of time. Perhaps this reflects an unconscious desire to keep the idea of "time" as being something more absolute than "space." But from the mathematical point of view, there seems to be nothing to object to in distortions of both the units of space *and* time in the mapping $\mathbf{R}^3 \times \mathbf{R} \to S^3 \times \mathbf{R}$. In fact, rather than simply choosing the identity mapping on the second component, one could argue that it is a sensible idea to require instead that the mapping be angle-preserving. This results in a distortion of time as well, giving the cosmological red shift of light.

Such models, of the form $K^3 \times \mathbf{R}$, where K^3 is a compact three-dimensional space, are doubly attractive for us, since also they allow us to easily construct many examples of strongly discrete, partially ordered sets: Practically any discrete p.o.s. that is constructed on the basis of such a model is also strongly discrete.

3.3. Why Four Dimensions?

The title of this subsection is a question of the greatest importance when it comes to the formulation of a new geometric framework for physics. This geometry with which we are familiar—three dimensions of space and one dimension of time—is strongly characterized by its dimensionality, and one could even argue that the main difference between *geometry* and *algebra* is the fact that the former is concerned with dimension. Now, following Einstein's "General Remark," we are looking for a "purely algebraic theory for the description of reality." Thus, it is necessary to think about how the idea of dimension can be sensibly applied to the class of discrete, partially ordered sets. From the outset it must be admitted that I have been unable to find any particularly satisfactory answer to this question.

Definition 1. The discrete, partially ordered set W is n-dimensional if there exists a mapping $\Psi: W \to \mathbf{R}^n$ that is one to one and order-preserving, where the Lorentz ordering on $\mathbf{R}^n$ is being taken, and where $n \in \mathbf{N}$ is the smallest integer with this property.

This definition undoubtedly establishes the simplest and most direct connection between abstract, partially ordered sets and the question of dimension. But it leaves largely open the question of why four dimensions are imporant. For example, it is not difficult to see that for each n, $\mathbf{Z}^n$ is n-dimensional according to this definition, and so the number four seems to remain rather mysterious. One might think that if we limit ourselves to *finite*, partially ordered sets, then the construction of n-dimensional examples for arbitrary n would be more difficult. But consider the following.

Example 1. Given $n \in \mathbf{N}$, let W_n be a p.o.s. with $2^n + n$ elements. We write

$$W = \{e_1, \ldots, e_n, u_1, \ldots, u_{2^n}\}$$

The first n elements $e_1, \ldots, e_n$ are such that there are no ordering relationships between them. That is, for all $i, j \in \{1, \ldots, n\}$, we have neither $e_i \leq e_j$ nor $e_j \leq e_i$. The remaining 2^n elements $u_1, \ldots, u_{2^n}$ are determined as follows. Clearly there are 2^n possible ways that an element of W_n can be beneath some collection of elements from $\{e_1, \ldots, e_n\}$ while not being beneath the others and also not being above any of the elements of $\{e_1, \ldots, e_n\}$. The

set $\{u_1, \ldots, u_{2^n}\}$ is chosen so that each of these possibilities is accounted for by precisely one of the elements of the set.

Theorem 3.2. The set W_n of Example 1 is n-dimensional according to Definition 1.

Proof. Given that there is a one-to-one, order-preserving mapping $\Psi: W_n \to \mathbf{R}^m$ for some $m \in \mathbf{N}$, then, since W_n is finite, we can find an $(n-1)$-dimensional hyperplane $\mathbf{R}_t^{n-1}$ of the form

$$\mathbf{R}_t^{n-1} = \{(x_1, \ldots, x_{n-1}, t) \in \mathbf{R}^n : t \text{ is fixed}\},$$

which is such that $\Psi(W_n)$ lies completely above $\mathbf{R}_t^{n-1}$. Now for each element $w \in \Psi(W_n)$, let B_w be the set of points in $\mathbf{R}_t^{n-1}$ that lie beneath w. Clearly B_w is an $(n-1)$-dimensional ball whose boundary is an $(n-2)$-dimensional sphere S_w. The set of all these balls represents W_n in the sense that it forms a partially ordered set under inclusion, which is isomorphic to W_n. We may assume that $S_w \cap S_v$ is either empty, or else it is an $(n-3)$-dimensional sphere, for any two different elements $w \neq u$ in W_n. Furthermore, if we look at the set of all $(n-3)$-spheres that are thus defined, then they intersect one another either in the empty set, or else an $(n-4)$-sphere, and so on. That is, the spheres are in "general position." At this stage we use a result that is of interest in its own right.

Lemma 3.3. Let S_i, $i = 1, \ldots, m$, be spheres of dimension $n-1$ embedded in $\mathbf{S}^n$, the standard n-dimensional sphere. If $m > n+1$, then the set $\mathbf{S}^n - \{S_1, \ldots, S_m\}$ contains fewer than 2^m connected components. If $m \leq n+1$, then there are at most 2^m connected components in the this set.

Proof. We use induction on n. For $n = 1$, the assertion is obvious. (Remember that a 0-dimensional sphere consists of two points.) Therefore assume it is true for some given n. Assume furthermore that it is not true for $n+1$. Thus, there exist m n-dimensional spheres S_i, $i = 1, \ldots, m$, in $\mathbf{S}^{n+1}$ where either (case 1) $m > n+1$, such that there are at least 2^m different connected componens in $\mathbf{S}^{n+1} - \{S_1, \ldots, S_m\}$, or else (case 2) $m \leq n+1$, such that there are at least $2^m + 1$ such components. Assume that m is the smallest such integer.

Now take the sphere S_m. The sphere S_m is an n-dimensional sphere. For each $i < m$ we have that $S_i \cap S_m$ is either empty or else it is an $(n-1)$-dimensional sphere (which might degenerate into a single point). Thus, applying the inductive hypothesis, we have that $S_m - \{S_1, \ldots, S_{m-1}\}$ contains fewer than 2^{m-1} connected components in case 1, or at most 2^{m-1} such components in case 2. However, each such component can at most split a component of $\mathbf{S}^{n+1} - \{S_1, \ldots, S_{m-1}\}$ into two further pieces. In case 1 this would mean that the sum of two numbers no greater than 2^{m-1}, and one being less than 2^{m-1}, is itself at least 2^m. This is impossible. In case 2 we

would have the sum of two numbers, both no greater than 2^{m-1}, being greater than 2^m. Again impossible. ■

By noting that the set $\mathbf{R}^n$ can be considered as a kind of degenerate sphere—and noting that the proof carries over to this case—we have:

Corollary. The lemma is also true for the set $\mathbf{R}^n - \{S_1, \ldots, S_m\}$.

Remark. If all of the spheres intersect one another in further nondegenerate spheres, then in case 2 of the proof of the theorem, the number 2^m can be achieved.

Now we continue with the proof of Theorem 3.2. The fact that there exists an order-preserving embedding of W_n in $\mathbf{R}^n$ follows from the Remark. That there is no such embedding of W_n in $\mathbf{R}^{n-1}$ follows from the Corollary. ■

Example 1 shows that it is hopeless to expect to find an explanation of the fact that space-time is four-dimensional in terms of the simple Definition 1. On the other hand, this definition reflects only the (flat) geometry of the special theory of relativity. In thinking about typical examples of discrete, partially ordered sets, it often appears that a slight relaxation of Definition 1 would allow a set of high dimension to be viewed sensibly as being, in fact, of lower dimension. Furthermore, when gravity is brought in, and with it the general theory of relativity, one finds that the light-cones beneath points of space-time—representing the set of points *less than* the given point—become curved.

Now, according to the general theory of relativity, objects that appear to be deflected by the force of gravity are, in reality, following straight paths through a curve space-time. This curvature of space-time can also be directly observed—in a more conventional sense—by looking at one of the recently discovered quasars whose light is so curved that it appears on the earth to be split into a number of separate images. One is led then to the following alternative definition.

Definition 2. The set $C_s = \{(x, y, z, t) \in \mathbf{R}^4 : t \leq 0 \text{ and } x^2 + y^2 + z^2 \leq t^2\}$ is called the *standard light-cone beneath* $(0, 0, 0, 0)$. Any set $\mathbf{H}_t = \{(x, y, z, t) \in \mathbf{R}^4 : t \text{ is fixed}\}$ is called a *horizontal hyperplane in* $\mathbf{R}^4$. Any diffeomorphism $\Psi : \mathbf{R}^4 \to \mathbf{R}^4$ that maps horizontal hyperplanes onto horizontal hyperplanes is called a *level-preserving* mapping. Then a subset $C \in \mathbf{R}^4$ is called a *proper light-cone* $\Leftrightarrow$ there exists a level-preserving mapping Ψ with the property that $C = \Psi(C_s)$. Two proper light-cones C_1 and C_2 will be said to *intersect normally* if either one is contained within the interior of the other, or else there exists a level-preserving mapping, mapping the standard light-cones beneath the points $(0, 0, 0, 0)$ and $(2, 0, 0, 1)$ onto C_1 and C_2. A discrete, partially ordered set W will be said to have a *proper*

representation in $\mathbf{R}^4$ if there exists a set of proper light-cones Q in $\mathbf{R}^4$ that are such that the interection of each pair of cones in Q is normal and a one-to-one, order-preserving mapping from W to Q (Q is naturally ordered through set inclusion.)

It seems reasonable to require that any discrete, partially ordered set that is to be considered as a model for space-time should have a proper representation in $\mathbf{R}^4$. By replacing $\mathbf{R}^4$ with $\mathbf{R}^n$ throughout Definition 2 for any $n \in \mathbf{N}$, we obtain the idea of a proper representation of a given discrete p.o.s. in $\mathbf{R}^n$.

Definition 3. The discrete partially ordered set W is n-dimensional if there exists a proper representation of W in $\mathbf{R}^n$ and n is the smallest such integer.

It is not difficult to find discrete p.o.s. that are n-dimensional, according to Definition 3, for $n = 1, \ldots, 4$. But I have been unable to find an example of a finite, discrete p.o.s. that is n-dimensional for some $n > 4$.

The question can be more easily resolved, however, if the definition of proper representations is strengthened somewhat.

Definition 4. The discrete p.o.s. W is *strongly n-dimensional* if there exists a proper representation of W in $\mathbf{R}^4$ such that there exists a horizontal hyperplane $\mathbf{H}_t$ with the property that each proper cone C in the set Q (as in Definition 2) intersects $\mathbf{H}_t$ in a three-dimensional ball. Furthermore, $\mathbf{H}_t$, together with all these intersections, is topologically equivalent to the space in the (four-dimensional case of the) proof of Theorem 3.2.

Now, if Definition 4 is chosen, we obtain an idea of dimensionality that, arguably, is appropriate for dealing with the curvatures required by gravitation, and yet the sets W_n of Example 1 are in each case n-dimensional according to this definition as well. Below I examine a completely different definition of dimensionality, based on the idea of "particle paths."

3.4. Particle Paths in Discrete Sets

Until now, I have made no use of the fact that the discrete model for space-time is to be based on Fokker's theory of classical electrodynamics. But the most superficial examination of this theory will reveal the fact that it is concerned exclusively with the behavior of particles—rather than points, as in the usual field theory—in $\mathbf{R}^4$. Now, if we are to adopt the idea of discrete, partially ordered sets with proper representations in $\mathbf{R}^4$ as our model for space-time, then it will be necessary to progress a step further and define "particle paths" within such sets.

There is, admittedly, a certain difficulty in this. In classical electrodynamics, particles are infinitely long. (I am excluding "big bang" cosmologies

here, since they are not usually discussed in the context of classical electrodynamics.) On the other hand, in quantum mechanics (a theory itself based on the idea of particles) particle creation and annihilation events play a role. Thus, here I define a concept of discrete, partially ordered sets with a particle structure in which the particles are infinitely long. But I certainly will not exclude the possibility of generalizing this definition to include the phenomena of creation and annihilation, vacuum loops, and in general all of the particle structures normally considered in Feynman diagrams.

Definition 1. The discrete, partially ordered set W has a *particle structure* if W is the disjoint union of a set of totally ordered subsets P_i, $i \in \mathcal{I}$, for some index set $\mathcal{I}$. For each i, P_i is isomorphic to the integers $\mathbf{Z}$, considered as a totally ordered set, and the ordering on P_i is the ordering inherited from W.

It is interesting to look for various examples of sets with particle structure. Clearly there can be no finite p.o.s. with particle structure; the most trivial true example must surely be $\mathbf{Z}$ itself. But perhaps it is best to combine the idea of particle structure with the concepts met in Section 3.2. For example, does there exist a strongly discrete, partially ordered set with particle structure?

Example 1. Let $W = \{(i, j) \in \mathbf{Z}^2 : |j| \leq |i|\}$. The set w is given the following ordering: $p = (p_1, p_2) \leq q = (q_1, q_2) \Leftrightarrow p_2 \leq q_2$. The set W has a simple particle structure given by the rule that $p = (p_1, p_2)$ and $q = (q_1, q_2)$ lie in the same particle $\Leftrightarrow p_1 = q_1$.

This example shows that such sets do indeed exist. It also illustrates a further interesting point: W is symmetric with respect to the ordering relation, i.e., the mapping $\Psi : W \to W$ given by $\Psi((p_1, p_2)) = (p_1, -p_2)$ is an order-reversing correspondence, which shows that W would be an equally good example if the symbol $\geq$ had simply been substituted for the symbol $\leq$. But it could hardly be said that W is evenly distributed throughout horizontal slabs of $\mathbf{R}^4$.

Example 2. Let W be the space of Example 2 in Section 3.2. The obvious particle structure here is given by the rule that two points in W lie on the same particle precisely when they lie on the same vertical line.

3.5. Another Approach to the Question of Dimension

In Section 3.3, the idea of "dimension" for discrete, partially ordered sets was discussed in terms of embeddings or representations in Euclidean space $\mathbf{R}^n$. But the fact that Fokker's theory is based so strongly on the idea of particles leads naturally to the thought that perhaps the dimensionality of space-time is determined in some way by this particle structure and the

requirement that the expression (14) should be an extremum. The first term in (14), corresponding to Newton's first law of motion, is of less interest in this regard, but the second term, which describes particle interactions, may have some relevance. In fact, after studying the expression, it becomes clear that it describes the interactions of *pairs* of particles. The behavior of a large number of particles is determined by the sum of the pairwise interactions, thus confirming the linear character of classical electrodynamics. Let us therefore take some pair $(\gamma, \xi) \in \Gamma$, with $\gamma \neq \xi$. The fact that both γ and ξ are timelike and infinitely long means that any point x in space-time is uniquely associated with four points on $\gamma \cup \xi$, namely the intersections of the light cones above and below x with the two particles γ and ξ. Now, if one is prepared to accept this philosophy—that is, say, that the points of space-time should really be thought of as being 4-tuples of points on typical pairs of particles—then it is clear that a conception of four dimensions of space-time must come about.

But one cannot, in general, simply identify the space of such 4-tuples with the usual 4-dimensional space $\mathbf{R}^4$. For, let us call the space of all such 4-tuples on (γ, ξ), $\mathbf{R}_{\gamma\xi}$. Then we have described a mapping $\Psi: \mathbf{R}^4 \to \mathbf{R}_{\gamma\xi}$. This mapping depends strongly on the paths γ and ξ, but it is clear that no matter how γ and ξ are chosen, Ψ cannot be everywhere one to one. For example, if γ is given by $\gamma(t) = (1, 0, 0, t)$, $\forall t \in \mathbf{R}$, and $\xi(t) = (-1, 0, 0, t)$, then we have $\Psi((x, y_1, z_1, t)) = \Psi((x, y_2, z_2, t))$ if $y_1^2 + z_1^2 = y_2^2 + z_2^2$. This example might serve to damp our enthusiasm for the idea, but on the other hand, observe that, in this special case, at least one dimension seems to be lost due to the extreme symmetry of the particle pair (γ, ξ). The following example is, however, even worse. Let $\gamma(t) = ([t^2+1]^{1/2}, 0, 0, t)$ and $\xi(t) = (-[t^2+1]^{1/2}, 0, 0, t)$, $\forall t \in \mathbf{R}$. It is not difficult to see that in this case, $\Psi^{-1}(\mathbf{R}_{\gamma\xi}) = \varnothing$! But it is difficult to imagine that such a situation could arise in any sensible solution to classical electrodynamics. Therefore, it is reasonable to make the following definition.

Definition 1. A particle γ in $\mathbf{R}^4$ is *regular* if for each point $x \in \mathbf{R}^4$, there exist $t_a, t_b \in \mathbf{R}$ with $\gamma(t_a)$ above x and $\gamma(t_b)$ below x.

From now on I assume that all particles considered are regular.

Definition 2. Let γ and ξ be two nonintersecting, regular particles in $\mathbf{R}^4$. The pair (γ, ξ) is in *general position* if the mapping $\Psi: \mathbf{R}^4 \to \mathbf{R}_{\gamma\xi}$ is at most two to one [i.e., $\Psi^{-1}(v)$ consists of at most two points for each $v \in \Psi(\mathbf{R}^4)$].

Theorem 3.4. There exist pairs of paths in $\mathbf{R}^4$ that are in general position.

Proof. Define the pair (γ, ξ) by means of the following rule: $\forall t$, $\gamma(t) = (0, 0, 0, t)$ and $\xi(t) = (1, t/2, 0, t)$. Now let $p_1 < q_1$ be two points on γ. Then

the intersection of the cone of points diagonally above p_1 and the cone of points diagonally below q_1 is a 2-sphere S^2 in some hyperplane H_T of the form $H_T = \{(x, y, z, t) \in \mathbf{R}^4 : t = T\}$. On the other hand, if $p_2 < q_2$ are two points on ξ, then the intersection of the cone of points diagonally above p_2 and the cone of points diagonally below q_2 is a two-dimensional ellipsoid E^2 that lies skew to H_T. Now, $E^2 \cap H_T$ is either empty, or else a single point, or else a one-dimensional sphere (circle) whose axis in H_T does not pass through $(0, 0, 0, T)$. Only the last case could lead to more than one point being in $E^2 \cap S^2$, but even here there can certainly be no more than two points in the intersection. ■

The use of the term "general position" suggests the idea that, given *any* pair of regular particles (γ, ξ) in $\mathbf{R}^4$ and a sensible idea of distances between particles, then there exists another pair of particles (γ_1, ξ_1) that are "near" to (γ, ξ), such that (γ_1, ξ_1) are in general position. However, this is not true. For example, one can construct pairs of particles that spiral about one another and cannot be near a pair in general position. But the properties of such examples are tedious to prove and unimportant for our further purposes. In any case, it might be thought that the points at which the condition of Definition 2 does not hold are rather special, being due to the properties of a continuous space, and thus one could imagine that when carrying these ideas over to the discrete case, nothing will be lost if it is assumed that all pairs of particles are, in fact, in general position.

3.6. Positions in Discrete, Partially Ordered Sets

Fokker's theory is concerned with particle paths, but beyond this, it is expressed in terms of a *variational principle.* Given any set of particle paths Γ, we are expected to assign some value J_Γ to Γ, and the idea is that a set with the property that J_Γ is an extremum can be thought of as representing a possible universe of particles. But this valuation function J_Γ is defined in terms of distances within $\mathbf{R}^4$. Our goal is to describe space-time in terms of discrete partially ordered sets, so it is clear that we will need to adopt some reasonable conception of "distances" within such sets. What possibilities are there?

One might, for example, take the fact that in a discrete, partially ordered set W there are at most finitely many points between any two points $a, b \in W$. Then a simple definition would be that the distance between a and b is given by this number. This definition only makes sense for pairs of points related to one another (either $a < b$ or $b < a$) within the ordering on W. However, it is usual to define distances in the theory of relativity by means of thought experiments that involve bouncing beams of light back and forth between spatially different points. (Practical modern surveying makes use

of this technique as well.) An analogous definition for distances between points $u, v \in W$ with neither $u \leq v$ nor $v \geq u$ could be established for discrete p.o.s. But this approach to the question of distances must surely be inappropriate. The reason is not difficult to see.

First of all, W—our discrete model for space-time—is not to be an abstract, unobservable backdrop for the processes of physics. On the contrary, we would like it to be defined *in terms of these processes themselves.* But this leads to the following thought. In the center of a very dense collection of particles—for example, in a star—we would expect there to be a great many points of W between any two given points $a_1 < b_1$. On the other hand, far away in interstellar space, if we are given two other points $a_2 < b_2$ that are, according to conventional ideas, approximately the same distance apart as a_1 and b_1, then we would expect to have much fewer points between a_2 and b_2 than between a_1 and b_1. Thus, distances in W would (comparatively speaking) become "compressed" in interstellar space and "extended" within the interior of stars. Now indeed, the general theory of relativity is based on the idea that the presence of matter alters the metrical properties of space-time, and in fact it also predicts such an "extension" of space-time in the presence of matter. But this effect in general relativity is very much smaller than what would be given by our proposed definition. In addition, gravity is felt over long distances, but this definition of distances in W would only produce a very short-range gravitational effect.

Therefore we are prompted to look for another idea, and, for a number of reasons, the following definition seems to suggest itself.

Definition 1. Given a discrete, partially ordered set W, a *position* in W is a subset $C \subset W$ that itself can be decomposed into two nonempty subsets $C = C^+ \cup C^-$ such that: (i) $x \leq y$ for all $x \in C^-$, $y \in C^+$, and (ii) C is *maximal* in the sense that there exists no set $D \subset W$, with $D = D^+ \cup D^-$, satisfying condition (i), such that $C^+ \subset D^+$ and $C^- \subset D^-$, and yet $D \neq C$.

It will prove convenient to work with the set of *positions* in a given partially ordered set W, rather than with the elements of W directly. In order to get some feeling for this concept, I will examine a number of examples, and present a new definition of discreteness.

Theorem 3.5. Let W be a discrete, partially ordered set, and let $a \in W$. Then $A = W^a \cup W_a$ is a position, where $W^a = \{w \in W: w \geq a\}$ and $W_a = \{w \in W: w \leq a\}$.

Proof. Property (i) of Definition 1 is trivially satisfied by A. Now let D be a set as in property (ii) of the definition. Let $z \in D^-$. Then $z \leq y$ for

all $y \in W^a$. In particular, $z \leq a$, and therefore $z \in W_a$. Thus, $D^- = W_a$. The fact that $D^+ = W^a$ follows by symmetry. ■

Remark. It may be interesting to consider the following definition. We have just seen that $W^a \cup W_a$ is always a position. But is, say, the set

$$X_a = \{w \in W: w \geq a\} \cup \{w \in W: w < a\}$$

also a position? Now, if W is finite, then clearly there must exist an a such that X_a is *not* a position. Let us call the set W *complete* if X_a is a position, $\forall a \in W$. Question: What properties do the complete sets have?

Definition 2. Let C be a position in the discrete, partially ordered set w. The element $a \in W$ lies *beneath* C if $a \in C^+$. The element a lies *above* C if $a \in C^-$. Finally, C lies *between* the two elements $a < b$ of W if a is beneath C and b is above C.

The definition can be extended to include positions as well. Thus, the position D lies *above* the position C if $D^+ \subset C^+$, etc.

Definition 3. The discrete well-ordered set W is *discrete with respect to positions* if, for any two elements $a, b \in W$, there are at most finitely many positions between a and b.

Clearly then, any finite p.o.s. is discrete with respect to positions. A further trivial example is $\mathbf{Z}$, the integers, with the usual total ordering. Also, $\mathbf{Z}^2$, with the Lorentz ordering, is discrete with respect to positions, since the only postions of $\mathbf{Z}^2$ are the points themselves. But, interestingly enough, $\mathbf{Z}^n$ is *not* discrete with respect to positions for all $n > 2$. Rather than proving this, I prove it for the following similar, but simpler, set.

Example 1. Let $W \subset \mathbf{R}^2$ be defined as follows. Let $\lambda \in \mathbf{R}$ be an irrational number. Then

$$W = \{(p, q) \in \mathbf{R}^2: q \in \mathbf{Z} \text{ and } p = n + q\lambda \text{ for } n \in \mathbf{Z}\}$$

W is to be given the Lorentz ordering inherited from $\mathbf{R}^2$.

To prove that this set is not discrete with respect to positions, I first prove:

Lemma 3.6. Let $a, b \in \mathbf{R}^2$ be such that neither $a \leq b$ nor $b \leq a$. Then there exist points p and q in W (the set of Example 1) such that $p \leq a$, $q \geq b$, but still p is not less than or equal to q.

Proof. We may assume, without loss of generality, that $a = (0, 0)$ and $b = (b_1, b_2)$, where $b_1, b_2 > 0$ and $b_2 < b_1$. The problem is then to find two pairs of integers (n_1, m_1) and (n_2, m_2) representing the two points $p = (m_1 + n_1\lambda, n_1)$ and $q = (m_2 + n_2\lambda, n_2)$ in W. We first find an appropriate q

and then look for p. That is, we search for an element q of W with $q \geq b$ and yet q not greater than a. But this problem can be formulated as follows. Let $\varepsilon > 0$ and $n_0 \in \mathbf{N}$ be given, and assume that $\lambda \in \mathbf{R}$ is irrational. Then the problem is to find an integer $n_2 \geq n_0$ such that $n_2\lambda - [n_2\lambda] < \varepsilon$ (where $[t]$ is the largest integer less than or equal to t, for $t \in \mathbf{R}$). Note that in this case we can take $\varepsilon = b_1 - b_2$. Now it is well known that the numbers $n\lambda - [n\lambda]$ for $n > n_0$ and λ irrational are dense in the interval $[0, 1]$. Therefore there must exist a solution to our problem, and we can take it to be q. But now we have the elements $a, q \in \mathbf{R}^2$ with neither $a < q$ nor $a > q$, and the problem is to find an element $p < a$ with p not less than q. This is essentially the same problem as before, but with the symbol $<$ substituted for $>$, and thus a solution exists for the same reason. ■

Now it is a simple matter to prove that the set W of Example 1 is not discrete with respect to positions. Let $a, b \in R^2$ such that neither $a < b$ nor $a > b$. Let C_a, C_b be positions in W such that $C_a^+ \subset W^a$ and $C_a^- \subset W_a$, and a similar condition holds for b, where W^a and W_a are defined by $W^a = \{w \in W: w \geq a\}$ and $W_a = \{w \in W: w \leq a\}$, respectively. Then Lemma 3.6 implies that $C_a \neq C_b$. But if $u, v \in W$ are any two elements such that $u < v$, then there must be infinitely many point pairs in $\mathbf{R}^2$ satisfying the above condition, and thus W cannot be discrete with respect to positions.

This example shows that even sets that are very much discrete can have a dense position structure. However, the sets in which we will be interested, namely the strongly discrete sets, are, in fact, discrete with respect to positions.

Theorem 3.7. Strongly discrete $\Rightarrow$ discrete with respect to positions.

Proof. Let W be a partially ordered set, and let C_1 and C_2 be positions in W. Assume $C_1^- = C_2^-$. But C_i^+ is completely determined by C_i^-, $i = 1, 2$, so therefore we must have $C_1 = C_2$. Thus, any position in W is determined by its lower cone. Now, assuming that W is strongly discrete, it follows that if $a, b \in W$ with $a < b$, then there are only finitely many possible lower cones between a and b, and therefore at most finitely many positions. ■

It might be thought that if we leave the subject of discrete spaces and return to the familiar framework of Euclidean spaces (with the Lorentz metric), then all positions have the familiar form $W_a \cap W^a$ for some $a \in \mathbf{R}^n$. That is to say, "positions" in $\mathbf{R}^n$ are really just points of $\mathbf{R}^n$, but formulated in an extravagant way. But this is by no means the case. In fact, there exist positions in $\mathbf{R}^n$ that depart strongly from this structure. I give an example of this phenomenon here.

Concentrate on $\mathbf{R}^3$ (three dimensions are the minimum for such examples). Take the line segment $L = \{(t, 0, 0) \in \mathbf{R}^3: t \in [-1, +1]\}$. Let C^+

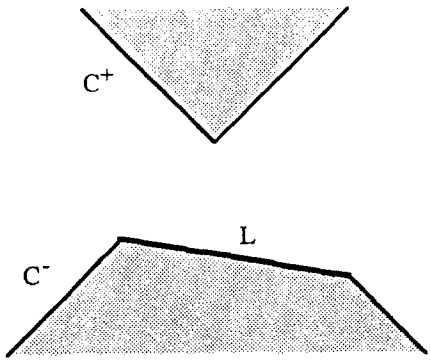

Fig. 3

be the set of points in $\mathbf{R}^3$ that are above all points of L. Define C^- to be the set of all points in $\mathbf{R}^3$ below at least one point of L (Figure 3).

Example 2. $C = C^+ \cup C^-$ is a position in $\mathbf{R}^3$.

To prove that C is a position, begin by noting that certainly $C^- < C^+$. Assume that the point $p \in \mathbf{R}^3$ is above all points of C^-. In particular, p is above all points of L; therefore, by definition, $p \in C^+$. Now assume that $q \in \mathbf{R}^3$ is below all points of C^+. Assume furthermore that q does not lie beneath any point of L. Let $q = (q_1, q_2, q_3)$. If $q_3 \geq 0$, then we could find a point in C^+, of the form $P = (0, \pm R, R + \delta)$, for some (perhaps small) $\delta > 0$, and some sufficiently large $R > 0$, such that q is not less than P, thus contradicting the assumption that $q < C^+$. Therefore, we assume that $q_3 < 0$. Now, if we examine the hyperplane

$$H_q = \{(x_1, x_2, x_3) \in \mathbf{R}^3 : x_3 = q_3\}$$

then $C^- \cap H_q$ consists of the union of the set of discs of radius q_3 with centers on points of the form $(t, 0, q_3)$, $t \in [-1, +1]$. In particular, this intersection is convex. Since q is not in the intersection, we can find a circle in H_q that has $C^- \cap H_q$ in its interior and q outside the circle. Let the center of the circle be (v, w, q_3) and let its radius be r. Then the point $P = (v, w, r)$ must be contained in C^+, and yet q is not less than P. This is again a contradiction. ■

3.7. Distances in Discrete, Partially Ordered Sets

It is now possible to formulate a definition of distances, using this idea of "positions." But before doing so, it is helpful to examine once again formula (14), describing Fokker's variational principle. The first term involves the determination of path lengths along the particles, using the usual

Lorentz distances. That is, given two numbers $t_1 < t_2$ in $\mathbf{R}$, we are required to determine the Lorentz distance between $p_1 = \gamma(t_1)$ and $p_2 = \gamma(t_2)$ for a typical particle $\gamma \in \Gamma$. But here we certainly have $p_1 < p_2$, so that it is sufficient just to be able to determine Lorentz distances between arbitrary points $a, b \in \mathbf{R}^4$ with $a < b$. For clarity, let us call these *distances of the first kind.* If we turn now to the second term in (14), we again notice an integral over such distances, but the Dirac δ-function gives the integral an entirely different character to that of the first term. In fact, the second term is really concerned with finding the advanced or retarded distances between two-points $a \neq b$ in $\mathbf{R}^4$, with the property that the *Lorentz distance* between a and b then vanishes. Let us call such advanced and retarded distances, *distances of the second kind.* It seems that these two different kinds of distances, which appear in Fokker's variational principle, must lead to two different definitions of "distance" in the context of discrete, partially ordered sets.

Let us begin with the distances of the first kind.

Definition 1. Let W be a discrete, well-ordered set with particle structure, which is also discrete with respect to positions. Let $P \subset W$ be a particle, so that we can write $P = \{p_i : i \in \mathbf{Z}\}$, where $p_i < p_j \Leftrightarrow i < j$. The particle P will be said to be *proper* $\Leftrightarrow$ for all $i, j \in \mathbf{Z}$ we have that the number of positions between p_i and p_{i+1} is the same as the number of positions between p_j and p_{j+1} (i.e., this number is a constant associated with the particle P).

Definition 2. Let W be as above. If all particles in W are proper, then W will be called an *admissible set.*

Do there exist admissible sets? Once again we have a number of standard trivial examples, such as $\mathbf{Z}$ itself. It is also not difficult to see that a discrete W, consisting of discrete vertical particles, constructed in a space of the form $K^3 \times \mathbf{R}$, where K^3 is compact, as in Section 3.2, must also be admissible. But nontrivial examples in $\mathbf{R}^4$ seem to be difficult to find.

What connection is there between the idea of proper discrete paths, and distances of the first kind, in $\mathbf{R}^4$? Consider the following observation.

Theorem 3.8. For $a < b$ in $\mathbf{R}^4$, define Ω_{ab} to be the standard Euclidean volume of the set $W^a \cap W_b \subset \mathbf{R}^4$. Furthermore, define $d_L(a, b)$ to be the Lorentz distance between a and b. If p, q, p', q' are four points in $\mathbf{R}^4$ with $p < q$, $p' < q'$, and $d_L(p, q) = d_L(p', q')$, then it follows that $\Omega_{pq} = \Omega_{p'q'}$.

Proof. The fact that $d_L(p, q) = d_L(p', q')$ implies that there must exist a Lorentz transformation that takes p to p' and q to q'. But then the fact that $\Omega_{pq} = \Omega_{p'q'}$ can be deduced as a simple consequence of Theorem 3.1. ■

Thus, in $\mathbf{R}^4$, it is possible to do away with distances of the first kind and instead, given two points $a < b$, one need only look at the Euclidean

volume of the set of points between a and b. This gives an equally good, relativistically invariant, conception of distances. A possible connection between these new Euclidean distances and the definition of proper discrete particles is described in the next section.

Definition 1 is, at most, related to the distances of the first kind in $\mathbf{R}^4$. What about the distances of the second kind?

Definition 3. Let W be a discrete, partially ordered set with particle structure, and let P and Q be two distinct particles in W. Let $P = \{p_i : i \in \mathbf{Z}\}$ and $Q = \{q_j : j \in \mathbf{Z}\}$, as above. Given a specific point p_r on P, then the point q_s on Q will be said to be *diagonally above* p_r if $p_r < q_s$, but p_r is not less than q_{s-1}. We also say that q_s lies *on the light-cone above* p_r. Similarly, p_r is diagonally below q_s, etc.

Thus, distances of the second kind are only to be defined between pairs of points of W such that one of the points is diagonally above the other.

Definition 4. Let $p, q \in W$, with q diagonally above p in the set W, which is strongly discrete and with particle structure. Then the *retarded distance* from p to q (and also the *advanced distance* from q to p) is the number of elements of W in the set $W_q - W_p$.

3.8. Possible Relationships with the Usual Idea of Space-Time

The basic assumption of this work is that it is possible to describe classical electrodynamics by means of a discrete geometrical framework. I have defined discrete, partially ordered sets with a proper particle structure that are discrete with respect to positions, and such sets will form the basis for my further reasoning.

The picture I have in mind is that the particles that are observed in nature are discrete, and the union of all the particles is a discrete set W. The particles themselves—the points of real, physical material—are the points of W. But what about the points of empty space? These are, in the present picture, the *positions* of W.

Now I would like to claim that it is possible to deduce the usual geometric properties of space-time purely in terms of the *algebraic* properties of W. Unfortunately I have been unable to succeed in this. Instead I have been forced make a number of rather vague assumptions of a geometric nature concerning the set W. What are these assumptions?

The main assumption has already been encountered in Section 3.3. This is that there should exist a proper representation of the set W in $\mathbf{R}^4$. But even this is not enough. To go further and justify all of the definitions

in terms of the usual ideas of physics, it will be necessary to assume that the elements of W, and also the positions of W, are more or less homogeneously distributed throughout $\mathbf{R}^4$. Specifically, I will assume that a certain proper representation of W in $\mathbf{R}^4$ is given, representing a certain embedding of the points of W in $\mathbf{R}^4$ (each point $u \in W$ is mapped onto the vertex of the cone that represents u). Thus, W becomes a subset of $\mathbf{R}^4$. I will further assume that the cones in this proper representation of W in $\mathbf{R}^4$ are, on the whole, nearly "straight" Lorentzian cones (or "standard light-cones" in the terminology of Section 3.3).

Given such an embedding of W in $\mathbf{R}^4$, then it is reasonable to go on and associate the *positions* of W with points in $\mathbf{R}^4$ as well. Thus, if C is a given position in W, then we will also think of C as being associated with a point in $\mathbf{R}^4$. This point is such that, according to the *Lorentz ordering* of $\mathbf{R}^4$, some of the points of W are above it and some other points of W are below it. Are these points just the elements of C^+ and C^- in W? My assumption is that this is—at least very nearly—the case. Then, finally, I will assume that in the neighborhood of any horizontal hyperplane H_t in $\mathbf{R}^4$ the points and positions of W are—at least on a sufficiently large scale—homogeneously distributed.

Now all of these assumptions, while admittedly being of an arbitrary nature, are really nothing more than the simplest analogies to the assumptions that cosmologists usually make. Therefore, I will loosely call all of these assumptions the *cosmological hypothesis* for W.

Does there exist a partially ordered set W, with proper particle structure, satisfying the cosmological hypothesis? This is certainly a difficult question to answer. On the other hand, if we are to base our cosmology on a space such as $S^3 \times \mathbf{R}$, as suggested in Section 3.2, then this question of existence reduces to a triviality.

To return to $\mathbf{R}^4$, it seems necessary at least to give some justification to the idea that the cosmological hypothesis can hold, and yet W can still be discrete with respect to positions. The following argument seems to be reasonable.

Imagine some discrete set of points W in $\mathbf{R}^4$ satisfying the cosmological hypothesis, and which is also such that the "density" of W in $\mathbf{R}^4$ increases exponentially with increasing "time." That is, given a small, rectangular box in $\mathbf{R}^4$ of volume V located near the point (x, y, z, t), then the expected number of points of W in V is $ue^{vt} \times V$ for some constants u and v. Is it reasonable to expect that such a W is strongly discrete? Consider two points $p < q$ in W; for example, $p = (0, 0, 0, 0)$ and $q = (0, 0, 0, 1)$. Let $\Lambda_{pq} = \{x \in \mathbf{R}^4 : x \leq q, \text{ but } x \text{ is not } \leq p\}$. Then for any of the horizontal hyperplanes H_t in $\mathbf{R}^4$ with $t < 0$ we have $H_t \cap \Lambda_{pq}$ being the three-dimensional space between the 2-sphere of radius t and the 2-sphere of radius $t + 1$. The volume

of this space is $\frac{4}{3}\pi[(t+1)^3 - t^3]$. But then the expected number of points of W in Λ_{pq} below the hyperplane H_0, say, is given by

$$\text{const} \times \int_0^{+\infty} [(t+1)^3 - t^3]\, e^{-vt}\, dt$$

where the integral is taken from zero to $+\infty$, and this is finite.

Such an exponential increase in the density of points of W in $\mathbf{R}^4$ could be expected to bring with it a corresponding exponential increase in the density of the *positions* of W in $\mathbf{R}^4$. This can be seen by a symmetry argument: Let $\Psi: \mathbf{R}^4 \to \mathbf{R}^4$ be the mapping that takes the point $(x, y, z, t) \in \mathbf{R}^4$ to the point (kx, ky, kz, kt) for some constant $k > 0$. Then the density of the positions of $\Psi(W)$ in $\mathbf{R}^4$ increases exponentially with time $\Leftrightarrow$ the density of the positions of W in $\mathbf{R}^4$ increases exponentially with time. Another mapping is $\Psi_1: \mathbf{R}^4 \to \mathbf{R}^4$, given by $\Psi_1((x, y, z, t)) = (x, y, z, t-h)$, where h is some appropriate constant. This can be chosen so that $\Psi_1(W)$ has the same expected density at a given point of $\mathbf{R}^4$ as $\Psi(W)$. But since the positions of W are determined by the points of W, we conclude that the density of the positions at each hyperplane H_t must be k times the density at the hyperplane H_{t-h} and thus this density of positions must increase exponentially as well.

The corresponding increase, both in the density of points *and* of positions, is important, for it allows us to assert that the particles of W can be taken to be proper (see Definition 1 of Section 3.7). Thus, one sees that all of the definitions that have been made up to this point are designed to apply to such a picture of a discrete set W embedded in $\mathbf{R}^4$, with an exponentially increasing density in the direction of the fourth coordinate axis. This picture is also in accord with some standard models of the expanding universe.

Now, it is interesting to note that one such model—the "steady-state universe"—is often criticized on the grounds that it requires the "spontaneous creation" of energy and matter to fill the expanding voids of space-time. In the "big-bang universe" this creation is imagined to have occurred in an instant. But, if one is prepared to accept the philosophical proposition that the measure of "space" is in some way to be *determined* by the matter contained in it, then it becomes clear that this question of the "creation" of matter must be considered in a different light. Space-time, according to this picture, simply cannot be thought of as being an abstract entity, having nothing to do with the matter contained in it. On the contrary, matter determines its own density in space-time! Thus, there would no longer be a need for cosmologists to debate the mechanism of creation—whether it should be instantaneous or continuous. This picture will perhaps be made more acceptable if I note that particles, in the present view, are

not unchanging entities, whose mass and charge is independent of space-time. As will be seen in the next section, these quantities can be thought of as being determined by the local properties of space-time—that is, the *positions* of the underlying discrete set.

One point deserves further attention. I have asserted that the integral

$$\text{const} \times \int_0^{+\infty} [(t+1)^3 - t^3]\, e^{-vt}\, dt$$

taken from 0 to $+\infty$, is finite. But then the corresponding integral

$$\int_0^{+\infty} t^3\, e^{-vt}\, dt$$

is surely also finite, which means that the expected number of points of W below, say, $p = (0, 0, 0, 0) \in \mathbf{R}^4$ is also finite. This conflicts with the assumption that W is to consist of proper particles! However, this objection can be countered by observing that the requirement that W consist of proper particles implies a very special property of the fourth coordinate axis in $\mathbf{R}^4$ with respect to this embedding of W: namely, all standard light-cones must contain infinitely many points. On the other hand, the sets of the form Λ_{pq} do not share this property; they can only meet proper particles along finite intervals.

The objection could also be countered with two further arguments. First, I expect that a better model would be found by getting away from the idea of proper particles and allowing instead particle creation and annihilation and vacuum loops, as in the theory of quantum electrodynamics. But the simplest way to counter the objection is to note that I expect that it will prove to be best to adopt Segal's ideas and base cosmology not on $\mathbf{R}^4$, but rather on a space such as $S^3 \times \mathbf{R}$.

3.9. A Discrete Formulation of Fokker's Theory

It remains to translate formula (14) into the terminology of discrete, partially ordered sets. For this purpose, change the formula to

$$\sum_{\substack{P, Q \in \mathcal{P} \\ P_n \in P}} \sum_{n \in Z} F(p_n, Q)[D_+(p_n, Q)^{-1} + D_-(p_n, Q)^{-1}] \tag{47}$$

where $\mathcal{P}$ is the set of all particles in W, $F(p_n, Q)$ is some function defined on the set of all pairs $Q \in \mathcal{P}$, and $p_n \in P$; $D_\pm(p_n, Q)$ is defined as follows. In case $p_n \notin Q$, we define $D_-(p_n, Q)$ to be the retarded distance from p_n to Q according to Definition 4 of Section 3.7, and $D_+(p_n, Q)$ to be the advanced distance from Q to p_n. In case $p_n \in Q$, we define $D_\pm(p_n, Q) = D(Q)$, a constant independent of the particular point on Q which is chosen.

To what extent can one claim that (47) is equivalent to the usual formulation (14)? To begin with, one can assert that if the ideas of Section 3.8 are accepted, then the distances $D_{\pm}(p_n, Q)$ will correspond with the usual advanced and retarded distances of classical electrodynamics. For, let p_n and q_m, say, be points such that q_m is diagonally above p_n, both in W and in $\mathbf{R}^4$. Assume that $p_n = p = (0, 0, 0, 0)$ and $q_m = q = (t, 0, 0, t)$. Then, for horizontal hyperplanes of the form H_T, $T < 0$, the set $\Lambda_{pq} \cap H_T$ is the three-dimensional space between a 2-sphere of radius T and a 2-sphere of radius $T + t$. Now, if we assume that T is very much greater than t, then this is proportional to t, i.e., the retarded distance.

How is this assumption that T is much larger than t to be interpreted? The idea is that the interaction terms of Fokker's theory are to describe *local* electromagnetic interactions, the word "local" being interpreted in the sense of cosmology. Thus, t is "small." On the other hand, most of the points of W in Λ_{pq} are "cosmologically distant."

Our variational principle now becomes the assertion that we will only consider sets W with the property that (47), evaluated on W, is an extremum with respect to "local" variations in the sense of Section 2.

Given that the density of positions of W in $\mathbf{R}^4$ varies slowly with time—again on a cosmological scale—then Theorem 3.8 implies that the process of counting points along the particles in the discrete theory is the same as taking an integral along the path length, as in the Fokker theory. In this way the summations in (47) can be seen to correspond with the integrals in (14). But of course this correspondence is very much determined by the local density of positions of W in $\mathbf{R}^4$. If this density is great, then the measure of path length becomes correspondingly shorter, and thus, carrying through the analogy between (47) and (14), the masses of the particles can be considered to become correspondingly greater.

Now it is also true that the *velocities* of the particles are important in (14), since the expression $\gamma'\xi'$ is used there. I simply imagine this product of the velocities to be expressed in (47) in some way within the term $F(p_n, Q)$. Perhaps the reader may feel that it would be better to look for a specific analog to $\gamma'\xi'$ directly in terms of the ordering structure of W and not to suppress the necessary definition as I am doing. It is possible to choose such an analog, but for the further development the precise form it is given will play no essential role. Thus, rather than burdening the development with further arbitrary choices, I prefer to leave this question open.

At this stage, then, the description of a discrete partially ordered set W for use in classical electrodynamics is complete. The fact that the set W itself has only been incompletely described and that many questions have been left open are things it shares in common with other descriptions of

classical electrodynamics. The fact that at many points the description could be altered to obtain other, perhaps more appropriate, discrete geometries should be self-evident.

It would of course be possible to go on and prove further results of an abstract nature concerning the theory of discrete, partially ordered sets. But surely the most important task is to try to justify the assertion that such discrete sets do indeed have a place in physics. Now the study of physics is concerned with the description and explanation of physical experiments. But I have not as yet given any *new* explanations. On the contrary, my goal has been to find an alternate, but *equivalent*, geometric framework for Fokker's theory of classical electrodynamics.

This is not to say, however, that Fokker's theory provides a satisfactory and complete description of all possible experiments to be observed in the physical world. Far from it. Almost all aspects of modern physics, from gravitation, to quantum mechanics, to elementary particle physics, simply have no place in the Fokker theory. Thus, one is prompted to look for possible connections between the present discrete theory and these other branches of physics as well.

4. A POSSIBLE CONNECTION WITH THE THEORY OF GRAVITY

4.1. Discrete versus Differentiable Structures

One of the most fundamental premises of modern physics is the idea that gravity is a purely *geometrical* phenomenon. My goal is the development of a discrete geometry for use in theoretical physics, and thus the question naturally arises as to whether such discrete geometries can be made to be compatible with the theory of gravity.

Now it is certainly impossible to establish a *complete* correspondence between the usual theory of gravity—the general theory of relativity—and some discrete theory one might wish to set in its place. The fact is that the general theory of relativity is concerned with the behavior of differential manifolds. Much research has recently gone into the subject of possible "singularities" in these manifolds and the question of whether or not such singularities can have physical relevance. But how can the idea of a singularity in space-time be translated into the framework of discrete spaces? It seems obvious that such singularities result from the use of differential equations, describing *manifold* structures, which are, by definition, arbitrarily fine. What possible meaning could be given to the idea of a "singularity" in a discrete, partially ordered set? Thus, one is forced to assume, as a working hypothesis—and in common with Einstein (1956,

General Remark B)—that it is possible to describe gravity without making use of singularities in differential manifolds.

However, a still greater obstacle is the fact that the theory of gravity—in contrast to classical electrodynamics—is a *nonlinear* theory. Thus, there can exist no simple action-at-a-distance formulation of the theory of gravity in the style of Fokker's theory of classical electrodynamics; it is not possible to simply examine all the different pairs of particles in our set Γ and then sum up the effects due to each pair, as was done in the classical theory. Instead, some more subtle, collective phenomenon must be found.

This is surely a great problem. After all, the present discrete formulation of classical electrodynamics was only made possible by means of Fokker's *action at a distance* theory.

But perhaps one can proceed in a more unconventional direction. In the present view physical material *determines* the structure of space-time. This is also true of general relativity; but the discrete theory can be thought of as providing a more *direct* mechanism for describing this relationship of physical material to space-time: the space of *positions*, which determines the metrical properties of space-time according to the present theory, is *defined* directly in terms of the particles. Thus, it might be possible simply to take the discrete framework as described in Section 3 and investigate the question of whether or not a gravitational theory is *already present* within the original framework. This would surely be a great advance, for then it would be possible to view gravitation as a phenomenon that arises naturally in connection with the phenomenon of electromagnetism.

Such a goal would certainly be most ambitious. Even if such a theory could be established, the fact is that, for the reasons outlined above, it would of necessity depart from the usual gravitational theory. Furthermore this usual gravitational theory has itself been extended and newly interpreted in ways that depart strongly from Einstein's original conceptions.

I will begin with a short sketch of some of the standard results of the general theory of relativity.

4.2. The Gravitational Equations

The general theory of relativity has as its main hypothesis the assumption that space-time can be represented as a four-dimensional pseudo-Riemannian differential manifold. The metric tensor g has signature (3,1). The Riemann curvature tensor in a given local coordinate system has the components R^i_{jkl}. Let $R_{jk} = R^i_{jki}$, where the usual summation convention (one sums over the index i from 1 to 4) is being used. R_{jk} are the components of the Ricci tensor. The scalar curvature R is given by $R = g^{jk}R_{jk}$, where the summation is here to be taken over both the indices j and k. Finally,

the Einstein tensor is defined to be $G_{jk} = R_{jk} - g_{jk}R/2$. The gravitational equations for empty space are now taken to be

$$G_{jk} = 0 \tag{48}$$

In case the space is not empty, one takes the energy-momentum tensor T_{jk} and equates it, multiplied by a constant, with the Einstein tensor, i.e.,

$$G_{jk} = -8\pi G T_{jk} \tag{49}$$

where G is the Newtonian gravitational constant. These are Einstein's equations of gravity. It is a standard result that the divergence of the Einstein tensor vanishes identically, $G^{jk}_{;k} = 0$ (the semicolon indicates covariant differentiation, and again one is to take the sum over k from 1 to 4). The divergence of the energy-momentum tensor must then vanish as well. This is the law of conservation of energy-momentum.

The energy-momentum tensor is usually taken to be

$$T^{jk} = \rho_m v^j v^k + \frac{1}{4\pi}\left(F^j{}_m F^{mk} + \frac{g^{jk} F^{mn} F_{mn}}{4}\right) \tag{50}$$

Here F_{jk} is the electromagnetic tensor defined in Section 2. I am assuming, as there, that the units of measurement have been so chosen that the speed of light is 1. The tensor describes a dustlike fluid of massive material whose density is given by the scalar field ρ_m and whose velocity is the four-dimensional vector field v. For the purposes of astronomical observations, it is usually assumed that the electromagnetic field is negligible, so that F_{jk} is taken to be zero, and we obtain

$$T^{jk} = \rho_m v^j v^k \tag{51}$$

The philosophical considerations that led to all of these choices, and the standard techniques for dealing with the theory, are described in innumerable textbooks. See, e.g., Narlikar (1978) and Yilmaz (1965). Einstein (1956) also gave a very succinct and readable account.

4.3. The Schwarzschild Solution

The gravitational equations (49) are, in general, difficult to solve, so that only a very few exact solutions are known. In empty space we have the equation (48), which clearly has as a trivial solution flat Minkowski space (i.e., $\mathbf{R}^4$, with the Lorentz metric)—that is, the Riemann tensor vanishes. There is also a class of solutions involving "gravitational waves" in empty space. (Although a number of experiments have been in progress for many years with the aim of detecting such waves, it is apparently safe to say that gravitational waves have still not been observed, and thus I shall

not pursue this question further.) Perhaps the simplest known, nontrivial solution is the Schwarzschild solution.

The Schwarzschild solution is concerned with the case of a stationary, spherically symmetric mass of material without electrical charge. It may be imagined that the material is centered on the point $(0, 0, 0) \in \mathbf{R}^3$ and that beyond a certain radius, space and time are empty. (In particular, the universe is assumed to be empty, except for this one material object.) Now we are really only looking for a solution in empty space, outside the object. Here the energy-momentum tensor must vanish, so that we need only search for some spherically symmetric solution to the gravitational equations for empty space (48) on the set $(\mathbf{R}^3 - B_R) \times \mathbf{R}$ say, where B_R is a ball of radius R, centered at $(0, 0, 0)$, containing the massive, gravitating object. The question of extending the solution through the interior of B_R will be put off for now and relegated to the sequel.

If we use the standard notation $ds^2 = g_{ij}\, dx^i\, dx^j$, then we can write the Schwarzschild solution as

$$ds^2 = -\frac{dr^2}{(1-2m/r)} - r^2(d\vartheta^2 + \sin^2 \vartheta\, d\varphi^2) + \left(1 - \frac{2m}{r}\right) dt^2 \tag{52}$$

Note that we are using spherical coordinates (r, ϑ, φ) for $\mathbf{R}^3$ and the fourth coordinate (time) is denoted by t. Here "m" is the total mass of the spherically symmetric, massive object centered on $(0, 0, 0)$. If m is 0, then the Schwarzschild solution degenerates to the trivially flat empty-space solution. The more usual situation is that the mass is nonzero, so that $m > 0$. The verification that (52) is in fact a solution to (48) can be found in almost every book on the theory of gravitation.

Now the solution (52) serves to represent the metric tensor in terms of a *single* coordinate neighborhood. Difficulties arise with this strategy when $r = 2m$, the so-called "Schwarzschild radius." Thus, if the ball B_R has radius $\leq 2m$, then it will become necessary to build up the differential structure in terms of a finer set of coordinate neighborhoods in the region near the point $(0, 0, 0)$, as is usual in differential geometry. But, for example, in the case of the sun, the Schwarzschild radius is approximately 2.7 km. On the other hand, the *actual* radius of the sun is much greater, about 70,000 km, so that, at least in this case, the solution (52) is valid through out $\mathbf{R}^3 - B_R$.

It will be convenient for later purposes to express (52) in terms of a slightly different coordinate system: the so-called "isotropic coordinate system." Let $p = (r, \vartheta, \varphi, t) \in \mathbf{R}^4$. Then in the isotropic coordinate system we can write $p = (r_{\#}, \vartheta, \varphi, t)$, where

$$r = r_{\#}(1 + m/2r_{\#})^2 \tag{53}$$

This coordinate transformation is certainly well defined throughout $\mathbf{R}^3 - B_R$ for $R > 2m$. It is a simple exercise to verify that

$$1 - \frac{2m}{r} = \frac{(1 - m/2r_{\neq})^2}{(1 + m/2r_{\neq})^2} \tag{54}$$

Thus, substituting $r_{\neq}$ for r in (52), we obtain

$$ds^2 = -A[dr_{\neq}^2 + r_{\neq}^2(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2)] + B\, dt^2 \tag{55}$$

where

$$A = \left(1 + \frac{m}{2r_{\neq}}\right)^4, \qquad B = \frac{(1 - m/2r_{\neq})^2}{(1 + m/2r_{\neq})^2}$$

The isotropic form (55) is convenient, due to the fact that the three-dimensional volume element $dr_{\neq}^2 + r_{\neq}^2(d\vartheta^2 + \sin^2\vartheta\, d\varphi^2)$ is similar to the usual expression for the Euclidean volume in $\mathbf{R}^3$ expressed in polar coordinates.

At this stage one would like to emulate the development in Section 2.2 and consider that the spherically symmetric gravitating mass centered at (0, 0, 0) should shrink down to a point, leading perhaps to a generalized function, and then on to a generalized field theory, as was done by Dirac in the case of classical electrodynamics. But this will not do! We are no longer dealing with fields defined on a *given* space. On the contrary, the metrical structure is a fixed property of that space, and thus each different solution to the gravitational equations defines a *different* space. I have assumed that the space is such that the metric tensor has the form given by (52) in $(\mathbf{R}^3 - B_R) \times \mathbf{R}$. It is possible to continue this solution smoothly across the ball B_R, thus giving a—more or less—realistic space, say for a star of radius R. Taking the limit as $R \to 0$ amounts to finding a certain incomplete differential manifold [see, for example, Hawking and Ellis (1973) for a more detailed treatment of these ideas], which has been described as a "black hole."

Now it is important to remember that the conventional energy-momentum tensor (50) describes a diffuse, fluidlike substance, in accordance with the ways of thinking of some physicists in the 19th century. On the other hand, the idea of pointlike, "atomic" particles leads as we have seen, to the idea of matter in terms of black holes. Thus, the simple classical picture suddenly changes! Instead of imagining conventional particles moving sensibly through a smooth space-time, we are now confronted with the idea of an empty space-time punctuated with countless tiny "black holes"—or "singularities"—which are to be interpreted in some way as "particles." In particular, the energy-momentum tensor, at least in its form (51), vanishes everywhere where the space-time differential manifold is well defined! Such

drastic consequences led Einstein to the conclusion that the theory can only be considered valid as long as these "singularities" can be excluded.

4.4. Particle Motion in a Gravitational Field

The law of conservation of energy-momentum mentioned in Section 4.2 implies that a flow of electrically neutral, diffuse matter tends to follow straight lines, i.e., geodesics, through the manifold. Now this result ties in well with our use of the Fokker theory of classical electrodynamics. Given some particle $\gamma \in \Gamma$ with vanishing electrical charge, the interaction terms in (14) involving γ vanish, and we are left with a single integral involving the path length of γ. Fokker's principle reduces then to the assertion that the path length is an extremal with respect to local variations. This is precisely the assertion that γ should describe a geodesic. Thus, according to this way of thinking, Fokker's theory seems to be completely in harmony with the general theory of relativity; it is only necessary to generalize Fokker's theory to the extent of allowing its validity also in the context of curved space-time manifolds.

Unfortunately, this way of going about things ignores the most basic premises of the theory of general relativity! To begin with, we have seen that a pointlike particle can only be represented as a "black hole" in the theory, and therefore the space-time manifold does not even exist at the particle! Indeed, the "particle" is nothing more than a "hole" in the manifold, so it must be nonsense even to consider geodesics in this case. Physicsts traditionally avoid this difficulty by speaking about the motions of so-called "test particles."

Now, a test particle is taken to be an object with a sufficiently small mass so as not to effect the general gravitational fields through which it moves. Given that it is meaningful to consider such entities, then the geodesic hypothesis might indeed be sensible. But if, on the other hand, one is willing to consider the "test particle" as being a *massive*—albeit small, diffuse, and fluidlike—object, then (1) being diffuse, the test particle will not be a black hole, but (2) being massive, it will alter the space-time manifold, and thus it is no longer a "test particle" whose motion can be calculated within some *fixed* space, such as that of the Schwarzschild solution.

One arrives then at the idea that particle *motion* is only defined in terms of a given space-time manifold, but the space-time manifold is itself *determined* by the particles contained within it. This connection between "motion" and "curvature" leads to the thought that perhaps it is improper to simply postulate that particles should follow geodesics. Perhaps the particle motion is really determined by the gravitational equations (49) themselves. Einstein and Grommer (1927) followed this line of reasoning.

To what extent is it reasonable to calculate particle motion as the limit of the motion of diffuse, fluidlike balls of massive material whose radius tends to zero? This is undoubtedly an interesting, but perhaps too theoretical question. In the end, one simply obtains the expected result that "test particles" do follow geodesics in the space-time manifold.

4.5. The Advance of the Perihelion of Mercury

Surely the most famous—and historically the most important—calculation of motion in a gravitational field concerns the orbit of the planet Mercury. The calculation is interesting in itself, and it leads to a number of results that will be important for the further development, so I will reproduce it here. I will follow, for the most part, the exposition in Yilmaz (1965).

The calculation is based on the assumption that, at least in the regions where Mercury is to be found, the gravitational field of the sun can be represented by the Schwarzschild solution. Furthermore, Mercury is considered to be a "test particle" which does not perturb the sun's gravitational field. Now of course this is a grave departure from reality, but still, as we will see, it *does* lead to a reasonable result. There is also a more compelling circumstance which forces us to view Mercury as a "test particle": the fact is that the two-body problem, within the general theory of relativity, has not been solved!

I will use the Schwarzschild metric expressed in terms of isotropic coordinates (55). The radial coordinate was denoted by the symbol $r_{\neq}$ there, with r being reserved for the expression (52). However, at the risk of a slight confusion, from now on I use the more usual r, rather than $r_{\neq}$, in (55).

The problem is to describe the geodesics in a given pseudo-Riemannian manifold whose metric satisfies (55). Let us call this manifold M^4. Let $\gamma: \mathbf{R} \to M^4$ be a smooth path through the manifold. I will assume that γ is timelike, in the sense that the square of the velocity of γ is positive, i.e.,

$$|\gamma'|^2 = -\gamma_1'^2 - \gamma_2'^2 - \gamma_3'^2 + \gamma_4'^2 > 0$$

where γ_i' is the ith component of the velocity of γ for $i = 1, \ldots, 4$ for each point on γ. It is possible to assume that γ is parametrized by means of the proper-time parameterization, so that in fact $|\gamma'| = 1$ everywhere.

Returning now to the Schwarzschild metric (55), it will prove to be convenient to express it in the form

$$ds^2 = e^{-2\alpha}\, dt^2 - e^{2\beta}(dr^2 + r^2\, d\vartheta^2 + r^2 \sin^2 \vartheta\, d\varphi^2) \tag{56}$$

where α and β are suitable functions of r (I assume, of course, that the mass of the sun remains constant).

Everything has spherical symmetry, so it is considered sensible to restrict one's search for geodesics to paths confined to the plane defined by $\vartheta = \pi/2$, and thus one can take

$$ds^2 = e^{-2\alpha}\,dt^2 - e^{2\beta}(dr^2 + r^2\,d\varphi^2) \tag{57}$$

or, expressed slightly differently,

$$1 = e^{-2\alpha}t'^2 - e^{2\beta}(r'^2 + r^2\varphi'^2) \tag{58}$$

where, for example, $t' = dt/ds = d\gamma_4(s)/ds$, and so forth.

Let us now take a variation, say between the points $\gamma(t_1)$ and $\gamma(t_2)$, where $t_1 < t_2$. The length of the path is given by

$$\int ds = \int [e^{-2\alpha}t'^2 - e^{2\beta}(r'^2 + r^2\varphi'^2)]\,ds \tag{59}$$

and so the Euler equations corresponding to this variational problem can be immediately written down [these are equations (17.27) in Yilmaz (1965)]:

$$\begin{aligned} r'' + \beta' r'^2 - (r + r^2\beta')\varphi'^2 - e^{-2\alpha-2\beta}\alpha' t'^2 &= 0 \\ \varphi'' + 2(1/r + \beta')\varphi' r' &= 0 \\ t'' - 2\alpha' t' r' &= 0 \end{aligned} \tag{60}$$

(note that α and β are functions of r, and so $\alpha' = d\alpha/dr$ and $\beta' = d\beta/dr$). The second and third equations in (60) can be solved, to give

$$\varphi' = he^{-2\beta}/r^2, \qquad t' = ke^{2\alpha} \tag{61}$$

Here h and k are constants, which are determined by the boundary conditions at $\gamma(t_1)$ and $\gamma(t_2)$ in (59). Substituting (61) into (57) yields

$$\left(\frac{dr}{d\varphi}\right)^2 + r^2 = \frac{r^4}{h^2}(k^2\,e^{2\alpha} - 1)e^{2\beta} \tag{62}$$

The fact that α and β are functions of r makes equation (62) difficult to solve. The standard procedure at this point is to linearize things by taking the Taylor series

$$e^{2\alpha} = 1 + \frac{2m}{r} + \frac{2m^2}{r^2} + \cdots \qquad e^{2\beta} = 1 + \frac{2m}{r} + \frac{3m^2}{2r^2} + \cdots \tag{63}$$

If r is much bigger than $2m$—which is certainly the case when it comes to the path of Mercury about the sun—then the error in restricting one's attention to the first three terms is small. We also have the slightly coarser approximation $e^{-2\alpha} \approx 1$. In addition, a further approximation is used. Since

Mercury has a velocity much less than that of light, we can take $t' \approx 1$. Combining these approximations, therefore, we obtain $k \approx 1$. All of this gives to second order, the equation

$$\left(\frac{dr}{d\varphi}\right)^2 + r^2 = \frac{r^4}{h^2}\left(\frac{2m}{r} + \frac{6m^2}{r^2}\right) \tag{64}$$

Next, Yilmaz substitutes $u = 1/r$ and differentiates with respect to φ, obtaining

$$d^2u/d\varphi^2 + \kappa^2 = m/h^2 \tag{65}$$

Here κ is taken to be the constant $\kappa = (1 - 6m^2/h^2)^{1/2}$. Equation (65) has the solution

$$u = (m/h^2)[1 + \varepsilon \cos(\kappa\varphi)] \tag{66}$$

Once again, ε is a constant, which is determined by the initial conditions. If $\kappa = 1$, then this is the equation of an ellipse (a circle if $\varepsilon = 0$). More generally, it is an ellipse with precession, and it turns out that if the appropriate numbers are substituted for Mercury and the sun, then the magnitude of the precession amounts to 43 sec of arc per century.

It is interesting to note that the actual precession of Mercury is over ten times this amount! Almost all of this can be accounted for by first-order Newtonian approximations involving the motions of the other planets. These observations and calculations of the orbit of Mercury must surely represent one of the most precise physical measurements ever made. The effect is so fine that some theorists have attempted to find different gravitational theories that differ from the standard theory in the sense that they are based on gravitational equations similar to, but slightly different from, (49) (e.g., Dicke, 1964). Such theories often predict slightly different values for the precession of Mercury, and this has led to debates concerning such extraneous factors as the oblateness of the interior of the sun and so forth. However, recently the double pulsar system PSR 1913+16 has been investigated and it has been found that when the calculation is applied, a much greater value for m/r (and hence for κ) should be chosen. It appears that—subject to the usual uncertainties that must always surround the observation of such distant objects—the standard theory continues to hold.

4.6. An Unusual Interpretation of the Preceding Calculation

Having derived equation (66) for the orbit of Mercury, Yilmaz makes the following observation. When dealing with the classical Newtonian gravitational theory, one may consider the field $\varphi(r) = m/r$, the potential

field in the neighborhood of the sun. Let us now simply write

$$ds^2 = e^{-2\varphi}\, dt^2 - e^{2\varphi}(dr^2 + r^2\, d\vartheta^2 + r^2 \sin^2 \vartheta\, d\vartheta^2) \tag{67}$$

A glance at (63) shows that, at least as far as the first few terms are concerned, this agrees with the Schwarzschild metric. In fact, it turns out that the agreement goes so far as to give the same value for the advance of the perihelion of Mercury. Of course, the standard first-order effects—the agreement with the Newtonian theory, the bending of light rays, etc.—would be the same for a space satisfying the metric (67) and for one satisfying the "exact" Schwarzschild solution.

Would it be possible to base a new theory of gravity on equation (67)? This would surely be somewhat far-fetched. After all, (67) only describes an approximation to the usual theory in the case of a single, slowly moving, spherically symmetric object. Even the most basic question of how it could be possible to define a *relativistically invariant* theory on the basis of (67) remains unanswered.

On the other hand, it is true that there have never been any truly reliable observations of the grativational fields of swiftly moving or non-spherically symmetric objects. The fact is that gravity can be thought of as being an extremely weak "force" and thus, *in practice*, all *direct* experiments that can be performed concerning the gravitational effect involve a very small value for m/r. (The interpretation of the astronomical observations of cosmologically distant objects, in which larger values of m/r may play a role, must always involve great uncertainties.)

I will therefore proceed by accepting (67) in the following light. It seems possible to say that any theory of gravity capable of explaining the most well-established empirical observations must, as a first approximation, agree with (67). Thus, if one can produce a theory that, in the case of slowly moving, spherical objects, is similar in form to (67), then it is reasonable to claim that it is also in agreement with experiment.

4.7. Determining the Geometry of Discrete Sets

In this section I will describe a method for determining the geometrical properties of discrete, partially ordered sets that satisfy the cosmological hypothesis as described in Section 3.8. According to the definition given there, the discrete set W satisfies the cosmological hypothesis if there exists an embedding of W in $\mathbf{R}^4$ that is *nearly* order-preserving ($\mathbf{R}^4$ is, as always, considered to have the Lorentz ordering). Furthermore, it is assumed that there is also a nearly order-preserving mapping of the set of *positions* of W in $\mathbf{R}^4$ (see Section 3.6). Thus, it can be imagined that W and the positions of W are really just discrete, ordered subsets of $\mathbf{R}^4$. Following the reasoning

in Section 3.8, one can assume that the metrical properties of W—as far as they were defined and used in the description of the formula (47)—are given, at least locally and approximately, by the usual Lorentz metric in $\mathbf{R}^4$. Imagine that the set of positions of W, as a subset of $\mathbf{R}^4$, corresponds with the idea of points in "empty space" in the usual picture of space-time. The points of W are thought of as representing material particles traveling through $\mathbf{R}^4$.

To continue with the analogy, imagine that the set of positions, although being discrete, is extremely dense in $\mathbf{R}^4$ when compared with the points of W. If, as is often said, there are perhaps 10^{80} particles to be seen in the observable universe, then, upon thinking about our definition of a "position" in W, it seems reasonable to assert that the density of positions in $\mathbf{R}^4$ should be something like 10^{80} times as great as the density of the points of W in $\mathbf{R}^4$. Furthermore, since most matter in the observable universe is cosmologically distant, one expects the local density of positions to be nearly constant, varying little with the local density of the points of W.

Given all of these assumptions, I will then go on to argue in the following manner. To begin with, we can think about how the conventional theory of gravity—as a purely *practical* matter—is used in the calculation of the paths of material objects. It is surely so that the practicing physicist begins by considering empty space—using the single, canonical coordinate neighborhood for $\mathbf{R}^4$—and then thinks of altering or perturbing the flat space, Lorentz metric by adding in material objects. These perturbations generally involve just the first one or two terms in the Schwarzschild solution. One ends up then with an approximate solution to the equations of general relativity, which is usually sufficiently good to account for all possible observations.

Now, the present idea is to proceed in a similar manner with the analysis of our discrete set W. Let us begin by imagining an "empty" space, as in the case of the conventional theory. What is an empty space in our framework? Obviously, it would be false to simply take $W = \varnothing$. If W were empty, then there could be no positions, and thus no "empty space," according to our assumptions! The best one can do is to take a region of $\mathbf{R}^4$ (considered together with the embedding of the points and positions of W) that contains few points of W and yet has a dense and homogeneous set of positions. Thus, this region of $\mathbf{R}^4$ corresponds with the idea of a region of space-time, say, in some distant interstellar space. Then, continuing with the analogy, one adds in particles to W one by one in this region.

As the particles are added in, one will expect that, automatically, *positions* will also be added in to W, and thus in to $\mathbf{R}^4$. After all, positions are defined in terms of the points of W: If there are no points, then there will be no positions. If there are many points, then there will be many

positions. Hence, one expects that, in some way, adding in particles to an existing W will tend to generate new positions, and we will be interested in the density and distribution (in $\mathbf{R}^4$) of the set of these new positions in W.

But then, having added new particles into W, and without changing W in any other way, one will suddenly find that the original particles of W are no longer proper (Definition 2 of Section 3.7). Thus, to "restore" the property of regularity (so to speak), it will be necessary to alter the metrical structure of W—represented by an alteration of the metrical structure of $\mathbf{R}^4$, as is usually done in practical calculations involving the general theory of relativity.

All this is admittedly vague. A more acceptable procedure would be to deal with a given discrete set W as a fixed entity fulfilling all of the assumptions in Section 3. Unfortunately, one has no practical methods for the exact analysis of such sets, just as the theoretical physicist has no exact method for dealing with the n-body problem in the theory of general relativity.

4.8. How Points Generate Positions in Finite Sets

Let W be a finite, partially ordered set with, say n elements. If W happens to be empty, then, as already noted, the set of positions of W is also empty. On the other hand, if W is nonempty, then, remembering Theorem 3.5, we have that the set of positions of W must also be nonempty. In this section imagine that the elements of W are numbered from 1 to n, so that we can write $W = \{\omega_1, \ldots, \omega_n\}$. Then, for each $i \in \{1, \ldots, n\}$, let $W_i = \{\omega_1, \ldots, \omega_i\}$. For completeness, let $W_0 = \varnothing$ as well.

Now, as far as these *sets* are concerned, we clearly have $W_{i-1} \subset W_i$ for all relevant i. But what about the positions? Does it make sense to say that the set of positions of W_{i-1} is contained in the set of positions of W_i? Consider the following ideas.

To begin with, $W_0 = \varnothing$, and the set of positions of W_0 is empty. One has $W_1 = \{\omega_1\}$, and the set of positions of W_1 consists of just the element ω_1. In general we expect that W_{i+1} contains more positions than W_i, for, let, say, C_1 and C_2 be two different positions in W_i. Then C_1 and C_2 are also subsets of W_{i+1}, and we certainly have $C_j^- \leq C_j^+$ in W_{i+1} for $j = 1, 2$. Perhaps C_j is maximal (see the definition in Section 3.6) and is thus already a position in W_{i+1}. If C_j is *not* maximal, then (since $W_{i+1} - W_i$ consists of the single point ω_{i+1}), it can be made maximal by adding in the element ω_{i+1} either to C_j^- or to C_j^+. Thus, in either case we obtain a unique position D_j in W_{i+1}, which contains C_j, for $j = 1, 2$. Furthermore, since $W_{i+1} - W_i$ consists of a single point, we must have either $D_j^+ = C_j^+$ or $D_j^- = C_j^-$ for both $j = 1, 2$. Thus, certainly $D_1 \neq D_2$. Letting $\mathscr{P}(W_i)$ denote the set of

positions of W_i, we obtain then a natural mapping $\varphi_i : \mathcal{P}(W_i) \to \mathcal{P}(W_{i+1})$, which is monomorphic.

Now it is reasonable to say that a partially ordered set with many elements contains many positions. Thus, it is obvious that, in general, the set $X_{i+1} = \mathcal{P}(W_{i+1}) - \varphi_i(\mathcal{P}(W_i))$ is not empty. We will say that the positions in X_{i+1} are *associated with the element* ω_{i+1}. Of course this a rather labile association: it depends on the way we count the elements of W. But still, such positions can be thought of as being closely related to ω_{i+1} within the "geometry" of W. It is also interesting to note that if the set

$$C = \{u \in W: u < \omega_{i+1}\} \cup \{u \in W: u > \omega_{i+1}\}$$

happens to be a position in W_i, then, according to this definition, the position

$$D = \{u \in W: u \leq \omega_{i+1}\} \cup \{u \in W: u \geq \omega_{i+1}\}$$

in W_{i+1} is *not* associated with ω_{i+1}!

What can be said about the positions associated with a given element $\omega_i \in W$? Let the position C be associated with ω_i. Can it be that $\omega_{i+1} \notin C$? If this were true, then C would be a position in both W_i and W_{i+1}. Thus, C would be contained in $\varphi_i(\mathcal{P}(W_i))$, a contradiction. We must conclude that C is either above or below ω_{i+1}.

Furthermore, C is on the "edge" of either the cone of points above or below ω_{i+1} in the following sense. Let $c_1, c_2 \in C$ be such that $c_1 < \omega_{i+1} < c_2$. Then $c_1 \in C^-$ and $c_2 \in C^+$. For, if we assume, say, that both c_1 and c_2 are contained in C^+, then we would have $D = C - \{\omega_{i+1}\}$ being a position in W_i. This can be seen by noting that if D were not maximal, then there would exist an element, say $u \in W_i$, with $u \notin D$, and either $u > D^-$ or $u < D^+$. But $u > D^-$ is impossible, since $D^- = C^-$. The other case, $u < D^+$, is also impossible, since we would have in particular that $u < c_1$, and thus $u < \omega_{i+1}$. But then it would follow that $u < C^+$, so that $u \in C^- = D^- \subset D$.

Thus, this idea of the "edge" of a position can be given both a geometric and an algebraic meaning. The geometric idea is clear: we think of positions as being "light-cones," i.e., double cones in $\mathbf{R}^4$, and the edge of such a cone is its topological boundary. The algebraic idea is expressed in the above paragraph: we can imagine it by saying that an element $u \in W$ is on the edge of the cone C in W if there are no other elements of W between u and the "vertex" of C.

4.9. Positions in Infinite Sets

If the partially ordered set W is infinite, then the ideas of the last section can no longer be applied. It is not possible to construct W by successively adding in new elements one after another, as was done there.

On the other hand, if W is strongly discrete, then it is possible to investigate finite regions of W, using the methods of Section 4.8.

Specifically, let us take two points $p, q \in W$ with $p < q$. That W is strongly discrete implies that there are at most finitely many elements of W in the set $\Lambda_{pq} = \{u \in W: u \leq q, \text{ but not } u < p\}$. Let us say that there are n elements of W in Λ_{pq}. We can write $\Lambda_{pq} = \{\omega_1, \ldots, \omega_n\}$ and then consider the sequence of partially ordered sets $W_{n-i} = W - \{\omega_1, \ldots, \omega_i\}$. In particular, $W_0 = W - \Lambda_{pq}$, $W_n = W$, and $W_i \subset W_{i+1}$, for all relevant i.

Now it is clear that there can be no positions of W_0 that are strictly between p and q. By adding in the elements of Λ_{pq} one after the other, we reconstruct W and also add in all of the positions of W that lie between p and q. Thus, as far as the finite space between p and q is concerned, it is possible to apply the same analysis as was used in Section 4.8, and thus we can associate positions with elements of W, as was done there.

The main problem, of course, is to try to pursue the program outlined in Section 4.7. The partially ordered set W is assumed to satisfy the various assumptions listed in 4.7. The set W—and also the positions of W—are to be thought of as being embedded in a (more or less) order-preserving way in $\mathbf{R}^4$. It is this embedding in $\mathbf{R}^4$ that will enable us to deal with the positions of W in a *geometric* (that is, distance-related) way.

Now, the simplest way to proceed is to assume that the embedding $\Psi: W \to \mathbf{R}^4$ is *strictly* order-preserving, so that for $u, v \in W$ we have $u < v \Leftrightarrow \Psi(u) < \Psi(v)$. Certainly this assumption is very restrictive—the class of partially ordered sets W for which there is an order preserving mapping of $W \to \mathbf{R}^4$ can be thought of as being much smaller than the class of sets having a proper representation in $\mathbf{R}^4$. As previously noted, such a condition on W is certainly too restrictive. But for now I simply accept this assumption as providing a practical working hypothesis.

We are interested in the geometry of the set of positions of W. But Example 2 of Section 3.6 shows that, in general, these positions do not correspond with the points of $\mathbf{R}^4$. Thus, strictly speaking, a new kind of geometry, other than that of $\mathbf{R}^4$, is necessary for a precise analysis. But rather than getting involved in such complicated questions, it seems best to expand our working hypothesis to include the assumption that the positions of W are just the subsets C and W that are of the following form. Given the embedding $\Psi: W \to \mathbf{R}^4$, we can take an arbitary point $x \in \mathbf{R}^4$, and then take $C = C^- \cup C^+$ to be such that $C^- = \{u \in W: \Psi(u) \leq x\}$. Then, given such a C^-, we take C^+ to be such that $C^+ = \{v \in W: v \geq u, \forall u \in C^-\}$. Thus, according to this way of thinking, the positions of W can be very closely associated with the points of $\mathbf{R}^4$: such a position can be thought of as corresponding with the point $x \in \mathbf{R}^4$. Note that since W is strongly discrete, it follows that in any compact region $K \subset \mathbf{R}^4$ there can be only finitely many

different lower cones of the form $C^- = \{u \in W: \Psi(u) \leq x\}$ for $x \in K$. We could define an equivalence relation on $\mathbf{R}^4$ using the rule that for $x, y \in \mathbf{R}^4$, $x \approx y \Leftrightarrow \{u \in W: \Psi(u) \leq x\} = \{u \in W: \Psi(u) \leq y\}$. It could then be imagined that a *position* in W is just an equivalence class, by this rule, in $\mathbf{R}^4$.

It is interesting to note that this use of the concept of "strongly discrete sets" would seem to imply a type of "retarded interaction." That is to say, imagine first that we have W_i as above, and then we add in ω_{i+1} to W_i to obtain W_{i+1}. Let C be a position in W_{i+1} that is associated with ω_{i+1}. Now our assumptions imply that the cone C is determined by its lower cone C^-. In particular, this means that if C is associated with ω_{i+1}, then C must lie *above* ω_{i+1}. Thus, according to this way of thinking, elements ω of W generate new postions above ω—or thinking in terms of space and time, an event at some point of space-time generates new positions in the *future.* If we return to the ideas of Section 4.7, then we can translate this into the thought that the assumptions there—the "cosmological hypotheses" and in particular the assumption that the density of points of W in $\mathbf{R}^4$ increases with increasing "time"—imply that the force of gravity is of a purely retarded nature, and therefore that it satisfies the law of cause and effect.

Finally, it will be useful to consider the idea that adding in a new element to a partially ordered set increases the density of *positions* of that set. Once again, let C be a position in W_{i+1} associated with the element ω_{i+1}. Now, if we return to the smaller set W_i, then we have the single position D, say, which is such that $D^- = C^- - \{\omega_{i+1}\}$. On the other hand, in the set W_{i+1} we have *two* positions, namely the position determined by C^- and also the position determined by $C^- - \{\omega_{i+1}\}$. The later is indeed a position, since, as proved in Section 4.8, C is on the edge of the cone above ω_{i+1}.

Now, as we have seen, positions in strongly discrete, partially ordered sets W can be associated with elements of W. The question of most interest in the study of the geometry of positions is, what are the distances between the elements of W and the positions with which they are associated? One would like to claim that an element ω of W is associated with many positions close to ω, but as one travels further and further away from ω, the probability of finding a position associated with ω decreases. In fact, as far as establishing a correspondence with the theory of gravity is concerned, one would like to show that the density of positions associated with ω is proportional to the inverse of the distance from ω.

4.10. The Distance to New Positions: First Derivation

Let the partially ordered set W satisfy the conditions of the preceding subsection. I now present an argument that shows, for a large class of

possible discrete geometries, that the density of the positions in $\mathbf{R}^4$ associated with a given element $\omega \in W$ is inversely proportional to the retarded distance to ω (where ω is considered as a point in $\mathbf{R}^4$).

The idea is analogous to the usual "gauge invariance" condition that physicists often invoke. Specifically, assume that the density of the points of W in $\mathbf{R}^4$ depends only on the fourth—"time"—component. Thus, one can imagine that there exists some positive function $d_1 : \mathbf{R} \to \mathbf{R}_+$ such that the density of the points of W in $\mathbf{R}^4$ around the point $(x, y, z, t) \in \mathbf{R}^4$ is given by $d_1(t)$. Since W is strongly discrete, and furthermore, since the positions of W are defined in terms of the elements of W, it seems reasonable to assume that the density of the positions of W in $\mathbf{R}^4$ also depends only on the time. Thus, one can write $d_2(t)$ to represent the density of the positions of W in $\mathbf{R}^4$ at the point (x, y, z, t). The gauge invariance principle to be used is then the assumption that for all $t \in \mathbf{R}$, the ratio $d_1(t)/d_2(t)$ is a constant.

What are the consequences of these assumptions? To begin with, recall from Section 4.9 that the positions of W can be associated with points of W; in particular, each position is associated with a point that comes *before* the given position in time. Furthermore, it was concluded there that adding in a new element ω to W increases the density of positions of W above ω. For the purposes of the present argument then, assume that

$$d_2(t) = \int_{-\infty}^{t} \text{const} \times d_1(s)\, ds \tag{68}$$

This expresses the idea that (1) points such as ω_s in W at the time s contribute to building positions P_t in W at time t, where $t > s$, and (2) the propensity of ω_s to build P_t depends only on the density of the already given positions around P_t. Admittedly, this formula is somewhat stronger than the arguments of Section 4.9 allow. Continuous functions like d_1 and d_2 in $\mathbf{R}^4$ describe a structure that is very different from the discrete structures considered there. But a more serious objection is that one has not given a precise definition of these density functions. Thus, the idea 2 above, while appearing to follow from the philosophy of "gauge invariance," has not been strictly established; how can one reconcile the discrete adding in of individual elements of W, as in Section 4.9, with the assumed homogeneity of W in spacelike directions in $\mathbf{R}^4$ required by the density functions? In addition, one is assuming that the probability that a new position P_t is generated by adding in the element ω_s depends only on the density of the already existing positions around P_t. Again this might be a reasonable idea in the framework of continuous distributions of "particles" and "positions," but it can at best be an approximation when applied to discrete spaces. Finally, it could also be objected that while gauge invariance plays an

important role in "local" descriptions of quantum field theories, in fact its *global* counterpart in the realm of cosmology—namely the "steady state universe"—appears for various reasons to have fallen into disrepute. Thus, if we are to adhere to the currently accepted interpretations of the astronomical observations, then we must be prepared to abandon the principle of gauge invariance, at least to the extent that it is applied to globally defined models such as the one we are considering.

But granted all of these limitations, it is still interesting to examine the consequences of (68). Now, assuming that d_2 is proportional to d_1, it follows that

$$d_i(t) = c_i\, e^{kt} \tag{69}$$

for $i = 1, 2$, with appropriate constants c_i and k. [Of course, the critical reader may at this point object—with reason—that (69) could itself have been asserted directly from a global gauge invariance condition!]

Granted (69), then consider a contraction of $\mathbf{R}^4$ of the form $\lambda : \mathbf{R}^4 \to \mathbf{R}^4$ given by $\lambda((x, y, z, t)) = (ax, ay, az, at)$ for some $0 < a < 1$. Let the embedding $\Psi : W \to \mathbf{R}^4$ be given, having the density functions $d_1(t)$ and $d_2(t)$, satisfying (69). Now the set $W_0 = \lambda\Psi(W) \subset \mathbf{R}^4$ is similar to, but "denser" than, the original embedded set $\Psi(W) \subset \mathbf{R}^4$. Thus, we also have density functions b_1 and b_2 for W_0. The similarity between W and W_0 is expressed by the conditions $b_i = ad_i$ for $i = 1, 2$. It is therefore a natural idea of think of removing some "homogeneous" set of points W^* from $\lambda\Psi(W)$, thus "thinning out" $\lambda\Psi(W)$, so to speak, and producing a subset $W_1 = W_0 - W^* \subset W_0$ similar to the original embedding $\Psi(W)$ and having the same density functions d_1 and d_2 as $\Psi(W)$ had.

The purpose of this contraction and then removal of the subset W^* is to allow us to use the ideas of Sections 4.8 and 4.9. On one hand, W_1 is similar to $\Psi(W)$, and thus to W_0. On the other hand, W_1 is a subset of W_0, so that it is possible to add in the points of W^* to W_1, gradually building up W_0, and allowing thereby the association of positions in W_0 with points of W^*. Let $v \in W^*$ be some arbitrary point. We are interested in the set of positions $\mathcal{P}(v)$ in W_0 that are associated with v. Now, it is being assumed that the positions of $\mathcal{P}(v)$ are to be found on the boundary of the light-cone above v. There is no preferred direction here, so it is reasonable to say that the density of the positions of $\mathcal{P}(v)$ depends only on the retarded distance to v. Thus, it is assumed that there exists some real function $\phi : \mathbf{R}_+ \to \mathbf{R}_+$ such that the density of $\mathcal{P}(v)$ at the retarded distance r from v is given by $\phi(r)$.

What form does this density function ϕ have? Since W_1 occurs in the process of constructing W (according to the way of thinking in Section 4.9), one can write $\phi(ar) = \phi(r)/a$, or $\phi(r) = a\phi(ar)$, for all real $a > 0$.

(This follows because the relationship between the positions of W and the elements that generate them is the same for the "dense" set W_0, the "thinned out" set W_1, and also for the original set W.) But then we need only choose $a = 1/r$ to see that $\phi(r) = \text{const} \times (1/r)$. Therefore, by this argument, we have reached the desired result that the density of positions associated with a given element of W is inversely proportional to the retarded distance.

This derivation follows from simple and general principles of symmetry, and yet despite this fact, it should be criticized. I have already listed some of the problems that must result from such a mixture of continuous and discrete ideas. In addition, the argument is based on the idea of establishing a correspondence between two similar embeddings of W in $\mathbf{R}^4$ related by means of a guage transformation. Thus, the argument loses its validity if applied to spaces modeled on a different geometry than $\mathbf{R}^4$: in particular, it is invalid if we assume that our model for space-time is compact in spacelike directions. For these reasons it seems appropriate to look for a more direct explanation of the relationship between retarded distances and the density of the positions associated with a given element of W.

4.11. The Second Derivation

For the purposes of this second derivation of the density-distance relation I will continue to make use of the assumptions of Section 4.9. But, in contrast to the argument of Section 4.10, I will go beyond the idea that the positions associated with elements v of W can be calculated using "density" functions with respect to a given embedding of W in $\mathbf{R}^4$. Now it is proposed to investigate in more detail the method by which the geometry of W influences the relationship of "association" between positions and elements of W. Of course, it is still necessary to work in $\mathbf{R}^4$; the density-distance relation is expressed in terms of Euclidean geometry. Thus, it is still necessary to make difficult assumptions regarding the problem of finding good embeddings of W in $\mathbf{R}^4$.

To begin with, it is important to remember that only the positions that are nearly of the form discussed in Section 4.9 are important: extreme examples, such as Example 2 of Section 3.6, can be disregarded. This follows from the way we are using the concept of positions to define distances in W. Only the positions between adjacent points of a particle are used to determine the distance between these points in the discrete version of Fokker's principle (47); all other positions play no further role in the argument of Section 3.9. Recall, finally, that in the discussion in Section 4 it was concluded if the density of points of W in $\mathbf{R}^4$ increases with increasing time, then a given position of W is determined by its *lower* cone.

Now one can think of this as follows. Choose two points $p, q \in W \subset \mathbf{R}^4$, with $p < q$, and consider the space between p and q. Since W is strongly

discrete, there are at most finitely many points beneath q, but not beneath p. Choose a hyperplane

$$H_\tau = \{(x, y, z, t) \in \mathbf{R}^4 \text{ such that } t = \tau\}$$

The constant τ is chosen so that H_τ is below all the points of W that are beneath q, but not beneath p. Then, according to the cosmological hypothesis, one can assume that the intersections of the cones beneath these points—in the proper representation of W in $\mathbf{R}^4$—are topological 3-cells that are nearly perfect geometrical 3-balls.

The emphasis here is on the word *nearly*. The 3-cells in H_τ—representing points and positions of W—can be thought of as having boundaries that depart somewhat from true sphericity. In fact, the picture I have in mind is that most of these 3-cells can be nearly represented as convex hulls spanning some other set of smaller 3-cells in H_τ (Figure 4).

What is the geometric relationship between a point of W in $\mathbf{R}^4$ and a position with which it is associated? Let $v \in W$ be such a point—beneath q but not beneath p—and let C be a position in W associated with v and lying between p and q. Now, both v and C are represented in H_τ by nearly spherical 3-cells. We will assume that C can be represented nearly as the convex hull of some set of smaller 3-cells in H_τ. To describe this situation, we can say that C is *supported* by the point $u \in W$ if $u \in C^-$, and u is associated with a 3-cell in H_τ that is on the edge of the 3-cell in H_τ representing C. Now C is associated with v. That is, according to the reasoning of Section 4.9, one should first imagine W with v removed, then, upon adding in v to W, find C occurring as a new position. But v itself is defined in terms of its relationships with the other points of $W - \{v\}$; i.e., v is also determined by its lower cone in $W - \{v\}$. Now this lower cone of v is represented in turn by the convex hull of a collection of nearly spherical

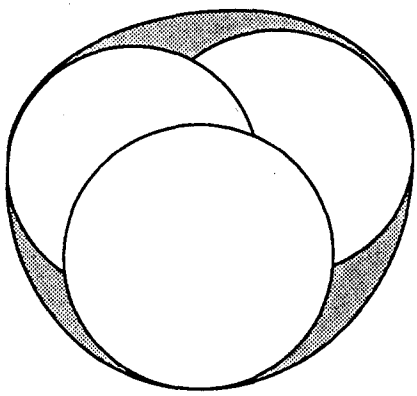

Fig. 4

3-cells in H_τ. Given all of these assumptions, then, we can assert that if there exists a point $u \in W - \{v\}$ that supports both C and v, then C can be associated with v.

What is the probability that two such 3-cells B_v and B_W in H_τ, representing a given point v and a given position C in W, have a common supporting point u? It is clear that at this stage one can bring in a geometric argument: two such 3-cells can have such a common supporting point only if their boundaries are nearly tangent. Furthermore, the probability of having such a common supporting point must be proportional to the two-dimensional area where ∂B_v and ∂B_W approach each other closely.

Thus, taking all of these thoughts into consideration, it seems reasonable to proceed on the basis of the following model. Let v, p, q, and H_τ be as above. Choose some point x between p and q in $\mathbf{R}^4$ lying on the Lorentz light-cone above v. Let $S_v \subset H_\tau$ and $S_x \subset H_\tau$ be the 2-spheres in H_τ that are the intersections of H_τ with the Lorentz light-cones beneath v and x, respectively. Then one may assert that the number of positions of W between p and q that are associated with v is proportional to the area of S_v that lies near S_x. More specifically, take some $\varepsilon > 0$ that is small in relation to the retarded distance from v to x and to the distance from v to H_τ. Then one can assert that the number of positions of W between p and q that are associated with v is proportional to the area of the subset of points of S_v that have a distance at most ε from S_x.

Figure 5 depicts the two spheres S_x and S_v. The radius of the inner sphere S_v is r, while the radius of the outer sphere S_x is R. The spheres S_x

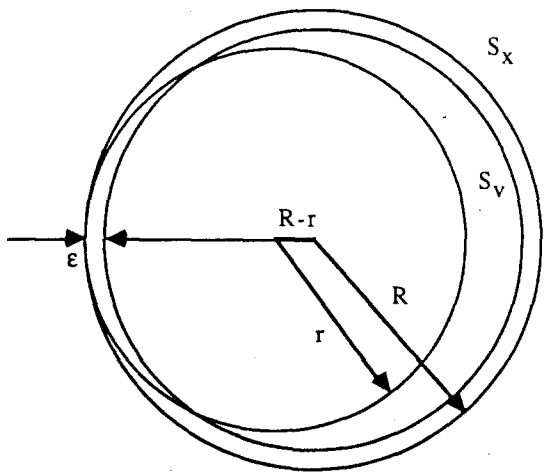

Fig. 5

and S_v are tangent to one another at some point P in H_τ. A sphere of radius $R-\varepsilon$ that is concentric to S_x has also been pictured. One may imagine that R is much greater than $R-r$ and that $R-r$ is much greater than $\varepsilon > 0$. The problem then is to determine the area of the portion of S_v that lies in the concentric region between S_x and the sphere of radius $R-r$. The area of this subset of S_v can be deduced by means of a simple geometrical argument (using the fact that we are working in four-dimensional Euclidean space) and is found to be proportional to $1/(R-r)$.

This is, once again, the desired answer—namely that the density of the positions associated with a given element of W is inversely proportional to the retarded distance to that element. It is obvious that once again I have made very many geometrical assumptions of a more or less arbitrary nature whose appropriateness is, at best, debatable. It can only be hoped that in the future better means can be found to describe the geometry of partially ordered sets.

4.12. A Connection with General Relativity

Let us return to the problem of calculating the geometry of space in the neighborhood of a massive body. In the framework of general relativity, one might say that the natural, "undisturbed" state of things is the flat space $\mathbf{R}^4$ with the Lorentz metric. Gravity comes about when one introduces matter into the flat space, thus altering it—producing a curvature. Now I will also make use of this way of thinking.

As in Section 4.7, consider the flat space of special relativity as represented by some region in $\mathbf{R}^4$ (with the partially ordered set W embedded in it) that is distant from any particular concentration of points of W. Thus, as we have seen, in such a region the positions of W can be expected to be distributed nearly homogeneously. It will be assumed that this set W satisfies the discrete version of Fokker's variational principle (47).

Now that we have settled on this flat, undistributed region of space-time, the next task is to introduce an accumulation of matter into the region, as one does in general relativity. Here, this means adding one by one a great number of discrete, nearly identical particles that pass through the region all following more or less the same path. For simplicity, and since I have not shown how to deal with the gravitational fields of moving objects, using the linearized gravitational equations [see, e.g. Narlikar (1978) for this] I assume that this path is straight and parallel to the fourth (time) coordinate axis of $\mathbf{R}^4$.

Denote the set of discrete particles to be added in to W by $\{L_1, \ldots, L_n\}$. The first thing to do, then, is to add in the particle L_1. According to the ideas of the last two subsections, this produces in the new partially ordered

set $W \cup L_1$ a new set of positions, additional to the original set of positions in W, namely the positions in $W \cup L_1$ associated with elements of L_1. The density of this new set of positions is approximately inversely proportional to the distance from L_1. Thus, if the density of positions of W in $\mathbf{R}^4$ (in the region we are considering) is given approximately by the constant D, i.e., D positions of W per unit volume in $\mathbf{R}^4$, then in the new set $W \cup L_1$ we have the density of positions at some point $(x, y, z, t) \in \mathbf{R}^4$ lying (cosmologically) near, but at a distance r from L_1, being given approximately by

$$D \times (1 + k/r) \tag{70}$$

where k is some appropriate small constant. Of course one should not take this formula too literally, and object that it implies that at L_1 the density must be infinite. Such an assertion would be nonsense, since we have assumed that W is strongly discrete. Instead, the formula can only be seen as providing a useful appproximation, making use of familiar concepts from the traditional *continuous* geometry (in particular, making use of the idea of "density").

We now have the set $W \cup L_1$. The next thing is to add in the particle L_2, giving the larger set $W \cup L_1 \cup L_2$. Once again we obtain an additional set of positions, which this time are associated with the elements of L_2. The increase in *density of positions* at a given point in $\mathbf{R}^4$ near L_2 is once again proportional to the inverse of the distance to L_2. Since L_1 and L_2 nearly coincide, we obtain that the total density of positions in $W \cup L_1 \cup L_2$ is given approximately by

$$D \times \left(1 + \frac{k}{r}\right) \times \left(1 + \frac{k}{r}\right) = D \times \left(1 + \frac{k}{r}\right)^2 \tag{71}$$

It is important to note the nonlinear character of this formula. It is definitely *not* the case that the same number of positions in $W \cup L_1 \cup L_2$ are associated with L_1 as are associated with L_2! If such were the case, then we would have the density of positions in $W \cup L_1 \cup L_2$ being given by $D \times (1 + 2k/r)$. The formula we have gives a somewhat greater value, namely $D \times (1 + k/r)^2 = D \times [1 + 2k/r + (k/r)^2]$. How can this be explained? I could refer to the argument in Section 4.9, which shows that adding in new elements of W increases the density of the existing positions of W. But it is also easy to see that when adding in L_2 there are more possibilities for creating new positions than there were in the case of L_1; these new positions associated with L_2 can utilize not only the original elements of W, but, in addition, also the elements of L_1.

It is now possible to add in the further particles $L_3, \ldots, L_n$. Using the same arguments and assuming that n is large, we obtain that the density

of the set $W \cup L_1 \cup \cdots \cup L_n$ is given approximately by

$$D \times (1 + k/r)^n \approx D \times e^{kn/r} \tag{72}$$

But kn is proportional to the total mass of the particles in $L_1 \cup \cdots \cup L_n$, and thus a connection with the formula (67) is established.

Is it possible to deduce from these ideas a theory of gravity that is similar to the usual theory? For this purpose a number of new ideas will be necessary. Relation (72) only describes the relationship of the density of the positions in W to the density of the positions in $W \cup L_1 \cup \cdots \cup L_n$. But, as we have seen, gravitation is a geometrical phenomenon, depending on the measurements of distance within space-time. Thus, it will be necessary to look at the effect of this change of density of the positions in our partially ordered set W on the metric structure of W.

4.13. The Metric Structure of W

The idea of distances in discrete, partially ordered sets was dealt with in Section 3.7. In particular, I assume that all the discrete particles in W are *proper* (see Definition 1 of Section 3.7). This definition is applied in the discrete formulation of Fokker's action principle (47). Since I am dealing with gravity, I assume that all of the particles have zero electrical charge, so that the interaction terms in (47) vanish. What remains is the assertion that in W all the particles are proper, and their lengths—measured by counting the number of elements of W along finite stretches of typical particles—are extrema with respect to finite variations.

Now, the present approach to dealing with this complex situation has been to imagine that W is embedded in the familiar space $\mathbf{R}^4$ in such a way that the ordering structure of W is nearly reflected by the Lorentz ordering in $\mathbf{R}^4$. This seems to be a reasonable idea in the "flat" area of W, which was the beginning position for the argument of Section 4.12. But, as we have seen, the introduction of the new particles $L_1, \ldots, L_n$ changes the density of the positions of W, resulting (through the definition of "proper" particles) in a change in the Lorentz distances to be associated with the distances between adjacent points on a proper particle. How can we deal with these changed Lorentz distances?

Let, say, p, q be adjacent elements on the particle L_1. That is, $p < q$, and there is no element of L_1 that lies properly between p and q. Since L_1 is assumed to be a proper particle, there is some constant that specifies the number of positions of W between p and q. (Note that there can be no position of $W \cup L_1$ between p and q that is associated with an element of L_1.) It can therefore be imagined that p and q according to the Lorentz ordering has some fixed four-dimensional volume, characteristic for the particle L_1.

Then the other particles $L_2, \ldots, L_n$ are gradually introduced into $\mathbf{R}^4$ one by one, bringing with them more and more particles and positions, according the formula (72). Now if we leave p and q and the basic metrical structure of $\mathbf{R}^4$ fixed, then we will suddenly find that there are too many positions of W between p and q. With the increased density of positions, it is to be expected that some of the new positions will be found in the space between p and q. But not only here. All of the (proper) particles of W, those moderately near L_1 and also those far away, will also receive new positions between their adjacent elements—according to the formula (72)—thus invalidating throughout W the property that the particles are proper!

The only way to repair the situation and restore the property that the particles should be "proper" is to alter somewhat the ordering structure of W. This can be achieved by (1) adjusting the embedding of W in $\mathbf{R}^4$ and also by (2) altering somewhat the basic ordering of $\mathbf{R}^4$, resulting in some departure from the Lorentz metric. A combination of these methods should be most appropriate. There is a certain freedom of choice, which results in a multiplicity of essentially different ordering structures. What choice should we make?

To answer this question it is best to remember that the choices here, involving various metrics for $\mathbf{R}^4$, are, from the point of view of physics, concerned with the measurement of the speed of light. The causal structure of W is, after all, to be thought of as representing the causal structure generated by the light-cones in the usual formulation of space-time. Now, the most basic principle of the theory of relativity is that all possible measurements of the speed of light must yield the same constant result. Thus, it is natural to adopt the same principle and assume that the alterations to the metrical structure of W resulting from the introduction of the new particles $L_1, \ldots, L_n$ must be made in such a way as to preserve the constancy of the measured speed of light in W.

This leads to the question, how should the speed of light be measured in W? Imagine a conventional experiment to measure this speed. The experiment begins with a pulse of light being emitted at the point $A \in \mathbf{R}^4$, say (Figure 6). An observer is stationed at A to keep track of events. The light travels to some point $B \in \mathbf{R}^4$, where it is reflected by a mirror back to the observer, whom it meets at the point $C \in \mathbf{R}^4$. Now it is assumed that the observer travels from A to C in a straight line, without acceleration. Let D be the halfway point on the line connecting A and C. Then the speed of light deduced from this experiment is to be calculated by taking for the *distance* the Lorentz distance between D and B, and for the *time* the Lorentz distance between A and D.

All of this can be made to make sense in an abstract, partially ordered set W (even without imagining an appropriate embedding in $\mathbf{R}^4$). Finding

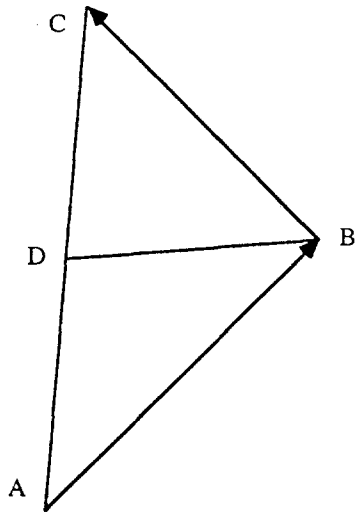

Fig. 6

appropriate points A, B, and C within such a set is just a matter of finding points on the edges of the appropriate cones of W—given by the ordering—above A and above B. The halfway point D can be found by specifying that the number of positions between D and A is equal to the number of positions between D and C, and this number is the largest possible.

This experiment allows the observer traveling from A to C to fix an idea of the local speed of light for that observer. The particular value chosen is unimportant for the purposes of the theory of relativity. The important thing is to have a constant, and thus *consistent* value for this speed, for all possible measurements. Now, one measure of consistency is to consider the following, slightly more complicated experiment, which allows the observer to measure the speed of light at some distance.

In this new experiment, the observer emits a pulse of light at A, then travels a short distance to A' and emits another pulse of light (Figure 7). The observer then continues in the same direction without acceleration. The first pulse is reflected by a mirror at B, while the second is reflected by a mirror at B', which is closer to the line taken by the observer than B. The two light pulses travel back and meet the observer at the points C and C', respectively. One may assume that things have been so arranged that the distance from A to A' is equal to the distance from C to C'. Once again, D is the halfway point between A and C. By this means, the speed of light around B and B' can be measured by an observer traveling from A to C.

Of course, the retarded distances, say from A to B or from B to C, are given by Definition 4 of Section 3.7. Thus, we have two different ways

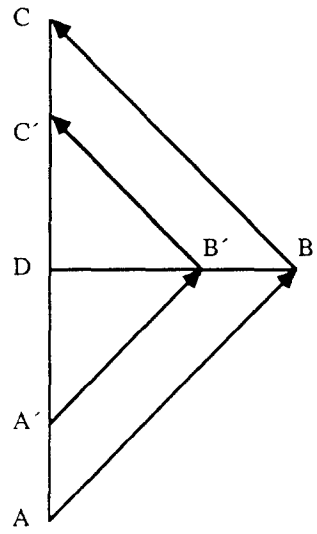

Fig. 7

of measuring distances within the abstract, partially ordered set W, and it is certainly conceivable that these various experiments might measure different speeds of light within W. Looked at another way, it seems reasonable to say that the condition that all such measurements must yield the same local speed of light is a very restrictive new condition to be imposed on all partially ordered sets that are to be considered as models for space-time. Thus, W is defined to be *consistent with respect to distances* if all measurements of the "speed of light" in W using these procedures and definitions yield a consistent value.

Can it be that this condition is what is needed to bring us from expression (72) to a geometrical model for gravity, similar to that described in Section 4.6? That fact that there are three spatial dimensions and one dimension of time should play an important role in such an investigation.

5. A POSSIBLE CONNECTION WITH THE THEORY OF QUANTUM MECHANICS

5.1. The Interpretation of Quantum Mechanics

The question of the interpretation of quantum mechanics must surely be one of the most difficult and controversial areas of philosophy. In what other subject can one find such vehement, even bitter criticism, as that

expressed by Landé (1965) or Schrödinger (1953)? It is well known that Einstein was never prepared to accept the theory, except insofar as it could be considered to be an "incomplete" version of some more reasonable, underlying theory (see, e.g., Schilpp, 1949). A summary of many of the possible points of view and of the positions maintained by famous physicists can be found in Jammer (1974). Particularly useful is Feynman *et al.* (1965).

Now the idea of the standard "Copenhagen interpretation" of quantum mechanics is that we should begin by rejecting all hopes of constructing a geometric model that has mathematical properties modeling the physical world as we experience it. Instead, it is necessary to construct an abstract model: a Hilbert space. In principle, there is nothing to object to in this: after all, the theory of Hilbert spaces is an interesting and widely studied area of pure mathematics. As far as Hilbert spaces are concerned, one normally considers such things as "points," "lines," "functionals," and so forth. The physicist prefers to use other names, such as: "experiments," "observers," "states." Once the Hilbert space framework is accepted, one can build up a consistent mathematical model for describing the physical world. For example, von Neumann (1955) was one of the first to do this.

Thus, the problem in the interpretation of quantum mechanics is not the lack of a good model, or even the question of whether or not there exist experiments that may be at variance with the model. Physics, when taken to be the study of Hilbert space, appears to pass all such tests. The real question is, why is it necessary for physicists to base their theory on a kind of axiom system whose basic, immutable, undefinable objects—denoted by the words "observer," "experiment," etc.—appear to signify concepts that any *sensible* person would think of as being very much mutable and definable. Another common objection, expressed particularly by mathematicians, is that the relationship of the Hilbert space construction to the more concrete models of the physical world that are used in mathematics (for example, in elementary geometry or classical probability theory) appears to be rather obscure. For example, the successful and practical new methods for analyzing the structures of large molecules can be understood within a very traditional geometrical framework. Is it possible that this is only an illusion, resulting from some obscure "correspondence principle"?

It is also true that physics, when considered purely in terms of Hilbert space, is, strictly speaking, a most restricted subject. Taken literally, it would seem to imply that "describable" phenomena only came into existence after the evolution of man supplied the world with suitable "observers." The prehistoric world must have been governed by a different set of physical rules! Such an idea seems difficult to accept, but logically speaking there is nothing wrong with it. It *is* true that we have no way of observing a world without people, and therefore such a world could have strange properties.

Thus, as with Schrödinger's famous "cat paradox," we are reduced to a simple banality.

The idea of "hidden variable" theories has received much attention. The idea is that perhaps quantum mechanics is not so very different from classical mechanics after all. The advocates of such theories argue that some unobservable processes within a classical framework could explain the apparently indeterminate nature of quantum mechanics. In his book, von Neumann presented an argument that purported to show that hidden variable theories could not be valid. But his argument has itself been the subject of much criticism. Thus, a certain amount of controversy has surrounded the whole question of hidden variables for many years. This undue attention that the subject has received seems to have led to a polarization of thought on the foundations of quantum mechanics. Thus, we are confronted with two equally disagreeable alternatives: either (1) we should pursue the study of hidden variable theories, or (2) we should accept the idea that the "observer" is a mystical entity beyond the reach of scientific inquiry. The first alternative seems to have led nowhere, and in any case it would appear to be in direct conflict with the experimental evidence. The second alternative amounts to an abandonment of the scientific method itself—surely a repulsive idea to anyone interested in science! But it seems to me that this polarization of thinking on quantum mechanics is, in fact, unnecessary. Are these two alternatives the only ones really available? Could it not be that the conventional assumption that continuous (Euclidean) space should form the basis for all models in physics leads to just these two alternatives? On the other hand, if we broaden our way of thinking and allow the possibility of discrete geometries—combined with action at a distance in physics—then it seems reasonable to expect alternative explanations to become possible.

Can it be that a discrete geometrical framework would allow a more satisfying interpretation of quantum mechanics? The fact is that quantum mechanics, which is most certainly a theory of *discrete* phenomena, is described in terms of *continuous*, even differentiable, structures. On the face of it, this appears to be obviously inappropriate. Furthermore, this description has been found to be unsatisfactory by many of the great physicists—even ones who themselves contributed to the original formulation of the theory. Thus, it is most natural to consider some speculations on the possible relevance of the present discrete model for the question of the interpretation of quantum mechanics.

The goal of such an investigation should not be the overthrow of the existing theory. Surely any sensible person will accept the validity of the quantum theory *as it is applied in practice.* But by the same token, even the most enthusiastic supporter of the standard interpretation of quantum

mechanics must admit that the "observer" is forced into a strange role apparently outside conventional understanding. This problem of the observer is a blemish on the theory that one must attempt to eliminate. Thus, far from seeking a new mathematical description of quantum mechanics, my goal is simply to find a more satisfying foundation for the theory, encompassing not only the external world, but also the observer as well.

In the rest of this section I will outline a number of ideas that appear to provide a basis for such a new foundation. As with any theory, it is based on a number of hypotheses that the reader may or may not be willing to accept. But even if the specific hypotheses are not accepted, the basic motivation for our theory must be taken seriously. The fact is that even now—more than 60 years after the initial formulation of the quantum theory and its probabilistic interpretation—many theorists still profess themselves to be unhappy with the foundations. If we leave aside all of the accumulated philosophical speculation of those 60 years, the basic problem remains: namely, does there exist a mathematical model for physics that exhibits the statistical properties required by the quantum theory? The mathematical model should be complete, in the sense that it contains within itself everything to be described—including possible "observers"! Within the framework of differential geometry, nothing sensible has yet been found. What is the situation with respect to discrete geometries? Strangely enough, this subject—the investigation of the statistical properties of possible discrete geometries—has apparently never been investigated. Thus, it would be a good idea to study such discrete structures first, before pursuing further the philosophical problem of the role of the observer in the theory of quantum mechanics.

5.2. Nonlocality and the Concept of Probability

It is necessary to think carefully about the concept of probability, how it is used in quantum mechanics, and its relationship with the concept of time. My thesis is that the "classical" way of thinking about time and probability is no longer appropriate when it comes to describing quantum systems. These concepts should be considered not from a "local," but rather from a "global" point of view.

Now the idea of "time" has meaning within some specific model of physical space-time. Within the present framework, this would be some specific partially ordered set W. On the other hand, "probability" should express relationships between possible *different* models for space-time. That is, probabilities should be calculated by taking the class of all possible (and "physically meaningful") partially ordered sets. The probability P that the experiment E has the outcome O is calculated by taking all possible sets

W that happen to contain E. The fraction of these sets that exhibit the outcome O is then *defined* to be the probability of O, namely P.

Now, in classical physics one has, very typically, a problem of the following sort. Let the state of a physical system be given at a certain time. Then the question is, what will the state of the system be at some time in the future (say, in 1 min)? Perhaps many traditionalists would say that this, and nothing else, should be the proper domain of physics. Thus, the idea of "time" plays a very important part, not only in the theory itself, but it even determines what we are allowed to *think about* as belonging to the theory!

For example, probability theory, or statistics, might be used to analyze the outcomes of throws of dice or of actuarial data for an insurance firm. One might throw the dice repetitively many times onto the same table, or perhaps many people could throw a dice onto many different tables simultaneously. One thinks of this as being a great number of similar, when not identical, experiments, which are all independent of one another, occurring at different, but essentially similar times.

The important point, though, is that all of these experiments occur within the *same* universe. Thus, if one is prepared to think of things from the global point of view (and if one accepts action at a distance, then there is no alternative!), then all of these different throws of the dice are *not* independent. On the contrary, they must be thought of as being constituent parts of an unchangeable entity, namely a partially ordered set W that satisfies an appropriate variational principle. One can only begin to talk about truly independent experiments when one compares *different* possible partially ordered sets.

Clearly, two different experiments E_1 and E_2 that occur within the same set W are not independent if one of the experiments follows the other in time. That is, if $E_1 < E_2$, where $E_1, E_2 \subset W$, then E_1 is not independent of E_2. (There might be obvious causal effects here: for example, wafts of air or vibrations of the table might influence the successive throws of the dice.) But when we consider the way W is defined, it will be realized that W must satisfy a *global* variational principle, and so even if E_1 and E_2 are "simultaneous" in the sense of relativity theory, they still cannot be independent of one another.

Now it might be thought that this discussion is nothing more than obscure philosophical hair-splitting. Every sensible person knows that different throws of the dice are, for all practical purposes, independent. It might be admitted that one throw could produce small effects that might persist long enough to influence, in some way, the next throw. But we are unable to keep track of all these effects, and this *lack of knowledge* of the fine details is precisely what is needed to make probability theory work in

the classical sense. Thus, one can think of the question of the independence of repetitive experiments as being one of degree, rather than of substance; obviously, if two dice are thrown almost together, so that they are in physical contact, then the outcome of one experiment influences profoundly the outcome of the other experiment. On the other hand, if they are widely separated, then it is difficult to imagine how one throw could influence the other.

It might be thought that it is easy to decide in an intuitive way—at least in the case of the dice game—whether or not two throws are sufficiently independent of each other to allow probability theory to be invoked with reason; one uses everyday, sensible experience for this. But what about the case of quantum mechanical experiments, where much runs completely counter to the usual intuition? It is perhaps best to cite a version of Einstein, Podolski, and Rosen's (1935) famous "thought experiment" [Bell (1966) is also relevant here].

This can be formulated as follows. Two observers A and B sit facing one another at a great distance apart. At the midpoint between them is a radioactive source, which occasionally emits pairs of electrons, one traveling to A and the other to B. If A's electron has spin up, then B's has spin down, and vice versa. According to the hypotheses of quantum mechanics, the individual electrons are at first in neither the spin-up nor the spin-down state. But then A decides at some time to observe its electron, observing, say, that it has spin up. At that moment in time, the "wave function" collapses instantly, thus apparently violating the principle of relativity and forcing B's electron into the spin-down state! The paradox here is that two seemingly widely separated and independent events turn out to be not as independent as the "everyday intuition" would lead one to believe.

Certainly other people have also emphasized the importance of considering quantum mechanics in terms of global, rather than local, phenomena. But it seems to me that the action-at-a-distance theory, combined with a framework of discrete, partially ordered sets, provides a means of understanding in a simple and practical way how these seemingly paradoxical effects can come about.

5.3. A Definition of "Probability" for Quantum Mechanics

5.3.1. The Definition of Probability (Finite Case)

How should probabilities be defined in this framework? To begin with, let Ξ be the set of all possible discrete partially ordered sets satisfying the various assumptions made so far [and especially including the variational principle (47)]. Now it is unclear how many elements Ξ contains. (Our

definitions have, in any case, been vague.) Perhaps Ξ contains infinitely many elements, or perhaps only finitely many, or it may even be empty! We would like to define "probabilities" by simply counting the numbers of elements of Ξ with various properties.

Let E be an "experiment." What does that mean? For us it means that a possible universe $W \subset \Xi$ contains the experiment E if a certain agreed upon pattern of elements can be found within W. If this pattern occurs more than once in W, then we will agree to consider each occurrence of the pattern as representing a separate universe within this set. Denote the set of all possible universes containing the experiment E by $\Xi_E \subset \Xi$.

Let us say that the experiment E can have different outcomes A, B, etc. In this case we will denote by $\Xi_{E(A)} \subset \Xi_E$ the set of universes containing the experiment E with the outcome A. In the case that Ξ_E is finite, the probability of the outcome A will be defined to be $|\Xi_{E(A)}|/|\Xi_E|$ (where the notation $|X|$ denotes the number of elements in the set X).

5.3.2. *Probabilities (Infinite Case)*

If Ξ_E is infinite, then things are more difficult. One way to proceed is to consider neighborhoods of E. Let the set $W \subset \Xi_E$, and take this to mean that there exists a finite subset $E \in W$ that represents the given experiment. How can the idea of "finite subsets" in strongly discrete, partially ordered sets be defined? We will say that for each $n \in \mathbf{N}$, the *ball of radius n with center E* is the set of points $a \in W$ such that there exists some $b \in E$ with $|W_b - W_a| + |W_a - W_b| \leq n$. (Here $W_a = \{u \in W: u \leq a\}$.) Denote this ball of radius n by $B_n(W, E)$. It seems reasonable to assume that $B_n(W, E)$ is finite and of limited size [that is, there exists an $m(n) \in \mathbf{N}$ such that $|B_n(W, E)| \leq m(n)$ for all $W \in \Xi_E$]. This could be taken to be part of the definition of E (i.e., an *experiment* will not be allowed to have some arbitrarily large gravitating body in the immediate neighborhood). For fixed $n \in \mathbf{N}$ we say that the sets W, $W^* \in \Xi_E$ are *n-equivalent* if there exists a one-to-one correspondence $B_n(W, E) \leftrightarrow B_n(W^*, E)$ that preserves order. This is an equivalence relation, and for each n, the number of equivalence classes must be finite.

Now let A be some possible outcome of the experiment E. For each $n \in \mathbf{N}$, the *probability of A of order n* is defined to be the ratio of the number of n-equivalence classes of E that have the outcome A to the total number of n-equivalence classes of E. Finally, the *probability of A* is defined to be the limit of the probability of A of order n as $n \to \infty$, if it exists. If the experiment E is such that all possible outcomes A have probabilities in this sense, then E will be called a *proper experiment*. In the sequel it will be assumed that all experiments under consideration are proper experiments.

This definition leads certainly to difficult, or even impossible to prove, conjectures concerning its relevance in physics. Are the usual experiments dealt with in most physics textbooks "proper" experiments or not? This is obviously a question of such complexity as to defy all hopes of answering it. But such a situation is not unusual in theoretical physics. For example, it is often claimed that classical mechanics represents a kind of "limiting behavior" of quantum mechanics as Planck's constant is allowed to approach zero. But what hope is there of ever proving this assertion: that is, of deriving the basic definitions of classical mechanics from the axioms of quantum field theory?

5.4. The Two-Slit Interference Experiment

The traditional two-slit interference "thought experiment" consists of the apparatus depicted in Figure 8. Particles—for example, electrons—are emitted from a pointlike source and begin to travel through the apparatus in the direction of an absorbing screen (e.g., a photographic plate). Most particles collide with an obstruction set up between the source and the screen and are thus lost. But some particles find their way through two narrow slits cut into the obstruction. These particles travel on to the screen and make a mark there. The source is allowed to emit many particles, and eventually a pattern emerges on the screen. The pattern shows an interference effect in the *statistics* of the particles, as if waves had gone through the two slits and produced interference with themselves. But the curious thing is that these are waves of *probability*!

How can we explain this? According to Feynman *et al.* (1965), this experiment illustrates the single, inexplicable, and essential mystery of quantum mechanics. We are doubly interested in an explanation, since the Feynman path integrals, which arguably provide a basis for understanding all quantum theory, are concerned with the analysis of idealized experiments like these.

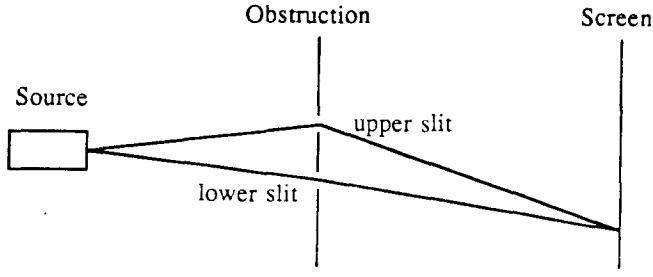

Fig. 8

Now, the first step is to stop thinking of the two-slit experiment as consisting of a large number of electrons passing through the apparatus, one after the other, to produce an averaged interference pattern on the screen. Instead, according to the present way of thinking, a single experiment consists of the passage of a *single* electron from the source to the screen in a single apparatus. The object of the experiment is to find the relative probabilities for the arrival of an electron at different points of the screen. We therefore choose two different points P and Q on the screen and compare the probabilities for the arrival of the electron at P and Q (Figure 9). Point P is chosen to give constructive interference in the sense of the quantum mechanical waves. Q is a point of destructive interference. P is such that the difference of the (three-dimensional) path lengths for the straightest possible paths from the source to P through the upper and lower slits is an exact multiple of the de Broglie wavelength for the electron. Q is such that this difference is an exact multiple plus one-half de Broglie wavelength.

Thus, it can be imagined that there are three classes of possible universes (or, in our terminology, discrete, partially ordered sets) to be compared here, namely the class Ξ_E of all possible sets that contain the two-slit experiment E, and also $\Xi_{E(P)}$ and $\Xi_{E(Q)}$, which are defined to be subsets of Ξ_E such that the electron lands at P and Q, respectively.

Now, this experiment is formulated in terms of three-dimensional Euclidean space plus time, whereas our partially ordered sets W have only been defined in terms of the four dimensional space $\mathbf{R}^4$ together with the Lorentz metric. Therefore, it is necessary to think about the experiment, and in particular the significance of the de Broglie wavelength, in terms of a relativistic formulation.

The most appropriate way of doing this is to use the path integral formalism (Feynman and Hibbs, 1965). I shall go into this in more detail in the sequel, but for the moment the relationship between the path integrals and the de Broglie wavelength can be thought of as follows. Let $W \in \Xi_{E(P)}$

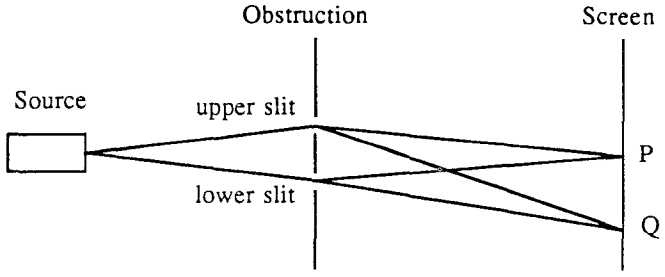

Fig. 9

and let $\gamma \in W$ be the path that corresponds to the electron in the experiment. In particular, assume that W satisfies the Fokker condition (47). For this specific W we have γ passing either through the upper or the lower slit. Let us say, for the sake of argument, that γ passes through the upper slit. Then we can alter W slightly to give a new partially ordered set $W_{\neq}$. The set $W_{\neq}$ is identical to W except that γ has been changed slightly to give the path $\gamma_{\neq}$. The path $\gamma_{\neq}$ in turn is identical to γ except for the section between the source and P. For this section, $\gamma_{\neq}$ takes the most direct path through the *lower* slit. Now, as is shown in Feynman and Hibbs (1965), the condition on the de Broglie wavelength is equivalent to the condition that

$$\int [\gamma(v)v' - \gamma_{\neq}(v)v'] \, ds = nh/2\pi$$

where $n \in \mathbf{Z}$, and h is Planck's constant. The point Q in the experiment is such that, instead of n, we would take $n+1/2$.

Of course I do not claim that the new set $W_{\neq}$ also satisfies (47). In general, one expects that it will not. On the other hand, there is *some* chance that (47) might be satisfied by $W_{\neq}$, and this will form the basis for further arguments.

5.5. The Concept of Clusters

To proceed further with the analysis of the two-slit interference experiment, it is necessary to introduce a new idea, namely the idea of "clusters" of sets. If one considers the set of all possible universes Ξ, then it is reasonable to imagine that some pairs of elements of Ξ might be similar to one another, while other pairs might be very different. One could try to describe this situation by looking for some appropriate definition of a "distance" between different elements of Ξ, thus making Ξ into an abstract metric space. But this would involve introducing more details than are really necessary. The important question is, given two elements W, $W' \in \Xi$, is W similar to W'?

Let W, W' be two elements of Ξ. I will say that W is *equivalent to* W' if there exist finite subsets $V \subset W$ and $V' \subset W'$ such that there is a one-to-one correspondence between $(W - V)$ and $(W' - V')$ that preserves order. That is, W and W' are equivalent if they are the same, except perhaps for some finite subsets. This is obviously an equivalence relation.

Definition 1. The equivalence classes of Ξ under this equivalence relation are called *clusters.*

Unfortunately, the idea of a cluster appears to be as difficult as it is important. For example, the question might be posed, do there exist any clusters with more than one element? I will in fact proceed on the assumption

that most clusters contain many elements; this assumption expresses some of the relationships I had in mind when formulating the framework of discrete, partially ordered sets. But how could one possibly prove anything concerning the expected number of elements in an average cluster? Clearly a great many simplifying assumptions would be necessary before any progress could be made in that direction.

If we now return to the classical, continuous case, namely Fokker's model for electrodynamics, then each "cluster" can contain only one element. (See Theorem 5.1, below.) Here Definition 1 must be changed to deal with this continuous case: one posssibility for doing this is to say that W and W' are in the same cluster if they differ at most in a *compact* subset. Theorem 5.1 shows that in the continuous case a localized perturbation must, if Fokker's variational principle is to be satisfied, also result in a perturbation that grows throughout the set.

The discrete case seems to be different. One can again have small perturbations, but now they must respect the discretization. This could perhaps best be thought of as being a variation with constraints: namely, if γ is a path and $\gamma_{\neq}$ is a variation of this path between two points, then in order to have the paths remaining proper (see Section 3.7), it is necessary that the arc length of the varied segment of $\gamma_{\neq}$ differ from the arc length of γ along this segment by an exact multiple of some finite discretization constant. Certainly there may be other statistical effects that make the discrete case different from the continuous case. But it seems reasonable to assert that clusters in Ξ have, in general, more than one element.

To summarize, then I assume that each cluster can be thought of as representing a single *classical* path, and thus it represents the result of a single experiment, considered in the classical physics of the continuum. On the other hand, I will argue that the strange new quantum statistical phenomena can be explained by assuming that some clusters in the *discrete* version of classical physics may contain many elements. One therefore has the correspondence

paths in the usual classical physics $\Leftrightarrow$ *clusters* in discrete physics

Theorem 5.1. Let Γ, Γ' be two infinite sets of sufficiently smooth, nonintersecting paths in $\mathbf{R}^4$ such that any compact region of $\mathbf{R}^4$ meets only finitely many elements of Γ and Γ'. Assume, further, that the assumptions of Section 2.6 hold for both Γ and Γ'. Let $B \subset \mathbf{R}^4$ be a bounded set such that only finitely many paths meet B, and furthermore Γ coincides with Γ' in $\mathbf{R}^4 - B$. On the other hand, Γ differs from Γ' in B. Then at least one of the sets Γ, Γ' does not satisfy Fokker's principle (14).

Proof. Clearly Γ and Γ' must contain the same number of particles in B. Subject to a certain restriction, it is possible to find a particle γ, not

meeting B, and a point $\gamma(t)$ on γ such that no points diagonally above $\gamma(t)$ are in B, and for only a single particle in Γ the contribution to F_{ret} at $\gamma(t)$ is different from that of the corresponding particle in Γ'. But for (14) to hold, we must then have γ at $\gamma(t)$ in Γ differing from the corresponding γ in Γ', a contradiction that proves the theorem. The exception is that there exist two (or more) particles ξ_1, $\xi_2 \in \Gamma$ meeting B and points $\xi_i(r_i)$ of ξ_i, $i = 1, 2$, that are the highest points such that ξ_i coincides with the corresponding particle in Γ' below $\xi_i(r_i)$, with the property that all of the points on the particles, not meeting B, that are diagonally above $\xi_1(r_1)$ fall on a straight line. Furthermore, $\xi_2(r_2)$ is also on this line. In this case, although the contribution of ξ_i to F_{ret} immediately above $\xi_i(r_i)$ differs for Γ and Γ', $i = 1$, 2, these differences could cancel one another at some points diagonally above those points of ξ_i. These points of cancellation may happen to lie on particles that do not meet B, thus invalidating the original argument. But such can only be the case for finitely many of those particles, and thus it is possible to find a particle γ not meeting B such that the total contribution to F is different at some point $\gamma(t)$ on γ for Γ and Γ', thus producing the same contradiction as before. ■

This theorem seems to justify the idea that each classical cluster contains only a single element. But it does not imply that for each classical universe there is a corresponding discrete cluster. On the contrary, one expects classical universes that correspond to no discrete clusters, and also discrete clusters that correspond to no classical universes, but examples of such sets—together with the appropriate proofs—must be very difficult to find.

5.6. An Explanation of the Two-Slit Experiment

Next I attempt to provide an explanation of the two-slit interference experiment in terms of clusters and the definition of probability in Section 5.3. In the notation of Section 5.4, two possible outcomes P and Q of the experiment E are considered. As far as classical physics is concerned, both P and Q are equally likely. If one accepts the reasoning of Section 5.5, then this would mean that in the discrete framework the number of clusters associated with P is the same as the number of clusters associated with Q. Half of the clusters in each case are associated with paths going through the upper slit and half through the lower slit. Thus, we retain the classical picture: our model contains geometric paths representing the paths of physical particles. Furthermore, these paths either go through the upper or the lower slit, just as classical physics suggests they should. The only difference is that we associate a *cluster* with each classical path.

Is this association of clusters with classical paths and the method of counting the clusters reasonable? The symmetries in the experiment seem

to suggest that they are. But any possible *proof* would certainly depend on a much sharper formulation than I have yet attempted.

However, if one is prepared to carry the reasoning further, then, by construction, the clusters associated with the outcome P can be expected to contain more elements than the clusters associated with Q. In the first case, the particle can experience a local perturbation taking it through the upper or lower slit. In the second case, the particle has only one choice, since the other would involve a perturbation giving a change in arc length of the particle not equal to an integral multiple of the discretization constant. That is, after experiencing such a local perturbation, the particle would not be able to get back "into phase" with itself and continue along as a discrete particle in an admissible set W' that is a finite variation of the original set W.

Therefore, to summarize the situation: while the number of clusters associated with the outcome P is the same as the number of clusters associated with Q, the P-clusters contain more elements than do the Q-clusters. Thus, according to the definition of relative probabilities in Section 5.3, the probability of P is greater than the probability of Q, in agreement with quantum mechanics.

As a final point, while discussing the two-slit interference experiment, it may be interesting to mention the Bohm-Aharonov effect. This is a modified version of the two-slit experiment, which has actually been performed. Behind the two slits, parallel to and between them, a long, thin solenoid is so placed that if the electron goes through the upper slit, then it must pass above the solenoid; if it goes through the lower slit, then it must pass below. According to the ideas of classical physics, the net charge of the solenoid is zero—thus, there is no electrical field. There is a nontrivial magnetic field inside the solenoid, but, assuming that the solenoid is sufficiently long, the field outside must be null. Neither of the two paths available to the electron passes through the solenoid, so it would appear that the electron could not be subjected to any new influences. However, in this instance the classical wisdom breaks down; it is found that the observed interference pattern shifts on the screen. This is thought to be a good demonstration of the fact that classical and quantum physics are indeed very different. Esoteric ideas from algebraic topology are often brought into the analysis to help explain the effect.

How does this fit in with the present seemingly classical explanation of the two slit interference experiment? If the present reasoning is followed, will it also predict no shift of the pattern, in conflict with the experimental evidence and the precepts of quantum mechanics? To answer this question, it is necessary to look further into the calculation associated with the Bohm-Aharonov effect. While it is true that the magnetic field outside the

solenoid is null, nevertheless the classical vector potential A is nonzero both inside *and* outside the solenoid. Now, the present theory is based on Fokker's theory, which uses the variational principle expressed in (14). As seen in (22), the electromagnetic part of this formula involves A directly, not the derived quantities: the electrical and magnetic fields. Thus, the evaluation function J_Γ is in fact very much different when the solenoid is introduced. We see then that the Bohm-Aharonov effect can be thought of as being a *classical* effect; it simply shows that the classical vector potential can be directly observed in interference experiments.

5.7. A Theorem Concerning Complex Numbers

Whether one is prepared to accept the reasoning in the previous subsection or not, an obvious question remains. How can these rather vague estimates of the sizes of certain sets be reconciled with the very exact formulas of quantum mechanics? The complex-valued exponential function plays a central role.

Now, if one accepts the reasoning of Alfred Landé (1965), then it could be concluded that we are already completely finished! Our discretization hypothesis gives, in some philosophical sense, "Duane's Third Rule of Quantum Mechanics," as Landé calls it, and therefore, following his reasoning one arrives by default at the whole quantum theory with its probability "amplitudes" and so on. But rather than taking this drastic step, I prefer to prove a simple result that might have some bearing on the question of quantum probabilities.

Theorem 5.2. Let $u:[-\pi, \pi)\to[-1,+1]$ be a continuous function. Assume that $u(\vartheta)=u(-\vartheta)$ for all $\vartheta\in[-\pi, \pi)$, and that $u(0)=1$, $u(\pi)=-1$, and u is monotone in the domain $[0, \pi)$. Assume also that if $J=\{j_1,\ldots,j_{2n}\}$ is some finite set of numbers with $j_i\in[-\pi, \pi)$ for all $i=1,\ldots,2n$ such that (1) $j_i=-j_{i+n}$ for each $i=1,\ldots,n$ and (2) if $\sum u(j_i)=0$, then $\sum u(j_i+\varphi)=0$ for any "phase angle" $\varphi\in[-\pi, \pi)$. If these conditions are fulfilled, then $u(\vartheta)=\cos\vartheta$.

Proof. To begin with, it is easy to see that the cosine function satisfies the conditions of the theorem. Just use the standard trigonometric formula

$$\cos(\alpha+\beta)+\cos(\alpha-\beta)=2\cos\alpha\cos\beta$$

Thus, if $\sum\cos j_i=0$, then

$$\sum\cos(j_i+\varphi)=\sum\cos(j_i+\varphi)+\sum\cos(-j_{i+n}+\varphi)$$

where the sum on the right-hand side is from 1 to n. But this is then

$$[\sum\cos(j_i)]\cos\varphi=0\times\cos\varphi=0$$

Is cosine the only function satisfying these conditions? I will show that the function u is uniquely determined at all points in $[-\pi, \pi)$ of the form $n\pi/2^m$, and thus the uniqueness of the function must follow. As a first step, we note that u corresponds, by definition, with the cosine function at the points 0 and π. Furthermore, one has $0 = 2u(0) + 2u(\pi) = 2u(\vartheta) + 2u(\vartheta + \pi)$ for all $\vartheta \in [-\pi, \pi)$, and thus $u(\vartheta) = -u(\pi + \vartheta)$. It follows that $u(\pi/2) = 0$. Now choose some small $\varepsilon > 0$, and choose the set J in the following manner:

$$J = \{j_1, \ldots, j_s, j_{s+1}, \ldots, j_{s+t}, \ldots, j_{2s+2t}\}$$

Here $j_i = 3\pi/4$, $i = 1, \ldots, s$, and $j_i = 0$, $i = s+1, \ldots, s+t$. Since u is monotone, $u(j_1) < 0$. The numbers s and t are chosen so that $|u(j_1) + t/2| < \varepsilon/2s$. That is,

$$|su(3\pi/4) + su(-3\pi/4) + 2tu(0)| < \varepsilon$$

or, using a somewhat more suggestive notation,

$$su(3\pi/4) + su(-3\pi/4) + 2tu(0) \approx 0$$

so that $u(3\pi/4) \approx -t/s$. But we may add the phase angle $\pi/4$ to all the j_i and thus we obtain $u(\pi/4) \approx s/2t$. It follows that $s/t \approx 2^{1/2}$, so that $u(\pi/4) \approx 2^{-1/2} = \cos(\pi/4)$. We can choose ε arbitrarily near to 0, and thus we conclude that $u = \cos$ for the numbers $\pm\pi/4$, and $\pm 3\pi/4$. In particular, u has been determined at these points by its value at the four known points 0, $\pm\pi/2$, and π.

This trick can be carried a step further. Let

$$J = \{j_1, \ldots, j_s, j_{s+1}, \ldots, j_{s+t}, \ldots, j_{2s+2t}\}$$

be such that $j_i = 7\pi/8$, $i = 1, \ldots, s$, and $j_i = 0$, $i = s+1, \ldots, s+t$. Assume that s and t are chosen so that

$$su(7\pi/8) + su(-7\pi/8) + 2tu(0) \approx 0$$

Thus, $u(7\pi/8) \approx -t/s$. Now add $\pi/8$ to obtain $u(\pi/8) \approx s(1 + 2^{-1/2})/2t$. Therefore,

$$t/s \approx [(1 + 2^{-1/2})/2]^{1/2} = \cos(\pi/8)$$

and, as before, we conclude that $u = \cos$ for the values $\pm\pi/8$ and $\pm 7\pi/8$, where only the previously known values of u were used in the argument. The values $\pm 3\pi/8$ and $\pm 5\pi/8$ can be also be checked if one starts with the set J such that $j_i = 5\pi/8$, $i = 1, \ldots, s$. Clearly, this method can be extended to include all $\vartheta \in [-\pi, \pi)$ that can be expressed in the form $\vartheta = n\pi/2^m$ for suitable integers n and m. But such numbers are dense in $[-\pi, \pi)$, and therefore $u = \cos$. ■

Does Theorem 5.2 have any relevance in our attempts at interpreting the two-slit interference experiment? One could argue as follows. The

problem is to determine the sizes of the various clusters associated with the experiment. As we have seen, the important idea is that the clusters will contain *many* elements if the various possible paths in the cluster have lengths that differ from one another by an integral number of de Broglie wavelengths. On the other hand, a cluster will contain *few* elements if the possible lengths vary by "random phases" with respect to the de Broglie wavelength. Thus, one could assign to each possible classical path γ_i in the experiment a "phase angle" $\vartheta_i \in [-\pi, \pi)$, where $i \in K$ and K is some index set. Given some specific classical path γ_i representing a specific outcome of the experiment, its contribution to the total probability of the outcome is determined by the number of other possible paths that have nearly the same phase angle as γ_i. On the other hand, the presence of a path that is completely out of phase with γ_i does not enhance the probability of that outcome. On the contrary, it has a negative effect, since we are assuming that the total number of classical paths is the same for all possible outcomes.

Thus, the only things of importance are the differences between the phase angles for different pairs of classical paths. That is, given two paths γ_i and γ_j, then the question of whether or not they will tend to enhance one another, i.e., increase the sizes of the clusters associated with the experiment, is determined by the differences in the phase angles $\vartheta_i - \vartheta_j$. We can denote the value of this enhancement by some function $u^*: [-\pi, \pi) \to \mathbf{R}$.

It seems reasonable to assume that u^* is continuous and that $u^*(\vartheta) = u^*(-\vartheta)$ by an obvious symmetry. Thus, the image of $[-\pi, \pi)$ under u^* is compact, and therefore u^* assumes a maximum and a minimum. It is reasonable to assert that $u^*(0)$ is the maximum and $u^*(\pi)$ is the minimum, and u^* declines monotonically between 0 and π. Now we can simply find two constants c_1 and c_2 such that $u^* = c_1 + c_2 u$, where u is a function u: $[-\pi, \pi) \to [-1, +1]$ with $u(0) = 1$ and $u(\pi) = -1$.

Let $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ be some "random" set of classical paths for a certain outcome R of the experiment E (by random, I mean that the phase angles are widely distributed). Denote by $\Xi_{E(\Gamma)}$ the set of possible universes that are such that the particle follows one of the classical paths in Γ through the experiment. Thus, $\Xi_{E(\Gamma)} \subset \Xi_{E(R)} \subset \Xi_E$. Since Γ is random, we might assert that

$$\sum_{i,j=1}^{m} u(\vartheta^i - \vartheta^j) = 0$$

Now let some other classical path $\gamma^* \notin \Gamma$ be given. The randomness of Γ would imply that $\sum u(\vartheta_i - \Theta) = 0$, where Θ is the phase angle of γ^*.

Thus, if one is prepared to accept these assumptions, then the relevance of Theorem 5.2 will also follow. The total probability of an outcome will

be determined by

$$\sum_{i,j\in K} u(\vartheta^i - \vartheta^j)$$

where the index set K includes all possible classical paths leading to a given outcome. We can write

$$u(\vartheta_i - \vartheta_j) + u(\vartheta_j - \vartheta_j) = 2\cos(\vartheta_i - \vartheta_j) = e^{i(\vartheta_i - \vartheta_j)} + e^{i(\vartheta_j - \vartheta_i)}$$

and thus the connection with complex numbers becomes apparent.

5.8. Observers, States, Uncertainty Relations

It is interesting to carry the reasoning further, and to think about how some of the standard, apparently paradoxical, rules of quantum mechanics could be explained in terms of discrete spaces.

For example, it is often stated that one can alter the two-slit interference experiment by introducing explicitly the concept of an observer. One can imagine that light is directed at the two slits, and the observer attempts to see if the particle actually did pass through the upper or the lower slit. But as soon as these attempts become successful, the quantum mechanical interference effects are destroyed! Thus, it would appear that, in some strange way, Nature succeeds in "hiding" the details of quantum mechanics, thereby frustrating the aspirations of an inquiring humanity. This is customarily explained in terms of the *principle of uncertainty.* It has also been asserted that the will of the observer is imposed on the system, forcing the wave function to "collapse" and the system to enter a given state.

Now, if one is prepared to accept the definition of relative probabilities given in Section 5.3, then it seems to be possible to do away with such complicated explanations. It is no longer necessary to philosophize about the relationship of mind and matter. Remember that, for us, each trial of the two-slit experiment represents a possible partially ordered set W. If we observe that the particle definitely went through one of the slits, say the upper one, then this is part of the *definition* of the experiment. Thus, to calculate the probabilities for the experiment, it is necessary to count up the set of all possible universes that contain an observer who sees that the particle is going through the upper slit. The set of these universes contains, by definition, only experiments in which the particle passes through the upper slit. It is hardly surprising that the statistics for *this particular* experiment are different from the statistics of the other experiment in which a similar observer fails to detect whether the particle passed through the upper or the lower slit.

I could go further into such matters, but surely it is obvious that, for example, an "uncertainty principle" must be associated with any discrete

model in physics—the uncertainty varying with the scale of the discretization. Rather than pursuing these questions, it seems best to refer simply to Landé (1965), where such criticisms are presented in great abundance.

5.9. Other Quantum Mechanical Experiments

I have still failed to account for many of the phenomena often thought of as being central to the "essence" of quantum mechanics: for example, the use of spin matrices. After all, the phenomenon of "spin" is considered to be intimately connected with the geometric structure (the group of symmetries) of Minkowski space. In fact, the standard equations of quantum mechanics (Klein-Gordon, Dirac, etc.) are generally thought of as being nothing more than simple consequences of this group of symmetries. Now, in principle, it would not seem to be appropriate simply to assume that the same symmetries hold in our discrete spaces. After all, Minkowski space—a perfectly homogeneous space, and thus empty, according to the general theory of relativity—can at best be used to calculate the behavior of a single quantum mechanical system in an otherwise empty, and thus highly symmetric, universe. As we have already seen, this abstract idea of "emptiness" cannot be sensibly carried over to the discrete description. But this should not deter us from expecting that the quantum theory—with it's "spinors" and so forth—can also be applied within the framework of discrete spaces. After all, even in the usual nonsymmetric, continuous geometric framework (the considerations of cosmology alone invalidate many of the symmetries) it is considered to be sensible to continue to apply the standard equations of quantum field theory. Certainly the general framework of discrete spaces allows enough latitude for many such developments. But for the present it may be worthwhile to speculate briefly and simply on one or two very typical quantum mechanical experiments.

5.9.1. Atomic Structure

Perhaps the most basic, and historically the most important, "experiment" was the structure of the atom. The attempts to understand how the negatively charged electron can be stably bound to the positively charged proton in a hydrogen atom, giving very sharply defined energy levels, led to the original formulation of quantum mechanics. It is possible to refer to the method of reconstructing the Schrödinger equation from the path integrals, and thus to claim that the atomic structure can be explained using the ideas of Section 5.6. But another effect may become important, depending on the details of the way we choose to define our discrete model. If it is imagined that the discretization is such that points along the particles occur only once every de Broglie wavelength, then we no longer have an

electron "circling" a proton, but rather an electron that appears only sporadically. If these sporadic appearances form a static pattern, then it might be expected that the atom is stable; if, on the other hand, the pattern tends to move (rotating, etc.), then this will result in radiation of energy from the system, just as would be expected in classical electrodynamics.

When thinking about this effect, it might be remarked that it is not necessary to adhere rigidly to the definition of "distance" as given in Section 3.7. Here is a possible alternative definition.

Definition 1. Let $P \subset W$, and p_i, p_{i+1} be as in Definition 1 of Section 3.7. The *distance* between p_i and p_{i+1} can now be defined as the largest number n such that there exists a chain of positions $p_i = C_0 < C_1 < \cdots < C_n = p_{i+1}$. Then an alternative to Definition 1 of Section 3.7 is that P is *proper* if all adjacent elements of P are the same distance n apart.

Now one might define a particle to include not only the elements $P = \{p_i\} \subset W$, but also the positions along such maximal chains. Thus, we could, even with a seemingly coarse discretization, still have a very fine structure. However, it hardly seems worthwhile to pursue such further speculations here.

5.9.2. Light

According to the theory of quantum electrodynamics, the propagation of photons of light should be dealt with similarly to the propagation of "normal" massive particles. But in the action-at-a-distance formulation of classical electrodynamics, the concept of "light" simply does not exist! How can these two completely contradictory viewpoints possibly lead to similar theories?

At best we can say that "light" expresses some relationships between different particle paths in the model. Electromagnetic waves arise in a complicated way, which can be seen in the correspondence between the Maxwell theory and the action-at-a-distance theory, which is demonstrated in Section 2. Now, the present ideas are certainly classical, and so it would follow that we should also try to explain "light" purely in terms of the Fokker model. The "wave" properties of light have perhaps been demonstrated, at least through the classical correspondence between the Fokker and Maxwell theories. What about the "particle" properties in the so-called "wave-particle duality"?

Imagine an atom in an excited state losing energy and dropping down to its ground state, thus emitting a photon of light. Now, according to the classical picture, the light waves would gradually spread throughout space in a spherical pattern, becoming weaker and weaker. They might eventually be completely absorbed by extremely tiny movements of the many particles

in the future absorbing universe. Thus, the original pulse of energy will become dissipated and absorbed by many other particles: the universe "runs down," so to speak.

The quantum picture appears to be completely different. Here we have the pulse of energy remaining whole, although, in some statistical way, wavelike. It travels randomly into the future universe, and then suddenly it collides with a single particle, thus annihilating itself and giving *all* its energy to the new particle.

But are these two pictures really so different? The problem is that in the classical picture the energy is supposed to become dispersed and dissipated among many particles. In the Maxwell theory this is certainly the case, but is it also true in Fokker's action-at-a-distance theory? It seems that even in the completely classical (i.e., continuous, not discrete) case, Fokker's theory might predict photons. For example, Hoyle and Narlikar (1974) show how spontaneous atomic transitions might be explained in terms of effects in the future absorbing universe. That is to say, it might be sensible to weaken some of the assumptions in Section 2.9 to allow possible small advanced electromagnetic fields. These might be explained in the following way. Imagine first that the energy given off by the atom is distributed among many other particles in the future universe. This situation gives an extremum to Fokker's expression (14). Assume, for the sake of argument, that the energy is distributed among the particles $p_1, p_2, \ldots, p_n$. Now, according to Section 2.7, the principle of conservation of energy holds also within the theory of action at a distance. One might imagine a small variation of the particle paths, giving, say, p_1 slightly more energy and p_2 slightly less, or vice versa. Assuming that p_1 and p_2 are widely separated—and therefore "independent" of one another—then it would seem plausible to assert that such variations should change the value of (14) more or less linearly, and thus assigning slightly more energy to one particle than another would lead to a "better" value for (14). This is similar to the discussion in Section 2.8, and we conclude that it may be possible that coherent pulses—corresponding to photons of light—resulting from "kinks" in the paths of typical solutions to (14) could arise naturally even in the classical Fokker theory.

5.10. Momentum and Energy As Operators

Much of the formalism of quantum mechanics follows by the standard quantization procedure: one begins with a formula from classical physics, then partial differential operators—multiplied with constants that are purely imaginary numbers—are substituted for the momentum and energy terms in the formula. There results an equation between complex differential

operators. Solutions to the equation—smooth complex functions—are then the "probability amplitudes."

Now it would seem that this formulation in terms of complex differential operators is, on the face of it, irreconcilable with a discrete description. One may attempt to approximate a differential structure with a sufficiently fine discrete space, or one could even contemplate retreating to a classical probability space, thus mixing discrete and continuous mathematics. But such methods seem to be inappropriate. Why seek a discrete description of the physical world in the first place? Surely the reason must be that the continuous description is unsatisfactory, and thus it is counterproductive to look for a complicated discrete theory that is, in the end, simply equivalent to the original continuous description.

The way out of this dilemma is to base things on the Feynman path-integral approach to quantum mechanics. Feynman and Hibbs (1965) show how the operator formulation of quantum mechanics can be derived from path integrals. Thus, it would be possible to generalize our explanation of the two-slit interference experiment to include all possible path integrals. Indeed, modern treatments of quantum field theory, going beyond the established ideas of quantum electrodynamics, are increasingly expressed within the "language" of the path integral. I will deal further with these questions in the next section. But for the moment, where we are concerned with questions of interpretation, it is appropriate to consider just how the path integrals could arise in a discrete framework.

Recall that, in principle at least, the path-integral approach is fully equivalent to the operator approach. One obtains the same functions in each case, representing the probability amplitudes for quantum mechanical processes. But the point is that *in practice* these probability amplitudes are difficult or impossible to calculate. The path-integral approach involves finding approximate solutions, which consist of sums over various types of "Feynman diagrams." Each given Feynman diagram—while arising from a certain method of thinking analogous to the Taylor expansion in mathematical analysis—has the appearance of a definite physical process in the normal Minkowski space. For example, the photon self-energy diagram in quantum electrodynamics involves a photon arriving at a given point in $\mathbf{R}^4$, creating an electron-positron pair there, which later collapses to create a further photon at some different point in $\mathbf{R}^4$.

How seriously is one to take the physical process that corresponds with a given Feynman diagram? The conventional attitude in mathematical physics seems to be that the Feynman diagrams have no real relevance to the problems of quantum field theory; although they are useful for practical calculations, they have little to do with possible solutions to the operator equations, which are the essence of the *exact* theory. But it seems that a

more pragmatic approach is called for. The Feynman diagrams do, in fact, give a good description of physical reality. Furthermore, we can plausibly interpret the calculations that result from the use of Feynman diagrams in terms of our discrete spaces; the operator approach seems to defy all attempts at a similar interpretation.

According to the ideas presented here, the statistical effects of quantum mechanics should be thought of as arising from the properties of a large ensemble of possible "universes" Ξ, which is the collection of all possible discrete partially ordered sets satisfying the various axioms given here. Now, I have, admittedly, specified that "particles" should be infinitely long. This axiom, along with the others as well, no doubt would have to be changed to allow short "segments" of particles that occur in creation and annihilation diagrams. But still, the idea should be clear. As in the case of the two-slit experiment, different possible "paths"—representing definite elements of Ξ—all contribute to building the probability of the experiment as a whole. The difference in the general case is that rather than just considering different paths for a single electron, one now considers all possible Feynman diagrams. Thus, each Feynman diagram represents, in the present model, a whole class of possible universes contained in the total set Ξ. A diagram such as the photon self-energy diagram results in a physical process that is indistinguishable from the same process, but without the pair creation and annihilation loop. Thus, according to the ideas of Section 5.5, both of these processes occur in different universes that belong to a common cluster. In this way one will find that the probabilities for the different outcomes of a given experiment can only be calculated by summing over all possible Feynman diagrams.

6. DISCRETE SPACES IN QUANTUM ELECTRODYNAMICS

6.1. The Role of Renormalization Theory

Undoubtedly the main reason for considering discrete structures in physics is that there are technical problems in certain field theories; some of the theories that have been proposed do not seem to admit a "renormalization theory" similar to what is used in quantum electrodynamics. In such cases it has often been found possible to impose a renormalization by "brute force," using the technique of replacing $\mathbf{R}^4$ by $\mathbf{Z}^4$. Now, it seems that much more can be expected of a discrete formulation. Nevertheless this should not deter us from asking whether or not our discrete framework does in fact allow the solutions to such problems.

For example, the Feynman integral involves a sum over sets of possible particle paths. This has created, in the framework of continuous spaces,

immense mathematical difficulties. But in our discrete spaces it seems obvious that one can, in principle at least, reduce things to a finite sum, and thus the Feynman integral could regain the simple and intuitive motivation it originally had.

Now there exists *in fact* a version of renormalization theory for quantum electrodynamics based on a discrete action-at-a-distance model. This theory seems to me to be elegant and perhaps capable of more general application. Thus in this section I shall briefly describe the theory, following the development in Hoyle and Narlikar (1974) and referring to Feynman (1962). I will only sketch the theory to a sufficient extent to show the possible relevance of Observation 1, below. I presume that this observation is equally relevant to other possible renormalization theories (those involving Yang-Mills theories, etc.).

6.2. Feynman Integrals

Let $a, b \in \mathbf{R}^4$. Assume that a particle passes through the point a. (Again, particles are paths in $\mathbf{R}^4$, but now the paths are not necessarily strictly timelike.) What is the probability that the particle (or at least some particle indistinguishable from the original particle) also passes through b? The Feynman integral provides the answer, and thus it can be considered as giving a kind of conditional probability for problems of this type.

The basic idea for path integrals was apparently first proposed by Dirac. He posed the question, why are variational principles so important in physics? The models used in physics are often of the following form. Some mathematical system is given, and the system can be described in terms of a certain class of possible "paths" $\mathscr{P}$. Some evaluation function $S: \mathscr{P} \to \mathbf{R}$ is given for this set of paths, and one specifies that the *actual* path taken is an extremum with respect to this function.

Dirac's idea was to suggest that this should be translated directly into the framework of the quantum theory, as it was then understood. Rather than taking the real function $S: \mathscr{P} \to \mathbf{R}$, one should take a complex function $\Psi: \mathscr{P} \to \mathbf{C}$. The most logical choice seemed to be simply $\Psi = \exp(2\pi i S/h)$, where h is Planck's constant. Now, according to the quantum philosophy, the result of the "experiment" involving the points a and b is no longer the assertion that a particle followed some given path from a to b. Rather, an experiment simply determines whether or not a particle appeared at a and a similar particle also appeared at b. The "probability" of this occurrence is calculated to be

$$\sum_{\gamma \in \{\text{all paths from } a \text{ to } b\}} \exp\left(\frac{2\pi i S(\gamma)}{h}\right) \tag{73}$$

Each of these terms is a "probability amplitude" for a given path, and thus

the "probability" has something to do with the set of all possible paths, rather than individual paths, as is the case in classical physics. Why does the "classical" path, representing the solution to the variational problem, give a result with high probability? The answer is that in the neighborhood of an extremum (given that S is reasonably smooth, etc.), many paths have almost the same value under the function S. Thus, under the complex exponential, these paths tend to add up to give a large sum.

Of course in the usual geometry of $\mathbf{R}^4$ there are uncountably many possible paths, so it is not clear how one should define the sum in this expression. It is possible to replace the summation symbol with an integration symbol, but perhaps this only serves to confuse matters. A number of researchers have attempted to give meaning to these ideas (see, e.g., Glimm and Jaffe, 1981), but in the end they have been forced to abandon this simple geometric idea of Dirac's and instead to formulate everything in terms of more complicated and less intuitive geometric frameworks.

Feynman was led to consider a "perturbation series" involving known solutions to the famous Dirac equation

$$\gamma_\mu(i\nabla_\mu - eA_\mu)\Psi = m\Psi \tag{74}$$

Here $\mu = 1, \ldots, 4$, and the γ_μ are given by

$$\gamma_4 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \gamma_i = \begin{pmatrix} 0 & \sigma_i \\ -\sigma_i & 0 \end{pmatrix}, \qquad i = 1, 2, 3$$

where the σ_i are the 2×2 "spin matrices" and the 1's in the expression for γ_4 are the 2×2 identity matrix. Ψ is a 4-vector at each point of $\mathbf{R}^4$, m is the mass of the particle, ∇ is the divergence, and A is the "vector potential" of classical electrodynamics.

In fact, Feynman only needed the "free particle" solution, in which a particle appears first at some point $a \in \mathbf{R}^4$ and then at another point $b \in \mathbf{R}^4$ and it experiences no electromagnetic interactions underway. He treats this solution in Feynman (1962, p. 81, Seventeenth Lecture). There he calculates the so-called "propagation kernel" $K_+(2, 1)$ for a free particle to travel from point 1 to point 2, and obtains

$$K_+(2, 1) = \int (E_p\gamma_4 - p\cdot\gamma + m)\frac{e^{-i(E_p t - p\cdot x)}d^3p}{2E_p(2\pi)^3} \tag{75}$$

Here the point 1 has the coordinates $(\mathbf{x}_1, t_1) \in \mathbf{R}^3 \times \mathbf{R} = \mathbf{R}^4$ and point 2 has the coordinates $(\mathbf{x}_2, t_2)$. Then $\mathbf{x} = \mathbf{x}_2 - \mathbf{x}_1 \in \mathbf{R}^3$ and $t = t_2 - t_1$. The integration is over $\mathbf{R}^3$ and the variable of integration p is thought of as being three-dimensional "momentum." Here $\gamma = (\gamma_1, \gamma_2, \gamma_3)$, i.e., a fixed vector whose components are matrices. Finally, $E_p = (p^2 + m^2)^{1/2}$ is the "energy" associated with the particle. Feynman reduces this integral to another expression, which is, however, itself equally complicated.

What is the meaning of the propagation kernel? To answer this question, it is necessary to first ask, what is the meaning of the probability amplitude? The probability amplitude is a function $\Psi: G^4 \to \mathbf{C}^4$ associating with each point of some region $G^4 \in \mathbf{R}^4$ a complex 4-vector that is a solution of the Dirac equation. Finally, the probability that an "observation" of the system will result in the particle being observed at a given point is the product of Ψ with its transpose: $\Psi^* \times \Psi$. Now let us imagine that two hyperplanes $\mathbf{R}_t^4$ and $\mathbf{R}_{t'}^4$ are given, where $t < t'$ and $\mathbf{R}_s^4 = \{(x_1, \ldots, x_4) \in \mathbf{R}^4 : x_4 = s\}$. Imagine further that an appropriate function Ψ is given on $\mathbf{R}_t^4 \cup \mathbf{R}_{t'}^4$. Then the definition of Ψ can be extended to include the space between $\mathbf{R}_t^4$ and $\mathbf{R}_{t'}^4$ by using the prescription (Feynman, 1962, formula 16-1)

$$\Psi(2) = \int K_+(2, 1)\gamma_4 \Psi(1)\, d^3\mathbf{x}_1 - \int K_+(2, 1')\mathbf{x}_1 \tag{76}$$

where 2 is a point between the hyperplanes, 1 is a point on the hyperplane $\mathbf{R}_t^4$ with coordinates $\mathbf{x}_1$, and $1'$ is a point on the hyperplane $\mathbf{R}_{t'}^4$ with coordinates $\mathbf{x}_{1'}$. This prescription works for a free particle.

The case of a particle that is not "free," but rather is moving in an electromagnetic field, is extremely complicated. The fact that the particle produces its own field, which becomes "infinite" at the particle, leads to further difficulties. (Also, the fact that a particle produces its own field shows that the free particle case, strictly speaking, cannot arise.) There is a huge amount of current research on these questions.

Nevertheless Feynman was able to find a simple and practical method, which also works well in the case of non-free particles. His idea was to associate (74) with the propagation kernel. We can write, following Hoyle and Narlikar (1974),

$$S(\gamma) = S_0(\gamma) - \int V(\gamma(\mathbf{x}, t))\, dt \tag{77}$$

where S is the classical action along a path γ that leads through a region with an electrical potential V. Here S_0 stands for the classical action that would be experienced along the path if there was no electrical potential. Then, in analogy to (73), one can write

$$K_V(2, 1) = \sum \exp\left(\frac{2\pi i S}{h}\right) = \sum \exp\left[\frac{2\pi i}{h}\left(S_0 - \int V\, dt\right)\right] \tag{78}$$

Here, $K_V(2, 1)$ is the propagation kernel for the particle to go from the point 1 to the point 2 through the electrical potential V and, as in (73), the sum is over all possible particle paths from 1 to 2. Next one

expands the exponential function as a Taylor series, obtaining

$$K_V = \sum \exp\left(\frac{2\pi i S_0}{h}\right)\left[1 - \frac{2\pi i}{h}\int V\,dt + \left(\frac{\pi i}{h}\int V\,dt\right)^2 + \cdots\right] \tag{79}$$

Now the first term is associated with $K_0(2, 1)$. The second term can also be expressed in terms of the free particle propagator if we reason as follows. The sum in equation (79) is a sum over all possible paths from 1 to 2. On the other hand, the integral over t is from t_1 to t_2. Therefore, it is sensible to consider each path γ as being composed of two paths γ_1 and γ_2. The path γ_1 is a path from the point 1 with coordinates $(\mathbf{x}_1, t_1)$ to an intermediate point 3 with the coordinates $(\mathbf{x}_3, t_3)$, where $t_1 < t_3 < t_2$. Then γ_2 is a path from the point 3 to the point 2. [Note that it is being assumed here that $\gamma_4(r) < \gamma_4(s)$ for $r < s$.] Thus one can write

$$-\sum \exp\left(\frac{2\pi i S_0}{h}\right)\left(\frac{2\pi i}{h}\int V\,dt\right)$$

$$= -\frac{2\pi}{h}\int\int K_0(3, 1)\,V(3)\,K_0(3, 2)\,d^3\mathbf{x}_3\,dt_3 \tag{80}$$

This term in the Taylor series is illustrated by means of the so-called "Feynman diagram" of first order, as shown in Figure 10. The idea is that this term in equation (79) actually has some physical significance. One imagines that the particle travels "freely" from 1 to 3, then at 3 it is deflected with some "probability" (or "probability amplitude") by a photon of light, and then it travels from 3 to 2.

This philosophy is carried further; one looks at more and more complicated Feynman diagrams, and the probabilities are calculated using (76),

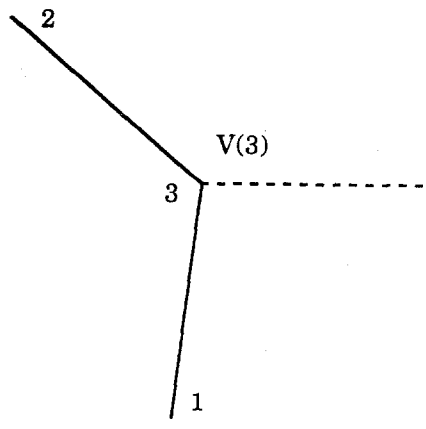

Fig. 10

but with the free particle propagators replaced by (79), truncated after the first few terms. Or often, only certain classes of terms are considered, depending on various intuitive ideas (see especially Mattuck, 1976). In addition, the electrical potential V is replaced by the classical electromagnetic "vector potential" A, and Dirac matrices are inserted at appropriate points. Everything works well, except for the unfortunate but hardly surprising fact that most diagrams lead to divergent integrals. These are diagrams with self-interactions, e.g., Figure 11, where the photon of light is considered to be exchanged between the points 3 and 4 along the particle path. The integration of this term must allow for the approach of 3 arbitrarily near to 4, and thus one sees that, once again, the fact that pointlike particles have infinite fields is causing difficulties. But Feynman found a way to circumvent these difficulties by means of the "renormalization theory," a version of which I will discuss below.

These ideas have often been criticized by mathematical physicists, who are unconvinced that the "perturbation series" has much meaning. Nevertheless, the Feynman integral approach works; it provides simple rules for calculating the results of real experiments, and the results have been verified to the highest degree of accuracy in many experiments.

Much research today is directed toward a search for a theory that gives true solutions to the Dirac equation, but another approach is possible. Why is it necessary to hang on, at all cost, to the notion of space in terms of a continuum? We have seen that this idea causes great difficulties, not only in quantum electrodynamics, but also in classical electrodynamics and in general relativity. Thus, it is strange that the view has become prevalent that diverging integrals are a special problem in quantum field theory which can be circumvented by some sort of "cleaning up" process (the work of

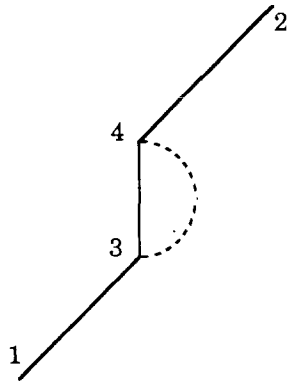

Fig. 11

Weierstrass is often quoted in this connection), affecting the basic methods of mathematical analysis.

Indeed, there is no need to look for "mistakes" in the basic definitions of analysis, as they are understood by most mathematicians. The question is simply whether or not these definitions are appropriate for application in physics. Thus, Feynman's methods should not be dismissed as being nothing more than vague mathematical approximations. On the contrary, the best approach might be to start with the Feynman diagrams, and then to see what reasonable (and presumably discrete) mathematical models are possible.

6.3. Hoyle and Narlikar's Renormalization Theory

I will now sketch the argument that shows how the integrals in the last subsection diverge and how they can be altered so that they are in fact well defined. This alteration reflects the kind of discretization that has been discussed in the previous sections and therefore it seems likely that any attempt to give meaning to an expression like (73) in terms of discrete paths would be likely to make use of this renormalization theory. I will consider only the simplest self-energy diagram: that in which a single photon is emitted and then reabsorbed by a single particle. The exposition in Hoyle and Narlikar (1974) will be closely followed.

To begin, it is necessary to recall from Section 2 that the Liénard–Weichard potential generated by the particle is given by equation (13). Thus, following the philosophy of Section 6.1, it would seem that the electrical potential V in the appropriate terms of equation (79) should be taken to be $\delta(s_{4,3}^2)$, where the four-dimensional Dirac δ-function is as in Section 2 and it is assumed that the Feynman diagram is such that the particle starts at point 1, travels to 3, where it emits a photon, then to 4, where it reabsorbs the same photon, then continues on to the point 2.

In treating these ideas, Feynman (1962, p. 120, 24th Lecture) begins by taking the one-dimensional Dirac δ and writes it in the following way:

$$\delta(X) = \int_{-\infty}^{+\infty} \frac{\exp(-iwX)\, dw}{\pi} \tag{81}$$

He then asserts that this is a representation of the δ-function (and thus the photon with which it is associated) in terms of both positive and negative frequencies. But, following his reasoning, a photon can only have positive energy. Thus, the "negative frequencies" should be disregarded, and we should instead take the new Dirac δ-function

$$\delta_+(X) = \int_0^{+\infty} \frac{\exp(-iwX)\, dw}{\pi} \tag{82}$$

which has no negative frequencies. Then the four-dimensional δ-function with positive frequencies $\delta_+(s^2_{4,3})$ should be defined by means of an equation similar to (19) in Section 2.

The formula for δ_+ used by Hoyle and Narlikar (1974, Appendix 3), formula (102), is

$$\delta_+(s^2_{4,3}) = 4\pi i \int \frac{1}{2K} \exp[-ik|t_4 - t_3| + i\mathbf{k}(\mathbf{x}_4 - \mathbf{x}_3)] \frac{d^3K}{(2\pi)^3} \tag{83}$$

This is an integral over $\mathbf{R}^3$. As before, the points $3 = (\mathbf{x}_3, t_3)$ and $4 = (\mathbf{x}_4, t_4)$ in $\mathbf{R}^4$ are of interest. Also, $K = |\mathbf{k}|$. The part of their formula (133) for the relevant term in the perturbation expansion for the case of a particle with vanishing spatial momentum can then be written

$$-ie^2 \int \int K_+(2,3)\gamma_i K_+(3,4)\gamma^i \delta_+(s^2_{3,4})\, U_0\, e^{-imt_4}\, dX_3\, dX_4 \tag{84}$$

The γ_i, $i = 1, \ldots, 4$, are the Dirac matrices, and one is to sum over them. The integration is over the four-dimensional slab between the two three-dimensional hyperplanes $\mathbf{R}^4_t$, $t = t_1, t_2$.

Now, it is well known that this integral "diverges" and is thus ill defined. Feynman's renormalization theory involves altering the definition of δ_+ so that the integral is "cut off" at high "frequencies." This approach certainly works. The main result is that, when considered as a whole, the "perturbation series" does not depend upon the form of this cutoff. What is more, the modified version of δ_+ also produces a convergent perturbation series. But do these modifications to δ_+ have any meaning in terms of an underlying geometric model?

Hoyle and Narlikar's renormalization theory also involves a modification to δ_+, and thus it is very similar to Feynman's approach. However, instead of considering a "frequency" cutoff, they imagine that there is a minimal possible interaction distance ε, so that if the points 3 and 4 have a vanishing Lorentz separation (thus allowing the exchange of a photon), and if they approach each other more closely than ε, then there is no longer an interaction, and δ_+ is null. Thus, the integral in (84) is to be carried out subject to the restriction that

$$|t_3 - t_4| > \varepsilon \tag{85}$$

for some small constant $\varepsilon > 0$.

In the case of free particles with nonzero spatial momentum, they take the cutoff rule given by

$$|X^i_3 - X^i_4| \geq \varepsilon p^i / m \tag{86}$$

for $i=1,\ldots,4$, and p^i being the components of the four-dimensional momentum of the particle. They obtain a renormalization theory similar to the usual one, with a "renormalization constant" given by

$$\frac{3e^2}{2\pi}\ln(m\varepsilon) \tag{87}$$

where e is the charge and m is the mass of the particle. The details can be found in Hoyle and Narlikar (1974). In particular, it should be noted that both the mass and charge renormalization follow from the same principle. The spatial cutoff given by (86) transforms in the same way as the momentum, and is thus independent of the coordinate system.

Hoyle and Narlikar base their theory on the idea of paths consisting of many short "null segments" and use this idea to justify the cutoff rule (85). They also mention the idea that this cutoff could have something to do with the gravitational Schwarzschild radius of the particles, although this seems to play no essential role in their theory. If indeed the concept of a manifold can be used on such small scales of distance, and furthermore if the particles create tiny "black holes" whose radius is of the order 10^{-40} cm, then it is difficult to see how the idea of particles consisting of null segments of about this length can be maintained.

I certainly do not claim to have a better picture, nor do I claim to have an adequate understanding of the calculations themselves. What relevance can a long and complicated calculation, involving many integrals and simplifying assumptions regarding those integrals, have to a finite combinatorial model? It seems clear that a true renormalization theory based on a discrete model of space-time has yet to be developed. But still, given these limitations, it might be interesting to point out a certain structure in the present model that does correspond to Hoyle and Narlikar's renormalization condition (86).

Recall Definition 1 of Section 5.9. We are given two adjacent points p_i, p_{i+1} on the particle $P \subset W$. Then a maximal chain of positions $p_i = C_0 < C_1 < \cdots < C_n = p_{i+1}$ between p_i and p_{i+1} is considered. The vector $p_{i+1} - p_i$ in $\mathbf{R}^4$ is proportional to the 4-momentum of the particle P (this follows from Theorem 3.1). Now it might be considered that the chain $\{C_j\}$ of positions between p_i and p_{i+1} is in some sense the smallest subdivision of this interval, and the observation we now make is that if each C_j is associated with a point in $\mathbf{R}^4$ (as in Section 3.8), then the vector $C_{i+1} - C_i$ in $\mathbf{R}^4$ is on average also proportional to the 4-momentum of the particle P.

Observation 1. Let p_i, p_{i+1} and q_i, q_{i+1} be two points in a set W satisfying our general assumptions. In particular, assume that W, together with the

set of positions of W, has an embedding in $\mathbf{R}^4$ that is nearly order-preserving (with respect to the Lorentz metric), and assume further that the density of positions near the images of p_j, q_j, where $j = i, i+1$, is nearly constant (for ease of notation call these points of $\mathbf{R}^4$ simply p_j, q_j also). Assume that $p_i < p_{i+1}$ and $q_i < q_{i+1}$ and that the Lorentz distance from p_i to p_{i+1} is equal to the Lorentz distance from q_i to q_{i+1}. Let $p_i = C_0 < C_1 < \cdots < C_n = p_{i+1}$ be a maximal chain of positions from p_i to p_{i+1} and let $q_i = C_0' < C_1' < \cdots < C_m' = q_{i+1}$ be a maximal chain of positions from q_i to q_{i+1}. Then $n \approx m$ and the points $s_0, s_1, \ldots, s_n$ associated with $C_0, C_1, \ldots, C_n$ run in nearly a straight line from p_i to p_{i+1}. The same holds for the positions between q_i and q_{i+1}.

Justification. It suffices to consider the case that $p_1 = (0, 0, 0, 0)$, $p_2 = (0, 0, 0, 1)$, and then q_j is in a more general position. Now it is reasonable, from the point of view of symmetry, to assert that the points associated with $C_0, C_1, \ldots, C_n$ lie nearly along the straight line from p_1 to p_2. Let $\Psi : \mathbf{R}^4 \to \mathbf{R}^4$ be a Lorentz transformation taking q_i to p_i. Now, Ψ produces a different faithful representation of W in $\mathbf{R}^4$ that no longer satisfies the cosmological hypothesis (Section 3.8). But since the original embedding of W in $\mathbf{R}^4$ satisfies the cosmological hypothesis, it follows that, for example, the point $\Psi(s_0)$ of $\mathbf{R}^4$ can be associated with a nearly Lorentzian cone in $\mathbf{R}^4$, representing the cone C_0 of W under the transformed embedding of W in $\mathbf{R}^4$. More generally, the density and distribution of the points of $\mathbf{R}^4$ associated with cones of W in the neighborhood of p_j, q_j can be expected to be invariant under Ψ since such "Lorentz mappings" as Ψ preserve Lesbegue measure on $\mathbf{R}^4$. Thus, a maximal chain of positions between q_i and q_{i+1} is represented by a maximal chain of positions between (0, 0, 0, 0) and (0, 0, 0, 1), with the same density and distribution of positions in $\mathbf{R}^4$ as under the original embedding. Hence the Observation follows.

One point upon which Hoyle and Narlikar do not dwell is the question of whether or not the "probability amplitudes" they use—that is, the limits of the (convergent) perturbation series—do in fact define a function that satisfies the Dirac equation. But for this it is only necessary to use the standard technique used by Feynman (1962, Fifteenth Lecture): one shows that the propagation kernel K^A, which is defined as the sum of the terms in the series that do not include self-interactions, satisfies the inhomogeneous Dirac equation [Feynman (1962), (15-9)]. But then, since Hoyle and Narlikar use a *distance* cutoff, it is possible to generalize K^A to include self-interactions. We can treat each such self-interaction as if it were an external potential, and thus we arrive at the inhomogeneous Dirac equation also in this case.

Of course this theory is not gauge invariant—the minimum interaction distance ε plays an important, though unobservable, role. Thus, the theory cannot claim to provide a solution to the axioms of the conventional field theory. But our basic assumption, namely that space is discrete, is already a denial of gauge invariance. Therefore, following this reasoning, one can say that if the reader is prepared to accept that discrete mathematics is a reasonable framework for theoretical physics, then the idea of gauge invariance—as a *fundamental* principle of physics, rather than simply as an aid to the calculation of certain quantities in quantum mechanics—should no longer be strictly adhered to.

7. CONCLUSIONS, PROBLEMS, AND FURTHER SPECULATIONS

7.1. Connections with Current Research

Any reader who has reached this point will have long since recognized that the ideas presented have very little to do with current research in theoretical physics. The treatment of gravitation theory was, for the specialist in the subject, very basic. I have neither touched on quantum gravity, and the hopes that subject has of establishing a connection between the theory of general relativity and quantum mechanics, nor on cosmology, except for a few simple criticisms. Also, the discussion of quantum mechanics mentioned nothing more than a very small number of the oldest and most established results in that subject. I have not dealt with the lattice theories that are currently being used.

But such investigations are the proper domain of specialists. It is unclear whether a discrete formulation is in each case possible or appropriate. Nevertheless, it is interesting to speculate on how these ideas might be applied.

7.1.1. Elementary Particle Physics

The simplest and most immediate thought is that discrete geometry might be of particular relevance in the theory of the elementary "particles" (that is, the phenomena investigated in high-energy collisions between "ordinary" particles). For example, two electrons can be made to collide with one another head on. If the collision is energetic enough, a shower of new particles is created. The pleasant idea that this shower represents the "elementary" constituent parts of the original electrons is soon dispelled by the observation that the shower can itself contain two or more electrons.

A great deal of effort has gone into the search for more and more "elementary" particles over the last 50 years. Certainly no one could claim

that the search has come to an end. At present, though, most people seem to agree that the electron is a truly elementary particle. There is a great flood of particles that, while at first thought to be elementary, are now thought of as resulting from combinations of simpler, though unobservable, particles—the "quarks." Could the quarks also be "truly" elementary particles? Many people today seem prepared to pursue the hypothesis that this is in fact true. In any case it would appear that, just as the discovery of "atomic" physics showed that the atoms of the 19th century were not really "atomic" after all, so many of the elementary particles of the 20th century are also not "atomic."

Consider once again a collision of two electrons. If we accept the picture of "truly" elementary particles as being discrete paths, then such a collision would be represented by two discrete particle paths coming together. The "shower" of particles would be a great many short paths, or even single points in the discrete structure. (Of course we must generalize our definitions to allow paths that are not necessarily infinitely long.) Now it seems obvious that some combinations of discrete paths will be more likely than others with respect to the underlying variational principle. Thus, the discrete structure could explain why simple combinatorial rules have been found to apply. It would also explain why we would never come to the end of the search for the "elementary" particles: higher and higher energies would allow ever new combinations of the points on the paths.

7.1.2. Cosmology and "Local" Physics

Recently much has been made of possible connections between certain cosmological speculations and some properties of the "elementary particles." For example, the idea has often been expressed that perhaps the particles exhibit certain properties that might have something to do with an "evolving" universe. Now it is unclear to what extent these ideas could be carried over into a discrete framework. On the other hand, every action-at-a-distance theory involves, by its very nature, a description of local phenomena in terms of a global formulation. Thus, a changing or evolving universe would generally imply changing local laws.

One area where this question of possibly changing local laws might be best investigated is gravitation. If we accept the reasoning of Section 4, then gravity could be explained as a complicated statistical effect. In particular, the relative densities of nearby and (cosmologically) distant matter would play an important role in determining the gravitational constant. In an evolving universe, these relationships might change, and thus the gravitational constant might change. But in the absence of any empirical evidence, it hardly seems worthwhile to pursue such thoughts.

7.2. Some Further Problems

7.2.1. Four Dimensions

Can the four-dimensional structure of space-time be explained in terms of the properties of discrete, partially ordered sets? This is the most important question that can be posed when it comes to discrete methods in physics. A satisfactory answer would enable us to do away with the idea of proper representations of partially ordered sets in $\mathbf{R}^4$. Thus, we would be able to find a more natural basis than that given in Section 3 for the association of geometrical and combinatorial ideas.

The examples of Section 3 show that four-dimensionality is a special property, characteristic of only a special class of discrete, partially ordered sets. How is it possible to characterize four-dimensionality for such discrete sets?

There are a number of ways of defining the "dimension" of a given space. Perhaps the most appropriate method could be based on the results related to Example 1 of Section 3.3. It was shown there that if a partially ordered set K is geometrically n-dimensional, then it can contain no subset of the form W_{n+1}. (Recall that W_n contains $n+2^n$ elements. The first n elements are all unrelated to one another in the ordering, while the last 2^n elements are either below or unrelated to these first elements. All 2^n combinations are represented here.) We can now take this result as our definition of dimension. Thus, a set W could be defined to be n-dimensional if it contains W_n, but not W_{n+1}, as a subset. Is it possible to prove that any set that is n-dimensional according to this definition also has a faithful representation in $\mathbf{R}^n$?

Another idea is to observe that not all discrete, partially ordered sets satisfy the variational principle given by (47). In addition, the considerations of Section 4.13 would seem to involve restrictions on the classes of sets that are to be allowed. Could it be that in some subtle way these restrictions also involve a condition on the dimension?

7.2.2. Quantum Mechanics

The treatment of quantum mechanics in Section 5 centered on the idea of clusters of sets. On the other hand, it was asserted that each experiment would, as in classical physics, involve a tremendous amount of detail. All of these details would contribute to the determination of the statistical properties of the experiment.

Thus, it would be a natural idea to investigate the details of "discrete" experiments within the framework of classical physics. As an example, one might consider the passage of a pointlike particle through a crystal consisting

of a regular lattice of pointlike particles. The discretization would consist of imagining that these particles only appear for a moment at discrete intervals of time. At each of these appearances the particle would experience a deflection, say, in accordance with an analog of the Coulomb law. A first approximation might be to consider just the "test particle" to be discrete and the particles of the crystal to be normal continuous classical particles in $\mathbf{R}^4$. What could one say about the classical probabilities of such a system?

7.2.3. Spin and Complex Numbers

It is thought that the connections among the Dirac matrices, spin, the use of complex numbers in physics, and the basic structure of physical space are subtle and profound (Wells, 1979). In particular, the well-known connection between spin and statistics needs explanation. Certainly we have not as yet dealt with these concepts in any very satisfactory way. At best we could remark that the range of the function F in equation (47) could be taken to be some appropriate set that might be more complicated than the integers.

But allowing F to have complex, or even real, values seems to go against the basic philosophy of a discrete space. Surely the assumption of complex values amounts, indirectly, to the assumption of Euclidean space, with all of the complicated axioms and assumptions that that brings with it. This is just what we want to get away from!

What alternatives are there? Perhaps the following idea might be worthy of investigation. Begin by considering that the electron and the other "normal" long-lived massive particles obey Fermi statistics. How does this come about? Consider an experiment involving two electrons e_1 and e_2 going from a point P to a point Q, where $P, Q \in \mathbf{R}^4$. The electrons are to be considered as paths in $\mathbf{R}^4$, or in our framework as discrete paths. Now the fact that electrons are fermions means that if e_1 and e_2 are indistinguishable (the same spin, etc.), then the probability of the experiment is small, at least in comparison with the case that the electrons are distinguishable.

How can we explain this in terms of the sizes of the clusters associated with the experiment? In the case of distinguishable particles, there is nothing new. However, if the particles are indistinguishable, then there is a new effect to consider. Imagine that e_1 and e_2 approach each other closely at two points (say the points P and Q above). Between P and Q, e_1 and e_2 follow different path segments. But then one could imagine another possible universe in which e_1 follows e_2's segment between P and Q, and vice versa. In the case of continuous paths, these two different universes are really different, but *not* when the paths are discrete. Here we would just have two rows of elements of a *single* discrete, partially ordered set, and we have the

choice of assigning the elements to the rows in a number of different ways. But, according to the definition of probability given in Section 5, this is a *single* universe, and so the probability for the experiment would be small, as the Pauli exclusion principle suggests it should be.

These thoughts should undoubtedly be associated more with "speculations" than "problems," since it seems unlikely that they would lead to any sensible mathematical results. Nevertheless, it is clear that if one is willing to accept the idea that quantum mechanical statistics can be explained according to the ideas in Section 5, then the strange recipes for combining the "probability amplitudes" for fermions and Bosons must also be explained in these terms.

REFERENCES

Bell, J. S. (1966). On the problem of hidden variables in quantum mechanics, *Reviews of Modern Physics*, **38**, 447.

Cox, D. R., and Isham, V. (1980). *Point Processes*, Chapman and Hall.

Dicke, R. H. (1964). *The Theoretical Significance of Experimental Relativity*, Gordon and Breach.

Dirac, P. A. M. (1938). Classical theory and radiating electrons, *Proceedings of the Royal London, Series A*, **167**, 148.

Einstein, A. (1956). *The Meaning of Relativity*, 5th ed., Princeton University Press.

Einstein, A., and Grommer, J. (1927). Allgemeine Relativitatstheorie und Bewegungsgesetz, *Sitzungsberichte Preussische Akademie der Wissenschaften*, **2**.

Einstein, A., Podolsky, B., and Rosen, N. (1935). Can quantum-mechanical discription of reality be considered complete?, *Physical Review*, **2**, 47.

Eyges, L. (1972). *The Classical Electromagnetic Field*, Addison-Wesley.

Feynman, R. P. (1962). *Quantum Electrodynamics*, Benjamin.

Feynman, R. P., and Hibbs, A. R. (1965). *Quantum Mechanics and Path Integrals*, McGraw-Hill.

Feynman, R. P., Leighton, R. B., and Sands, M. (1965). *The Feynman Lectures on Physics*, Vol. III, Addison-Wesley.

Fokker, A. D. (1929). Ein Invarianter Variationssatz fur die Bewegung Meherer Elektrischer Massenteilchen, *Zeitschrift für Physik*, **58**, 386.

Gauss, C. F. (1845). Letter to W. Weber (19 March, 1845), in C. F. Gauss, *Werke*, Vol. 5, p. 629.

Glimm, J., and Jaffe, A. (1981). *Quantum Mechanics: A Functional Integral Point of View*, Springer-Verlag.

Gödel, K. (1949). An example of a new type of cosmological solutions of Einstein's field equations of gravitation, *Reviews of Modern Physics*, **21**, 447.

Hawking, S. W., and Ellis, G. F. R. (1973). *The Large Scale Structure of Space-Time*, Cambridge University Press.

Hogarth, J. E. (1962). Cosmological considerations of the absorber theory of radiation, *Proceedings of the Royal Society of London, Series A*, **267**, 365.

Hoyle, F., and Narlikar, J. V. (1974). *Action at a Distance in Physics and Cosmology*, Freeman.

Jammer, M. (1974). *The Philosophy of Quantum Mechanics*, Wiley.

Jauch, J. M. (1968). *Foundations of Quantum Mechanics*, Addison-Wesley.

Landé, A. (1965). *New Foundations of Quantum Mechanics*, Cambridge University Press.

Llosa, J., ed. (1981). *Relativistic Action at a Distance: Classical and Quantum Aspects*, Springer-Verlag.

Mattuck, R. D. (1976). *A Guide to Feynman Diagrams in the Many-Body Problem*, McGraw-Hill.
Messiah, A. (1970). *Quantum Mechanics*, North-Holland.
Narlikar, J. V. (1978). *Lectures on General Relativity and Cosmology*, Macmillan.
Nelson, E. (1967). *Dynamical Theories of Brownian Motion*, Princeton University Press.
Rebbi, C. (1982). Lattice gauge theories and Monte Carlo simulations, in *Non-Perturbative Aspects of Quantum Field Theory*, World Scientific, Singapore.
Roe, P. E. (1969). Time-symmetric electrodynamics in Friedmann universes, *Monthly Notices of the Royal Astronomical Society*, **144**, 219.
Schilpp, P. A., ed. (1949). *Albert Einstein, Philosopher-Scientist*, Library of Living Philosophers.
Schrödinger, E. (1953). The meaning of wave mechanics, in *Louis de Broglie Physicien et Penseur*, A. George, ed., Albin Michel.
Segal, I. E. (1976). *Mathematical Cosmology and Extragalactic Astronomy*, Academic Press.
Von Neumann, J. (1955). *Mathematical Foundations of Quantum Mechanics*, Princeton University Press.
Wells, R. O. (1976). Complex manifolds and mathematical physics, *Bulletin (NS) of the American Mathematical Society*, **1**, 296.
Wheeler, J. A., and Feynman, R. P. (1945). Interaction with the absorber as the mechanism of radiation, *Reviews of Modern Physics*, **17**, 157.
Wheeler, J. A., and Feynman, R. P. (1949). Classical electrodynamics in terms of direct interparticle action, *Reviews of Modern Physics*, **21**, 425.
Yilmaz, H. (1965). *Introduction to the Theory of Relativity and the Principles of Modern Physics*, Blaisdell.