I

General introduction

1. Standpoint and purpose

As mathematical science has evolved, the natural tendency toward the differentiation of labor led to the separation of mathematics and physics, and to the organization of the whole subject along craft and technical lines, rather than along integrated and externally motivated ones. This has resulted in the development of relatively objective and uniform standards of professional work, and its more effective and precise communicability, in mathematics, experimental physics, observational astronomy, and elsewhere. On the other hand, this very useful and natural technical clarification, proliferation, and standardization increases significantly the difficulty of appropriately treating and exposing any single coherent idea or problem of both mathematical and empirical relevance. The present book attempts to steer a sound course between the Scylla of wishful speculation, which may result from attempts to bypass the rigors of conformity to the fundamental scientific disciplines, in the interest of dealing with issues of large relevance, and the Charybdis of elaborate technical refinement, which may result from the determination to ignore questions of relevance, in the interest of achieving an ultimate technical perfection.

This work treats a single coherent conception of space, time, causality, and related notions, and so is presented as an entity, although it depends on the utilization of highly developed areas in the fields of mathematics, physics, and astronomy. Following the present general introduction, it will be treated seriatim from the standpoints of these respective fields; however, a certain underlying terminological and notational uniformity will be employed. The work is primarily a synthesis of foundational developments in each of the fields, coupled with the observation of the existence of a remarkable new space–time model. It involves few special features of unusual technical difficulty; rather, it leads to important and interesting technical problems, by virtue of its crystallization of a new scientific outlook, and its proposal of new notions of time, space, and energy applicable both to cosmology and to microphysics.

Our present notions of these matters are inevitably based on and colored by anthropomorphic perceptions and experiences of them. On the other hand, it is neither inevitable nor desirable that purely theoretical notions of these physically crucial matters should, on the basis of their apparent cogency in anthropomorphic or limited professional realms, be judged binding on theories attempting to deal with the physics of extreme distances. Of course, the subject is inherently very difficult, in that significant observation or experimentation seems possible only on the basis of a substantial theoretical framework, in view of the highly indirect accessibility of objects at these distances; yet any such framework is necessarily rather tentative. A priori, it might well appear doubtful whether a physically conservative and mathematically well-founded treatment of the concepts of time, space, energy, causality, etc. could be sufficiently incisive to attain verifiable quantitative conclusions, and be thereby subject to empirical validation. A methodologically interesting aspect of this book is the demonstration that this is the case.

2. Causality and geometry—historical

When Einstein questioned the absolute nature of simultaneity, and developed a theory of time and communication based on the propagation of light signals, causality considerations were implicitly introduced into the theory of space and time. These emerge more clearly in the work of Minkowski. However, causality was treated in a largely philosophical and intuitive way, as a marginal feature of quantitatively more central matters, e.g., the addition of large velocities. Indeed, the latter feature is identified by Bridgman as the main one in relativity, and one that is logically unrelated to causality. None of these authors, nor their immediate successors, attempted an axiomatic treatment; nor made a consistent explicit separation between mathematical and physical considerations; and to this day (with the exceptions noted below), the notions of "causality," "observer," "clock," and "rod," are commonly used in quite intuitive, if not subjective, senses in work in relativity theory.

The explicit and cogent significance of causality for relativity theory was first recognized and emphasized by Robb (1911–1936) and developed by him into a deductive mathematical theory in which special relativity is effectively derived without any use of such notions as " clock " or " simultaneity " (at different points of space). As recognized by Fokker (1965), Robb thereby founded the subject of *chronogeometry*, in which considerations of temporal order are merged with geometry in a mathematical way, but with a presumption of applicability to physical space–time. A central notion in Robb's theory was a partial ordering in a given space, representing physically the relation of temporal precedence, in the world's space–time medium. Many mathematical axioms, in significant part motivated by optical considerations, with relatively objective physical interpretations (not requiring notions such as observer, clock, or rod at *different* points of space–time) lead after extensive analysis in this theory to the conclusion that the given causality-endowed space is isomorphic to Minkowski space–time, the partial ordering being the usual notion of temporal precedence in this manifold. By modern standards, while Robb's work was quite original and exhibits a high order of mathematical clarity and coherence, it was isolated, unsophisticated, and apparently terminal in intent. Its main significance seems to lie in its formulation of the causality point of view, and demonstration of its power to lead to a more objective and at the same time philosophically satisfying treatment of relativity.

Since the war, the subject of chronogeometry has attracted the attention of a number of mathematicians, including notably A. D. Alexandrov and E. C. Zeeman, and in a modified form, J. L. Tits. In work beginning in the early 1950's, Alexandrov developed a school of work on mathematical relativity, very much from the chronogeometric outlook (explicitly so in Alexandrov, 1967, a key work), which has been contributed to by Busemann (1967) and Pimenov (1970), among others. In 1964 Zeeman rediscovered and exposed cogently the theorem (due originally to Alexandrov and Ovchinnikova, 1953, a work which seems not to have been widely disseminated outside the Soviet Union), that a causality-preserving transformation on Minkowski space is necessarily a Lorentz transformation, within a scale factor. In 1960 Tits, in a key work, published a summary of a classification of all four-dimensional Lorentzian manifolds enjoying certain physically natural transitivity properties.

Chronogeometry has also emerged, in quite a different although related way, from the needs of the general theory of hyperbolic partial differential equations, and our initial acquaintance with the subject was derived from the fundamental work of Leray (1952), which correlated in a very general way the infinitesimal and finite notions of causality. A given hyperbolic equation defines an infinitesimal notion of temporal order, in the form of a proper convex cone in the tangent space at each point of the space–time manifold. Prewar work by Zaremba and Marchaud was completed and applied with cogency in Leray's work. His work, and particularly its chronogeometric side, has been further developed by Choquet-Bruhat (1971), who has made applications to general relativity; somewhat related work is due to Lichnerowicz (1971). Partially similar but more intricate and specialized ideas have been applied to the problem of the structure of space–time in general relativity by Hawking, Penrose, and a number of collaborators (cf. Ellis and Hawking, 1973), as well as by other recent writers on the problem of singularities in general relativity.

The subject of hyperbolic partial differential equations in the large can be considered in large part as falling under the general heading of causality and evolutionary considerations in functional analysis. This is not a question of pure geometry, of course, but rather of function spaces built on the space–time manifold; nevertheless there are some essential geometrical aspects, and causality plays a crucial role. This is the case, for example, for the key notions of domain of dependence, region of influence, and of causal propagation. Indeed, hyperbolicity may well be necessary as well as sufficient for causal propagation, as evidenced in part by recent work of Berman (1974). This shows in particular that in the Klein–Gordon equation $u_{tt} = \Delta u + cu$ (c = constant), it is impossible to replace $\Delta + c$ by any other semibounded self-adjoint operator in L_2 over space if propagation is to remain both causal and Euclidean-invariant in Minkowski space.

The latter work continues an extensive line of work on the implications of causality for temporally invariant linear operators. The treatment of the dispersion of light by Kramers and Kronig was among the earliest and most influential in this general direction. The work of Bode on the design of wave filters applied a similar idea in a nonrelativistic context, that of linear network theory. Mathematically, the work of Paley and Wiener on complex Fourier analysis, and of Kolmogoroff, and later Wiener, and many others on linear prediction theory, in part relate to causality considerations in a context of temporal development and invariance. The Paley–Wiener theory was extended to a more general setting, applicable to relativistic cases, by Bochner. This was used in the postwar development of the general theory of linear hyperbolic equations due to Gårding and Leray, and thereby connected with causality features.

A partial synthesis of the causality ideas involved in this line of work is involved in the abstract study of linear systems by Fourès and Segal (1955). A general conclusion which is relevant to the present considerations and which emerges from this work is that the "future" may be represented by an essentially arbitrary nontrivial closed convex cone in the underlying linear manifold, without any fundamental loss of scientific cogency in the treatment of global questions. Furthermore, the convexity of the cone is both physically natural and technically crucial.

3. Conformality, groups, and particles—historical

Several different lines of physical mathematics, in addition to the chronogeometric and hyperbolic partial differential equation ones are involved in the present work. Indeed, the model proposed here originated in a study from the vantage points of group deformations and particle classification of the conformal space proposed as a cosmos forty years ago by O. Veblen. Chronogeometry supplied only the decisive final clue and a perspicuous and natural framework.

The rough idea bringing in the deformation of transformation groups was clearly enunciated by Minkowski, who pointed out—admittedly, ex post facto—that the displacement of Galilean relativity by special relativity amounted to a change from one group of transformations to a more sophisticated (and in his view, more attractive) one, of which the first is in a sense a limiting case. This is the limit as the velocity of light becomes effectively infinite, for the phenomena under consideration.† Twenty years later it was found that classical (unquantized) mechanics was similarly a limiting case of a more accurate theory, quantum mechanics. Actually, Planck's constant h, which is involved here, and the velocity of light c, involved in the deformation of the Lorentz into the Galilean group, are fixed constants, unvarying in Nature; but a precise mathematical meaning for the notion of limiting case corresponding to the intuitive physical idea was given in Segal (1951). This concept of group deformation has since been explored in slightly different settings in both the physics and the mathematical literature.

In the light of Minkowski's idea and persistent foundational difficulties in relativistic quantum mechanics, it was natural to raise the question of whether this theory is not in itself a limiting case of a more accurate theory. A model with a discrete space and an associated fundamental microscopic

† Minkowski wrote: "If we now allow c to increase to infinity, and $1/c$ therefore to converge toward zero, we see . . . that the group G_c (the Lorentz group) in the limit when $c = \infty$, i.e. the group G_∞, becomes no other than that complete group which is appropriate to Newtonian mechanics (i.e. the Galilean group). This being so, and since G_c is mathematically more intelligible than G_∞, it looks as though the thought might have struck some mathematician, fancy-free, that after all, as a matter of fact, natural phenomena do not possess an invariance with the group G_∞, but rather with a group G_c, c being finite and determinate, but in ordinary units of measure *extremely great*." ("Space and Time," H. Minkowski, translation of address delivered at 80th assembly of German Natural Scientists and Physicists, Cologne, 21 September 1908; in "The Principle of Relativity," H. A. Lorentz, A. Einstein, H. Minkowski, and H. Weyl, 1923, pp. 78–79; reprinted Dover, New York.)

length, involving a species of approximation of the Lorentz by the de Sitter group, was proposed by Snyder (1947); the set of fundamental dynamical variables did not form a Lie algebra. It was noted by Segal (1951) that the Lie algebras of certain pseudo-orthogonal groups, namely $O(5, 1)$, $O(4, 2)$, and $O(3, 3)$ were deformable into that of the fundamental dynamical variables (momenta, boosts, and space–time coordinates) in relativistic quantum mechanics; a parallel heuristic observation had been independently made by Yang (1947). These Lie algebras were themselves terminal, in the sense that, unlike the Lie algebras of Galilean and special relativity, or of classical mechanics, they were not limiting cases of any other (nonisomorphic) Lie algebras. In physical terms, this indicated a relatively terminal property for a corresponding physical theory, for such a theory based on commutation relations (i.e., a Lie algebra) could be a limiting case of another such theory only if the latter was of higher dimension. While a slight increase in dimension is always a possibility, any large increase would produce many more invariants ("constants of the motion," or mathematically, number of generators of a maximal Abelian subalgebra of Lie algebra, in the relevant representations) than appear compatible with the limited number of states and selection rules observed in elementary particle experiments. Furthermore, groups of larger finite dimension rarely operate effectively on a four-dimensional space–time.

Of the cited pseudo-orthogonal groups $O(5, 1)$ and $O(4, 2)$, the groups of de Sitter and conformal space respectively, have been the most studied. As indicated by Segal (1967a) and Philips and Wigner (1968), $O(5, 1)$ is difficult to reconcile with the fundamental principle of positivity of the energy in quantum mechanics; more specifically, in no nontrivial unitary representation of $O(5, 1)$ does any self-adjoint generator correspond to a nonnegative self-adjoint operator. The group $O(4, 2)$ is free of this failing, and a variety of physical desiderata have pointed to it as a likely candidate for a more accurate higher symmetry group. As essentially the conformal group, it contains the Lorentz group as a subgroup; as shown by Bateman and Cunningham sixty years ago and extended by L. Gross (1964), it is the invariance group of the Maxwell equations—a statement which is mathematically fully meaningful only when Minkowski space is extended to the conformal space treated by Veblen. More recently, experimental indications of scale invariance in elementary particle interactions have led to renewed studies toward the utilization of the conformal group (cf., e.g., Carruthers, 1971).

There have been two major obstacles to the use of the conformal group in foundational theoretical physics, which are roughly macro- and microscopic in nature, respectively. Macroscopically, conformal space is acausal in the sense that at a fixed point x, the limit of the space–time point (x, t) as $t \to +\infty$ is identical with its limit as $t \to -\infty$; these limits exist, the space being closed (i.e., "compact"). This is contrary to physical intuition, leads to serious difficulties of physical interpretation, makes it impossible to distinguish between the advanced and retarded elementary solutions of Maxwell's equations in conformal space, etc. From an elementary particle viewpoint, the fundamental symmetry group is probably more important than the geometrical space serving as particle medium, but the conformal group suffers from a corresponding lack of invariant temporal orientation.

Microscopically, relevance to physical elementary particle observations requires an explicit correlation of representations, and a set of generators of a maximal abelian subalgebra of the enveloping algebra, with observed particles and their quantum numbers. This is a highly vertical and complex process; relatively small differences in the initial aspects of this correlation may ramify and produce gross differences in the implications subject to empirical validation. For example, it is not clear a priori whether the energy and other conventional dynamical variables should remain unchanged, as is possible because of the inclusion of the Lorentz group in the conformal group, and is assumed implicitly in most of the theoretical physical literature (but which leads to difficulties because of the lack of conformally invariant wave equations for massive particles, among other reasons); or whether the energy, etc. require modification so as to involve the full conformal group in a more essential way, as originally proposed by Segal (1951). In Segal (1967a), qualitative evidence for such a new generator was adduced: (a) unlike the conventional generator P_t, which cannot lead to mass splitting, according to a theorem due in infinitesimal form to O'Raifeartaigh (1965) and in global form to Segal (1967b), the new energy operator P_t' (which corresponds to a generator of the conformal group which is essentially different from, i.e., nonconjugate to, $\partial/\partial t$) may have a discrete spectrum; (b) the idea that temporal displacement, as dynamically fundamental, should be definable in a mathematically unique and natural way is substantiated by P_t', which has such definitions, in terms of $O(4, 2)$ as the generator of the corresponding $O(2)$ subgroup, and in terms of the twofold covering group $SU(2, 2)$ as the correspondent to its simplest generator. However, a quantitative check on the validity of this definition via microscopic observation appears difficult except in conjunction with a number of additional assignments or correspondences between apparent quantum numbers and theoretical operators, required to identify the particles whose energy spectrum should be correlated with an appropriate representation of P_t'.

In Segal (1971) it was observed that the acausality of conformal space-time could be remedied in the present connection through its replacement by its locally identical universal covering space; this covering has an infinite number of sheets, and is thereby suggestive of large-scale macroscopic phenomena, e.g., those of large-distance astronomy. Theoretical exploration of this infinite-sheeted covering space from the standpoint of more objective notions of observer, clock, and rod, in the conservative spirit of the foundations of geometry, leaves little doubt that the appropriate notion of time is different in the large from the special relativistic one, although microscopically nearly identical to it. This new time τ is identical with that with the new energy P_t' just mentioned is associated (i.e., $P_t' = -i(\partial/\partial\tau)$ essentially); it leads directly to physical implications which can be checked against observation—in astronomy, rather than in microphysics. This book details the basic theory involved; the astronomical implications; and their successful and interesting confrontation with observation.

4. Natural philosophy of chronogeometric cosmology

As earlier indicated, when the underlying space–time is linear (i.e., a linear vector space), it is rather well established, in a variety of ways, that an appropriate general starting point for a notion of causality is a given closed convex cone in the space–time manifold, representing physically the "future."

The general process of nonlinearization of a theory, that is, the transference to an arbitrary sufficiently regular n-dimensional manifold of a theory established for n-dimensional vector spaces, then suggests as a starting point for causality considerations in a (nonlinear) manifold a structure consisting in the assignment to each point of a closed convex cone in the tangent space at the point.

In physical terms, this is the specification of infinitesimal future, i.e., the set of all future directions at the point.† A given linear hyperbolic partial differential equation provides a particular such assignment, which we shall call a causal orientation. However, from foundational and philosophical viewpoints, there is no particular reason to assume that the causal structure of space–time arises in this way from a hyperbolic equation; rather, hyperbolicity should be an expression of compatibility of propagation with the given causal orientation.

It thus appears—from other standpoints as well—that a natural starting point for the study of temporal order and associated developments consists of a smooth manifold together with a causal orientation, in the sense

† This specification can be regarded as a mathematical formulation of "time in its most primitive form," in the sense of Maxwell, who wrote: "The idea of Time in its most primitive form is probably the recognition of an order of sequence in our states of consciousness." ("Matter and Motion," London, 1877, reprinted Dover, New York.) One of our aims will be to show that in space–time manifolds with realistic features, this apparently minimal physical structure already suffices to determine much of the physical interpretation—the notions of clock, rod, energy, momentum, etc.

of a smooth assignment to each point of the manifold of a nontrivial closed convex cone in the tangent space at the point. Initially, such a causal orientation might appear too qualitative for technical cogency, in comparison with the familiar differential-geometric structures. However, the notion of causally oriented manifold appears to be one of considerable economy and naturalness for the analysis of temporal order, both from a philosophical and a mathematical standpoint.†

All this is not to say that it would not be interesting or possible to have a treatment of causally structured spaces which did not depend on the local smoothness of the space. (For example, there is no essential difficulty in extending the notion of causal orientation to the genre of arbitrary topological spaces.) But until one has a better understanding of causality matters in the more accessible context of smooth manifolds, it might well be mathematically foolhardy as well as physically irrelevant to attempt to obtain results for such general spaces comparable to what may be expected to be available in the smooth manifold case.

Indeed, even the concept of causally oriented manifold is highly qualitative, from a physical standpoint. Although timelike and spacelike directions in the manifold are determined, the notions of time and space, in the precise senses associated with the ideas of "clock" and "rod" are elusive in this context. It is difficult to see how physically cogent results can be obtained without a "clock," or an equivalent structure. For in physics one observes, largely, the change in the state of a system from one time to another. To give meaning to a statement concerning the change of state, one needs an objective parametrization of states which is time-independent, in addition to an objective notion of time itself. Moreover, the key physical notions of energy and scattering appear uniquely and effectively definable only when there is a notion of temporal invariance.

In an arbitrary causally oriented manifold, there may well be many different types of "world lines" (mathematically, maximal chains relative to temporal precedence as order relation); and different, possibly topologically distinct or causally inequivalent spacelike surfaces (i.e., submanifolds such that neither of any two of its points precedes the other, and which are maximal with respect to this property). The notion of a clock as an additive functional on intervals of world lines is conceptually acceptable, but is much

† One rough indication of the cogency of a causal orientation is the existence of evidence for, and the lack of evidence against, the conjecture that the automorphism group of a causally oriented manifold is finite dimensional (i.e., a Lie group in the classical sense), provided the cones in question are proper. Another is evidence that an analogue to the Alexandrov–Ovchinnikova–Zeeman theorem holds; any one-to-one transformation of a *globally causal* manifold of the indicated type is automatically smooth, and so a causal automorphism (cf. Choquet-Bruhat, 1971).

too limited to provide an adequate basis for correlation of the theory with empirical physics. The stronger assumption, that a hyperbolic pseudo-Riemannian structure is given in the space–time manifold, is likewise insufficient, e.g., to determine a fully viable notion of energy, whose precise definition and nature is still controversial in general relativity.

A conceptually natural way to introduce notions such as observer, clock, and rod, a way which generalizes special relativity and is closely related to elementary particle considerations, is to assume and exploit group invariance properties. The plausible and widespread if partially implicit idea that a certain temporal stability underlies the possibility of describing the dynamics of real physical systems is appropriately formalized by the assumption that the causal manifold in question admits a nontrivial class of "temporal displacements," these being automorphisms of the manifold (qua causal manifold) which carry each point into one which is either later or earlier than the given point. *Clock* may correspondingly be defined, essentially as a continuous one-parameter group of such temporal displacements. *Time* is then uniquely determined, within a scale factor, as the additive parameter t of this group, normalized (partially) so that $t > 0$ corresponds to a forward displacement (i.e., one carrying each point into a later point). Given any such clock, one may of course construct other clocks by conjugating the given one by other automorphisms of the causal manifold; physically, any such automorphism leaving a point fixed can be interpreted as a change in the frame of reference of an observer at the point. In a similar way the important although less fundamental notion of *rod* can be associated with the assumed homogeneity and isotropy of *space*. *Observer* then corresponds, in operational terms, to a splitting of the space–time manifold into respective space and time components, in such a way that the groups of the "clock" and "rod" act only on the corresponding component, the temporal action T_t being simply the transformation $\tau \to \tau + t$. Quantum mechanics relative to a given causally oriented manifold is naturally taken to involve a representation of the fundamental symmetry group of all causal automorphisms, i.e., causality-preserving transformations on the manifold. The *energy* for a given observer is then definable as the infinitesimal generator of the corresponding one-parameter group representing his clock. The *spatial momenta*, generalizing the usual linear and angular momenta, are correspondingly describable in terms of the generators of the "spatial displacement" group, consisting of those causal automorphisms that affect only the space component of the observer. Spatial homogeneity means that this group acts transitively; spatial isotropy means that the group acts transitively on the directions at any point. The assumptions of temporal and spatial homogeneity, and of spatial isotropy, are tantamount to the conservation of energy, of linear and angular momentum, respectively. Without these laws, the correlation of theoretical and empirical physics as they exist today would appear impossible.

The physical validity of these notions is confirmed by the observation that in special relativity the usual notions of observer, clock, etc. are in essential conformity with the general ones indicated here. Moreover, it follows from the structure of the causal automorphism group of Minkowski space earlier indicated that there are no other observers or clocks; thus, all observers are conjugate within the Lorentz group augmented by the group of scale transformations. Physically, it seems clear that the Cosmos is four dimensional, and that absolute simultaneity does not exist, i.e., no mode of communication or interaction has infinite velocity. Mathematically, these are readily formulated, the first assumption without change, the second as the assumption that the future cones contain no full straight lines (as they do in primitive Newtonian mechanics). Together with the existence of an observer, these seem to form a physically conservative and intuitive set of axioms for the Cosmos.

It should be interesting to determine all mathematical cosmos in this sense, particularly those which approximate locally, in the vicinity of a point, the Minkowski cosmos. But already a certain ambiguity in the Minkowski cosmos itself appears. While globally all observers are equivalent, locally this is not the case; the concepts of local observer, time, space, etc. can be introduced in entirely parallel fashion to the global concepts, by replacing the global transformation groups involved by local ones. The theoretical concept of local observer seems physically quite relevant since direct measurement of the entire Cosmos is impossible. Indeed, in Minkowski space, considered as a causal space–time continuum, there exist invariant local observers that are nonconjugate to special relativistic ones; and these local observers are applicable to regions which when scaled in accordance with physical parameters are far larger than those accessible to direct observation. The question of which of the local observers is physically correct is a real one; it cannot be eliminated by a mathematical transformation; while subject to various theoretical considerations, it must ultimately be weighed against observation, as we shall do later.

The question arises in particular of whether the same clocks are appropriate, in the sense of yielding a convenient notion of energy, including energy conservation, etc., at all distance levels of physics (or for all types of interactions). The conventional standard relativistic model is very well established at the middle distance levels, but its applicability at the extremes (i.e., extragalactic and fundamental particle physics) is largely a matter of extrapolation in the absence of any other established theory. Since dynamical theories primarily describe transitions from one approximate stationary state to another, such states at the middle distance level may appear complex in terms of inequivalent observers, and in particular nonstationary; conversely, simple descriptions of systems at the extreme distance levels may depend on the analysis of their dynamics in terms of states that are approximately stationary relative to clocks nonconjugate to any Minkowski clock. This abstract possibility is exemplified in the treatment of cosmology later in this book, in terms of the model briefly indicated in the next section.

All of the foregoing has been independent of dynamical assumptions, apart from the implicit one that observed fields and particles are appropriately described by functions defined on the Cosmos, with values in a suitable spin space; and that the equations determining temporal development should imply compatibility with the causal orientation in the Cosmos (in particular, finite propagation velocity), as well as enjoy invariance with respect to the causal automorphism group (or at least the subgroups earlier indicated). These requirements probably essentially imply that the dynamical equations should be partial differential equations which are hyperbolic relative to the given causal orientation (cf. the related discussion earlier).

5. The universal cosmos—sketch

There exists a cosmos that is locally identical to the Minkowski cosmos, and has a certain theoretical universality, in being apparently applicable in a fundamental sense at all distance levels. It may be described as the universal covering space of the conformal compactification of Minkowski space. For these reasons, and by virtue of applications made below to large-distance astronomy, it seems appropriate to designate this model as the *universal cosmos*. Its essential ideas were summarized in a preliminary account in Segal (1972).

The mathematical origin of this cosmos may be briefly indicated as follows. As earlier indicated, conformal space $\bar{M}$, obtained from Minkowski space M roughly by the adjunction of a light cone at infinity, is highly symmetrical, but is acausal. Being non-simply-connected, it admits nontrivial coverings, which are locally identical to $\bar{M}$, and hence locally Minkowskian. The finite coverings of $\bar{M}$ are likewise acausal, but the universal covering manifold $\tilde{M}$, is globally causal with respect to its inherited causal orientation, and defines an admissible mathematical cosmos. The space–time conformal group G acts only locally on $\tilde{M}$, but its universal covering group $\tilde{G}$ operates globally on $\tilde{M}$. Both the covering of $\bar{M}$ by $\tilde{M}$ and that of G by $\tilde{G}$ are infinite-sheeted, and indeed the group $\tilde{G}$ is not a linear group. The center of $\tilde{G}$ is $Z_2 \times Z_\infty$; and the Z_∞ component precludes a faithful finite-dimensional linear representation; however $\tilde{G}/Z_2$ acts faithfully, as a group of conformal transformations, on $\tilde{M}$.

The universal cosmos $\tilde{M}$ is locally identical chronogeometrically to

Minkowski space, and is essentially[†] the only other cosmos with this property enjoying physically natural symmetry and causality properties. Its validation as a realistic model is discussed below in terms of quantitative applications at the extragalactic level of distance. The physical interpretation is fixed by the distance scale, which may be determined from redshift measurements; a convenient equivalent physical constant may be described informally as the radius R of the universe. Three fundamental physical units are determined in a geometrical way; this is impossible in Minkowski space. $\tilde{M}$ is invariant not only under the 11-parameter Lorentz group extended by scale transformations which acts on Minkowski space, but the full 15-parameter conformal group (more precisely, universal covering group thereof); and any two global physical observers are conjugate with respect to this group. The one-parameter subgroup of this group representing temporal evolution—again unique within conjugacy—is essentially distinct from, i.e., nonconjugate to (within the causal automorphism group) that in special relativity. However, as $R \to \infty$, the universal cosmos deforms locally into Minkowski space, and the universal covering group of the conformal group deforms correspondingly into the Lorentz group together with scale transformations, four of the generators deforming into zero; and arbitrarily large bounded regions in Minkowski space can be approximated arbitrarily closely by the universal cosmos, by taking R sufficiently large. In particular, the universal energy deforms into the special relativistic energy as $R \to \infty$. One thus obtains a particularly concrete form of deformation of one Lie group into another, involving in addition a type of deformation of certain representations of one group into representations of the other.

For any given global observer O on the universal cosmos, and any point P in the cosmos, there is a unique local relativistic observer O'_P (and corresponding Lorentz frame) which is tangential to O at P; and O and O'_P agree near P within terms of third and higher order in the distance from P. Thus O'_P is locally nearly P-independent; however, if Q is remote from P, O' and O'_P are physically quite different; their Lorentz frames are related by the product of a scale transformation with a Lorentz transformation. In particular, the Lorentz frame of Q is in motion relative to that of P, which may be regarded as a virtual Doppler effect; however, from the standpoint of the globally more fundamental universal time, the situation is static.

[†] An open orbit in $\tilde{M}$ under the action of $SO(2, 3)$ enjoys the most crucial properties; but the regions of influence of compact regions in space may ultimately become noncompact. In any event, the predicted relations between the primary observable quantities (redshifts, magnitudes, number counts, etc.) would not differ from those for $\tilde{M}$. The orbit decomposition of $\tilde{M}$ under $SO(2, 3)$ was determined by B. Kostant, who noted also the existence of invariant Lorentz metrics in each of the two open orbits. The global causality of this space was noted by Wigner (1950), and it is naturally included in the list of Tits (1960).

The natural energy operator $-i(\partial/\partial\tau)$ for the universal cosmos is however not scale covariant†; built into its structure is a fundamental length, the radius of the universe. On the other hand, measurements of microscopic phenomena taking place wholly within a laboratory (i.e., excluding gravitation and redshifting phenomena) are so far as is now known, and may well be in fundamental principle, *scale covariant*, i.e., based on units of time and distance that are wholly conventional. This would suggest that the observable representing local measurement of frequency is not $-i(\partial/\partial\tau)$, but rather the conventional, scale-covariant operator $-i(\partial/\partial t)$, which as it develops is precisely the scale-covariant component of $-i(\partial/\partial\tau)$, and locally unobservably different from it (in natural units). Despite the very small local difference between these operators, their noncommutativity in the large implies that the relativistic energy operator is not at all conserved under universal propagation over a lengthy period of time. It is in fact a purely mathematical deduction that the apparent frequency of light, propagated in accordance with Maxwell's equations by universal-time displacements, is shifted to the red.

6. The chronometric redshift theory

More specifically, the redshift z is found to vary with the distance r from the point of emission in accordance with the law: $z = \tan^2(r/2R)$, where R is the "radius of the universe." This is in itself not a relation between observable quantities; but a variety of relations between observed quantities, such as redshifts, apparent luminosities, number counts, and apparent angular diameters, are readily deduced from this law. These predictions from the theory have been found to be in much better agreement with actual raw observational data than would a priori have been expected for an astronomical theory. Some of the outstanding predictions, and their relations to observation, are as follows:

(1) For small r, z varies as r^2, in accordance with the observational analyses of Hawkins (1962) and G. de Vaucouleurs (1972), and as preferentially indicated by the complete sample of radio galaxies due to Schmidt (1972c), the list by Arakelyan *et al.* (1972) of Markarian galaxies at substantial redshifts, a sample of Seyfert-like Markarian galaxies studied by Sargent (1972), a small sample of N-galaxies treated by Sandage (1967), and rather definitely, the large sample of G. and A. de Vaucouleurs (1964). The very good fit of the chronometric theory to the m–z data for the de Vaucouleurs

† Analytically, a generator X of the fundamental symmetry group is scale-covariant if $[X, K] = X$, where K is the infinitesimal generator of scale transformations, $K = \sum_j x_j(\partial/\partial x_j)$.

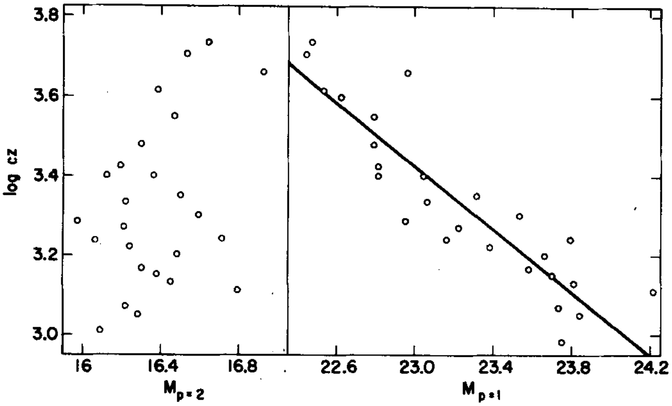

Figure 1 The redshift–absolute magnitude relations for the tenth brightest galaxies in bins of size 20 galaxies, ordered by redshift, included in the de Vaucouleurs tape.

All galaxies having m–z–θ data, 742 in all, were used. The absolute magnitudes for a square-law redshift–distance dependence, shown on the left, differ negligibly from those based on: (a) the maximum-likelihood power law fitted to the data; (b) the chronometric theory. Those shown on the right for the linear redshift–distance law have a trend that differs imperceptibly from that predicted by the chronometric theory, of slope 2.5, which is shown here as a solid line.

galaxies, together with the strong trend of the deviations from the expanding-universe hypothesis, is shown by Figure 1.

The apparent deviation from the law of the sample of bright cluster galaxies studied by Sandage is probably primarily a selection effect. This is evidenced by the extremely irregular $N(< z)$ distribution of this sample; this distribution is moreover highly deviant, even for $z < 0.04$, both from that to be expected in an expanding universe and the observational relation for all such galaxies with published redshifts, as compiled by Noonan (1973). Some of these circumstances are shown in Figure 2. It is evidenced also by an apparently very large dispersion in the intrinsic sizes of the galaxies. No objective statistical criterion for the sample has been published, and in fact its origin appears to be lost in early decisions of Humason. In addition, the deviation is augmented by the model-dependency of the apertures of observation which introduces a z-dependent trend, and possibly by the inherent tendency of an established theory to influence difficult observations.

(2) The apparent magnitude V depends on redshift z according to the relation

$$V = 2.5 \log z - 2.5(2 - \alpha) \log(1 + z) + c,$$

where α is the spectral index and c is a parameter representing the intrinsic luminosity of the source; corrections for absorption, aperture, intrinsic

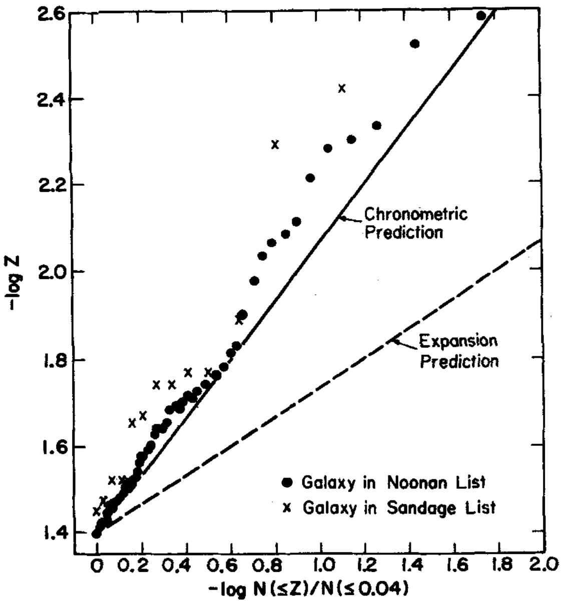

***Figure 2** The N–z relation for two samples of brightest cluster galaxies, in the range $z < 0.04$.*

Shown are 56 galaxies from Noonan (1973) and 13 galaxies from Sandage (1972b), Table 2. The $N(<z)$ curve for the Sandage sample differs even more from the expansion prediction than from the Noonan curve, even for very small redshifts. The respective values of $\partial \log N(<z)/\partial \log z|_{z=0}$, which should be unaffected by observational magnitude cutoff by virtue of the evaluation at $z = 0$, are ~ 1 for the Sandage sample, 1.45 for the Noonan sample, 1.5 for the chronometric prediction, and 3.0 for the expansion prediction.

motion of the source, if any, are not included. In particular, as z increases from 0.4 to 4.0, V should increase by ~ 1.1 mag (for $\alpha = 1$; for $\alpha = 0.7$ the increase is ~ 0.7 mag), as contrasted with the increase of ~ 5 mag on a typical expansion-theoretic hypothesis.

Quasar observations are in excellent agreement with the new law, with an overall dispersion for all quasars of less than 1 mag, and of 0.3 mag for the "locally brightest" fifth of the quasars, where "local brightness" is a model-independent measure defined as the difference of the magnitude of the object and the average magnitude of the six quasars at the nearest redshifts (three at greater and three at lower redshifts than that of the object). These dispersions are much less than those from the Hubble law (by more than a factor of 3 in the case of the last sample, of 32 quasars). In all substantial previously identified samples of quasars, including complete samples of radio sources due to Schmidt (1968) and Lynds and Wills (1972) and of optically selected quasars due to Braccesi *et al.* (1970), the chronometric dispersion is less than the Hubble-law dispersion, generally by factors of the

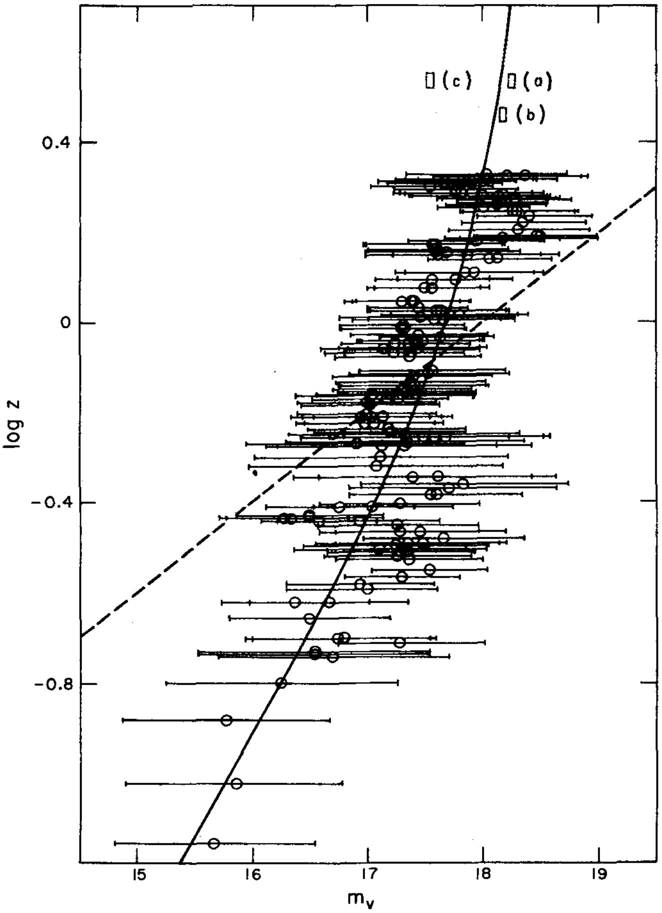

***Figure 3** The smoothed redshift-magnitude relation for quasars.*

○, average magnitude of seven quasars at approximately the same redshift. The bar indicates the unbiased probable error of the group. Chronometric (—) and expansion (---) ($q_0 = 1$) predictions, with zero points fitted to the average magnitude of the sample, include all quasars in the list of DeVeny *et al.* (1971) having unquestioned data. The three quasars of maximal presently published redshifts are denoted as (a), (b), and (c). On the expansion theory, (c) (at redshift 3.53) is ~ 50,000 times as bright as the average nearby galaxy, and is moving with > 0.9 times the velocity of light. On the chronometric theory, (c) has about the same intrinsic brightness as nearby bright galaxies, and need not be moving at all.

order of 2. This is true also of radio and infrared luminosities, when available. For the Einstein–de Sitter model, the disparity in dispersion is typically ~ 10% greater. The nearly optimal fit of the chronometric theory to the uncorrected quasar data, together with the pronounced deviation of the Hubble law, is shown by Figure 3.

(3) The redshift distribution of quasars that would be expected from a spatially uniform distribution in the present model conforms extremely well to observed redshift distributions, whether for the complete list of known quasars given by DeVeny *et al.* (1971), for complete samples limited in various magnitudes, or over specified redshift intervals. In contrast, in the expansion theory, strong evolutionary effects are required to explain the distribution, as documented by Schmidt (1968, 1972b), Lynds and Wills (1972), and others, and further strongly confirmed in the present analysis employing systematically the Kolmogorov–Smirnov test for a variety of substantial quasar samples, as well as Schmidt's V/V_m test when applicable. The latter test rejects the hypothesis of spatial homogeneity for the Peterson and de Vaucouleurs lists of galaxies on the expansion hypothesis, but accepts the chronometric hypothesis at substantial probability levels.†

(4) Perhaps the major anomaly in quasar phenomena from the expansion-theoretic standpoint, their apparent unprecedently large energy output, is fully resolved by the change in the relative luminosities of quasars and galaxies implied by the chronometric theory. When analyzed on this basis, observations indicate that the average quasar is ~ 0–1 mag fainter than the average brightest cluster galaxy, and quite comparable to an average N-galaxy, or average Seyfert galaxy. The hypothesis that quasars are the cores of certain bright galaxies whose outer regions are not seen at larger redshifts is somewhat supported by this and other consequences of the chronometric hypothesis.

(5) A second major anomaly regarding quasars, the apparent relative cutoff in quasars above $z \sim 2.5$, is well resolved without the use of fairly drastic hypotheses, hardly subject to independent verification, required to this end in the expansion theory. Specifically, the theory predicts that for any object uniformly distributed in space, the expected number in the redshift range $2.25 < z < 3$ will be $\sim 8\%$ of that in the redshift range $0 < z < 2.25$. In the light of probable and partially documented changes in the spectra of quasars at higher frequencies and spectroscopic selection effects making their identification more difficult at higher redshifts, this is quite compatible with quasar observations; the corresponding figure of $\gtrsim 40\%$ on the unembellished expansion theory is not.

(6) Apparent superluminal velocities of large redshift objects are eliminated on the chronometric hypothesis by the reduction in the theoretical distance to large redshift objects, typically by an order of magnitude.

† The Peterson galaxies were observed at given expansion-theoretic apertures, and their magnitudes consequently require correction to chronometric apertures for a valid test of the chronometric hypothesis, just as they must be used as given for a test of the expansion hypothesis. The de Vaucouleurs galaxies were measured at apertures determined by observational rather than theoretical criteria, and were used in the statistical analysis without correction for both hypotheses.

(7) The $N(m)$ relation for quasars is very well fitted by the chronometric curve for a single luminosity class, convolved with a normal law luminosity function of dispersion equal to that observed. This explains the apparent cutoff in quasar identifications at faint magnitudes ~ 20.5 noted by Bolton (1969), Braccesi *et al.* (1970b), and others, a phenomenon that appears anomalous from the expansion-theoretic standpoint. The expansion-theoretic $N(m)$ relation for a complete sample is in fact in disagreement with the observational relation for quasars in the DeVeny list even when limited to relatively bright magnitudes such as 18.0.

(8) The index $-\partial \log N/\partial \log S$ for a single luminosity class and spectral index $\alpha < 1$ is ~ 1.5 for very bright sources but eventually becomes infinite as S decreases, according to the chronometric theory, following which it drops to zero. When convolved with a luminosity function of about one decade width, the theoretical curve rises quite moderately above 1.5 for fairly bright sources and eventually declines to a value $\lesssim 1$, in qualitative agreement with observations of Pooley and Ryle (1968) and Kellermann *et al.* (1971).

(9) Predictions regarding the angular diameter in relation to redshift are in satisfactory agreement with the data on double radio source quasars as given by Miley (1971). The angular diameter measured here is properly identifiable with the metric diameter treated theoretically, and all dispersion in the observations attributable to variation in the redshift, as measured in a model-independent fashion, is removed by the chronometric relation. The situation is similar as regards double radio galaxies listed by Legg (1970).

The theoretical deductions involved here are obtained in a quite direct and objective manner, and involve no free parameters, other than the distance scale, which is determined by R. Further observational confirmation and predictions, standard statistical significance tests, and a discussion of corrections and selection effects, are given in detail in Chapter IV.

7. Theoretical ramifications; the cosmic background radiation

We close this section with comments on theoretical aspects which seem likely to be raised in the minds of certain groups of readers. First, on the general mathematical side, the question arises of whether the methods involved here are ad hoc and entirely particular, or whether the theory can be understood as an individual instance of a general type of theory. Indeed, the latter is the case. There is an analogous theory for general classes of causally oriented manifolds, in relation to corresponding flat manifold tangential to them. The Cayley transform being causal, there is no chronogeometric local distinction between the two different manifolds, but in the large there are topological and other differences. Our basic physical assumption is that local measurement (e.g., of frequency) is in terms of the flat tangential causally oriented manifold—roughly that microscopic observation is based on a Minkowski clock; but that true free temporal evolution is as given on the global curved manifold, i.e., runs on the universal clock. Thus, the universal energy of a free wave or particle is conserved, but the apparent frequency of a photon emitted from an atom changes noticeably after a long time because it is stationary with respect to the Minkowski rather than universal clock. Such a limitation on local phenomena and measurement is a priori plausible because of the absence of any absolute distance scale for measurements of entirely microscopic phenomena. Normalization of the commutation relations of quantum mechanics fixes the values of $\hbar$ and c as unity but leaves unspecified one of the fundamental units. In the universal cosmos the natural convention $R = 1$ fixes the distance scale and completes the specification of units.

The causal manifolds involved are all globally hyperbolic and have defined on them analogues of Maxwell's and Dirac's equations. In addition they are extremely symmetrical, being universal covering manifolds of Shilov boundaries of classical Siegel domains, whose automorphism groups are closely connected with the presently relevant physical symmetry groups. In the case of dimension 4, however, there is essentially only one known instance of the general theory, viz. the universal covering manifold of the conformal compactification of Minkowski space in relation to Minkowski space (or equivalently, the universal covering group of $U(2)$ in relation via the Cayley transform to the Lie algebra of $U(2)$ as identified with the 2×2 Hermitian matrices with their usual ordering). This seemingly purely mathematical aspect has in our view a certain physical relevance, in diminishing the selection effect involved in formulating any theory designed to explain previously observed phenomena, and thereby enhancing the significance of whatever agreement is found between theory and experiment. Indeed this is in essence no more than the broadly recognized distinction between correlating data by curve-fitting and the like, and the formulation of a true theory based on general ideas and principles.

Second, the relation to dynamical theories—general relativity, the question of the origin and "age" of the universe, elementary particle dynamics—is likely to be raised. Since conformal space is conformally locally identical to Minkowski space, the present model for space–time stands in essentially the same relation to general relativity as does special relativity. As a variant of special relativity, it is essentially a purely kinematical structure, on which one is free to impose interactions as in the case of Minkowski space. In particular, the ideas of general relativity carry over bodily and its applications to local gravitational phenomena (e.g., within a galactic cluster) appear unaffected.

On the other hand, material dynamical content resides in the postulate concerning local observations of dynamical quantities, to the effect that these are represented not necessarily by generators of true, global symmetries, but rather by generators of corresponding symmetries in the tangential flat model. While angular momenta, for example, are unchanged, the energy and linear momenta are altered in essential ways. The true energy is no longer represented by $-i(\partial/\partial t)$, but by an operator which while extremely simple and natural from the standpoint of universal space, appears complicated in terms of Minkowski space. It may be put in the form $-i\,\partial/\partial\tau$, where τ is the universal time. This differs from $-i\,\partial/\partial t$ by an operator that is virtually negligible up to galactic distances, and so as an interaction Hamiltonian should not be responsible for any readily observable microscopic processes. Moreover, as the radius of the universe becomes infinite, this interaction operator $i(\partial/\partial t - \partial/\partial\tau)$ tends to zero, in accordance with Minkowski's concept of limiting case. It is relevant to note that the difference between the universal and special relativistic energies, the "superrelativistic energy," is represented by a positive Hermitian operator in all physical (hence positive-energy) representations of the fundamental symmetry group (for example, in the representation defined by Maxwell's equations).

The redshifting process may be regarded in the chronometric theory as a conversion of relativistic into superrelativistic energy, which inevitably accompanies the delocalization of a photon wave function, the superrelativistic energy being negligible for a localized photon. The conversion becomes in classical theory total at redshift $z = \infty$, but at low frequencies and high redshifts the quantum-theoretic dispersion in frequency (which arises from the noncommutativity of the operators representing the relativistic and superrelativistic energies) will significantly broaden the spectrum of the radiation. It should then appear as background radiation, the totality of which would be in a state of equilibrium, assuming that the temporal homogeneity of the universal cosmos is dynamically as well as kinematically valid. By conservation of energy and maximization of entropy, this radiation should have a blackbody spectrum, as is consistent with observations of the microwave background, which is thereby theoretically predicted.