1975

Does the Causal Structure of Space-Time Determine its Geometry

David B. Malament

DOES THE CAUSAL STRUCTURE OF SPACE-TIME DETERMINE ITS GEOMETRY

A thesis submitted to the Faculty of The Rockefeller University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

by

David B. Malament

April 1, 1975
The Rockefeller University
New York

ACKNOWLEDGMENTS

Most of the ideas in this essay arose in conversation with John Earman, Clark Glymour, and Tony Martin. I am grateful for their assistance and encouragement. I also want to thank Robert Geroch for his help during the final stages of my research. He patiently criticized flawed versions of my proof of Theorem 3.28.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION -- CAUSAL THEORIES OF TIME AND SPACE-TIME

This essay explores the following problem: to what extent is it possible to construe the causal structure of space-time as basic and from it to reconstruct the topological and metrical structure of space-time? The problem is first examined within the context of Minkowski space-time (Chapter II) and then generalized to the class of space-time models considered in the general theory of relativity (Chapter III).

Much of the essay is technical. But it is motivated by philosophical debate over the viability of causal theories of time and of space-time. By way of introduction, this initial chapter presents a brief discussion of these "causal theories." Nothing approaching a thorough analysis is intended. Rather the goal is merely to state what a causal theory time or space-time is supposed to be, outline several objections that have been raised against such theories, and most importantly, suggest why interest in them leads to the questions considered in Chapters II and III. This should set the stage.

Put bluntly, the root issue is this: in talking about the temporal or spatio-temporal structure of the universe just what are we referring to? On one account time or space-time is to be thought of as a non-material "scaffolding" with which the universe comes equipped, against whose background the events and processes of physics unfold. Events stand in spatio-temporal relation to one another by virtue of their relative positions on this utterly pervasive scaffolding. So, for example, when we say that the emission of a photon precedes its absorption we are referring to the relative position of the "temporal moments" or "space-time points" at which the two events occur; we are not referring to the internal constitution of these events or any physical interaction (or possible physical interaction) between them. Furthermore, when we ascribe topological and metrical structure to time or space-time (saying, for example, that space-time is non-Euclidean in the presence of massive celestial bodies) we are really attributing this structure to the scaffolding itself; we are not, at least in the first instance, reporting on the behavior of clocks, rigid-rods, light rays, etc.

Many people have balked at this conception thinking it unintelligible or at least unnecessary. They believe in events, of course, but are loath to countenance some "thing" in which or against which they occur. Thus they contend that if talk about temporal or spatio-temporal structure makes any sense at all, it must be conceived as but

abbreviated talk about patterns of relations among events. Time or space-time for them is nothing above and beyond (is constituted by) these patterns of relations in much the same way that a "family tree" is nothing more than a pattern of genealogical relations among the members of a family.

Such an account is called a relational theory of time or space-time. The basic relations among physical events entering into a relational theory may or may not be themselves spatio-temporal. A causal theory of time or space-time is a species of relational theory in which the primitive relations are "physical" but not spatio-temporal. Thus the causal theorist not only denies the existence of space-time points populating the universe (-the stuff of which non-material scaffolds are built-) as do all relationalists, but he goes on to deny the existence of any irreducibly temporal or spatio-temporal relations between events as well.

Originally the causalist's primitive relation was that of causal influence or possible causal influence of one event on another. The goal of the causal theory of time was to use this relation to give an analysis of the temporal relation "earlier than." Leibniz entertained a theory of this form and so apparently did Lechalas.

Recent versions have taken different "causal relations" as primitive and have undertaken to analyze different temporal (or spatio-temporal) relations in terms of them. The development from Reichenbach through Carnap, Mehlberg, Grünbaum, van Fraasen, Winnie et al is a convoluted story with ins and outs, objections and revisions. (Critical historical surveys can be found in Lacey [11], and van Fraasen [19]).

There seem to be two reasons for the resurgence of interest in causal theories of time following the emergence of Einstein's special theory of relativity. For one thing the theory itself seemed to support the idea that the temporal relations "earlier than" and "simultaneous with" are defined by causal relations. This interpretation drew support from Einstein's own operationalist presentation of the theory written when he was still heavily under the influence of Mach. Reichenbach certainly read him this way as is clear from his essay in the Einstein-Schilpp volume.

> ...The concept of causal chain can be shown to be the basic concept in terms of which the structure of space and time is built up. The spatio-temporal order thus must be regarded as the expression of the causal order of the physical world. The close connection between space and time on the one hand and causality on the other hand is perhaps the most prominent feature of Einstein's theory, although this feature has not always been recognized in its significance. Time order, the order or <u>earlier</u> and <u>later</u>, is reducible to causal order; the cause is always earlier than the effect, a relation which cannot be reversed. That Einstein's theory admits of a reversal of time order for certain events, a result known from the relativity of simultaneity, is merely a consequence of this fundamental fact. Since the speed of causal tramsmission is limited, there exist events of such a kind that neither of them can be the effect of the other. For events of this kind a time order is not defined, and either of them can be called earlier or later than the other.

Second, the special theory of relativity brought with it the prospect of a causal theory of <u>space-time</u>, not merely one of time. Previously it had been assumed that time and space had their separate and distinct relational bases. Causal or signal relations would suffice for the one; but an analysis of the other would require, instead, congruence relations among rigid rods or something along those lines. Now, it appeared, signal relations would suffice for

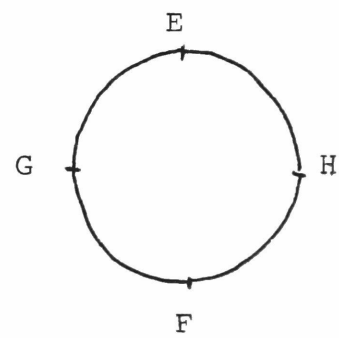
the whole job and this would be in accord with Minkowski's conception of a single, unified space-time structure of which temporal structure is but a (non-unique) projection. In fact, as early as 1914 the British mathematician Robb [16] showed that the geometry of Minkowski could be built up from a single two-place "after" relation. (Much more about this in chapter II.)

That a unified causal theory of space-time was impossible within the framework of classical (Newtonian) space-time structure can be demonstrated in the following way. Within that classical framework, because there is no bound to the velocity with which causal signals can propagate, the relation of non-signal connectibility factors the space-time into a one-parameter family of "simultaneity classes." Causal relations, however, provide no information whatsoever about the structure of these classes. In particular, it is not possible to define the Euclidean congruence relation in terms of them. In contrast, Robb was able to give an explicit definition of the Minkowski congruence relation in terms of his single causal priority relation.

Reichenbach's original "causal definition" of the relation "earlier than" was of the following form [14]: event E_1 is <u>earlier than</u> event E_2 if and only if it is possible that there be events S_1 and S_2, where S_1 is the cause of S_2 and S_1 coincides with E_1 while S_2 coincides with E_2. Actually the definition was more complicated allowing for the interpolation of a finite causal chain of events between S_1 and S_2, but that is secondary. The sense of coincidence intended was that of spatio-temporal coincidence.

Several objections were raised to Reichenbach's analysis and subsequent attempts at causal theories of time or space-time sought to circumvent them. First it was maintained that Reichenbach's asymmetric relation 'S_1 is the cause of S_2' is unavailable to a would-be causal theorist. According to the objection, the relation is itself of temporal character, there being nothing to distinguish initial and terminal events in a causal interaction or signal propagation except for their temporal order. Reichenbach had recognized the problem and had attempted to distinguish cause from effect (without prior recourse to spatio-temporal relations) with his "mark-method." But, as Mehlberg and Grünbaum argued, the mark-method just did not work.

Reichenbach and those that followed him accepted the objection. Subsequent versions of the causal theory (including Reichenbach's own revision in [15]) adopted symmetric relations of causal connection and separated-off the goal of giving a causal account of the "direction of time." They contented themselves with the project of giving an analysis of temporal order up to a specification of the direction of time. For the later Reichenbach this meant an analysis of the relation of "temporal betweeness." Grünbaum wanted a causal theory of time that did not prejudge the issue of whether time is "open" or "closed" i.e. whether it has the order structure of the line or the circle. He took it as his goal not just to give an analysis of the three place betweeness relation (-which collapses when time is "closed"-), but also that of the four place relation of "temporal separation." This relation holds between events E, F, G, and H, intuitively, if their relative position in time is as in the figure.



A second objection to Reichenbach's definition was that it made use of the relation of spatio-temporal coincidence between events. It thus failed, on this ground too, to be a full causal account of temporal priority, i.e. one given completely in non spatio-temporal terms. A recent version of van Fraassen [19] seeks to avoid this circularity by giving a causal analysis of spatio-temporal coincidence between events.

This move by van Fraassen has not been discussed in the literature (except by Earman [4] in a different context) and so it will not be inappropriate to make a short detour and examine it here more closely. In his set-up the relation of "causal connectibility" is taken as primitive. As a primitive it is never defined, but in an informal explication van Fraassen states: "X is causally connectible with Y if and only if a signal emitted with X would arrive coincidently with Y, or conversely." ([19], pg. 196) Using the relation he gives the following formal definition: "two events are <u>coincident</u> if and only if: any event is causally connectible with the one if and only if it is causally connectible with the other." ([19], pg. 184).

One is naturally inclined to object that the problem of circularity remains. The relation of spatio-temporal coincidence is built into the relation of causal connectibility. But it is in terms of the latter that the former is to be defined. van Fraassen anticipated the objection, but insists that it misses the mark. He maintains that a cross definition of the relations is all that one needs in

providing a causal theory of time or space-time. By way of explanation he says the following ([19], pg. 195):

> ...Our position here is that within natural language there is no defining-defined hierarchy and that there is no such thing as "the" meaning of a term, although there are meaning relations (inclusion, equivalence) among terms. Within a specific formulation, some terms are defined and others undefined, but the status of being defined is not invariant under transitions to other formulations of the theory. The claim of the causal theory of time is not that spatio-temporal terms are defined, but that they are definable, in terms of causal connectibility... Formulations of theories are, in a sense, artificial, since they must rely on a choice of primitive terms (and of axioms) that is to some extent arbitrary. But a dictionary (say, of English) is circular and should be for in natural languages there are no inherent definitional hierarchies. (italics in original)

The point seems to be that with terms like 'causal connectibility' and 'coincidence,' as with the basic terms of natural language, there is no clear subordination of one to the other as primitive and defined. It is sufficient that a formal definition of coincidence can be given in terms of causal connectibility. That a definition could also be given the other way around is harmless; so is the fact that one's informal interpretations of the relations reveal how they are intertwined.

But this will not do and it is hard to understand why van Fraassen did not recognize as much. The causal theorist is not just claiming that spatio-temporal relations are definable in terms of some formal two place relation called "causal connectibility." This is of no support to his reductionist ideals whatsoever if that relation is itself spatio-temporal in character. Nor will it do to deny that there is any such thing as the character of the relation (i.e. spatio-temporal or not) as there is no such thing as the meaning.

For then it is not clear what if anything is left to the reductionist thesis in the first place.

One can avoid a long debate with van Fraassen over the nature of "reduction" by simply turning his defense against him. If his defense against the charge of circularity were adequate it would serve equally well to justify the use of an asymmetric relation of causal connectibility. At arm's length he would have a snappy, no work solution to the problem of giving a causal theory of the direction of time, one which he somehow overlooked. If critics (like Mehlberg, Grünbaum, or van Fraassen [19]) objected to his use of this relation as primitive on the grounds that it is itself spatio-temporal in character (since it has built into it the temporal relation "later-than"), van Fraassen would only have to remind them that he also gives a formal definition of "later than" in terms of the asymmetric causal connectibility relation. After all, there is no such thing as "the" meaning to relations like "later than" and the asymmetric causal connectibility relation, etc.

It seems then that van Fraassen does not succeed in giving a "causal definition" of spatio-temporal coincidence. Without the use of that relation of spatio-temporal coincidence the causalist's program cannot succeed. But there are more important objections to be made which stand even if the causal theorist is conceded both that relation and the asymmetric relation of causal connection or connectibility. At least three such can be culled from the critiques of Lacey [11], Earman [4], and Sklar [18].

The first objection concerns the very motivation for a causal

theory of time or space-time. As mentioned earlier, the relationalist balks at the idea of an irreducible spatio-temporal structure to the universe. The prospect of non-material scaffoldings, to use that metaphor again, is too much for him. But it certainly is not clear that this conception is any more objectionable than others which play important roles in theoretical physics such as, for instance, a field of force. Furthermore, there are many physicists who are relatively comfortable with the concept of space-time but not so with that of causation, causal signal, or localized event. The latter are by no means secure, ultimate, first concepts of physical theory. Thus even the desirability of the causalist's reduction program, quite apart from its success, may be called into question.

A second basic line of objection does challenge the possibility of its success. It contends that in the end the causalist's primitive relations are at least in part irreducibly spatio-temporal, this quite apart from the problem of giving a causal analysis of spatio-temporal coincidence. As Earman puts it: "...there is the suspicion that 'causal connectibility' is just another name for a spatio-temporal relation, a relation which must be understood in terms of spatio-temporal structure." ([4], pg. 74)

Sklar [18] gives an argument to this effect. Though the causal theorists speak of causal connection or connectibility, they are thinking quite explicitly of signal connection or connectibility.

van Fraassen, in fact, explicitly denies that he wants or needs any recourse to the "general notion of causality." ([19], 194) He does this by way of defense to the charge that the "general notion" is an unworthy primitive relation. In one informal explication of the non-modal relation he states: "'X is causally connected with Y' is equivalent to 'Either X and Y belong to the history of one and the same object, or belong to the history of one and the same signal, or are coincident with some pair of events thus connected." ([19], 194)

Now Sklar would argue that even without the last clause involving coincidence, the relation reveals itself as irreducibly spatio-temporal in so far as it uses the notion of "one and the same signal (or body)." Our implicit criteria of signalhood or bodyhood, the argument runs, require of a set of events that they exhibit spatio-temporal continuity. He further argues that if anything like a Humean analysis of self-identity and causation is correct, then spatio-temporal continuity of events will be an irreducible component of their causal connection, or connectibility. Humean analysis leads, that is, to what is "at least partially a spatio-temporal theory of cause. " ([18], 341)

> ...Is it plausible to claim that we directly apprehended the genidentity, causal continuity, and causal order of a set of events and only inferred to the temporal or spatio-temporal structure of the set? or is it, rather, the other way around: We directly apprehend certain spatio-temporal features of a set of events and on the basis of these features attribute, by inference or definition, to the set of events certain "causal" features such as the set being a set of genidentical events, being a causally continuous set, and being a causally ordered set. ([18], 339)

It may well be that both of these uni-directional accounts are terribly naive. But Sklar is not here insisting on the latter. He is rather calling into question the adequacy of the former.

The force of this second line of objection is to deny to the causal theorist his basic relation of causal connection or causal connectibility. As such it does not cut against a relational theory of time or space-time but only against a causal species of one. Even if it were accepted one might still maintain that there are events and spatio-temporal relations (eg. causal connectibility) among events and nothing else - in particular, no temporal instants or space-time points. This position requires that there be enough events to go around, that the universe be an "event plenum"; but that does not worry the relationalist. He is prepared to countenance "null events" or think of events merely as specifications of (possibly zero) values to assorted matter fields.

A third basic line of objection to causal theories of time or space-time applies even if the causalist (or relationalist) is conceded the relation of causal connectibility. The objection is that the reconstruction program, even if it works for Minkowski space-time, breaks down for some of the space-time models envisaged in the general theory of relativity. One of these, the argument goes, just may correctly model the spatio-temporal structure of our (i.e. the real) universe; if so, the causalist is stuck.

Earman [4] levels this attack against van Fraassen. He claims that the criterion of event coincidence that van Fraassen gives (cited above) breaks down, for example, in any space-time which admits

of closed causal curves. (This is of course a very different objection to van Fraassen's causal definition than the one previously given. That one, in contrast, applies even if space-time structure were Minkowskian.) In particular, it collapses in space-times like "rolled-up" Minkowski space-time (i.e. Minkowski space-time where for some $k>o$, the points (t, x, y, z) and $(t+nk, x, z)$ are identified for all integers n) in which all points are pairwise connected by a causal curve. That Earman's objection should be telling is striking since van Fraassen wants his theory to apply even if "time is closed." And if it closed anywhere, it is in rolled-up Minkowski space-time!

Earman's claim is clearly sound if one accepts what might be called the "standard interpretation of causal structure" in relativity theory. (This is discussed at greater length in chapter III.) Accordingly, whenever two points in a space-time are connected by a future directed causal curve, that curve itself represents the trajectory of a possible signal, or causal interaction - possible in the sense that a particle or signal traversing this path would be moving less than or equal to the maximal possible velocity. For suppose X and Y are events occuring at distinct points on a closed causal curve and suppose Z is any third event whatsoever. If Z is causally connectible with, say, X then it must also be causally connectible with Y via a linked causal chain. And conversely. So by the criterion of event coincidence, X and Y must be coincident. In rolled up Minkowski space-time, all events everywhere turn out to be coincident under the criterion. This is of course absurd. (Earman, incidentally, might just as well have pointed out that van Fraassen's "Postulate V" is false

in rolled up Minkowski space-time.)

van Fraassen wrote a reply [20] in which he defended himself against Earman's charges and denied any incompatibility between his "Postulate V" and the possibility that time be closed. Not surprisingly, his reply is based on a rejection of the "standard interpretation." The reply, again, seems clearly inadequate.

van Fraassen claims that as applied to, say, rolled up Minkowski space-time, there are two ways to interpret the relativistic principle that there is an upper bound to the velocity with which signals propagate:

> ...Under the first interpretation, the principle holds only locally: light signals are the fastest among signals in restricted spatio-temporal regions. Under this interpretation any two events are causally connectible, but may not be 'locally causally connectible'...It would seem that this is the correct approach if one wants to provide a philosophical foundation for the General Theory...
>
> In van Fraassen (1970), however, I chose the second interpretation, a global interpretation of the 'first signal' principle, which is that arbitrary pairs of events are not causally connectible. That rules out actual trajectories 'all the way around time.'...The interpretation of 'causal curve' must now be that, if this curve is given by a function f (t) = (x_t, y_t, z_t, t), for all t, then only its <u>proper</u> segments can be the paths of possible causal signals. ([20], 91-2, italics in original)

The first interpretation supports Earman's objection. The second is contrived and, even from van Fraassen's standpoint, leads to unacceptable conclusions. If a signal (say a light signal) can be sent from a point P_1 to P_2, and if coincident with its arrival another signal can depart P_2 for P_3, then van Fraassen would certainly want to hold the departure event at P_1 causally connectible with the arrival event at P_3. Indeed this would seem a paradigm case of causal connectibility, the kind one encounters in Einstein's discussions of

simultaneity. But he cannot do so and still hold on to his second interpretation - just take $P_1 = P_3$. Of course van Fraassen cannot avoid this objection by revising his new principle to state: only very proper segments of causal curves can be the paths of possible causal signals. We would only have to choose n points $P_1,\ldots,P_n$ in order on a closed causal curve so that $P_1 = P_n$ and for each i the segment from P_i to P_{i+1} is appropriately small. Nor may van Fraassen simply assert that there is an interval on any closed causal curve such that no events occurring at any two of its points, no matter how close, are causally connectible. Apart from the lack of justification, this move would undercut the possibility of giving a causal account of local spatio-temporal structure (of that region in which the interval occurs).

Earman's criticism that causal theories of time and space-time break down in a large class of models studied in general relativity is really the starting point for this essay. Suppose one does concede to the causal theorist the relation of causal connectibility or some other primitive causal relation. Just what can he do with it? Chapters II and III are devoted to this question. Part of the work is just to formulate precise questions which cut across the lines of particular theories. The other part, of course, is to answer them. Both tasks are accomplished by appropriating recent work by mathematical physicists (Hawking, Penrose,Geroch, Carter, et al) on the "causal structure" of space-time.

Earman notes that one can define a topology explicitly in terms of the asymmetric causal connectibility relation, the Alexandrov

topology, but that this topology is equal to the manifold topology only if the space-time is strongly causal. (The technical terms used here are defined in Chapter III.) First, one can ask whether there might not be some other topology definable in terms of this relation which is equal to the manifold topology under weaker conditions than strong causality. John Winnie [21] holds this question up to Earman in his recent paper. Second, one can ask whether the Alexandrov topology, or others which are definable from the asymmetric causal connectibility relation, are also definable from the symmetric relation. It is the latter, after all, which recent versions of the causal theory employ. Third, one can ask if there might not be some sense weaker than explicit definability in which the topology of space-time is determined by its "underlying" causal structure. This weaker sense, at least, might be realized in non-strongly-causal space-times.

Fourth, one can ask if the metrical structure of space-time can be fully or partially reconstructed from the causal connectibility relation (symmetric or asymmetric). Robb did show that the reconstruction is in some sense (actually two different senses) possible in the special case of Minkowski space-time. The question is whether his results carry over to the larger class of space-times considered in general relativity. If not, is there some other interesting sense in which metrical structure is determined by causal structure.

All these questions are posed with respect to the primitive relation of causal connectibility (asymmetric and/or symmetric). One can also ask, fifth, whether there is not a richer level of structure

determined by some other "causal relation" which might serve to determine spatio-temporal structure. In particular, one can repose the first four questions in the case where a kind of "genidentity" relation is taken as primitive. In such a set-up one starts with the point set trajectories of causal curves (i.e. future directed causal curves), rather than simply the fact of their existence or non-existence, and tries to use them in reconstructing topological and metrical structure. It certainly is not clear that Earman's objections stand if causal structure is construed in this more generous spirit. Indeed, as it turns out, from the class of future directed causal curves one can, in any space-time, always reconstruct topology uniquely.

CHAPTER II

THE CAUSAL STRUCTURE OF MINKOWSKI SPACE-TIME -- WHAT ROBB DID AND DID NOT PROVE

Minkowski space-time provides the paradigm case for causal theorists of space-time. As noted above, Robb [16] proved in 1914 that one can in a sense reconstruct its topological and metrical structure from its causal structure. But after all this time there still seems to be a great deal of confusion about what he did and did not prove (and could not have proven). Part of the problem is that Robb's results are buried in a terribly long book whose flat style of exposition makes it easy to lose the line of argument. More important, there are at least two different senses in which the geometric structure of Minkowski space-time is (at least partially) determined by its causal structure. One has to do with "explicit definability," the other with "categorical axiomatizability." In discussing Robb's work it is important to distinguish them, though this is rarely done.

The goal of this chapter is to give a unified presentation of Robb's two theorems (--one corresponding to each of the two senses--) and related results by Zeeman and Latzer on the causal structure of space-time. In some cases new, simpler proofs are given. A few elementary results on the definability of non-standard topologies and congruence relations

are also established. Hopefully the rather detailed (perhaps laborious) exposition will resolve some of the confusion to which I referred and also provide a clear basis for comparison when the causal structure of general relativistic space-times is considered in the next chapter.

A. Basic Definitions and the Interdefinability of Causal Relations

At the outset there is a question about how to characterize Minkowski space-time. It can be taken to be an inner product space $(M,(\ ,\ ))$ where M is a 4 (or n) dimensional vector space over the reals, and $(,)$ is an indefinite (non-degenerate) inner product on M of Lorentz signature; it can also be taken to be a space-time (M,η) where M is a 4 (or n) dimensional, smooth differentiable manifold diffeomorphic to Euclidean space E^4 (or E^n) and η is a flat (non-degenerate) Lorentz metric on M. Of course there is a natural identification of the one with the other, but strictly speaking things can be said about Minkowski space-time in the one guise that are not well formed with respect to the other. So for example, the claim that the topology of Minkowski space-time is explicitly definable in terms of causal relations does not make sense, again strictly speaking, under the first construction. On the other hand, the claim that parallelism between

vectors is so definable does not make sense under the second.

Having made the point I shall ignore it for the most part, though it will have to be recalled occasionally. I shall go back and forth from vector space structure to manifold structure and take for granted the natural identification. Also I shall consider Minkowski space-time in dimension $n (n \geq 2)$, even though only the case $n = 4$ is of physical interest presumably. For the cases $n \geq 3$ nothing in the proofs turns on dimensionality and so the added generality comes for free. The case $n = 2$ is quite different from all the others. But seeing exactly why some of the proofs break down in this case is instructive. That dimension 2 should be different from all others is not surprising since only here are there finitely many (two) rather than continuum many null times through every point. To keep track of things, all propositions will be marked $(n \geq 2)$ or $(n \geq 3)$ to specify the range of applicability.

The Lorentz signature of the inner product $(,)$ means that a basis can be chosen with respect to which $(u,v) = u_o v_o - u_1 v_1 - \ldots - u_{n-1} v_{n-1}$ if $u = (u_o,\ldots,u_{n-1})$ and $v = (v_o,\ldots,v_{n-1})$ in that basis. Such bases and their corresponding coordinates are called standard. I shall assume one is fixed once and for all.

First some standard terminology and some not altogether standard notation is given. A vector v is timelike, (null

or *lightlike*, *causal*, *spacelike*) according as $|v| > 0$, ($|v| = 0$, $|v| \geq 0$, $|v| < 0$). Here $|v| = (v,v)$. Given two points a and b:

a is *temporally* (or *chronologically*) *prior* to b ($a << b$)
if $|b-a| > 0$ and $a_o < b_o$

a is *causally* *prior* to b ($a < b$)
if $|b-a| \geq 0$ and $a_o \leq b_o$

a is *null* (or *horismos*) *prior* to b ($a \rightarrow b$)
if $|b-a| = 0$ and $a_o \leq b_o$

The notation is that of Kronheimer and Penrose [10]. The first two relations are interpreted to mean that a signal can be sent from a to b traveling less than (resp. less than or equal to) the velocity of light. The third relation is taken to mean that a signal can be sent from a to b traveling at the velocity of light but that no signal can be sent which (at any time) travels less than that maximal velocity. These definitions depend on our prior choice of coordinates. Another choice of standard coordinates would systematically preserve or reverse the three. Hence the choice of one serves as a specification of "temporal orientation." Note that $a < a$ and $a \rightarrow a$, but not $a << a$. It is convenient to work with all three of these relations, but we may think of any one of them as basic since each can be defined from each of the others ($n \geq 2$):

$$a << b \Leftrightarrow \exists c,d[a < c < b \;\&\; a < d < b \;\&\; \neg(c < d) \;\&\; \neg(d < c)]$$

$$a \rightarrow b \Leftrightarrow a < b \;\&\; \neg(a << b)$$

(1)

$$a < b \Leftrightarrow \forall c[b << c \Rightarrow a << c]$$

$$a \rightarrow b \Leftrightarrow a < b \;\&\; \neg(a << b)$$

$$a < b \Leftrightarrow \exists c(a \rightarrow c \rightarrow b)$$

$$a << b \Leftrightarrow a < b \;\&\; \neg(a \rightarrow b) \quad .$$

Robb's relation 'after' is slightly different from all three. In the present notation, a after b obtains if $b \neq a$ and $a < b$.

Corresponding to each of the three causal order relations is a symmetric relation whose definition is independent of any prior specification of temporal orientation:

a is <u>temporally related</u> to b ($a\tau b$) if $a << b$ or $b << a$

a is <u>causally related</u> to b ($a\kappa b$) if $a < b$ or $b < a$

a is <u>null related</u> to b ($a\lambda b$) if $a \rightarrow b$ or $b \rightarrow a$.

The notation here follows Latzer [12]. The relation a is <u>spacelike related</u> to b ($a\sigma b$) if $|a-b| < 0$ is also common. Of course, the interpretation of $a\sigma b$ is that no signal whatsoever can possibly connect a and b. It is furthermore convenient to define "past and future sets" corresponding to the different causal order relations. In the notation of Kronheimer and Penrose [10] again:

$I^+(a) = \{b/a << b\}$ $\quad I^-(a) = \{b/b << a\}$ $\quad I(a) = \{b/a\tau b\}$

$J^+(a) = \{b/a < b\}$ $\quad J^-(a) = \{b/b < a\}$ $\quad J(a) = \{b/a\kappa b\}$

$E^+(a) = \{b/a \rightarrow b\}$ $\quad E^-(a) = \{b/b \rightarrow a\}$ $\quad E(a) = \{b/a\lambda b\}$

$I^+(a)$ is called the temporal or chronological future of a; $J^-(a)$ is called the causal past of a, etc. $E(a)$ is just what is usually called the null cone or light cone of a. Note that the I sets are open, while the J and E sets are all closed. The definitions which stipulated that $a < a$, $a \rightarrow a$, but not $a << a$ were chosen with this topological convenience in mind.

Only slightly less immediate than the interdefinability of the asymmetric relations (1), is the interdefinability of their symmetric counterparts:

$$(n\geq 2)\quad a\tau b \Leftrightarrow a\kappa b \;\&\; a \neq b \;\&\; \cap\{J(c): c\varepsilon J(a)\cap J(b)\} = \{a,b\}$$
$$a\lambda b \Leftrightarrow a\kappa b \;\&\; \neg(a\tau b)$$

$$(n\geq 2)\quad a\lambda b \Leftrightarrow \forall c,d[c,d\varepsilon \cap\{I(\ell): \ell\varepsilon I(a)\cap I(b)\} \Rightarrow \neg c\tau d] \qquad (2)$$
$$a\kappa b \Leftrightarrow a\tau b \vee a\lambda b$$

$$(n\geq 3)\quad a\tau b \Leftrightarrow \neg(a\lambda b) \;\&\; \forall c[a\lambda c \Rightarrow \exists d(d\varepsilon E(a)\cap E(c) \;\&\; d\lambda b)]$$
$$a\kappa b \Leftrightarrow a\tau b \vee a\lambda b$$

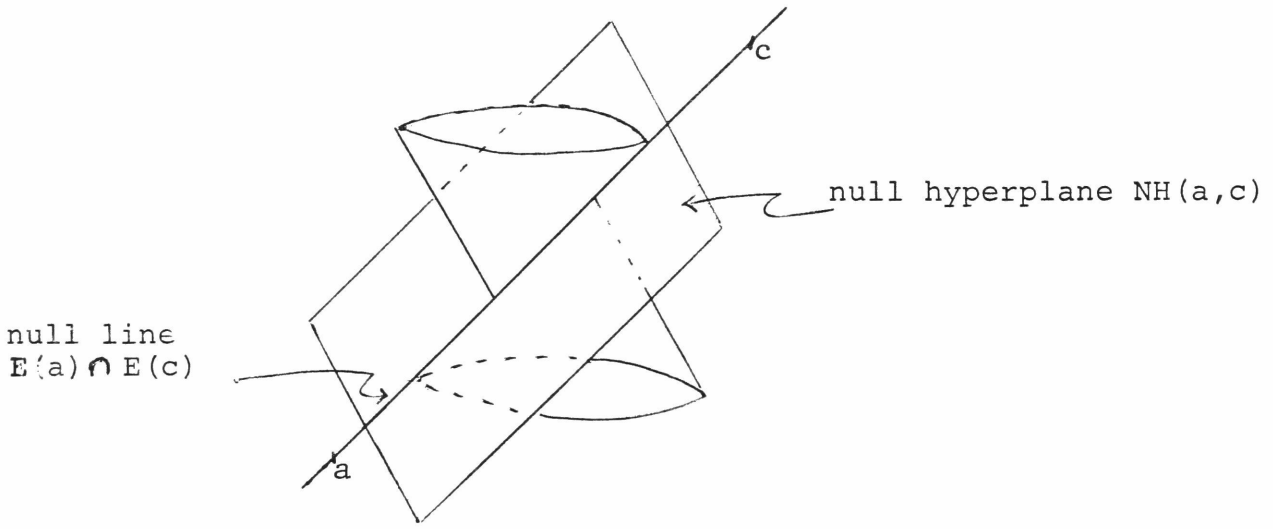
These are, clearly, first order sentences; set notation is used only to increase readability. It follows that (at least for $n \geq 3$) any of the three symmetric relations may be construed

as basic and the others as derived. To convince oneself of the validity of (2) it suffices to draw the right pictures. As a sample this will be done for the third equivalence which is just a bit trickier than the other two. A full computational proof will also be given to show, at least once, how such proofs run. Thereafter, computational verifications will be omitted.

First we check that for $n = 2$, τ and κ are not definable from λ. For this it suffices to exhibit a bijection $\phi: M_2 \to M_2$ of two dimensional Minkowski space-time onto itself which preserves λ (i.e. for all a,b: $a \lambda b \Leftrightarrow \phi(a)\lambda\phi(b)$) but which preserves neither τ nor κ. Let ϕ be the 90° counterclockwise rotation which turns the light cone at the origin onto its side. Expressed in our favorite standard coordinates, ϕ maps $a = (a_o,a_1)$ to $a' = (a'_o,a'_1) = (a_1,-a_o)$. Trivially, $\phi(a)\lambda\phi(b) \Leftrightarrow a'_o b'_o - a'_1 b'_1 = 0 \Leftrightarrow a_1 b_1 - (-a_o)(-b_o) = 0 \Leftrightarrow a\lambda b$. But ϕ takes timelike vectors to spacelike vectors (e.g. $(1,0) \to (0,-1)$), so it preserves neither τ nor κ.

Assume now that $n \geq 3$. Suppose a and c are distinct and $a\lambda c$. Then $E(a) \cap E(c)$ is just the null line passing through a and c. There is a unique null hyperplane, $NH(a,c)$, containing this line. (A hyperplane is an $(n-1)$ dimensional subspace. It is null if through each point there passes a unique null line contained in the hyperplane.)

It is tangent to the null cones of all points on the null line $E(a) \cap E(c)$. Intuitively it is clear that it consists of the line together with all points which are not λ related



to any points on the line. In fact this serves as its definition:

Def: For distinct a,c where $a\lambda c$

$$NH(a,c) = [E(a) \cap E(c)] \cup \{b: \forall d(d \varepsilon E(a) \cap E(c) \Rightarrow \neg b\lambda d\} \quad .$$

Still thinking of the case $n = 3$, it is intuitively clear that for a given point a, as c sweeps over $E(a)$, the sets $NH(a,c)$ sweep out all points except those in $I(a)$, i.e.

$$(3) \qquad (n \geq 3) \qquad a\tau b \Leftrightarrow b \notin \cup\{NH(a,c) : c\lambda a\} \quad .$$

This will be verified in detail in a moment. But first notice

that it is logically equivalent to:

$$a\tau b \iff b \notin \cup\{E(a) \cap E(c) : a\lambda c\} \;\&\; \forall c[a\lambda c \Rightarrow \exists d(d\varepsilon E(a) \cap E(c) \;\&\; d\lambda b)]$$

which in turn is equivalent to the third assertion of (2), since $E(a) = \cup\{E(a) \cap E(c) : a\lambda c\}$. Also note that, as had better be the case, (3) is false if $n = 2$ for in that case $NH(a,c)$ collapses to the line $E(a) \cap E(c)$, and for any particular a, $\cup\{NH(a,c) : c\lambda a\}$ collapses to $E(a)$.

To verify (3) we need a computational lemma.

2.1 Lemma: If u is non-zero then:

(1) ($n \geq 2$) if for some non-zero null vector v $(u,v) = 0$, it follows that $|u| \leq 0$ and $|u| = 0$ iff $u = kv$ for some constant k.

(2) ($n \geq 3$) $|u| \leq 0$ implies there is a non-zero null vector v where $(u,v) = 0$.

Proof: Choose a favorite standard coordinate system. Then for (1) we have:

$$v_o \neq 0, \quad v_o^2 = \sum_1^{n-1} v_i^2, \quad u_o v_o = \sum_1^{n-1} u_j v_j \quad .$$

If $|u| > 0$ we would also have $u_o^2 > \sum_1^{n-1} u_i^2$. Squaring the middle equality and comparing it with the product of the first and third relations yields:

$$\left(\sum_{1}^{n-1} v_i^2\right)\left(\sum_{1}^{n-1} u_i^2\right) < \left(\sum_{1}^{n-1} u_j v_j\right)^2 \qquad \text{which entails}$$

$$\sum_{i\neq j}^{n-1} (u_i v_j - u_j v_i)^2 < 0 \qquad \text{where the sum is taken over}$$

$i,j \geq 1$, which is impossible. So it must be the case that $|u| \leq 0$. If $|u| = 0$ the last inequality is replaced by an equality which entails $u_i v_j - u_j v_i = 0$ for all $i,j \geq 1$. Now there is an $i \geq 1$ such that $v_i \neq 0$ (for otherwise $v_o = 0$ since v is null). So for all $j \geq 1$, $u_j = (\frac{u_i}{v_i})v_j$ and since

$u_o v_o = \sum_{1}^{n-1} u_j v_j = (\frac{u_i}{v_i}) \sum_{1}^{n-1} v_j^2 = \frac{u_i}{v_i} v_o^2$ it follows that $u_o = \frac{u_i}{v_i} v_o$

as well. Of course if $u = kv$ for some constant k, then $|u| = k^2|v| = 0$. This completes the proof of part (1).

To check part (2) it is simplest to choose a standard coordinate system in which u takes the form $(u_o, u_1, 0, \ldots, 0)$. (If $\{e_i\}$ is the standard basis with which we began, take $e'_o = e_o$ and choose e'_1 to be the projection of u onto the space-like hyperplane spanned by $e_1, \ldots, e_{n-1}$ normalized so that $|e'_1| = -1$. Then $(e'_o, e'_1) = 0$ and $u = u_o e'_o + u_1 e'_1$ for some u_o, u_1. Next choose $e'_2, \ldots, e'_{n-1}$ in the hyperplane spanned by $e_1, \ldots, e_{n-1}$ so that $e'_1, \ldots, e'_{n-1}$ form an orthonormal basis of that hyperplane. The resulting set $e'_o, \ldots, e'_{n-1}$ is a standard basis for the entire space.) Since $|u| \leq 0$ we have $u_o^2 \leq u_1^2$ and since u is non-zero,

$u_1 \neq 0$. Take $v = (1, u_o/u_1, \sqrt{1-(u_o/u_1)^2}, 0, \ldots, 0)$ in the new coordinate system. (Here is where we use the fact that $n \geq 3$.) Then clearly $|v| = 0$ and $(u,v) = 0$. qed

As an immediate corollary we have:

2.2 Corollary ($n \geq 2$): For all distinct a,c where $a\lambda c$

$NH(a,c) = \{b : (b-a,c-a) = 0\}$

Assume first that $b\varepsilon NH(a,c)$. If $b\varepsilon E(a) \cap E(c)$, then $|b-a| = 0 = |b-c|$ and so $(b-a,c-a) = 1/2(|b-c|-|b-a|-|a-c|) = 0$. Here $|c-a| = 0$ since we are assuming $a\lambda c$. If $b\notin E(a) \cap E(c)$ then for all d, if $d\varepsilon E(a) \cap E(c)$ it must follow that $\neg\, b\lambda d$. For any constant k, if $d = k(c-a)+a$, then $|d-a| = |c-a| = 0$, i.e. $d\varepsilon E(a) \cap E(c)$. So $b\notin E(a) \cap E(c)$ entails $0 \neq |b-d| = |(b-a)-k(c-a)|$ for all k. Hence $|b-a| - 2k(b-a,c-a) \neq 0$ for all k. This is only possible if $(b-a,c-a) = 0$.

Conversely assume $(b-a,c-a) = 0$ and suppose there is a $d\varepsilon E(a) \cap E(c)$ such that $b\lambda d$. Since $|a-d| = 0 = |c-d|$ it follows that $(d-a,c-a) = 1/2(|d-c|-|d-a|-|a-c|) = 0$. So by the first part of the lemma (taking $u = c - a$ and $v = d - a$) we have $c - a = k(d-a)$ for some $k \neq 0$. Therefore $0 = |b-d| = |(b-a) - \frac{1}{k}(c-a)|$ and hence $|b-a| = 0 = |b-c|$. Thus $b\varepsilon E(a) \cap E(c)$ and so $b\varepsilon NH(a,c)$. qed.

Finally we can easily check that (3) follows from this corollary. Suppose first that $a\tau b$, but $b\varepsilon NH(a,c)$ for some c where $c\lambda a$. Then we would have $|c-a| = 0$, $(c-a,b-a) = 0$

and $|b-a| > 0$. This is impossible by the first part of the lemma again (taking $u = b - a$ and $v = c - a$). Conversely assume $\neg(a\tau b)$, so $|a-b| \leq 0$. If $a = b$, then trivially for any c where $a\lambda c$, $b \varepsilon E(a) \cap E(c)$, so $b \varepsilon NH(a,c)$. If $a \neq b$ then by the second part of the lemma (taking $u = b-a$) there is a non-zero null vector v where $(u,v) = 0$. If we take $c = v + a$ then $a\lambda c$, $(b-a,c-a) = 0$, and therefore $b \varepsilon NH(a,c)$. This proves (3).

Using the equivalences of (2) we can give a much simplified proof of a result of Latzer's [12] to the effect that the relation of "temporal betweeness" defined by:

$$B_\tau(a,b,c) \Longleftrightarrow a << b << c \vee c << b << a$$

is definable solely in terms of the symmetric relation λ. It sufficies to check that:

$$B_\tau(a,b,c) \Longleftrightarrow a\tau b \;\&\; b\tau c \;\&\; a\tau c \;\&\; I(b) \subseteq I(a) \cup I(c)$$

for we now know that τ is explicitly definable in terms of λ. (Similarly one can check that each of the three betweeness relations B_τ, B_κ, B_λ is definable from each of τ, κ, and λ ($n \geq 3$); for $n = 2$, B_τ and B_κ are definable from each of them, but B_λ is not definable from any.)

Interest in the betweeness relations arises because they can be used to recover the asymmetric causal structure of Minkowski space-time from its symmetric structure -- at

least up to a specification of temporal orientation. Given any two points a and b such that $a\tau b$ we can arbitrarily stipulate "a << b" and then project this orientation onto all other points c and d where $c\tau d$. The relation thus projected, which we write as $c \underset{ab}{\lesssim} d$ following Latzer [12], will systematically agree with or reverse the original relation <<. The definition can be given this way:

Def: For a,b where $a\tau b$,

$$c \underset{ab}{\lesssim} d \iff \exists e[B_\tau(a,b,e) \,\&\, B_\tau(c,d,e) \,\&\, \neg(I(b) \cup I(d) \subseteq I(a) \cap I(c))] \,.$$

Our claim, stated for future reference in the following lemma, is easily checked by considering the few possible cases:

2.3 Lemma ($n \geq 2$): If $a\tau b$, then either:

(i) for all c,d $c \underset{ab}{\lesssim} d \iff c << d$; or

(ii) for all c,d $c \underset{ab}{\lesssim} d \iff d << c$.

With this lemma, results which are cast in terms of asymmetric causal relations can be recast into a form using only symmetric relations. For example, suppose we can show that a certain predicate P is explicitly first order definable from asymmetric relations; schematically, $P \iff \ldots << \ldots$ where the expression on the right is formulated in a first order language whose only non-logical constant is '<<'. If that predicate P is itself symmetric, i.e. insensitive to reversal

of temporal orientation, then we also have $P \Leftrightarrow \exists a,b(a \tau b \ \& \ldots \overset{\leq\leq}{ab} \ldots)$ so that P is explicitly first order definable from symmetric relations. The idea to prove something like lemma 2.3 and use it in this way is due to Latzer [12], and also found in Winnie [21].

B. Causal Automorphisms and Zeeman's Theorem.

In this section a simple proof will be given of Zeeman's Theorem, characterizing the group of causal automorphisms of Minkowski space-time. The theorem is a consequence of several lemmas on the explicit definability of linear structure which constitute the first steps in the proof of Robb's definability theorem. The proof of that latter theorem will be completed in the next section. Zeeman's result, published in 1963, is only a rediscovery of what was at least implicit in Robb.

Generically, a causal automorphism of n dimensional Minkowski space-time M_n $(n \geq 2)$ is a bijection $\phi : M_n \to M_n$ which preserves one of the causal relations defined above. For example, a <<-isomorphism preserves << (i.e. for all a,b $a<<b \Leftrightarrow \phi(a)<<\phi(b)$) and a κ-isomorphism preserves κ (i.e. for all a,b $a \kappa b \Leftrightarrow \phi(a) \kappa \phi(b)$), etc. Just to avoid endless specification of cases, let us say that an

asymmetric causal automorphism (or just plain causal automorphism) is one which preserves any one (and hence all) of $<<$, $<$, and $\rightarrow$. Further we shall say a symmetric causal automorphism is one which preserves τ, κ, λ. It follows of course that symmetric causal autmorphisms preserve the symmetric relation $\lessdot_{ab}$, for all a and b where $a\tau b$. Thus lemma 2.3 implies that a symmetric causal automorphism is either a causal automorphism or it systematically reverses causal order (i.e. for all a,b $a << b \iff \phi(b) << \phi(a)$, etc.). Call such a map an order-reversing causal automorphism. This fact too we record for reference:

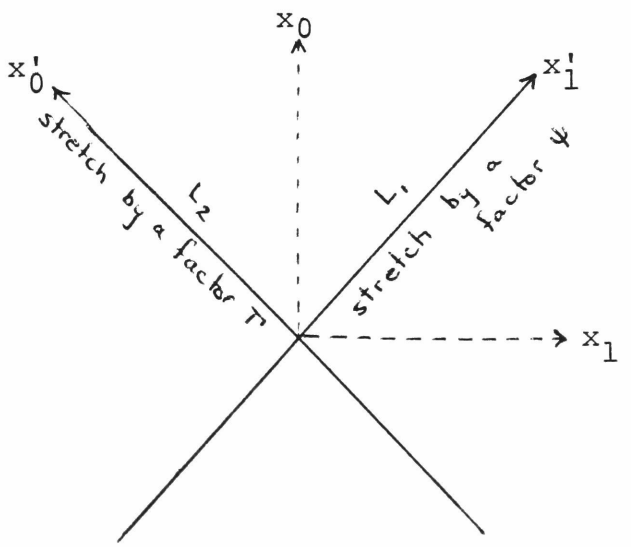
2.4 Corollary ($n \geq 2$): Symmetric causal automorphisms are either causal automorphisms or order-reversing causal automorphisms.

Now let $\phi : M_n \rightarrow M_n$ be a bijection. We say ϕ is a translation if there is a b such that for all a, $\phi(a) = a + b$; it is a dilation if there is a positive real k (called the dilation factor) such that for all a, $\phi(a) = ka$. If ϕ is both linear (i.e. for all a,b and reals k,ℓ $\phi(ka+\ell b) = k\phi(a)+\ell\phi(b)$) and norm preserving (i.e. for all a, $|\phi(a)| = |a|$), then ϕ is a (homogeneous) Lorentz transformation. Homogeneity just means that $\phi(\theta) = \theta$ where θ is the zero vector and follows automatically from linearity. Clearly, homogeneous Lorentz transformations are symmetric causal automorphisms (since $|a-b| = |\phi(a) - \phi(b)|$,

$a \kappa b \iff \phi(a) \kappa \phi(b))$. They are called <u>orthochronous</u> or <u>anti-orthochronous</u> depending on whether they systematically preserve or systematically reverse causal order.

All maps composed from translations, dilations, and orthochronous Lorentz transformations are causal automorphisms. This is trivial. Zeeman's Theorem states that for $n \geq 3$ no others are, i.e. the group of causal isomorphisms is generated by the translations, dilations, and orthochronous Lorentz transformations. Thus in a sense (up to translations and dilations), "causality implies the Lorentz group," which is the title of Zeeman's paper. The heart of the proof is the demonstration that homogeneous causal automorphisms must be linear.

This is not true for $n = 2$ and it is geometrically intuitive why not. Consider the light cone at the origin with its null lines L_1 and L_2. If we stretch the entire plane in the L_1 direction, or the L_2 direction, or both,



we induce a causal automorphism of the plane whether or not the component stretchings are themselves linear. [If x_o, x_1 are standard coordinates introduce new "null coordinates" x'_o, x'_1 where $x'_o = x_o - x_1$ and $x'_1 = x_o + x_1$. Let ψ and Γ be any order preserving bijections of $\mathbb{R}$ onto itself which leave 0 fixed. Then the map defined by $\phi : (a'_o, a'_1) \to (\psi(a'_o), \Gamma(a'_1))$ is a causal isomorphism since $|a| = a'_o a'_1$ and for all a and b, $\phi(a) < \phi(b) \Leftrightarrow [\psi(a'_o) - \psi(b'_o)] \cdot [\Gamma(a'_1) - \Gamma(b'_1)] \geq 0$ & $\psi(a'_o) \leq \Psi(b'_o) \Leftrightarrow (a'_o - b'_o)(a'_1 - b'_1) \geq 0$ & $a'_o \leq b'_o \Leftrightarrow a < b$.] In the presence of a third dimension the light cone is "rigidified" and stretchings along null lines, unless produced by dilations, will contort the null cones and tamper with causal relations.

In what follows let ϕ be a homogeneous symmetric causal automorphism. Trivially ($n \geq 2$) ϕ takes null related points to null related points ($a \lambda c \Leftrightarrow \phi(a) \lambda \phi(c)$). The first step in Zeeman's proof is to show ($n \geq 2$) that parallel pairs of null related points are taken to parallel pairs of null related points, i.e. if $a \lambda c$, $a' \lambda c'$, and $c - a$ is parallel with $c' - a'$ then $\phi(c) - \phi(a)$ is parallel with $\phi(c') - \phi(a')$. His argument involves the classification of doubly ruled quadric surfaces. It seems simpler to use the null hyperplane construction from the previous section. It is intuitively clear that $c - a$ and $c' - a'$ will be parallel iff their associated null hyperplanes NH(a,c) and NH(a',c')

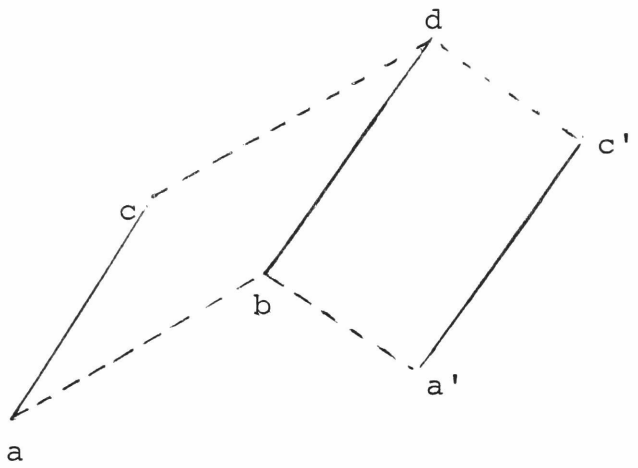
are either equal of disjoint. The equivalence can be checked with a straightforward computation using Corollary 2.2. The condition involving null hyperplanes is formulated solely in terms of symmetric causal relations. So parallelism between the null pairs a,c and a',c' will certainly be preserved under ϕ.

It follows ($n \geq 2$) that *if a and a' are null vectors, then* $\phi(a+a') = \phi(a) + \phi(a')$. This is because the null parallelogram with vertices $0,a,a',a+a'$ is necessarily taken to a null parallelogram with vertices $0,\phi(a), \phi(a')$, and $\phi(a+a')$.

Suppose a,c and a',c' are again pairs of distinct null related points. Suppose further that q is any rational number. Then the next claim we make is that ($n \geq 3$) the condition $c'-a' = q(c-a)$ is (explicitly, first order) definable from symmetric causal relations. Since the condition is symmetric (it holds for a,c,a',c' iff it holds for -a,-c,-a',-c') we may make full use of asymmetric causal relations in giving a "causal definition" and then avail ourselves of lemma 2.3.

First assume $a \to c$ and $a' \to c'$ and $a \sigma a'$. Then $c'-a' = c-a$ (the case q=1) is equivalent to the assertion that there exist points b,d such that $b \to d$ and such that the six points a,c,a',c',b,d form two null parallelograms as in the figure, i.e. the null vectors c-a,c'-a',d-b

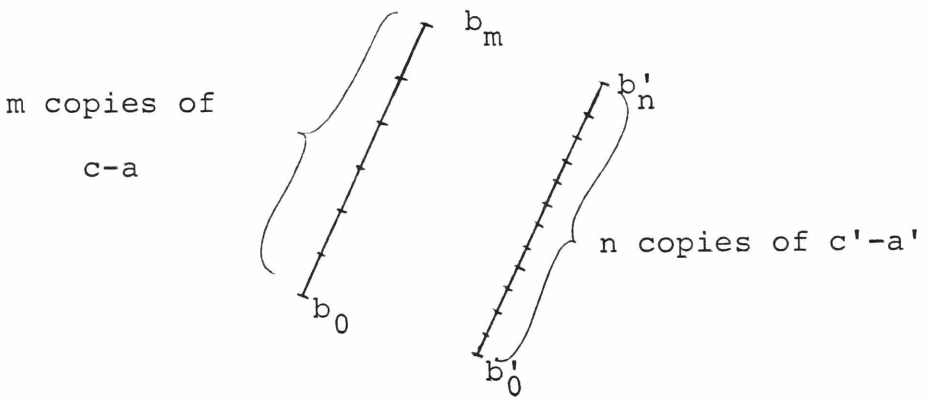
are parallel; and $d - c$, $b - a$ are parallel; and $d - c'$, $b - a'$ are parallel. Of course we already know that the



component assertions of parallelism between null vectors can be formulated using only (symmetric) causal relations. Note that this causal definition only works if a third dimension is available.

If we continue to assume that $a \rightarrow c$ and $a' \rightarrow c'$ but consider the more general case where a and a' are not necessarily space-like related, vector equality of $c - a$ and $c' - a'$ is defined by the interpolation of a third pair of null related points a'',c'' where $a'' \rightarrow c''$, where a'' is spacelike related to both a and a', and where $c'' - a''$ is equal (according to the condition just stated) to both $c' - a'$ and $c - a$. Finally, if we drop the assumption that $a \rightarrow c$ and $a' \rightarrow c'$ then $c - a = c' - a'$ is equivalent to a disjunction: $(a \rightarrow c \;\&\; a' \rightarrow c' \;\&\; \ldots) \vee (a \rightarrow c \;\&\; c' \rightarrow a' \;\&\; \ldots) \vee \ldots$ etc. This takes care of the case where $q = 1$.

If $q = m/n$ where m,n are positive integers, the causal definition of $c' - a' = q(c-a)$ amounts to the assertion that the vector result of laying out n copies of $c' - a'$ end to end is equal to the result of laying out m copies of $c - a$. If $a \to c$ and $a' \to c'$ the defining condition, more precisely, is that there exist $(m+1)$ points $b_0, \ldots, b_m$ and $(n+1)$ points $b'_0, \ldots, b'_n$ such that for each $i = 0, \ldots, m - 1$ and each $j = 0, \ldots, n - 1$ the vectors $c - a$ and $b_{i+1} - b_i$ are equal, while the vectors $c' - a'$ and $b'_{j+1} - b'_j$ are equal; and such that $b_m - b_0$ and $b'_n - b'_0$ are equal. Again by moving to a



disjunction the assumption $a \to c$ and $a' \to c'$ can be erased. (I shall not continue to repeat this each time in the future.) Finally if $m = 0$, then $c' - a' = m/n(c-a)$ is equivalent to $c' = a'$; and if m is negative, say $m = -p$, then $c' - a' = \frac{m}{n}(c-a)$ is equivalent to $c' - a' = \frac{p}{n}(a-c)$.

Thus, as claimed, for distinct null related pairs of points a,c and a',c' the condition $c' - a' = q(c-a)$ is (explicitly, first order) definable from symmetric causal relations for any rational q. It follows ($n \geq 3$) that <u>if a is null and k is any real, then $\phi(ka) = k\phi(a)$</u>. Certainly this must be true if k is rational. Now ϕ is either a causal automorphism or an order reversing causal automorphism by corollary 2.4. Suppose the former. Then for all rationals q,q' if $qa \to ka \to q'a$ it follows that $q\phi(a) \to \phi(ka) \to q'\phi(a)$. This implies $\phi(ka) = k\phi(a)$. If ϕ is order reversing then $qa \to ka \to q'a$ implies $q'\phi(a) \to \phi(ka) \to q\phi(a)$ and again $\phi(ka) = k\phi(a)$.

Putting together the two underlined facts we have:

2.5 <u>Lemma</u> ($n \geq 3$): If ϕ is a homogeneous, symmetric causal automorphism, ϕ is linear.

This follows immediately for we can choose a basis for the space consisting of n null vectors $e_1, \ldots, e_n$. Then for any $a = \Sigma r_i e_i$ and $b = \Sigma s_j e_j$ and any real $k\ell$, repeated application of the two facts yields:

$$\phi[k(\Sigma r_i e_i)+\ell(\Sigma s_j e_j)] = k[\phi(\Sigma r_i e_i)] + \ell[\phi(\Sigma s_j e_j)]$$

i.e. $\phi(ka+\ell b) = k\phi(a) + \ell\phi(b)$.

Before completing Zeeman's Theorem we need one more simple lemma about inner products. Recall that an inner product $(,)$ on a vector space V is positive (negative)

semi-definite if $(v,v) \geq 0$ for all v ($(v,v) \leq 0$ for all v) and indefinite if neither the one or the other.

2.6 Lemma: If $(,)$ is an indefinite inner product on a vector space V and $(,)'$ is any other inner product on V such that for all $v \in V$ $(v,v) = 0$ iff $(v,v)' = 0$, then for some real k, $(v,w)' = k(v,w)$ for all $v,w \in V$.

Proof: Fix some v where $(v,v) < 0$ and let w be any vector for which $(w,w) > 0$. Consider the equation $|v+\lambda w| = 0$, i.e. $(v,v) + 2\lambda(v,w) + \lambda^2(w,w) = 0$. Since $4(v,w)^2 - 4(v,v)(w,w) > 0$, the equation has real solutions λ_1 and λ_2 where $\lambda_1\lambda_2 = \frac{(v,v)}{(w,w)}$. Now λ is a solution of $|v+\lambda w| = 0$ iff it is a solution of $|v+\lambda w|' = 0$, so we have $\lambda_1\lambda_2 = \frac{(v,v)'}{(w,w)'}$ as well. Thus for any w such that $(w,w) > 0$ we have $(w,w)' = k(w,w)$ where $k = \frac{(v,v)'}{(v,v)}$.

Now let u,w be arbitrary vectors where $(u,u) < 0$ and $(w,w) > 0$. Repeating the very same argument we have $\frac{(w,w)'}{(w,w)} = \frac{(u,u)'}{(u,u)}$ and hence that $(u,u)' = k(u,u)$ for any u such that $(u,u) < 0$. Trivially if $(u,u) = 0$ then $(u,u)' = k(u,u)$.

Thus for all u whatsoever, $(u,u)' = k(u,u)$. But now for all u,w: $(u,w) = \frac{1}{2}[(u+w,u+w) - (u,u) - (w,w)]$, and hence $(u,w)' = k(u,w)$. qed.

Zeeman's Theorem [22] now follows:

2.7 Theorem ($n \geq 3$): If ϕ is a causal automorphism of Minkowski space-time, ϕ is the composition of a translation, a dilation, and a homogeneous orthochromous Lorentz transformation.

Proof: Let T be a translation taking $\phi(\theta)$ to θ and take $\phi_1 = T\circ\phi$. Then by lemma 2.5 ϕ_1 is linear. Define an inner product $(,)'$ by $(a,b)' = (\phi_1(a),\phi_1(b))$. Since $(a,a)' = 0$ iff $(a,a) = 0$ it follows by lemma 2.6 that there is a k such that $(a,b)' = k(a,b)$ for all a,b and k must be positive since ϕ_1 takes timelike vectors to timelike vectors. If $\ell = \sqrt{1/k}$ and D_ℓ is the dilation which takes a to ℓa, then $D_\ell\circ T\circ\phi$ is a homogeneous Lorentz transformation. It must be orthochronous since each of ϕ, T, and D_ℓ preserve causal order. So $\phi = T^{-1}\circ D_\ell^{-1}\circ(D_\ell\circ T\circ\phi)$ and the result follows. qed

If ϕ is only a symmetric causal automorphism, then everything in the proof goes through to the step where $D_\ell\circ T\circ\phi$ is pronounced a homogeneous Lorentz transformation. Then by lemma 2.4 ϕ must be either orthochronous or anti-orthochronous. Thus we also have:

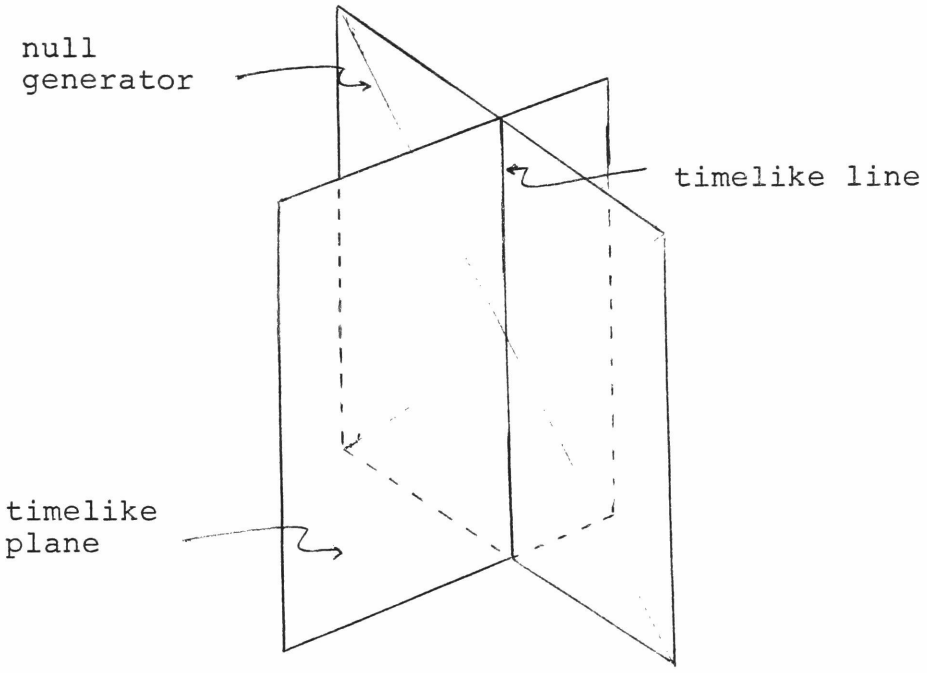
2.8 Theorem ($n \geq 3$): If ϕ is a symmetric causal automorphism of Minkowski space-time it is the composition of a translation, a dilation, and a homogeneous orthochronous or anti-orthochronous Lorentz transformation.

C. Robb's Definability Theorem

In the course of proving Zeeman's Theorem we showed that the relations of parallelism and equality between null vectors are explicitly first order definable from symmetric causal relations. In this section we show ($n \geq 3$) that these relations and also those of orthogonality and congruence are so definable for quite arbitrary vectors. (Vectors u,v are orthogonal if $(u,v) = 0$ and congruent if $|u| = |v|$.) The construction is that of Robb. In the last section causal definitions were given in complete detail to show at every stage that they were explicit and first order. Some of that detail will be omitted this time around so as not to obscure the intuitive line of the construction. Again, computational verifications will be omitted.

First parallelism. A plane (i.e. a two dimensional subspace) is called timelike if through every one of its points there pass two null lines each fully contained in the plane. Each of these null lines is called a generator of the timelike plane. Clearly, the generators fall into two classes of mutually parallel null lines. Two planes are parallel if their respective families of generators are parallel. Now ($n \geq 3$) all lines, whether timelike, spacelike, or null, can be realized as the intersection of two

timelike planes. (The figure shows a timelike line realized



this way.) Two general lines are parallel iff they can be realized as the intersections of respectively parallel timelike planes. This is the basis of the causal definition of vector parallelism.

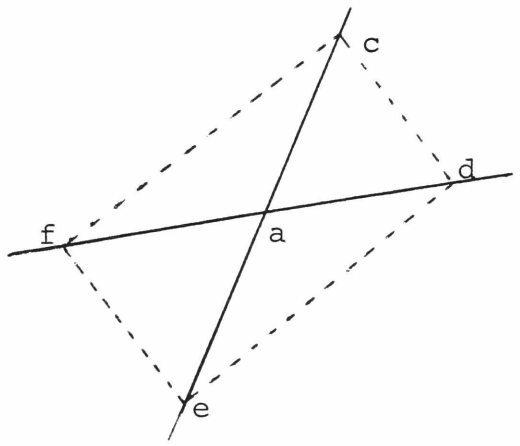
Given pairs of distinct points a,c and a',c' we want to assert that $c - a$ and $c' - a'$ are parallel iff there exist four timelike planes P_1, P_2, P_1', P_2' such that a,c belong to $P_1 \cap P_2$, a', c' belong to $P_1' \cap P_2'$, and P_1 and P_2 are parallel, while P_1' and P_2' are parallel. Rather than quantifying over planes, we need only quantify over triples of points which determine them. A timelike plane P is determined by a specification of three distinct points

d,e,f in P where $d\lambda e$, $d\lambda f$, and the null vectors e - d and f - d are not parallel. To assert that a point, say a, lies in P we in effect say that a lies in the vector span of e - d and f - d, i.e. that a and d form opposite vertices of a null parallelogram whose sides are respectively parallel to e - d and f - d. To assert that two timelike planes are parallel we say that their respective null vectors generators are parallel. Thus the defining condition that there exist four timelike planes ... can be written as the condition that there exist twelve points ... . As a first step then, we have ($n \geq 3$) that parallelism between arbitrary vectors is explicitly definable from symmetric causal relations. Note that the definition breaks down for $n = 2$ where there is only one timelike plane to work with. In fact, general parallelism is not definable for $n = 2$ even though, as we established, null parallelism is.

Next vector equality. Given pairs of distinct points a,c and a',c' we want to "define" the condition that c' - a' = c - a. This proceeds by a straightforward parallelogram construction similar to the one used in the previous section to define equality between null vectors. The only difference is that now we can avail ourselves of arbitrary parallelograms, not simply ones that are null.

Next orthogonality. It is easiest, again, to talk in terms of lines. There are several cases to consider. Two null

lines through a common point are orthogonal iff they coincide. A spaceline is orthogonal to a null line with which it intersects iff the spaceline lies in the null hyperplane associated with the null line. A timeline is never orthogonal to any other timeline or null line. A timeline is orthogonal to a spaceline sharing a common point iff they form the diagonals of a null parallelogram as in the figure.

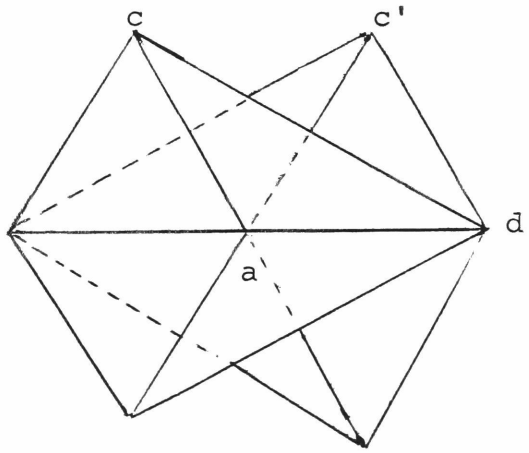


Two spacelines sharing a common point are orthogonal iff there is a timelike plane through one having the property that every timeline in it running through the common point is orthogonal to the other spaceline.

Now suppose a,c and a',c' are pairs of distinct points. We want to "define" the condition that $(c'-a',c-a) = 0$. In the special case where $a' = a$ we need only use the above sequence of cases, systematically replacing reference to lines and timelike planes with expanded reference to pairs or triples

of points which generate them. In the general case where $a \neq a'$ we extrapolate, asserting that there exists a point c'' such that $c'' - a = c' - a'$ and such that $c'' - a$ is orthogonal to $c - a$.

Suppose $a \tau c$ and $a \sigma d$. We say that $c - a$ and $d - a$ are conjugate if, as in the previous figure, they form half diagonals of a null parallelogram (i.e. there exist points e and f such that c,d,e,f form a null parallelogram, $c - a = a - e$, and $f - a = a - d$). The term is Robb's. With this relation we can give the causal definition of congruence between pairs of distinct points a,c and a',c'. Suppose first that $a = a'$. There are several cases to consider. If $a \tau c$ and $a \tau c'$ then the pairs are congruent iff there is a point d spacelike related to a such that $c - a$ and $c' - a$ are both conjugate to d (see figure).



If $a\sigma c$ and $a\sigma c'$ then the pairs are congruent correspondingly iff there is a point d timelike related to a such that $c' - a$ and $c - a$ are both conjugate to $d - a$. If $a\lambda c$ and $a\lambda c'$ then of course a,c and a,c' are always congruent. In all other mixed cases they are never congruent. Finally, in the general case where $a \neq a'$, we again extrapolate, asserting $|c'-a'| = |c-a|$ iff there is a point c'' such that $c'' - a = c' - a'$ and $c'' - a$ is congruent with $c - a$.

Thus we have at least sketched the proof of what we shall call Robb's Definability Theorem ($n \geq 3$): the relation of congruence, which holds between points a,c,a',c' just in case $|c-a| = |c'-a'|$, is explicitly first order definable from symmetric causal relations.

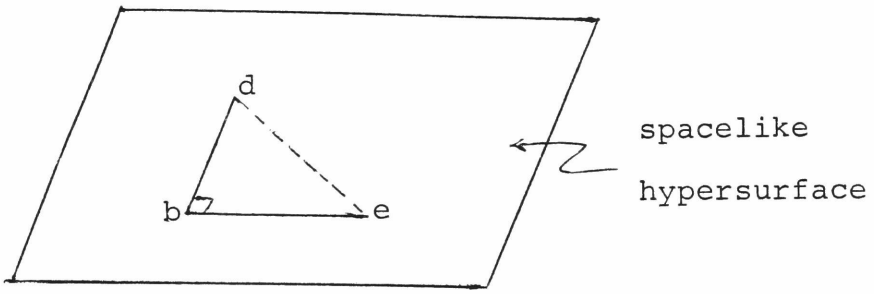
Actually we have at our disposal the means to prove something stronger still which should be recorded for future reference. First of all, if q is any rational number we can "casually define" the relation $|c'-a'| = q|c-a|$. This proceeds in two steps. We define the relation $c' - a' = q(c-a)$ just as we did in the previous section in the special case of null related vectors. Again we use a simple "end-to-end" construction, now availing ourselves of the causal definition of equality between arbitrary vectors. Then we define the relation $|c'-a'| = q|c-a|$ to obtain just in case there exist points a'',c'' such that $c''-a'' = q(c-a)$ and $|c''-a''| = |c'-a'|$

(when q is non-negative) or c"-a" = q(c-a) and c"-a" and c'-a' are conjugate (when q is negative).

Let us say that a real number k is first order definable from symmetric causal relations or just definable if the relation $|c'-a'| = k|c-a|$ is explicitly first order definable from symmetric causal relations. We have then that all rationals are definable. The problem naturally arises to characterize the full class of definable reals.

The Euclidean field over the rationals is that field which results from closing the rationals under the four rational operations (taking additive and multiplicative inverses, sums, and products) and the operation of extracting the square root of the sum of two squares. If $q_1, \ldots, q_n$ are rationals then the field contains $\sqrt{q_1^2+q_2^2}$ and $\sqrt{q_1^2+q_2^2+q_3^2} = \sqrt{(\sqrt{q_1^2+q_2^2})^2 + q_3^2}$, and ... $\sqrt{q_1^2+\ldots+q_n^2}$; it also contains $q_3\sqrt{q_1^2+q_2^2} + \frac{1}{q_4}$ etc. It is easy tc check that all reals in the Euclidean field over the rationals are definable. Suppose that k is definable. Then -k is definable since $|c'-a'| = -k|c-a|$ is equivalent to the assertion that there exist points a", c" such that c"-a" is conjugate to c'-a' and $|c"-a"| = k|c-a|$. Similarly if k_1 and k_2 are definable, so must be $\sqrt{k_1^2+k_2^2}$. A projection onto a spacelike hypersurface is involved (--given any points a,b where aτb, the set of points c such that c-a is orthogonal to b-a is a

spacelike hypersurface). The idea is that when restricted to a spacelike hypersurface the Minkowski orthogonality and congruence relations reduce to the Euclidean. Suppose first that k_1 and k_2 are non-negative. Then $|c'-a'| = \sqrt{k_1^2+k_2^2}\,|c-a|$ is equivalent in the case $c\sigma a$ to the assertion that there exist points b,d,e all in any one particular spacelike hypersurface such that (see figure) d-b and e-b are orthogonal, $|d-b| = k_1|c-a|$,



$|e-b| = k_2|c-a|$, and $|d-e| = |c'-a'|$. The case $c\tau a$ is similar, but the relation of conjugacy must be brought in. Now suppose that k_1 is negative, but k_2 is non-negative. Then we can define $|c'=a'| = \sqrt{k_1^2+k_1^2}|c-a|$ to hold just in case $|c'-a'| = \sqrt{(-k_1)^2+k_2^2}\,|c-a|$. And so forth.

It seems plausible (?) that the Euclidean field over the rationals exhausts the class of definable reals. (Certainly that field exhausts the class of reals constructible from ruler and compass in Euclidean geometry.) But I do not see a proof.

D. Robb's Categoricity Theorem

In his A Theory of Time and Space [ ], Robb presented an axiom system involving the single two place relation "after" (--in our notation, a after b iff $a < b$ & $a \neq b$--), proved an enormous number of theorems, and eventually sketched how (real valued) coordinates could be associated with elements of any model of his axioms in such a way that the standard coordinate structure of four dimensional Minkowski space-time M_4 is reproduced. He proved in effect, though without stating so, that his axioms are categorical: that any model of his axioms (S, after) must be isomorphic to M_4 in its natural causal structure, i.e. there must be a bijection $\phi : S \to M_4$ such that for all $s,s' \in S$, s' after $s \iff \phi(s) \neq \phi(s')$ & $\phi(s) < \phi(s')$.

Because of the way in which Robb sketched the "introduction of coordinates" it has been unclear to some authors whether in the end he really did prove the categoricity of his axioms. Domotor, for example, while thinking it obvious that such a theorem "could be shown," regrets that "no physicist or mathematician responded effectively to Robb's challenge [to find it]." ([3]).

In this section we give a simple categoricity proof, not of Robb's axioms, but of a related finite axiom set involving the single relation << (or whatever other asymmetric relation is construed as primitive). Thereafter, using lemma

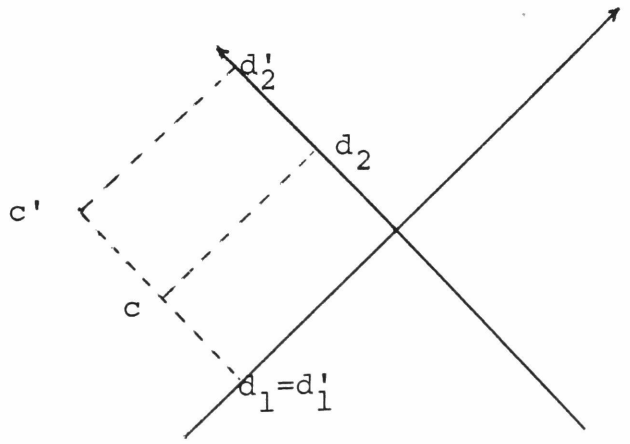
2.3, we extract a corresponding result for axioms formulated only in terms of one of the symmetric causal relations. Skolem-Löwenheim considerations alone, of course, require that not all of the axioms be first order. Our one second order axiom is a type of Dedekind completeness axiom. The others are all first order. Robb uses second order formulation (and more) quite freely in his system not only in his Dedekind axiom but in many others too. But it is unnecessary since, as already explained, talk about lines, planes, hypersurfaces, etc., can be replaced by talk about finite sets of points which determine them. Were one so inclined, one could formalize our axioms in a second order language whose only non-logical constants are the relation symbols '<<' and 'e.'

There is an important difference in approach between Robb's axiomatization and ours. From our point of view the goal is to establish "finite first order and Dedekind" categoricity, not to present a simple, minimal axiom system. We take anything we can get our first order hands on! As was shown in the previous section that is quite a lot. In effect we simply take over as axioms causal properties of Minkowski space-time which are so strong that it is possible to establish their categoricity with little work at all. Certainly these axioms could be derived from a simpler set.

This is more or less what Robb does, painstakingly but tediously. But so far as the logical problem of categoricity is concerned, little is achieved in the process.

Once again in this context, the case $n = 2$ is special. A particularly simple axiomatization is available here which does not carry over to the case $n \geq 3$. In fact it uses none of the material about general parallelism, orthogonality, and congruence developed in the previous section. Indeed, as we already know, these relations are not definable from causal relations for $n = 2$ and so they are not available for use in an axiomatization. (That they are not definable follows from the fact that they are not, in general, preserved under non-linear causal automorphisms of M_2. The existence of such automorphisms was established in section B.)

Consider the two null lines passing through some point in M_2. Every point c in M_2 is uniquely determined by its null projections onto these two lines.



This sets up a bijection between M_2 and pairs of points on them. The bijection has the special property that given points c and c', $c\lambda c'$ iff they share a common projection onto at least one of the lines; and $c \rightarrow c'$ iff $c\lambda c'$ and the projections of c are respectively $\rightarrow$ related to the projections of c'. Thus the causal structure of M_2 is fully coded in the order structure of the null lines.

One of the axioms we use to categorize the causal structure of M_2 simply asserts the possibility of setting up such a correspondence between points and their respective projections onto two arbitrarily chosen null lines. The others guarantee that null lines, under their inherited $\rightarrow$ order relation, are isomorphic to the reals. They assert that the induced order is linear, dense, without first or last element, and complete. Nothing more is needed.

Suppose $a \rightarrow b$ for distinct points a,b. Then for an arbitrary point c, the expression "d is the projection of c onto ab" abbreviates: $a\lambda d$ & $b\lambda d$ & $d\lambda c$. With that expression the axioms can be stated as follows:

Axioms for M_2:

(1) For all distinct a,b where $a\lambda b$: restricted to $E(a) \cap E(b)$ the relation $\rightarrow$ is linear, dense, and without first or last element.

(2) For all distinct a,b where $a\lambda b$: restricted to $E(a) \cap E(b)$ the relation $\rightarrow$ is Dedekind complete,

i.e. if X and Y are disjoint subsets of $E(a) \cap E(b)$ whose union is $E(a) \cap E(b)$, and if no point of either set lies between two points of the other, then there is a point of one set which lies between every point of that set and every point of the other set.

(3) There exist distinct points a, b_1, b_2 where $a \to b_1$, $a \to b_2$, but not $b_1 \lambda b_2$ such that:

(i) For every c, there exist unique points d_1 and d_2 where d_1 is the projection of c onto ab_1 and d_2 is the projection of c onto ab_2.

(ii) For all points $d_1 \in E(a) \cap E(b_1)$ and $d_2 \in E(a) \cap E(b_2)$ there is a unique point c such that d_1 and d_2 are its respective projections onto ab_1 and ab_2.

(iii) For all c, c' if d_1, d_2, d_1', d_2' are their respective projections onto ab_1 and ab_2 then:

$$c \to c' \text{ iff } d_1 \to d_1' \ \& \ d_2 \to d_2' \ \& \ (d_1 = d_1' \vee d_2 = d_2').$$

(see figure above).

To prove the categoricity of these axioms we need only rework the sketch above. We pose it with $<<$ construed as primitive.

2.9 Theorem: If $(S,<<)$ is a model of axioms (1), (2), (3), then $(S,<<)$ is isomorphic to M_2.

Proof: Choose points a,b_1,b_2 in S as guaranteed by (3). By (1) and (2), the sets $E(a) \cap E(b_1)$ and $E(a) \cap E(b_2)$ will be order isomorphic to $\mathbb{R}$. Let $\phi_1 : E(a) \cap E(b_1) \to \mathbb{R}$ and ϕ_2 $E(a) \cap E(b_2) \to \mathbb{R}$ be isomorphisms. Suppose x_o,x_1 are standard coordinates on M_2 and x'_o,x'_1 are corresponding null coordinates of the sort used before (i.e. $x'_o = x_o-x_1$ and $x'_1 = x_o+x_1$). In lieu of (3i,ii) we can set up a bijection $\phi : S \to M_2$ by associating with each point $s \in S$ that point in M_2 whose x'_o,x'_1 coordinates are $(\phi_1(s),\phi_2(s))$. Now by (3iii), for all s,s' in S, $s \to s'$ iff either $\phi_1(s) = \phi_1(s')$ or $\phi_2(s) = \phi_2(s')$, and both $\phi_1(s) \leq \phi_1(s')$ and $\phi_2(s) \leq \phi_2(s')$. These conditions on the x'_o,x'_1 coordinates of $\phi(s)$ and $\phi(s')$ are equivalent to $\phi(s) \to \phi(s')$. qed

Clearly the theorem could have been posed with either $<$ or $\to$ construed as primitive. Let $\ldots << \ldots$ abbreviate the conjunction of the three axioms. Then consider the axiom $\exists\, a,b[a\tau b \;\&\; \ldots \underset{ab}{\ll} \ldots]$. With exactly the same proof we can show that any model (S,τ) of this axiom must be isomorphic to M_2. So we have a symmetric version to Theorem 2.9.

In turning to consider the case $n \geq 3$, things are not so simple. Axioms (1) and (2) could be carried over

intact. We could also write down an axiom parallel to the first two clauses of (3) prescribing that points be uniquely determined by projections onto n linearly independent null vectors. Together these axioms would guarantee that any model naturally inherited the structure of an n-dimensional vector space over the reals. The problem is with clause (iii) of (3) which is specifically two dimensional in character. With $n \geq 3$, given points c and c' there does not appear to be any ready condition involving their respective projections onto n null lines which is equivalent to $c \rightarrow c'$ (or $c << c'$, etc.).

Robb's line of construction, and ours, is quite different. First, we stipulate axiomatically that it is possible to realize the model as the product of a timeline and a spacelike hypersurface orthogonal to it. (Certainly every point in Minkowski space-time is uniquely characterized by its projections onto any two such **subspaces** .) Then we impose axioms which guarantee that the spacelike hypersurface itself is Euclidean in structure. Finally we impose axioms which effectively code causal structure into the spacelike hypersurface. In the axiomatization which follows these stages are captured, respectively, in clauses I, II, and III.

The first stage is straightforward. It corresponds exactly to 3(i),(ii) in the axiomatization for M_2. Given

points a,b where a << b, let $\mathcal{L}(a,b)$ be the set of points collinear with a and b (i.e. those points c where a - c and b - c are parallel), and let $\mathcal{H}(a,b)$ be the spacelike hypersurface through a orthogonal to $\mathcal{L}(a,b)$ (i.e. those points d where a - d and a - b are orthogonal). Then the expression "c is the projection of e onto $\mathcal{L}(a,b)$" naturally abbreviates: $ce\mathcal{L}(a,b)$ and c - e is orthogonal to b - a; and similarly for "d is the projection of e onto $\mathcal{H}(a,b)$." Clause I, then, is certainly first order in the relation '<<' (or '<', etc.).

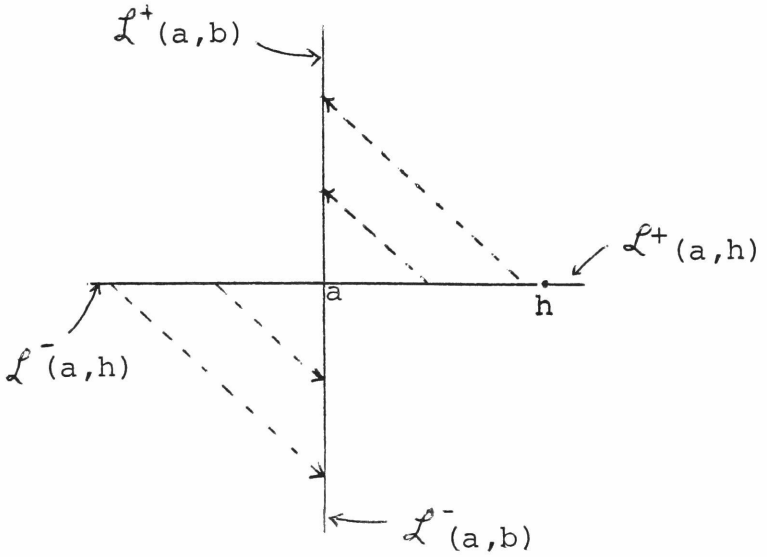
There are different ways to force the desired outcome that $\mathcal{H}(a,b)$ have the structure of (n-1) dimensional Euclidean space. One is to consider an axiom set $\mathcal{E}_{n-1}$ which is known to categorize that space and then add such axioms to one's own system as to make it possible to prove $\mathcal{E}_{n-1}$ as relativized to a space-like hypersurface. This is Robb's way. Our no work alternative is simply to adopt wholesale the relativized versions of the $\mathcal{E}_{n-1}$. This is possible so long as the primitive relations entering into the $\mathcal{E}_{n-1}$ are explicitly definable from causal relations.

One rather elegant finite categorical axiomatization of n-dimensional Euclidean space ($n \geq 2$) due to Tarski uses a three place betweenness relation Bxyz, interpreted to mean that y is collinear with and between x and z (with the case of y coinciding with x or z

not excluded), and a four place congruence relation Cxyzw, interpreted to mean that the distance from x to y equals that from z to w. We already know that the latter is ($n \geq 3$) explicitly definable from causal relations. So clearly is the other ($n \geq 3$). Given a,b,c, B(a,b,c) is the condition that the points are either not all distinct or they are collinear and either (i) $a < b < c$ or $c < b < a$, or (ii) $\neg(a \kappa c) \;\&\; J^-(a) \cap J^-(c) \subseteq J^-(b)$. Without bothering to enumerate the list or rehearse the proof, we now assume the categoricity of Tarski's axioms $\mathcal{E}_{n-1}$ (for $n \geq 3$) as a lemma, i.e. we assume that any model of them must be isomorphic (with respect to B and C) to (n-1) dimensional Euclidean space. [Actually some explanation here is in order. Tarski's system is stated for two dimensional Euclidean space. But as he notes, it can be adapted to higher n by changing the dimensionality axioms A11, A12 and leaving all others fixed. Also it is posed as a first order axiomatization of "elementary geometry." For our purposes his infinite axiom scheme A13 must be replaced by a second order completeness axiom -- such as the one cited above. This will be the one second order axiom.]

Clause III is less transparent than the other two. So far we have put in enough to conclude that any model can be realized as the product of some set $\mathcal{L}(a,b)$ with (n-1) dimensional Euclidean space. Clause III(i) allows

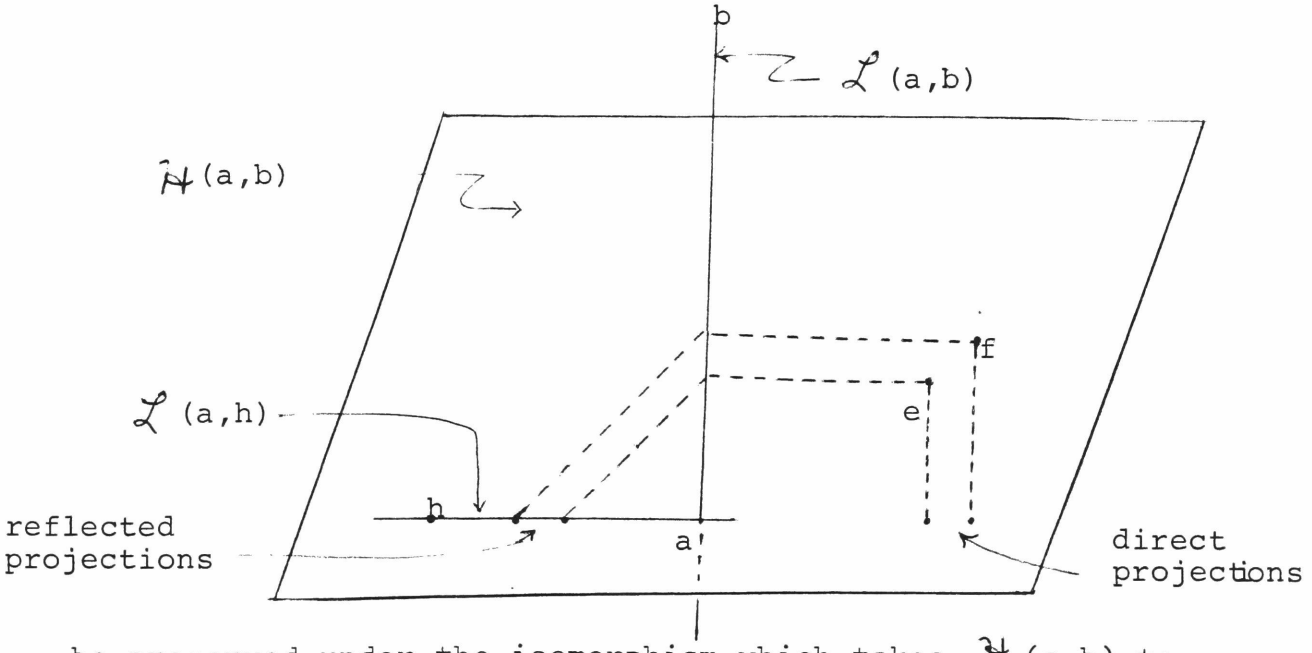
us to project onto $\mathcal{L}(a,b)$ and conclude that it must be isomorphic to $\mathbb{R}$. This can be done in different ways. One way is just to choose any one line in $\mathcal{H}(a,b)$ passing through a and <u>assert</u> that it can be put into one-to-one correspondence with $\mathcal{L}(a,b)$ via the $\rightarrow$ relation as in the figure. To formulate the assertion we must be able to



distinguish two "sides" of the line in $\mathcal{H}(a,b)$, one to project up, the other down. But this is easily done by specifying some point h on the line and using the betweenness relation. More precisely, if $h \in \mathcal{H}(a,b)$ with $a \neq h$, and if $\mathcal{L}(a,h)$ is the line through a and h, then we take $\mathcal{L}^+(a,h)$ to be that portion of the line on h's side (excluding a), i.e. those points $d \in \mathcal{L}(a,h)$ such that $\neg B(d,a,h)$. Similarly we can causally define $\mathcal{L}^-(a,h)$, $\mathcal{L}^+(a,b)$, and $\mathcal{L}^-(a,b)$. With these abbreviations, clause

III(i) is easily formulated.

Finally, III(ii) gives a condition for $e \to f$, where e and f are quite arbitrary, in terms of their respective projections onto $\mathcal{L}(a,b)$ and $\mathcal{H}(a,b)$. The points, of course, have their direct projections onto $\mathcal{H}(a,b)$. They also have reflected projections, first onto $\mathcal{L}(a,b)$ and from there onto $\mathcal{L}(a,h)$ in $\mathcal{H}(a,b)$ -- see figure. The condition imposed by III(ii) for $e \to f$ to hold is that their direct and their reflected projections onto $\mathcal{H}(a,b)$ be congruent and that the projection of e onto $\mathcal{L}(a,b)$ be less than or equal to the projection of f onto $\mathcal{L}(a,b)$. Congruence between vectors in $\mathcal{H}(a,b)$, we know, will necessarily



be preserved under the isomorphism which takes $\mathcal{H}(a,b)$ to (n-1) dimensional Euclidean space. The expression "d' is the reflected projection of e onto $\mathcal{L}(a,h)$" can be cashed

in, of course, as the assertion that d' is on $\mathcal{L}(a,h)$, and there exists a point c on $\mathcal{L}(a,b)$ such that c is the projection of e onto $\mathcal{L}(a,b)$ and either $d' \to c$ or $c \to d'$ depending on whether $ce\,\mathcal{L}^+(a,b)$ or $ce\,\mathcal{L}^-(a,b)$. Of two points $d_1', d_2' e\,\mathcal{L}(a,h)$, we can say that "$d_1'$ is equal to or before d_2'" meaning that:

$d_1' e\,\mathcal{L}^-(a,h)$ & $d_2' e\,\mathcal{L}^+(a,h)$, or $d_1', d_2' e\,\mathcal{L}^+(a,h) \cup \{a\}$ &

$B(a,d_1',d_2')$, or $d_1', d_2' e\,\mathcal{L}^-(a,h)$ & $B(d_1',d_2',a)$.

So clause III is certainly first order in the relation $<<$ (or whatever).

Axioms for M_n ($n \geq 3$):

There exist points a,b where $a << b$ such that:

(I) (i) for every e, there exist unique points c and d which are the respective projections of e onto $\mathcal{L}(a,b)$ and $\mathcal{H}(a,b)$;

(ii) for all points $ce\,\mathcal{L}(a,b)$ and $de\,\mathcal{H}(a,b)$ there is a unique point e such that c and d are the respective projections of e onto $\mathcal{L}(a,b)$ and $\mathcal{H}(a,b)$;

(II) restricted to $\mathcal{H}(a,b)$, the finite axiom set $\mathcal{E}_{n-1}$ obtains;

(III) there is a point $he\,\mathcal{H}(a,b)$ with $h \neq a$, such that:

(i) for all d in $\mathcal{L}^+(a,h)$ there is a unique c in $\mathcal{L}^+(a,b)$ such that $d \to c$

for all d in $\mathcal{L}^-(a,h)$ there is a unique c in $\mathcal{L}^-(a,b)$ such that $c \to d$

for all c in $\mathcal{L}^+(a,b)$ there is a unique d in $\mathcal{L}^+(a,h)$ such that $d \to c$

for all c in $\mathcal{L}^-(a,b)$ there is a unique d in $\mathcal{L}^-(a,h)$ such that $c \to d$

(ii) for all e,f: $e \to f$ iff if d_1,d_2 are the respective projections of e and f onto $\mathcal{H}(a,b)$, and if d_1',d_2' are their respective reflected projections onto $\mathcal{L}(a,h)$, then $C(d_1,d_2,d_1',d_2')$ and d_1' is equal to or before d_2'.

These axioms pay for their clumsiness with a trivial categoricity proof. All one has to do is extract in sequence what one has, with brute force, put in. Again we pose it in terms of $<<$ with the other relations understood as defined.

2.10 Theorem ($n \geq 3$): If $(S,<<)$ is a model of axioms I, II, III, then $(S,<<)$ is isomorphic to M_n.

Proof: Choose points a,b where $a << b$ which satisfy the requirements of I, II, and III. Let $\Gamma : \mathcal{H}(a,b) \to E_{n-1}$ be a bijection which preserves B and C. If $\{x_i\}$ are standard coordinates on M_n, then the image of Γ may be taken to be the set $x_o = 0$ in M_n and $\Gamma(a)$ the

origin. Let h be a point in $\mathcal{H}(a,b)$. Then $\Gamma[\mathcal{L}(a,h)]$ is of course a line in the $x_o = 0$ surface passing through $\Gamma(a)$, and it, together with the x_o line, satisfies clause III in M_n. Using III(i) we can then then set up a bijection $\psi : \mathcal{L}(a,b) \to$ the x_o axis in M_n. Given c on $\mathcal{L}(a,b)$, $\psi(c)$ results from first projecting c onto $\mathcal{L}(a,h)$, taking the image of that point under Γ, and then finally projecting onto the x_o axis.

The two maps ψ, Γ, in view of clause I, define a natural bijection $\phi : S \to M_n$. Given a point e, with respective projections c,d on $\mathcal{L}(a,b)$ and $\mathcal{H}(a,b)$, $\phi(e)$ is taken to be that point whose x_o coordinate is given by $\psi(c)$ and whose x_i coordinates $(i=1,\ldots,n-1)$ are given by $\Gamma(d)$. Clearly, ϕ takes projections and reflected projections in $\mathcal{H}(a,b)$ onto corresponding projections and reflected projections in the $x_o = 0$ surface. But this is all we need to conclude that ϕ preserves $\to$. By III(ii), these projections and reflected projections determine, for all e and f, whether $e \to f$. And since the right-hand side of the equivalence in III(ii) is posed in terms of B and C, it must be preserved under Γ and Γ^{-1}.
So $e \to f$ iff $\phi(c) \to \phi(f)$. qed.

Just as with the case $n = 2$, a symmetric version of the categoricity theorem follows from lemma 2.3. If $\ldots << \ldots$ is the conjunction of the several clauses, then

the axiom $\exists r,s\ (r\tau s\ \&\ \ldots \underset{rs}{<<} \ldots)$ is categorical, i.e. any model (S,τ) must be isomorphic to M_n.

Latzer [12] was the first to prove that Robb's categoricity theorem can be recast in symmetric form. (He construed λ as primitive and the other symmetric relations as defined.)

E. The Definability of Non-Standard Congruence Relations

Having distinguished Robb's definability theorem from his categoricity theorem and proven both, I want to mention two specific confusions about them and their relation to one another. The latter confusion motivates a problem which will also be considered in this section -- that of classifying all non-standard congruence relations which are definable from causal relations.

First, one sometimes hears that any model of Robb's axioms must be isometric (or conformally isometric) to Minkowski space-time. Strictly speaking such a claim makes no sense because the models satisfying his axioms, or ours, are simply sets with a two-place relation on them. They have neither topological nor metrical structure which could be or fail to be the same as that of Minkowski space-time. Of course these claims are meant to refer to the geometric structure which the models may "inherit." But here some care is in order because they inherit simultaneously geometries of different

type. Minkowski space-time may be construed as a vector space over the reals together with a non-definite inner product, or as a space-time (i.e., as a pseudo-Riemannean manifold), or, if one likes, as a Weyl space or some other more general geometric space. If the question is whether an inherited geometry in one of Robb's models is constrained to be Minkowskian, the answer depends on which geometry one has in mind. This is a simple point, but it is worth making precise.

Suppose (S,τ) is a model of Robb's axioms, or ours, and $(V, (,))$ is an inner product space. Then we say that $(V, (,))$ is compatible with (S,τ) iff there is a bijection $\phi : S \to V$ such that for all $a,b \in S$: $a\tau b$ iff $|\phi(b) - \phi(a)| > 0$. We know Minkowski space-time construed as an inner product space is compatible with (S,τ). Suppose (S,τ) is compatible with some other inner product space $(V, (,))$. Then there will certainly exist a composed bijection $\Gamma : V \to M_n$ such that for all v in V, $|\Gamma(v)| > 0$ iff $|v| > 0$. It follows by a variation on the proof of lemma 2.6, that there is a $k > 0$ such that for all v,v' in V: $(\Gamma(v),\Gamma(v')) = k(v,v')$. We have then that up to a dilation factor, Minkowski space-time is the only inner product space compatible with (S,τ).

Anticipating several definitions from the next chapter, we say that a model (S,τ) is compatible with a

space-time (M,g) iff there is a bijection $\phi : S \to M$ such that for all a,b in S: $a\tau b$ iff there is a piecewise smooth timelike curve connecting $\phi(a)$ and $\phi(b)$. As we shall see, if (M,g) is compatible with (S,τ), then (M,g) must be conformally isometric to Minkowski space-time. So, up to a conformal factor, Minkowski space-time is the only space-time compatible with (S,τ)

At this point we could formulate notions of compatibility appropriate to more general geometric spaces and state corresponding characterization theorems. The additional elements of geometric structure that are superimposed would be underdetermined by the initial causal structure. The more general the geometry, the weaker the characterization theorem.

A second confusion has to do with the definability of non-standard congruence relations from causal relations. I do not think it unfair to attribute this confusion to John Winnie based on his remarks in the discussion of his paper [21] at the May 1974 Conference on Space-Time held at the University of Minnesota. It is clear that, in the sense of categorical axiomatizability, the causal structure of Minkowski space-time does not distinguish it from those space-times to which it is conformally isometric. Winnie seemed to suggest that a parallel situation exists with respect to the definability of congruence relations.

This is not the case. Let (M,g) be an arbitrary space-time conformally isometric with Minkowski space-time other than Minkowski space-time itself. (Here we are again anticipating Chapter III and, in particular, the discussion on pp. III, 58-9.) Presumably the claim is that the metrical congruence relation in (M,g) is explicitly definable from causal relations. But it is by no means clear, first of all, that (M,g) carries any general notion of distance or congruence. Any two points will be connectible by curves of length zero, of arbitrarily large (positive) length, and of arbitrarily small (negative) length. One can naturally define a kind of distance function over the domain of points causally related to one another. Of all causal curves connecting two points there will always be a unique one (a geodesic) of maximal (positive) length. In the special case of Minkowski space-time this "maximizing" distance function will agree with that determined by the Minkowski metric. But no counterpart "minimizing" definition is available for the case of spacelike related points. They will always be connectible by spacelike curves of arbitrarily small (negative) length.

It is the case that in (M,g) two arbitrary points are connectible via a unique geodesic. (This is not true in arbitrary space-times.) While the motivation for doing so is not clear, one might just define the distance between arbitrary points to be the length of this unique geodesic. Relative to this notion of distance (M,g) does carry a congruence relation and one can

meaningfully assert that the relation is definable from causal relations. But now the claim, while meaningful, is flatly false. Indeed, the only congruence relation (of this sort) definable from causal relations (-- in any sense of definable, no matter how weak--) is the standard Minkowski relation. This will follow from a general classification result proven later in this section. Quite apart from Winnie's claim, the problem of classifying those "congruence relations" definable from causal relations is of interest because it helps to illuminate Robb's definability theorem.

A congruence relation on Minkowski space-time M_n is at the very least an equivalence relation between pairs of points. If nothing more than this is meant, there are clearly a great slew of rather uninteresting relations definable from causal relations which pass muster. One could render "congruent" only pairs of points whose associated vectors are parallel, or parallel and of equal Minkowski length, etc. Another hybrid relation congruence would agree with the standard congruence relation for, say, time-like related points, but would render congruent pairs of spacelike related points only if their associated vectors were paralletl, etc.

Rather than consider all of these, we will classify a more restricted class of relations. Let us say that a map $d: M_n \times M_n \to \mathbb{R}$ is a (generalized) distance function on

M_n if: (i) $d(a,a) = 0$ for all a; (ii) $d(a,b) = d(b,a)$ for all a,b; and (iii) d is continuous. By a congruence relation in what follows we shall mean an equivalence relation between pairs of points which derives from some such distance function. This characterization does rule out the "congruence relations" mentioned in the previous paragraph. But it is still quite general. Certainly any congruence relation derived from a space-time conformal to Minkowski space-time will be included.

The problem is to classify all congruence relations (in this sense) on M_n which are definable from causal relations. Rather than distinguishing between weaker and stronger senses of explicit definability (according to the richness of the language in which definition is given), we will classify all congruence relations that satisfy a minimal, necessary condition of definability from causal relations. We say, quite generally, that a relation is <u>implicitly definable</u> (<u>from causal relations</u>) if it is preserved under all causal automorphisms of Minkowski space-time.

It is first off quite easy to think of a number of congruence relations which are definable:

$$(1) \qquad \{a,b\} \sim \{c,d\} \text{ iff } |a-b| = |c-d|$$

(We use $\sim$ as a general symbol for congruence between pairs of points.) This is just the standard relation. By

Robb's theorem we know it is explicitly, first order definable ($n \geq 3$).

(2) $\{a,b\} \sim \{c,d\}$ iff $a = a$

(3) $\{a,b\} \sim \{c,d\}$ iff $a\tau b \;\&\; c\tau d \;\&\; |b-a| = |d-c|$ or $\neg(a\tau b \;\&\; c\tau d)$

(4) $\{a,b\} \sim \{c,d\}$ iff $a\sigma b \;\&\; c\sigma d \;\&\; |b-a| = |d-c|$ or $\neg(a\sigma b \;\&\; c\sigma d)$

(2) simply renders congruent all pairs; (3) renders congruent all spacelike or null pairs, but renders timelike pairs congruent iff they are congruent in the standard sense; (4) is like (3) except that the distinguished role played by timelike pairs in (3) is played by spacelike pairs in (4). Clearly these are explicitly, first order definable ($n \geq 3$) too. More interesting is:

(5k) $\{a,b\} \sim \{c,d\}$ iff $|b-a| = |d-c|$ or $a\tau b \;\&\; c\sigma d \;\&\; |b-a| = -k|d-c|$

where k is some positive real. (5) renders pairs congruent iff they are Minkowski congruent or if one pair is timelike, the other spacelike, and the absolute value of the Minkowski length of the timelike pair is k times that of the absolute value of the Minkowski length of the spacelike pair. It follows immediately from Zeeman's characterization

of the causal automorphisms of M_n that this congruence relation is implicitly definable ($n \geq 3$). We know, further, that if k belongs to the Euclidean field over the rationals, it is even explicitly, first order definable ($n \geq 3$).

As we shall now prove, these few possibilities which readily come to mind are the only ones!

2.11 Theorem ($n \geq 2$): If a congruence relation is implicitly definable, then it must be one of those listed.

It follows that for the case $n = 2$, the only implicitly definable congruence relation is (2). The underlined claim above also follows since none of the listed congruence relations is one derived from a space-time conformal to M_n.

In what follows let us suppose that $\sim$ is a congruence relation associated with a distance function d and is implicitly definable. Then for all causal automorphisms ϕ and points a,b,c,d: $d(a,b) = d(c,d)$ iff $d[\phi(a),\phi(b)] = d[\phi(c),\phi(d)]$. The argument proceeds, quite simply and laboriously, by repeated use of this fact. The following notation is useful: for any distinct a and b, $(ab \ldots)$ is the open half line beginning at a and containing b.

Step 1 ($n \geq 2$): Let a and b be distinct. Suppose there are distinct points $c,de(ab \ldots)$ such that $d(a,c) = d(a,d)$. Then $d(a,e) = 0$ for all $ee(ab \ldots)$.

Suppose that c lies between a and d. Then take ϕ to be the dilation which leaves a fixed and takes d to c. Then since $d(a,c) = d(a,d)$, we have $d[a,\phi(c)] = d[a,\phi(d)] = d(a,c)$. Repeating the argument we have for every n, $d[a,\phi^n(c)] = d(a,c)$. Since $\phi^n(c) \to a$ as $n \to \infty$ we have by continuity that $d(a,c) = d(a,a) = 0$. If $e \in (ac \ldots)$, let ϕ' be a new dilation which takes c to e and leaves a fixed. Then $d(a,e) = d(a,\phi'(c)) = d(a,\phi'(d))$ Repeating the very same argument (with e playing the role of c, and $\phi'(d)$ playing the role of d) we have $d(a,e) = 0$.

Step 2 ($n \geq 2$): Suppose a and b are distinct and $d(a,b) = 0$. Then:

(i) if $a \to b$, then for all c and d: $c \to d \Rightarrow d(c,d) = 0$

(ii) " $a << b$, " " " " " ": $c << d \Rightarrow d(c,d) = 0$

(iii) " $a \sigma b$, " " " " " ": $c \sigma d \Rightarrow d(c,d) = 0$.

The argument is the same in all three cases so we may as well prove it for the relation α which stands for any one of them. For any a, let $d_a : M_n \to \mathbb{R}$ be defined by $d_a(b) = d(a,b)$. Since $d_a(a) = d_a(b) = 0$ and d_a is continuous there must exist distinct points $b', b'' \in (ab \ldots)$ such that $d(a,b') = d(a,b'')$. Hence, $d(a,e) = 0$ for all $e \in (ab \ldots)$.

Now let $(cd \ldots)$ be any half line of the same

"αtype" as $(ab \ldots)$. There will exist a causal automorphism taking a to c, and $(ab \ldots)$ to $(cd \ldots)$. It will be the composition of a translation and rotation or pseudorotation in the case $n \geq 3$. In the special case $n = 2$, ϕ will arise from a translation and the right combination of null stretchings. So $d[c,\phi(b')] = d[c,\phi(b'')]$ for distinct $\phi(b')$ and $\phi(b'')$ on $(cd \ldots)$, and hence $d(c,f) = 0$ for all $f \in (cd \ldots)$. In particular $d(c,d) = 0$.

Step 3 ($n \geq 2$): (i) For all a,b: $a \to b \Rightarrow d(a,b) = 0$

(ii) For all a,b,c,d where $a \ll b$ and $c \ll d$: $d(a,b) > 0 \Leftrightarrow d(c,d) > 0$

(iii) For all a,b,c,d where $a \sigma b$ and $c \sigma d$: $d(a,b) > 0 \Leftrightarrow d(c,d) > 0$.

We first prove (ii) -- the argument for (iii) is parallel -- and then use it to prove (i). Suppose first that $a \ll b$ and $a \ll d$. Consider the line segment connecting b and d within $I^+(a)$. Suppose $d(a,b) > 0$ while $d(a,d) \leq 0$ or $d(a,d) > 0$ while $d(a,b) \leq 0$. Then there would have to be a point e on the line segment such that $d(a,e) = 0$, by continuity again. Since $a \ll e$ this would entail $d(a,b) = 0 = d(a,d)$ by the previous step in the argument. And this is a contradiction.

Now assume $a \ll b$ and $c \ll d$. Let ϕ_k: $e \mapsto e + k(a-c)$ define a one parameter family of translations

for $0 \leq k \leq 1$. Since $\phi_1(c) = a$ and $\phi_1(c) << \phi_1(d)$ it follows from the previous paragraph that $d(a,b) > 0$ iff $d[\phi_1(c),\phi_1(d)] > 0$. But the function $d : \mathbb{R} \to \mathbb{R}$ defined by $k \to d[\phi_k(c),\phi_k(d)]$ is continuous. So by the same considerations as in the previous paragraph, $d(0) > 0$ iff $d(1) > 0$, i.e. $d(c,d) > 0$ iff $d[\phi_1(c),\phi_1(d)] > 0$. Hence, $d(a,b) > 0$ iff $d(c,d) > 0$.

Next we prove (i) and give the computational details as a sample. Suppose there were points a and b with $a \to b$ but $d(a,b) \neq 0$. Without loss of generality we assume $d(a,b) > 0$. It follows then that for all c, if $a \sigma c$ then $d(a,c) > 0$. (If $\{c_i\}$ is a sequence of points converging to c where $a \sigma c_i$ for each i, then $d(a,c_i)$ will eventually be positive. Hence the claim follows by (iii) of this step.)

Consider first the case $n \geq 3$. Choose standard coordinates in which $a = (0, \ldots, 0)$ and $(ab \ldots)$ consists of the points $(k,k,0, \ldots, 0)$ for $k > 0$. Let each such point be b_k and let c_k be the point $(0,0,k,0, \ldots, 0)$ for $k > 0$. Since for all k, $d(a,b_k) > 0$ and $d(a,c_k) > 0$, by continuity we can find $k,\bar{k} > 0$ where $d(a,b_k) = d(a,c_{\bar{k}})$. Now there is a Lorentz transformation ϕ (a "pseudo-rotation") which leaves a and $c_{\bar{k}}$ fixed but moves b_k to some other point on $(ab \ldots)$. We shall check this in a moment. But then it follows by Step 1 that $d(a,b) = 0$ which is a contradiction. [It suffices to define ϕ by:

$$\phi:(x_o, \ldots, x_n) \to (5/4\ x_o+3/4\ x_1, 3/4\ x_o+5/4\ x_1, x_2, \ldots, x_n).$$

Then $|\phi(e)| = |e|$ for all e, ϕ is linear, and ϕ preserves temporal orientation. So it is indeed a Lorentz transformation and as claimed $\phi(a) = a$, $\phi(c_{\bar{k}}) = c_{\bar{k}}$, but $\phi(b_k) = (2k, 2k, 0, \ldots, 0) \neq b_k$.]

Finally consider the case $n = 2$. Take standard coordinates $\{x_i\}$ in which $a = (0, \ldots, 0)$ and $(ab \ldots)$ consists of the points $b_k = (k,k)$ for $k > 0$. If $c_k = (0,k)$ for $k > 0$ then again there exist $k, \bar{k}$ such that $d(a,b_k) = d(a,c_{\bar{k}})$ and again to establish a contradiction it suffices to exhibit a causal automorphism which leaves a and $c_{\bar{k}}$ fixed but which moves b_k to some other point on $(ab \ldots)$. No linear map, in general, will do this. But we can find a null stretching map which will. [If $x'_o = x_o - x_1$ and $x'_1 = x_o + x_1$ are null coordinates, define ϕ by: $(x'_o, x'_1) \longmapsto [x'_o, (x'_1)^3/\bar{k}^2]$. ϕ is a causal automorphism and we have $\phi(a) = a$ and $\phi(c_{\bar{k}}) = c_{\bar{k}}$ -- since in x'_o, x'_1 coordinates $c_{\bar{k}} = (-\bar{k}, \bar{k})$. But $\phi(b_k) \neq b_k$ though $\phi(b_k) \in (ab \ldots)$ -- since in x'_o, x'_1 coordinates $b_k = (0, 2k)$ and $\phi(b_k) = (0, 8\ k^3/\bar{k}^2)$.]

Putting together the first three steps we have established that the character of d (whether less than, equal to, or greater than zero) is the same for all timelike pairs and the same for all spacelike pairs. Next we show that if d is not identically zero over all timelike

pairs (resp. spacelike pairs), then within that class, $\sim$ agrees with the standard Minkowski congruence.

Step 4 ($n \geq 2$): If there exist points a and b where $a << b$ (resp. $a\sigma b$) and $d(a,b) \neq 0$ then for all c,d,c',d' where $c << d$ and $c' << d'$ (resp. $c\sigma d$ and $c'\sigma d'$) we have: $d(c,d) = d(c',d')$ iff $|d-c| = |d'-c'|$.

Again the cases $a << b$ and $a\sigma b$ are parallel and we do the former. The argument runs more or less as in the previous steps. Suppose $a << b$ and $d(a,b) \neq 0$. Without loss of generality assume $d(a,b) > 0$. Let c,d,c',d' be arbitrary points with $c << d$ and $c' << d'$.

Suppose first that $d(c,d) = d(c',d') > 0$ although $|c-d| \neq |c'-d'|$, say $|c-d| < |c'-d'|$. Then there would be a causal automorphism ϕ taking c' to c, d' to d, and having the property that for all e,f : $|\phi(e) - \phi(f)| = 1/k|e-f|$ for some $k > 1$. ϕ is the composition of a translation, a pseudo-rotation and a dilation with dilation factor $1/k$. Now we have $d(c,d) = d(c',d')$ and hence $d(c,d) = d[\phi(c'),\phi(d')] = d[\phi(c),\phi(d)] = \ldots = d[\phi^n(c),\phi^n(d)]$. Since $|\phi^n(c) - \phi^n(d)| = 1/k^n|c-d| \to 0$ as $n \to \infty$ it follows that $d(c,d) = 0$ which is a contradiction. So $d(c,d) = d(c',d') \Rightarrow |d-c| = |d'-c'|$.

Conversely suppose $d(c,d) \neq d(c',d')$. Say $d(c',d') > d(c,d) > 0$. Then by continuity there is a point

e'e(c'd' ...) between c' and d' such that d(c',e') = d(c,d). So by the previous paragraph it follows that |c'-e'| = |c-d|. But since e' is between c' and d' we have |c'-e'| > |c'-d'|. Therefore |c-d| > |c'-d'|.

At this point we are almost done with the characterization. We know that $\sim$ must render timelike pairs congruent either always, or precisely when they are congruent with respect to the Minkowski congruence. Similarly for spacelike pairs. It follows that if $\sim$ does not render any timelike pairs congruent with spacelike pairs, then $\sim$ can only be (1), (3), or (4) above. (If all timelike (spacelike) pairs are rendered congruent, d must assign them all 0 and hence they must be congruent with all null pairs too.) It remains to determine that if $\sim$ does render some timelike pair congruent with some spacelike pair then $\sim$ must be of form (2) or (5k) for some k.

Step 5 ($n \geq 2$): If there exist points a,b,c,d such that $a \ll b$ and $c \sigma d$ and such that $d(a,b) = d(c,d) \neq 0$, then there is a $k > 0$ such that for all e,f,g,h if $e \ll f$ and $g \sigma h$ then $d(e,f) = d(g,h)$ iff $|f-e| = -k|h-g|$.

Take k such that $|b-a| = -k|d-c|$. Let ϕ be the causal automorphism which takes e to a and f to b with dilation factor $\ell > 0$ i.e. $|\phi(r) - \phi(s)| = \ell|r-s|$ for all r,s. Then we have: $d(e,f) = d(g,h)$ iff $d(a,b) = d[\phi(g),\phi(h)]$ iff $d(c,d) = d[\phi(g),\phi(h)]$ iff $|d-c| = |\phi(g)-\phi(h)|$

(by the previous step) iff $|b-a| = -k\ell|h-g|$ iff $\ell|f-e| = -k\ell|h-g|$ iff $|f-e| = -k|h-g|$.

This completes the proof of the theorem. (There should be a faster, more direct argument.)

F. Definability of Standard and Non-Standard Topologies

In this section we discuss the definability of standard and non-standard topologies from causal relations as the definability of congruence relations was discussed in sections C and E.

To define a topology is to specify which subsets are "open." For the standard Minkowski topology this is easily done in the following way:

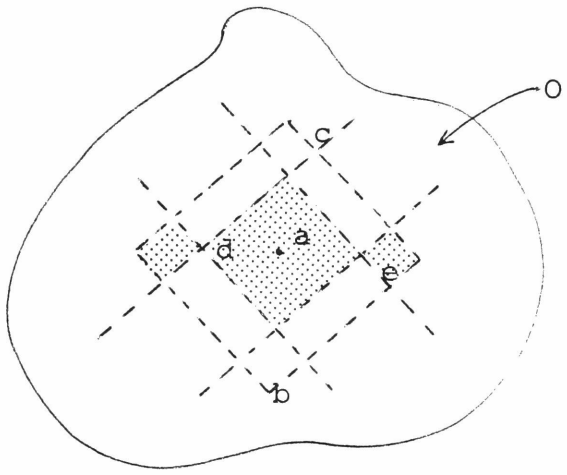
(Euclidean topology) O is open $\Leftrightarrow \forall a[ae0 \Rightarrow \exists b\ \exists c(be0 \ \& \ ce0 \ \& \ b \ll a \ll c \ \& \ I^+(b) \cap I^-(c) \subseteq 0)]$.

The equivalence is proven with a trivial computation which establishes that "chronology diamonds" of the form $I^+(b) \cap I^-(c)$, where $b \ll c$, can always be inscribed in (Euclidean) open balls. Recall that a basis for a topology is a set of sets $\mathcal{B}$ such that: (i) $\cup\{B : B\in\mathcal{B}\}$ = whole space; and (ii) given any B_1 and B_2 in $\mathcal{B}$, $B_1 \cap B_2 = \cup\{B : B\in\tilde{\mathcal{B}}\}$ where $\tilde{\mathcal{B}}$ is some subset of $\mathcal{B}$. The topology generated by the basis $\mathcal{B}$ is simply the collection of sets which are (possibly empty) unions of sets in $\mathcal{B}$. In this language we can say that the sets $I^+(b) \cap I^-(c)$ form a

basis for a topology, the Alexandrov topology, and that the Alexandrov topology is equal to the Euclidean topology.

The defining condition above can also be given in terms of the symmetric causal relations.

(Euclidean topology) O is open $\Leftrightarrow$ $\forall a[aeO \Rightarrow \exists b,c,d,e(beO \,\&\, ceO \,\&\, deO \,\&\, ceO \,\&\, B_\tau(b,a,c) \,\&\, B_\tau(b,d,c) \,\&\, B_\tau(b,e,c) \,\&\, a \notin J(d) \cup J(c) \,\&\, I(b) \cap I(c) - (J(d) \cup J(e)) \subseteq O)]$ See diagram



The shaded area is the set $I(b) \cap I(c) - (J(e) \cup J(d))$

Here the sets $I(b) \cap I(g) - (J(d) \cup J(e))$, which are open because the J sets are closed, form a basis of the symmetric Alexandrov topology, which is equal to the Euclidean topology.

Extrapolating from the form of these causal definitions, we say that a topology is explicitly first order

definable from causal relations if it satisfies an equivalence of the form: 0 is open $\Leftrightarrow$ _____ where the condition on the right is formulated in a first order language with non-logical symbols '0', and 'e' in addition to '<', 'κ' (or whatever). We also say that a topology is <u>implicitly definable</u> from causal relations iff openness is preserved by all causal automorphisms, or, equivalently, if the causal automorphisms form a subset of the homeomorphisms of Minkowski space-time onto itself with respect to that topology.

Quite clearly there are many non-standard topologies which are explicitly first order definable from causal relations. The following are a few examples:

(Indiscrete Topology) 0 is open $\Leftrightarrow \forall a(ae0) \vee \forall a(a \not e 0)$

(Discrete Topology) 0 is open $\Leftrightarrow \exists a(ae0 \vee a \not e 0)$.

(I^+ Topology) 0 is open $\Leftrightarrow \forall a[ae0 \Rightarrow \exists b(be0 \ \& \ b << a \ \& \ I^+(b) \subseteq 0)]$.

This is the topology generated by the basis sets $I^+(b)$. It is strictly coarser than the Euclidean topology and is T_0 but not T_1 (and hence not Hausdorff).

(J^+, I^- Topology) 0 is open $\Leftrightarrow \forall a[ae0 \Rightarrow \exists b \exists c(be0 \ \& \ ce0 \ \& \ b<a \ \& \ a<<c \ \& \ J^+(b) \cap I^-(c) \subseteq 0)]$.

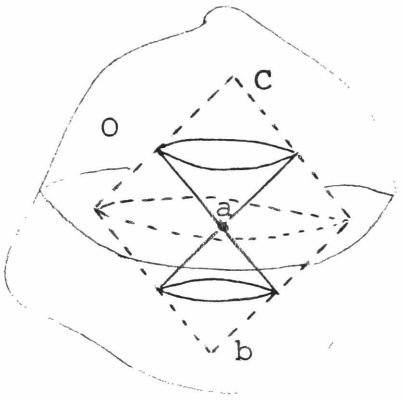
This is the topology generated by basis sets $J^+(b) \cap I^-(c)$. It is finer than the Euclidean topology. It induces the

discrete topology on all spacelike sets, and induces the half open interval topology on time lines and null lines. One can easily add to this list by choosing other combinations and iterations of the $I^+, I^-, J^+, J^-, E^+, E^-$ sets for bases.

It seems quite natural to ask what is so special about the Euclidean topology from the causalist's point of view. Some definable topologies might be ignored because they are not even Hausdorff (e.g. I^+ topology). But others are even finer than the Euclidean topology (e.g. J^+, I^- topology). Nor will it do to dismiss all non-standard topologies as mathematical curiosities devoid of physical significance. From the causalist's point of view, at least, it would seem that some of them are of more obvious physical significance than the Euclidean topology.

If $\mathcal{T}$ is a topology on Minkowski space-time it is implicitly definable from causal relations, we said, if the causal automorphisms form a subset of the $\mathcal{T}$ -homeomorphisms. If the causal automorphisms fully exhaust the set of $\mathcal{T}$ -homeomorphisms, we say that $\mathcal{T}$ is a Zeeman topology (Zeeman [23].) One such is the following:

(Zeeman Topology): O is open $\Leftrightarrow \forall a[ae0 \Rightarrow \exists b\ \exists c(be0\ \&\ ce0\ \&\ b<<a<<c\ \&\ (I^+(b) \cap I^-(c) - E(a)) \cup \{a\} \subseteq 0)]$

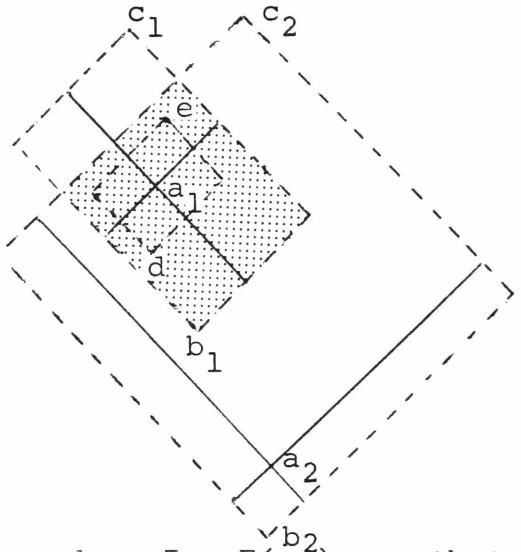


$[I^+(b) \cap I^-(c) - E(a)] \cup \{a\}$

The Zeeman topology is generated by basis sets which consist of the usual $I^+(b) \cap I^-(c)$ sets, where $b \ll c$, from which the points on the null cone of an interior point, except for the point itself, have been deleted. This topology is not the one investigated by Zeeman in [23] which he calls the "fine topology," but is equivalent to another which he mentions in passing. Obviously it is explicitly first order definable from causal relations. We now describe some of its properties and verify that it is, in fact, a "Zeeman topology." This should show, by example, what one can do with nonstandard topologies.

First, it should be clear that sets of the form $[I^+(b) \cap I^-(c) - E(a)] \cup \{a\}$ where $b \ll a \ll c$ (call them "Z sets") do indeed form a basis for a topology which is strictly finer than the Euclidean topology. Given any Euclidean open set O ("E open" as against "Z open") and a point $a \in O$, we can find points $b,c \in O$ with $b \ll a \ll c$ such that $I^+(b) \cap I^-(c) \subseteq O$. Hence $a \in (I^+(b) \cap I^-(c) - E(a)) \cup \{a\} \subseteq O$. So every E open set is the union of Z sets. If

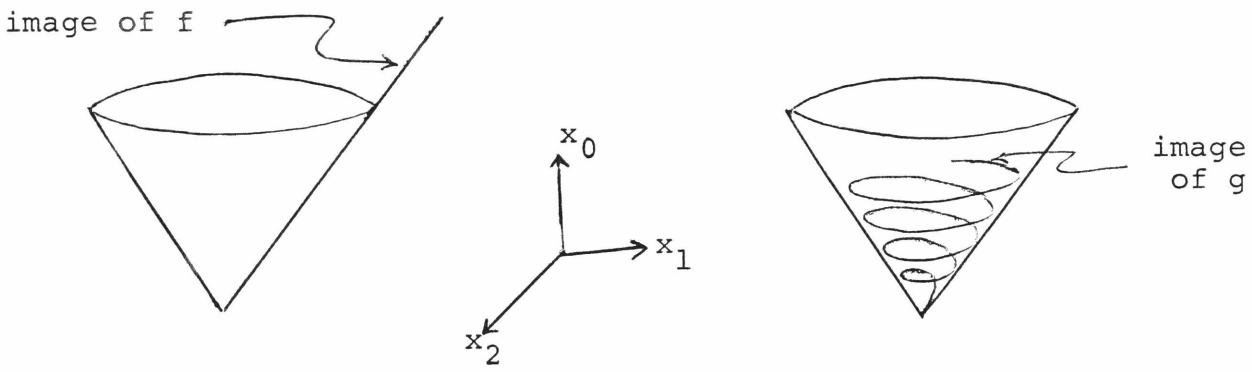
$Z_1 = [I^+(b_1) \cap I^-(c_1) - E(a_1)] \cup \{a_1\}$ and $Z_2 = [I^+(b_2) \cap I^-(c_2) - E(a_2)] \cup \{a_2\}$, then too it follows that $Z_1 \cap Z_2$ is the union of Z sets. Let $I = I^+(b_1) \cap I^-(c_1) \cap I^+(b_2) \cap I^-(c_2)$. If neither a_1 nor a_2 is in I, $Z_1 \cap Z_2 = I - E(a_1) \cup E(a_2)$ is E-open and so certainly must be the union of Z sets. If a_1 or a_2 (say a_1) is in I (see figure), then $Z_1 \cap Z_2 = (I-E(a_1)) \cup \{a_1\}$.



$Z_1 \cap Z_2 = (I-E(a_1)) \cup \{a_1\}$ is shaded

By choosing d, eeI - $E(a_1)$ so that $d << a_1 << e$, we have that $Z_1 \cap Z_2 = (I-E(a_1)) \cup [(I^+(d) \cap I^-(e)-E(a_1)) \cup \{a_1\}]$ and so in this case too $Z_1 \cap Z_2$ is the union of Z sets. If both a_1 and a_2 are in I, then $Z_1 \cap Z_2 = [I- E(a_1) \cup E(a_2)] \cup \{a_1,a_2\}$ and by inscribing <u>two</u> small Z-sets one can recapture both a_1 and a_2 as a_1 was recaptured in the previous case. So the Z-sets do form a basis for a topology which is strictly finer than the Euclidean topology.

Its fineness exhibits itself in the following way. In M_3 consider the two curves $f : [0,\infty) \to M$ and $g : [0,\infty) \to M$ defined by $f : t \to (t,t,0)$ and $g : t \to (t, t \sin t, t \cos t)$. Both curves are E-continuous but neither is Z-continuous at $t = 0$. This is because no Z-set



"centered" at the origin will include any points in the image of the curve other than the origin itself. More generally, an E-continuous curve will be Z-continuous at a point a iff no (E connected) segment of the curve containing a lies in the null cone of a.

The special topological significance given to curves falling within null cones is what is crucial here. Zeeman's idea is to code the null cone structure of Minkowski space-time into a hybrid topology. It should not be surprising that he is able to recover what he, with hindsight, so carefully encoded. The recovery theorem, stated with proof in Zeeman [23], is:

2.12 Theorem ($n \geq 3$): If $\phi : M_n \to M_n$ is a bijection then ϕ is a homeomorphism with respect to the Z-topology iff ϕ is a symmetric causal automorphism.

It follows, of course, by the symmetric version of the Zeeman theorem above, that the group of Z-homeomorphisms is generated by the translations, dilations, orthochronous homogeneous Lorentz transformations, and anti-orthochronous homogeneous Lorentz transformations.

Proof: If ϕ is a symmetric causal automorphism, then ϕ, ϕ^{-1} will take Z-sets to Z-sets; so they will certainly be Z-continuous.

Assume conversely that ϕ is a Z-homeomorphism. We note first that ϕ must be an E-homeomorphism as well. [This is so because of the fact that a Z-open set O is E open iff for every point $a \in O$ there is a Z-open set $O' \subseteq O$ containing a which has the property that $O' - \{a\}$ is Z-connected.] It follows that given any point a there is an E-open set O containing a such that for all b in O, $a \lambda b$ implies $a' \lambda b'$. (Primes will denote images under ϕ.) This is a local version of the claim to be established. To see this, suppose to the contrary that we could find a nested sequence of E open sets O_i whose intersection is $\{a\}$, and we could find points $b_i \in O_i$ with $a \lambda b_i$ but not $a' \lambda b_i'$. Since the b_i E-converge but not Z-converge to a, the b_i' E-converge but not Z-converge to a'. But since

$\neg(a'\lambda b_i')$ for all i, this is impossible. If there were a Z-open set $(I^+(c) \cap I^-(d) - E(a)) \cup \{a\}$ which did not eventually contain all the b_i', then the E-open set $I^+(c) \cap I^-(d)$ would not eventually contain all the b_i' either.

Suppose now that $a \to b$ but $\neg(a'\lambda b')$. Consider the null segment [ab] connecting them. By the previous paragraph we know that $a'\lambda d'$ for all points d in some initial (half E-open) subsegment [ac). Let c_o be the first point in [ab] such that $\neg(a'\lambda c_o')$. If $\{c_i\}$ is a sequence in $[ac_o)$ E-converging to c_o, then the c_i' must all be in E(a') though they E-converge to a point not in E(a'). This is impossible.

We conclude that $a\lambda b$ implies $a'\lambda b'$. The reverse direction is, of course, symmetrical. So ϕ is a symmetrical causal automorphism. qed.

The "recovery theorem" is only valid for dimension $n \geq 3$. The argument to see that it is false for $n = 2$ is precisely the same as that employed above to establish that for $n = 2$ the relations $\to$ and κ are not definable from λ. Turning M_2 on its side by rotating 90° takes Z-sets to Z-sets, but destroys the relations τ and κ.

We have established that the group of homeomorphisms of the Zeeman topology is the group of symmetric causal automorphisms. We can push this line still further and explicitly

define from causal relations a "super Zeeman topology" whose associated group of homeomorphisms is the group of (asymmetric) causal automorphisms. In effect we code into the topology not only the cone structure, but the temporal orientation as well.

(super Zeeman topology) O is open $\Leftrightarrow \forall a[ae0 \Rightarrow \exists b\ \exists c(be0\ \&\ ce0\ \&\ b<<a<<c\ \&\ (I^+(b)\cap I^-(c)-E^+(a))\cup\{a\} \subseteq 0)]$.

The basis sets for this topology are sets of the form $(I^+(b) \cap I^-(c) - E^+(a)) \cup \{a\}$. The future null cone of a (except for a itself) is removed from $I^+(b) \cap I^-(c)$, rather than the entire cone as in the Zeeman topology. With almost precisely the same arguments as just used, we can show that if ϕ is a homeomorphism of the super Zeeman topology, then $a \to b$ implies $\phi(a) \to \phi(b)$. So in dimension $n = 2$ as well as $n \geq 3$ ϕ must be an (asymmetric) causal automorphism.

2.13 Theorem ($n \geq 2$): If $\phi : M_n \to M_n$ is a bijection, ϕ is a homeomorphism with respect to the super Zeeman topology iff ϕ is a causal automorphism.

The Zeeman and super-Zeeman topologies do have properties which from the usual viewpoint are pathological. The former induces the discrete topology on all null lines so that the trajectories of photons are nowhere continuous. The latter induces the half open ( ] interval topology on all

null lines so that the trajectories of photons are still nowhere continuous, but are everywhere lower semi-continuous. [Photons depart discontinuously, but arrive continuously?] On the other hand, both do induce the standard Euclidean topology on all timelines and they single out via their homeomorphisms the group of symmetric (resp. asymmetric) causal automorphisms. That might be considered of great physical significance to causal theorists. Zeeman's own starting point is one questioning the physical significance of the Euclidean topology for Minkowski space-time. His claim is that the Euclidean topology fails to reflect the non-isotropy of space-time and that its group of homeomorphisms is unwarrantedly large.

CHAPTER III

CAUSAL STRUCTURE IN GENERAL RELATIVISTIC SPACE-TIMES

The notion of "causal structure" is readily extended from the special case of Minkowski space-time to the larger class of space-times considered in general relativity. Once this is done Robb-type questions about the recoverability of topological and metrical structure from causal structure suggest themselves. Several such questions will be formulated and answered in this chapter.

The starting point of the investigation is a well-known result of Kronheimer and Penrose [10] that in all strongly causal space-times (as in Minkowski space-time) the Alexandrov topology is equal to the manifold topology and hence the latter is explicitly definable from (asymmetric) causal relations. An unpublished theorem of Hawking [7] extends the result establishing that any (asymmetric) causal isomorphism between strongly causal space-times must not only be a homeomorphism, but a smooth conformal isometry as well. These results will here be extended in several directions. We show that: (1) Hawking's theorem generalizes to the larger class of past and future distinguishing space-times; (2) the result of (1) is in a sense best possible since it cannot be extended to the class of space-times which are past or future distinguishing but not both; (3) it falls out of the proof of (1) that in any space-time topological, differential, and conformal structure are uniquely determined by the class of (point set images of) causal curves.

Of necessity the chapter presupposes a knowledge of basic topology and differential geometry.

A. Space-times and their Physical Interpretation

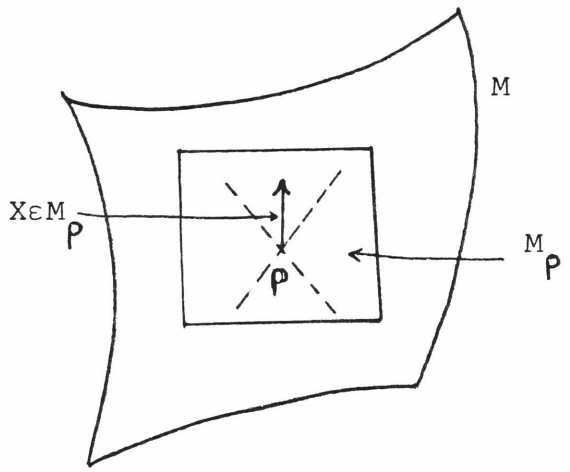
Relativistic space-times, or just plain space-times,

are geometric models which in a quite natural way generalize the structure of Minkowski space-time. They are models for the spatio-temporal structure of the universe.

Space-times are, first of all, differentiable manifolds. They are standardly taken to be smooth (C^∞) though this is largely a matter of convenience; nothing beyond being C^2 or C^3 ever seems to enter into proofs. They are standardly taken to be connected and without boundary. The first condition is reasonable because even if the universe could be disconnected it is natural for us to limit investigation to "our component." Indeed it is not clear how we could ever obtain knowledge of, or even just establish the existence of, any other. The condition that space-times be without boundary is again largely a matter of convenience. But it can also be defended with the vague intuition that there is no "edge" to space-time. God's work is not so untidy. The same intuition motivates other conditions that are often imposed on space-times, e.g. inextendibility, local inextendibility, the condition of being "hole-free," geodesic completeness, etc. Some of these will be discussed later.

As differentiable manifolds (of dimension 4, or more generally dimension $n \geq 2$), space-times carry both topological and differential structure. They also carry a metrical structure which arises, intuitively, from the assignment of a Minkowski type metric at each point. Space-times are taken to be endowed with a smooth, non-degenerate, pseudo-Riemannian metric of Lorentz signature (+1, -1, -1, -1) or, in general,

(+1, -1, ..., -1). Such a metric assigns to each point p in the manifold M a symmetric bilinear map g_p: $M_p \times M_p \to R$, where M_p is the tangent space to the manifold at p. Most important for our purposes, the metric defines a "null cone" in each M_p. A vector $X \in M_p$ is said to be <u>timelike</u> (<u>null</u>, <u>causal</u>, <u>spacelike</u>) if $g_p(X,X) > 0$ ($g_p(X,X) = 0$, $g_p(X,X) \geq 0$, $g_p(X,X) < 0$). The condition that the metric be non-degenerate i.e. that for all points p there is no vector $X \in M_p$ such that $g_p(X,Y) = 0$ for all $Y \in M_p$, can be understood as asserting that the null cone at each point is not a collapsed or "degenerate" cone. The assumption that the metric be smooth is again adopted for convenience and simplicity only.



A non-degenerate pseudo-Riemannian metric of signature (+1, -1, ..., -1) has the property that for each point there is a coordinate frame $\{x_i\}$ on an open neighborhood of the point such that <u>at the point itself</u> the metric assumes the form $ds^2 = dx_0^2 - dx_1^2 - \ldots - dx_{n-1}^2$. In this sense the metric is <u>pointwise</u> or "infinitesimally" Minkowskian. (In general pointwise or infinitesimal properties are ones which obtain in the tangent space of a point.) This is not to be confused with the very much more stringent condition that the metric

be locally Minkowskian in the sense that there be a coordinate frame $\{x_i\}$ on an open neighborhood of each point such that at all points in the neighborhood the metric assumes the form $ds^2 = dx_0^2 - \ldots - dx_{n-1}^2$. This condition, equivalent to the vanishing of the Riemann curvature tensor, is flatness.

One final condition that is usually built into the definition of a space-time is temporal orientability i.e. the condition that M admit a continuous non-vanishing timelike vector field. At every point of a space-time manifold the null cone has two lobes. Temporal orientability, intuitively, is the condition that it be possible to assign labels of "future lobe" and "past lobe" to cones over the entire manifold in a continuous manner -- that it be possible to assign a globally consistent "direction of time." A temporal orientation is a specification of just such a direction to time. It is a specification of a continuous, non-vanishing timelike vector field. It is an interesting question whether the spatio-temporal structure of the universe really must be temporally orientable or not. Certainly in the context of much (local) work on relativity theory, the assumption plays no role at all. But there does not seem to be any interesting theory of global causal structure which does not presuppose temporal orientability. For this reason we will assume it. On the other hand, it will emerge that work on causal structure can be developed without

presupposing a specification of temporal orientation. In particular, as will be shown in section D, results concerning the recoverability of spatio-temporal structure from causal structure can be cast in a form involving only symmetric causal relations. Until then, however, we will assume that a space-time does come equipped with a temporal orientation. Summarizing what we have said so far, we have the following definition:

3.1 Definition: A (relativistic) space-time is a connected, smooth, four (or n) dimensional differentiable manifold without boundary M, together with a smooth, non-degenerate pseudo-Riemannian metric of Lorentz signature g on M, together with a temporal orientation of (M,g).

A space-time inherits from its pseudo-Riemannian structure component conformal and projective structures. The former is that structure given by the null cones at each point. Two metrics g and $\bar{g}$ on M are said to be conformally equivalent or conformally isometric if at each point p and for all vectors $X \in M_p$, X is causal with respect to g_p if and only if X is causal with respect to $\bar{g}_p$, i.e. $g_p(X,X) \geq 0 \leftrightarrow \bar{g}_p(X,X) \geq 0$. [One could equally well require that X be timelike (null, spacelike) with respect to g if and only if it be timelike (null, spacelike) with respect to $\bar{g}$. The conditions are equivalent] Conformal structure on a manifold can be thought of as being given by an equivalence class of conformally isometric pseudo-Riemannian metrics. There are other forms in which the condition of conformal equivalence between metrics g and $\bar{g}$ can be stated. By the argument in the proof of Lemma 2.6, it is

equivalent to the condition that at each point p there is a real, positive k such that for all $X,Y,\varepsilon M_p$: $\bar{g}(X,Y) = k\, g_p(X,Y)$. This is the pseudo-Riemannian version of the claim that metrics are conformally equivalent if and only if they agree on all "angles" between vectors. Given a differentiable curve γ: $I \to M$ where I is a connected subset of $\mathbb{R}$, we say that <u>the curve is causal</u> (timelike, null) if its tangent vector at each point is non-zero and causal (timelike, null). Clearly, then, g and g' are conformally isometric if and only if they admit the same causal (timelike, null) curves. With a simple computation one can also verify that they are conformally isometric if and only if they admit the same null geodesics. Finally, two smooth metrics g and $\bar{g}$ on M are conformally isometric if and only if there is a smooth, non-vanishing map Ω: $M \to \mathbb{R}$ such that at all p, $\bar{g}_p = \Omega^2(p) g_p$, or, to abbreviate, $\bar{g} = \Omega^2 g$. Two space-times (M,g) and (M',g') are, naturally, said to be conformally isometric if there is a smooth diffeomorphism $f = M \to M'$ such that g' and $f_* g$ are conformally equivalent on M.

Projective structure, in contrast, is that structure given by the class of geodesics. Two metrics g and $\bar{g}$ on M are projectively equivalent if for every differentiable curve γ: $I \to M$, where I is a connected subset of $\mathbb{R}$, γ is a geodesic of g if and only if it is a geodesic of $\bar{g}$. Correspondingly, we say two space-times (M,g) and (M',g') are projectively equivalent if there is a smooth diffeo-

morphism f: $M \to M'$ such that γ is a geodesic in (M,g) if and only if $f \circ \gamma$ is a geodesic in (M',g'). It is a standard result of differential geometry that two pseudo-Riemannian metrics g and $\bar{g}$ are projectively equivalent iff in any coordinate system the components of their unique derived affine connections satisfy:

$$(*) \qquad \bar{\Gamma}^{i}_{jk} - \Gamma^{i}_{jk} = \delta^{i}_{j}\phi_{k} + \delta^{i}_{k}\phi_{j}$$

for some vector field with components ϕ_k.

Using (*) we can give a simple computational verification of the fact that <u>if two metrics g and $\bar{g}$ on M are both conformally and projectively equivalent then they can differ by at most a constant factor</u>, i.e. $\bar{g} = \Omega^2 g$ where Ω is constant. Recall that Γ^{i}_{jk} can be expressed as:

$$(**) \qquad \Gamma^{i}_{jk} = 1/2 g^{il}\left[g_{lj,k} + g_{lk,j} - g_{jk,l}\right]$$

where $g_{lj,k} = \frac{\partial g_{lk}}{\partial x_k}$, etc. Substituting $\bar{g} = \Omega^2 g$ into (**) yields:

$$(***) \qquad \bar{\Gamma}^{i}_{jk} - \Gamma^{i}_{jk} = \Omega^{-1}\left[\delta^{i}_{j}\Omega_{,k} + \delta^{i}_{k}\Omega_{,j} - g^{il}g_{jk}\Omega_{,l}\right]$$

where $\Omega_{,k} = \frac{\partial \Omega}{\partial x_k}$, etc. Now given any i, choose $j = k \neq i$. From (*) and (***) jointly we obtain: $g^{il}g_{kk}\Omega_{,l} = 0$. We can certainly choose coordinates so that (at any point) $g_{kk} \neq 0$. Hence $g^{il}\Omega_{,l} = 0$ for all i. Finally, multiplying by g_{ik} and summing yields $\Omega_{,k} = 0$ for all k. So Ω must be constant as claimed.

The underlined fact is usually cited in connection with the

physical interpretation of conformal and projective structure. It is taken to be a fundamental assumption of relativity theory that: i) light rays travel along (piecewise) null geodesics, and also that: ii) free particles (i.e. particles subject to no forces other than gravity) travel along (piecewise) timelike geodesics. In a highly idealized, schematic sense, then, we can determine the space-time metric up to a conformal factor by observing the trajectories of light rays. (Part of the idealization enters in the assumption that there happen to be sufficiently many light rays around to provide adequate data.) Similarly, by observing the paths of free particles we can determine the metric up to a projective factor. Thus, given a background assumption that conformal and projective structure are derived from a pseudo-Riemannian metric, we can determine the metric up to a constant factor, a factor which is usually conceived as a choice of units.

Without wishing to enter into an extended discussion of this standard gloss, I want to make a few comments. First, the fact that two metrics agree on the subclass of timelike geodesics does not <u>on the face of it</u> entail that they agree on all geodesics. And we certainly do not have observational access to the class of spacelike geodesics, not even in the highly idealized sense in which we do have access to timelike and null geodesics. Thus

without further argument it is not clear that, as claimed, by observing the paths of free particles, or with any observations whatsoever, we can determine the metric up to a projective factor. Further argument, however, is not difficult to provide. One can prove that if two space-time metrics agree on which curves are timelike geodesics, then they must agree on all geodesics i.e. if given all differentiable curves γ: $I \to M$, γ is a timelike geodesic of g if and only if γ is a timelike geodesic of $\bar{g}$, then g and $\bar{g}$ must be projectively equivalent. One proves that agreement on timelike geodesics so constrains the components of the respective affine connections of g and $\bar{g}$ that (*) above must hold.

But agreement on timelike geodesics certainly also constrains g and $\bar{g}$ to be conformally equivalent. This is immediate. Just the fact that all timelike geodesics of g are timelike curves of $\bar{g}$ (geodesic or not), and symmetrically, forces this conclusion. So the condition that metrics agree on all timelike geodesics forces them to agree up to a constant factor. In a sense, then, fundamental assumption (ii) that free particles traverse timelike geodesics, together with our idealized set of observations, determines the metric (up to a constant factor). Any further invocation of assumption (i), that light rays traverse null geodesics, is redundant. This is the second point.

Third, it is misleading to construe (i) as a fundamental

assumption of relativity theory. That light does travel along null geodesics is a consequence of the particular equations which govern the electromagnetic field. One could make good sense of relativistic physics even if electromagnetic theory were quite different so that, for example, electromagnetic waves, like sound waves, traveled at less than the maximal possible velocity. What is essential to relativity theory is that there is such a bound to the velocity with which physical bodies, signals, or interactions can propagate. So it is appropriate to formulate the first conformal-structural-fixing-assumption in the following form which makes no particular reference to light or electromagnetic waves: (i') all physical bodies, signals, and interactions traverse causal curves in space-time. This principle does rule out the existence of "tachyons" but is certainly embraced by relativity theory in its standard (classical) form.

Fourth, there is a question whether (ii) really can be used to empirically determine the space-time metric up to a projective factor, even in the highly idealized manner described. Here I do <u>not</u> have in mind the point that (ii) as it stands is false insofar as gravitationally dipole test particles will not traverse timelike geodesics. Quite apart from this there are difficulties

reflecting a significant difference between (ii) and both (i) and (i'). The application of (ii) to determine the projective structure of space-time presupposes the availability of test particles, i.e. it presupposes the absence of non-gravitational forces. This is an assumption above and beyond the idealizing assumptions needed for the application of (i) or (i') -- that there happen to be many particles and/or signals around. We can imagine that in some chaotic region of space-time signals are beamed in all directions from all points, yet all are being perturbed in their trajectories by some cluster of forces which are present. Now it might be the case that from a knowledge of the dynamic effects of these forces on different sorts of bodies and signals we could correct for perturbation and reconstruct the paths they would otherwise traverse. In this way we could at least indirectly determine projective structure. The point I want to make here, however, is that this indirect procedure requires much more than an acceptance of relativity theory. It requires in addition a knowledge of what forces happen to be present and a knowledge of their dynamical effects. In contrast, to invoke (i) or (i') one need not know anything about what forces happen to be present or about their dynamic effects. One may ignore such considerations altogether.

A fifth point is that when discussing the observational determination of conformal and projective structure one must

be sensitive to the difference between "curves" and their images. Straightforwardly, to assert that two metrics g and $\bar{g}$ are, say, conformally equivalent one must establish that for all differentiable curves γ: $I \to M$, γ is a causal curve with respect to g if and only if it is a causal curve with respect to $\bar{g}$. But even in our wildly idealized laboratory situation we have, at best, observational access to the ordered point set images of causal curves. We can only check-off which space-time points are occupied by a given signal or particle, and in what order. For this reason we would like to be able to construe the underlined mathematical result in a manner which refers only to images of null geodesics (or causal curves) and timelike geodesics.

This can be done but things are just a bit delicate. Given an ordered point set on a manifold there is a clear sense in which we may say that it forms a continuous curve. But it is not on the face of it meaningful to assert that the ordered point set forms a differentiable curve. Differentiability is parametrization dependent. The best one can do is say that it admits of differentiable parametrization. Accordingly, to say a continuous curve-image is a (differentiable) causal curve, null geodesic, timelike geodesic, etc. is to say that it admits a parametrization with respect to which it is a causal curve, null geodesic, timelike geodesic, etc.

On this construction it is clear what it means to assert

that a mapping between space-times preserves causal curves, etc. while referring only to images of curves. So the following formulation is unambiguous.

3.2 Theorem: If (M,g) and (M',g') are space-times and $f: M \to M'$ is a smooth diffeomorphism where both f and f^{-1} preserve causal curves and timelike geodesics (or just the latter), then g' and $f_* g$ agree up to a constant factor.

A sixth and final point will help to motivate a technical question considered in section C (Theorem 3.30). Suppose we are committed to the belief that our universe has the structure of a space-time in the sense of the definition above and suppose we are committed to principles (i') and (ii). Further let us grant the idealizing assumptions needed for the invocation of (i') and (ii) that in any region of space-time where we want to perform our measurements physical bodies and signals are being sent in all possible directions at all possible velocities. Let us also overlook the problem mentioned in connection with the invocation of (ii) -- that in our laboratory region of space-time non-gravitational forces may be present.

Even with all these assumptions does it follow that we can "read-off" the metrical structure of space-time (up to a constant factor)? The assumptions are meant to be so strong that we have a kind of empirical access to which ordered point sets are causal curves and which timelike geodesics. Still it is not clear that this in conjunction with the theorem is enough to uniquely determine metrical structure. It is not because the manifold structure of space-time is not

empirically "given" the way it is in the formulation of the theorem. It remains to be established that the manifold (topological + differential) structure is uniquely determined by the specification of causal curves, timelike geodesics, or both.

What we want to know is whether the following is true:

> Claim: If (M,g) and (M',g') are space-times and if $f: M \to M'$ is a bijection where f and f^{-1} preserve causal curves and timelike geodesics, then f is a smooth diffeomorphism and (hence) g' and $f_* g$ agree up to a constant factor.

In section C we do establish precisely this. It follows that the standard gloss about the empirical recoverability of space-time structure can be rendered in a new strong form.

The notion of a "space-time" as defined above is an exceedingly general one. No mention is made of Einstein's equation (which imposes constraints on the space-time metric g), though traditional accounts of general relativity give it a central role. Further, almost no constraints are placed on the character of the underlying manifold M; all sorts of mathematical curiosities are admitted. [The qualification "almost" is needed since the very admissibility of a smooth (or just continuous) non-degenerate pseudo-Riemannian metric of Lorentz singature is a constraint. A manifold will admit such a metric, in general, if and only if it is paracompact and either non-compact, or compact with vanishing Euler characteristic. For example, the two sphere S^2 does not admit one. Hawking and Ellis [8], pg. 39-40]. We close

this section with a few remarks about further constraints on the structure of space-time which are often imposed.

First, Einstein's equation. So far as general relativity goes, space-time is only half the story. The other concerns its material content, where matter is so understood as to include any repository of energy-momentum such as, for example, an electromagnetic field. It is assumed that the "matter fields" which populate space-time are described by a number of tensor fields on the underlying manifold M and that these tensor fields satisfy "field equations" involving the space-time metric g. It is further assumed that the various matter fields contribute to an aggregate energy-momentum field described by a symmetric, second order covariant tensor T. Einstein's equation:

$$R_{ij}-1/2g_{ij}R+\lambda g_{ij} = T_{ij}$$

correlates the space-time metric g with this energy-momentum tensor T. The comparison with classical electromagnetic theory is instructive. In that theory the subject matter is the electromagnetic field and its dynamical behavior is correlated with the distribution of charge and current sources by Maxwell's equation. In general relativity the subject matter is the space-time metrical field and its dynamical behavior is correlated with gross energy-momentum sources by Einstein's equation.

One line of research in relativity theory, traditionally

and still today, is to find expressions for the energy-momentum tensor when it derives from simple matter sources, plug them into Einstein's equation, and look for solutions. These solutions corresponding to specified matter sources are called "exact solutions." They have been found where energy-momentum derives from: a) a vacuum; b) a homogeneous fluid of "matter dust;" c) an electromagnetic field; and d) a combined source of dust fluid together with electromagnetic field. Because of the great complexity of the Einstein field equation, solutions are found only after very strict symmetry conditions on g are imposed.

Quite naturally there is interest among relativity theorists not only in finding and classifying a relatively small number of highly idealized exact solutions, but in generic space-time structure as well. In a sense, <u>any</u> space-time as defined above can be regarded as satisfying Einstein's equation for <u>some</u> energy-momentum distribution or other. Given a metric g we can simply compute the corresponding value of the left-hand side of the equation and use it to <u>define</u> the right-hand energy-momentum tensor. The resulting energy-momentum distribution may turn out to be physically unreasonable in that it has negative density somewhere or other (though it automatically satisfies a conservation principle). But this can be avoided by stipulating that the metric, whatever it happens to be like, at least satisfies one of several hierarchly ordered "energy

conditions." That energy-momentum density be everywhere non-negative is the weakest of these conditions and is called the "weak energy condition." (See Hawking and Ellis [8], sec 4.3 for details)

So at a second level of research in relativity theory a space-time is taken to be a space-time in the sense above which also satisfies some weaker or stronger energy condition. Then the idea is to see what general results one can establish concerning space-time structure, results which by the very way things are set up will not depend on particular assumptions about the material content of space-time. The most striking success here has been the "singularity theorems" of Penrose, Hawking, and Geroch.

At yet a third level of research even the constraints imposed by the energy conditions are dropped. There are two reasons for doing so. First, the added assumptions are not needed for some theorems that one wants to prove. In particular, many results about causal structure (e.g. those proven in this chapter) make no use of them. Second, it is thought <u>desirable</u> to avoid dependence on Einstein's equation if possible. Authors such as Penrose have expressed much greater confidence in the assumptions discussed above -- that space-time has the structure of a pseudo-Riemannian manifold of Lorentz signature and that axioms (i') and (ii) obtain -- than in the particular form of Einstein's equation. And there are other gravitational

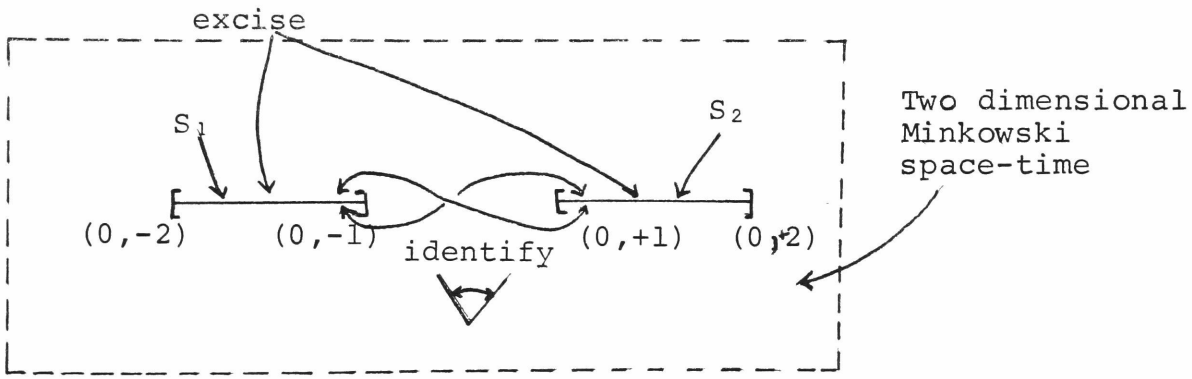
theories around which hold on to these central assumptions while rejecting the equation. The Brans-Dicke theory is an example. Understandably one is interested in proving results which apply to all of these theories.

Finally a word on constraints that are sometimes imposed to exclude seemingly arbitrary, contrived space-time manifolds. As the definition stands, if one starts with a space-time (M,g) and removes a closed set S from M, then the resulting manifold M-S together with the restricted metric $g|_{M-S}$ jointly from a perfectly good space-time. One can, for instance, just remove a single point. But there seems to be something rather unsatisfactory about such punctured space-times. Perhaps it violates our sense that the universe should satisfy a Leibnizean "principle of plenitude." God, after having created M-S, would not have stopped but rather would have continued to create all of M, that being the more perfect and total universe. However compelling the theology, a condition of "inextendibility" or "maximality" is often imposed on space-times. A space-time (M,g) is <u>inextendible</u> if for all space-times (M',g'), if there is a smooth isometric imbedding f: M→M' of M into M' then f must be onto, i.e. (M,g) cannot be extended smoothly and isometrically to some large space-time (M',g').

Even after inextendibility is imposed as a constraint still many seemingly pathological space-times slip through.

The following simple example is (essentially) given in Hawking and Ellis [8], pg. 58-59. Start with two dimensional Minkowski space-time in standard (t,x) coordinates and remove two closed slits:

$$S_1=\{(t,x):\ t=0\ \&\ -2\leq x\leq -1\}\ \text{and}\ S_2=\{(t,x):\ t=0\ \&\ 1\leq x\leq 2\}.$$



A space-time which is inextendible but not locally inextendible

But now identify the upper edge of slit S_1 with the lower edge of slit S_2 except for their respective endpoints. And similarly identify the lower edge of S_1 with the upper edge of S_2 (see figure). The resulting space-time is inextendible but still the points (0,-2), (0,-1), (0,+1), (0,+2) seem to be "missing." One can capture one sense in which they are missing as follows. One can find an open set (e.g. an open disk of unit radius centered at the point (0,3)) such that the set has non-compact closure in the given space-time, but which can be smoothly isometrically imbedded in another space-time, say two-dimensional Minkowski

space-time, in such a way that its closure in this new space-time is compact. This gives the definition of "local inextendibility." A space-time (M,g) is locally inextendible if for all open $O \subseteq M$ and all space-times (M',g'), if $\mathcal{C}[O]$ is non-compact in M and if $f: O \to M'$ is a smooth isometric imbedding of O into M', then $\mathcal{C}(f[O])$ must be non-compact in M'.

Imposing the condition of inextendibility or of local inextendibility on space-time structure may or may not be reasonable. The point to be made here is that insofar as one is dealing only with problems of conformal structure or causal structure (which is derivative from the former) the imposition effects no restriction whatsoever. Indeed one has the following result:

3.3 Theorem: Every space-time is conformally isometric to a space-time which is locally inextendible.

(This point arose in conversation with Robert Geroch.)

To prove this it suffices to appropriate a proof used by Clarke [2] to show that every strongly-causal space-time is conformally isometric to one which is null geodesically complete. [The definition of strong causality is given in the next section. A space-time is null (time-like, causal) geodesically complete if every null (timelike, causal) geodesic through a given point can be extended in either direction to arbitrarily large values of (any one of its) affine parameters.] We skip here both Clarke's argument and the

demonstration that it can be adapted to prove the theorem. Instead we turn to the description of causal structure in relativistic space-times.

B. Causal Structure of Space-times and the Hierarchy of Causal Conditions

So far five levels of structure in space-times have been discussed: topological, differential, pseudo-Riemannian, conformal, and projective. Our primary interest is with a sixth level, causal structure, and the extent to which it uniquely determines the other levels. In this section we develop a fragment of the theory of causal structure and present results on the hierarchy of causality conditions. Most of the material is needed for later sections but not all. A few other results are proven (e.g. Hawking's Theorem that stable causality is equivalent to the existence of a universal time function) in the hope of rounding-out a serviceable exposition of work in causality conditions. The material is not new. The exposition is certainly not comprehensive, nor even is it fully self-contained (-- though it is relative to the assumption of a few technical lemmas from differential geometry --). But hopefully it may be more readable than the expositions in Hawking and Ellis [8] and Penrose [13] from which it is largely drawn. It differs somewhat from both.

First, a preliminary word about what causal structure

has to do with "causality." According to assumption (i') above, all bodies, signals, and physical interactions in space-time traverse causal curves. If causal interaction is a species of physical interaction then, no matter what else is the case, it follows that an event at one space-time point can enter into causal interaction with an event at another point only if there is a causal curve connecting those points. And conversely, if the points are connected by a causal curve then it is possible that some signal travel from one to another (along the curve) while at all times moving less than or equal to the maximal possible velocity. It is possible in the sense of being compatible with the principles of relativity theory though, to be sure, it may not be possible in some other sense which also takes into account particular conditions present at the two points (e.g. absence of an energy source). Certainly the emission and absorption of a signal is a paradigm case of causal interaction. So one has that two events can possibly stand in causal connection to one another if and only if the points at which they occur are joined by a causal curve. On the assumption (-- not one to be discussed here --) that causal interaction is future directed, an event can exercise causal influence on a second if and only if there is a future directed causal curve γ: $[0,1] \to M$ with initial point x at which the first occurs and terminal point y at which the second occurs. When there is such a curve we say that x is causally

prior to y and write $x \prec y$. At the level of its causal structure a space-time is just its underlying point set together with this abstracted two place relation of causal priority (or with one of the several other "causal relations" that are studied).

The formal theory can be set up in different ways. We have so far taken curves γ: I→M to be causal if they are differentiable with everywhere non-vanishing causal tangent vectors, and taken them to be future (past) directed if their tangent vectors everywhere fall in the future (past) lobe of the null cone. One can, however, not only work with: a) differentiable curves, but also with: b) smooth curves; c) piecewise differentiable curves; d) piecewise smooth curves; e) piecewise geodesic curves; and f) continuous curves. In working with piecewise differentiable (smooth, or geodesic) curves one must add to the defining condition of being causal that at points where differentiability fails at least the curve is upper and lower differentiable and that <u>both</u> upper and lower tangent vectors are causal, falling in the same lobe of the null cone. One can work with merely continuous curves taking them to be causal if they can be arbitrarily well approximated by causal curves in one of the other senses. More precisely, a continuous curve γ is taken to be causal if given any two points a and b on γ and any open set O containing the section of γ stretching from a to b, O must contain a causal curve (say, in sense (a)) connecting a and b as well.

Now there are reasons, technical and interpretive, why one might want to choose any one of these different kinds of "causal curves" in setting up one's definitions. But so far as the resulting causal structure is concerned it makes no difference which one chooses. With the aid of technical lemmas on smoothing of curves one can show that two points are causally connectible in one sense if and only if they are causally connectible in each of the other senses. [The differences between the different kinds of curves cannot be ignored, however, when it comes to considering the claim made above in section A.] For specificity we shall continue to construe causal curve in the sense of (a). We shall not bother to explicitly invoke the smoothing lemmas each time we need them, however. Also we shall lapse into referring to ordered point sets as causal curves. How such talk is fleshed out was explained in section A.

Similar remarks apply to timelike curves. These are curves (in any of the senses above) whose tangent vector or vectors at every point are non-vanishing timelike (and, where distinct, fall in the same lobe of the null cone.) Here again the six different senses of connectibility by a timelike curve collapse.

The basic causal relations $<$, $<<$, and $\rightarrow$ defined in Minkowski space-time can now be defined so as to apply to the general space-times being considered. It is convenient to relativize them to arbitrary open sets. Given a space-time (M,g)

and an open subset $O \subseteq M$ with points a,b in O we define:

$a<b(O) \Leftrightarrow$ a=b or there is a future directed causal curve from a to b lying wholly within O.

$a<<b(O) \Leftrightarrow$ there is a (non trivial) future directed timelike curve from a to b lying wholly within O.

$a \rightarrow b(O) \Leftrightarrow a<b(O)$ but not $a<<b(O)$.

In the special case where O=M, we write simply a<b, a<<b, and $a \rightarrow b$. These relations are referred to, respectively, as causal, timelike (or chronological), and null priority. In Minkowski space-time they agree with the definitions given previously.

The corresponding symmetric relations: $a \kappa b\ (O)$, $a \tau b\ (O)$, and $a \lambda b\ (O)$ are defined in the obvious way, e.g. $a \kappa b \leftrightarrow$ a=b or there is a future or past directed causal curve from a to b. To define these relations one need not presuppose a specification of a temporal orientation. Furthermore, just as before, one defines future and past sets corresponding to each of the basic relations:

$$J_0^+(a) = \{b \mid a<b(0)\} \qquad J_0^-(a) = \{b \mid b<a(0)\} \qquad J_0(a) = \{b \mid a \kappa b(0)\}$$

$$I_0^+(a) = \{b \mid a<<b(0)\} \qquad I_0^-(a) = \{b \mid b<<a(0)\} \qquad I_0(a) = \{b \mid a \tau b(0)\}$$

$$E_0^+(a) = \{b \mid a \rightarrow b(0)\} \qquad E_0^-(a) = \{b \mid b \rightarrow a(0)\} \qquad E_0(a) = \{b \mid a \lambda b(0)\}$$

Where 0 = M we simply drop the subscript.

Many facts about these basic causal relations carry over intact from the case of Minkowski space-time. The following are immediate from the definitions. In any space-time (M,g) if a,b,c are in M then: a<a; $a<<b \rightarrow a<b$; $a<b \,\&\, b<c \rightarrow a<c$; and $a<<b \,\&\, b<<c \rightarrow a<<c$. But in order to establish that other relations

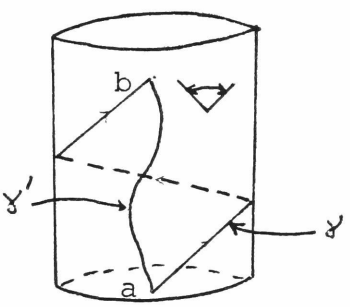
carry over, certain basic "working lemmas" from differential geometry are needed. These concern the local behavior of geodesics in Lorentzian manifolds. We will not prove them, but will state them for reference and then show how they are used to develop the theory of causal structure.

The first working lemma applies quite generally to all differentiable manifolds carrying an affine connection whether or not the connection is derived from a pseudo-Riemannian manifold. Suppose M carries a smooth affine connection. Recall first the exponential mapping $\exp_a: M_a \supset D_a \to M$ defined on a subset D_a of the tangent space M_a at each point a. For every vector $X \in M_a$ there is a unique (affinely parametrized) geodesic segment through a whose tangent at a is X. If there is a point on that geodesic (extended in the direction of the tangent vector) whose affine length from a is 1, then the vector X is in the domain of $\exp_a$ and $\exp_a(X)$ is defined as that point on the geodesic. D_a, it turns out, must be an open subset of M_a and $\exp_a$ must be a smooth diffeomorphism of D_a on to its image (Hicks[], pg. 131). An open set O containing a is called a normal neighborhood of a if there is an open subset $D \subset D_a$ in the tangent space M_a such that $\exp_a[D] = O$. An open set O is called a convex neighborhood if it is at once a normal neighborhood of every one of its points, or equivalently, if every two of its points are joined by a unique geodesic segment fully contained in O. The first lemma we need is the following theorem of Whitehead:

3.4 Basic Lemma: For every point a in a space-time (M,g) and every open set U containing a, there is a convex neighborhood O such that $a \in O \subseteq U$.

A proof is given in Hicks [8], pg. 134.

The next lemma concerns causal structure within convex neighborhoods. In any space-time there are two different "null cones" which can be associated with any point a. One is simply E(a). The other consists of those points which are connected to a via a null geodesic, i.e. the image under $\exp_a$ of the null cone in M_a. The former is always a subset of the latter (as we shall prove in a moment), but only in certain space-times (eg. Minkowski space-time) will the two coincide. For example, consider a copy of two-dimensional Minkowski space-time rolled-up into a vertical cylinder as in the figure. Here there is a future directed null geodesic γ from a to b, but still $a \ll b$ since there is a future directed timelike curve γ' from a to b as well. As one would intuite, at least the cones must always coincide locally.



3.5 Basic Lemma: If (M,g) is a space-time with $O \subseteq M$ a convex neighborhood, then for all points b,c in O:

(i) $b<c$ (O) <=> b=c or the unique geodesic in O from b to c is causal

(ii) $b \ll c$ (O) <=> the unique (non-trivial) geodesic in O from b to c is timelike; and hence:

(iii) $b \rightarrow c$ (O) <=> b=c or the unique geodesic in O from b to c is null.

Hawking and Ellis [8], pg. 103 give a proof of the theorem using "Gauss' Lemma." As an immediate corollary one has:

3.6 Corollary: If (M,g) is a space-time then for all $a \in M$, $I^+(a)$ and $I^-(a)$ are open.

Proof: Suppose $a \ll b$ and $\gamma: I \to M$ is a future directed timelike curve from a to b. Let O be a convex neighborhood containing b but not a and let c be any point on $\gamma \cap O$ other than b. So $c \ll b$ (O). By the theorem there must be a future directed timelike geodesic in O from c to b. If $D_c \subseteq M_c$ is the (open) domain of $\exp_c$, and if D_c' is the intersection of D_c with the (open) set of future directed timelike vectors in M_c, then $\exp_c (D_c')$ is an open subset of $I^+(c)$ containing b. But $I^+(c) \subseteq I^+(a)$. So $I^+(a)$ must be open. Symmetrically for $I^-(a)$. qed.

Lemma 3.5 has other simple consequences which are used repeatedly in proofs. One is that given any point a and any open set O containing a, we can find a point b in O such that $a \ll b$ (O) [$b \ll a$(O), $a \kappa b$(O), etc.]. To find such a b we need only pass to a convex neighborhood $O' \subseteq O$ containing a, choose a future directed timelike vector X in the domain of $\exp_a$ and consider $\exp_a(X)$. Either $\exp_a(X)$ itself falls in O' or some point on the future directed timelike geodesic from a to $\exp_a(X)$ does. Another simple consequence of this last mentioned fact together with Corollary 3.6 is that in all space-times, for all points a, b, c we have: $a \ll c$ & $b \ll c \Rightarrow (\exists d)$ [$a \ll d$ & $b \ll d$ & $d \ll c$]; and symmetrically. Yet another consequence is the fact that if $a \ll b$, then there are open sets O_a and O_b containing a and b respectively such that $O_a \ll O_b$. [In general we take $S \ll T$ to abbreviate: $\forall s\ \forall t$ ($s \in S$ & $t \in T \Rightarrow s \ll t$); $s \ll T$ abbreviates:

$\forall t\ (t \varepsilon T \Rightarrow s << t)$, etc.]

The final basic working lemma establishes that another property true in the tangent space of a point remains true under projection onto a local neighborhood of the point.

3.7 Basic Lemma: If (M,g) is a spacetime, with $O \subseteq M$ a convex neighborhood, then for all points a,b in O, $a \rightarrow b$ (O) implies that the only future directed causal curve from a to b in O is the unique null geodesic connecting them.

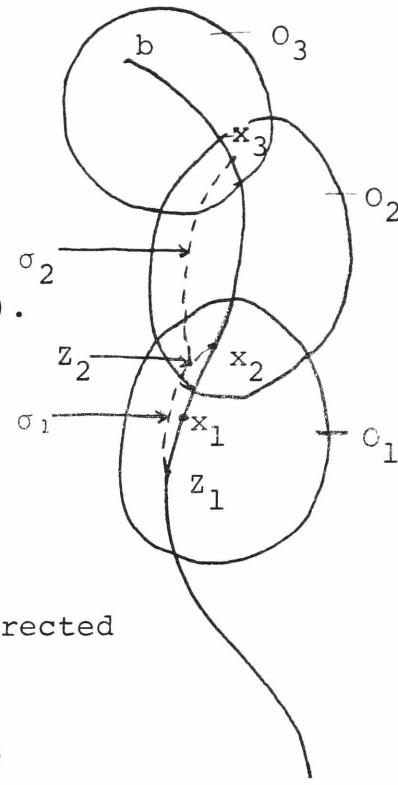
Proofs along different lines are given in Hawking and Ellis ([8], pg. 112) and in Penrose ([13], pg. 13). It follows immediately from this that within a convex neighborhood O "strong transitivity" holds: $a<b$ (O) & $b<<c$ (O) $\Rightarrow a<<c$ (O) and $a<<b$ (O) & $b<c$ (O) $\Rightarrow a<<c$ (O). Also if $a \rightarrow b$ (O) & $b \rightarrow c$ (O) then $a<<c$ (O) unless the two null geodesics from a to b and from b to c are mutual null geodesic extensions of one another.

One now shows that this result can be extended to the global case.

3.8 Corollary: If (M,g) is a space-time, then for all points a, b in M, $a \rightarrow b$ implies that any future directed causal curve from a to b must be a null geodesic.

Pf: The proof follows from the fact that if a causal curve λ from a to b is not a null geodesic, it fails to be a null geodesic at some point.

By compactness of (the image of) γ, we can cover it with a finite set of convex neighborhoods. Suppose γ is not a null geodesic at a point x_1 and suppose x_1 falls in one of the neighborhoods O_1. Let z_1 be some point on $\gamma \cap O_1$ to the past of x_1. If γ remains in O_1 we have $z_1 << b$ (O_1). Otherwise there will be a point on $\gamma \cap$ Bnd (O_1) to the future of x_1 which falls in some neighborhood O_2 (see picture). Then if x_2 is some point on $\gamma \cap O_1 \cap O_2$ to the future of x_1, we have $z_1 << x_2$ (O_1). Let τ_1 be a future directed timelike curve from z_1 to x_2. Let z_2 be some point on $\tau_1 \cap O_2$. If γ remains in O_2 we have $z_2 << b$ (O_2) and hence $z_1 << b$. Otherwise there will a point on $\gamma \cap$ Bnd (O_2) to the future of x_2 which falls in some convex neighborhood O_3. If x_3 is some point on $\gamma \cap O_2 \cap O_3$ to the future of x_2 we have $z_2 << x_3$ (O_2) and hence $z_1 << x_3$. Continuing in this way we generate a future directed timelike curve from z_1 to b .



In a parallel fashion we can work our way downward in a finite sequence of steps. Eventually we shall have a future directed timelike curve from a to b thus contradicting our assumption that $a \rightarrow b$. qed.

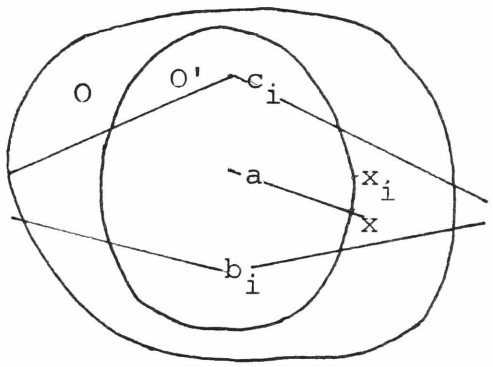
Using some of the last results we can establish the existence of neighborhoods of a particularly convenient type.

3.9 Corollary: Given any point a in a space-time (M,g) and a convex neighborhood O of a, there are points b,c in O where $b<<a<<c$ (O) and

$$J_O^+(b) \cap J_O^-(c) = \mathcal{A}[I_O^+(b) \cap I_O^-(c)] \subseteq O$$

Proof: The idea is to show that if b,c satisfying these conditions could not be found then the metric at a would have to be degenerate.

First we can find an open subset $O' \subseteq O$ containing a such that $\mathcal{A}(O') \subseteq O$ is compact and hence Bnd $(O') \subseteq O$ is compact. Now let $\{b_i\}$ and $\{c_i\}$ be sequences in O' converging chronologically to a, respectively, from below and above, i.e. for all i, $b_i<<b_{i+1}<<a<<c_{i+1}<<c_i$ (O').



Suppose for all i, it is not the case that $\mathcal{A}[I_O^+(b_i) \cap I_O^-(c_i)] \subseteq O$. Then for each i there will be a point x_i on Bnd (O') such that $b_i<<x_i<<c_i$ (O). Let x be a point of accumulation of the x_i on Bnd (O'). There must be a geodesic segment ax in O from a to x.

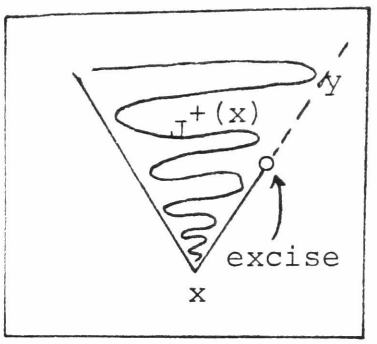
Using Lemma 3.5 and the fact that at all points the map exp is continuous, we can now prove that ax is both past and future directed causal. This implies that its tangent in M_a is both past and future causal which is impossible. Consider any b_i. There must be a geodesic segment from b_i to x. By

the continuity of $\exp_{b_i}$ the tangent to this segment in M_{b_i} must be future directed causal. This is so because $b <\!\!< x_{i'}$ (0) for all $i' \geq i$ and the $x_{i'}$ accumulate at x. So $b_i <\!\!< x$ (0) for all i. But now applying the same argument with respect to $\exp_x$ we may conclude that $a <\!\!< x$ (0). The argument to establish $x <\!\!< a$ (0) is symmetrical.

So there must exist points b,c in O with $b <\!\!< a <\!\!< c$ (0) and $\mathcal{C}\ell[I_O^+(b) \cap I_O^-(c)] \subseteq O$. Now quite generally we have $J^+(x) \subseteq \mathcal{C}\ell[I^+(x)]$ and $J^-(x) \subseteq \mathcal{C}\ell[I^-(x)]$ for all x. This follows by "strong transitivity." Certainly then we have here that $J_O^+(b) \cap J_O^-(c) \subseteq \mathcal{C}\ell[I_O^+(b) \cap I_O^-(c)]$. But now inside any convex neighborhood it follows by Lemma 3.5 and the continuity of $\exp_y$ that for all x,y $x \in \mathcal{C}\ell[I^+(y)]$ implies $y <\!\!< x$ (0); and symmetrically. So $\mathcal{C}\ell[I_O^+(b) \cap I_O^-(c)] \subseteq J_O^+(b) \cap J_O^-(c)$ q.e.d.

In the proof we built in enough to insure that $J_O^+(b) \cap J_O^-(c) \subseteq \mathcal{C}\ell(O')$. As the closed subset of a compact set, $J_O^+(b) \cap J_O^-(c)$ must itself be compact. Sets of the form $J_O^+(b) \cap J_O^-(c)$ will be called <u>local closed diamonds</u> if O is a convex neighborhood, if $b <\!\!< c$ (O), and if they are compact (which itself entails $J_O^+(b) \cap J_O^-(c) = \mathcal{C}\ell[I_O^+(b) \cap I_O^-(c)]$). The corresponding sets $I_O^+(b) \cap I_O^-(c)$ will be called <u>local open diamonds</u>.

As noted in the previous proof, $J^+(x) \subseteq \mathcal{C}\ell[I^+(x)]$ holds quite generally and the converse holds inside convex neighborhoods. But in cases as in the figure $y \in \mathcal{C}\ell[I^+(x)]$ even though $x \not< y$; hence $J^+(x)$ is not closed. One does have the following simple



characterization of $[I^+(x)]$ to work with:

$y \in \mathcal{C}[I^+(x)] \Leftrightarrow I^+(y) \subseteq I^+(x)$.

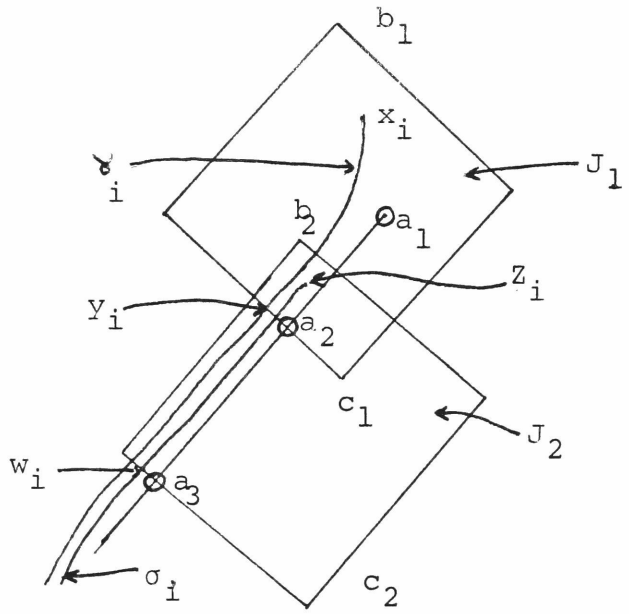
Similarly: $y \in \mathcal{C}[I^-(x)] \Leftrightarrow I^-(y) \subseteq I^-(x)$.

This is easy to check. Also whenever $y \in \mathcal{C}[I^+(x)]$ it must be the case, as in the figure, that there is a past directed null geodesic from y which is fully contained in $\mathcal{C}[I^+(x)]$. We prove a somewhat more general version of this claim. First we need a definition. If (M,g) is a space-time and $\gamma : I \to M$ is a future directed causal curve, then a point b is a <u>future endpoint of</u> γ if for every open set O containing b, γ eventually enters and thereafter remains in O, i.e. there is a $t_0 \in I$ such that for all $t > t_0$, $\gamma(t) \in O$. Similarly one defines a <u>past endpoint</u>. Note that endpoints of γ need not belong to γ. We say that a causal curve is <u>future</u> (<u>past</u>) <u>inextendible</u> if it does not contain a future (past) endpoint. Also some notation: for any set S, $I^+[S] = \cup\{I^+(s) : s \in S\}$. Similarly for $I^-[S]$, $J^+[S]$, etc.

3.10 <u>Corollary</u>: If (M,g) is a space-time with $S \subseteq M$ a closed subset, and if $a \in \mathcal{C}(I^+[S]) - J^+[S]$, then there is a past directed null geodesic from a which lies entirely within $\mathcal{C}(I^+[S])$ and which is past inextendible; and symmetrically.

<u>Proof</u>: Since $a \notin J^+[S]$, $a \notin S$. Hence for some convex neighborhood O_1 there must be a closed local diamond $J_1 = J^+_{O_1}(b_1) \cap J^-_{O_1}(c_1)$ containing $a = a_1$ such that $J_1 \cap S = \phi$. Since $a_1 \in \mathcal{C}(I^+[S])$ we

can find an infinite set of points x_i in the interior of J_1 accumulating at a_1, all of which are in $I^+[S]$. So for each i, there will be a point $s_i \in S$ and a future directed timelike curve γ_i from s_i to x_i. To reach x_i, γ_i will have to cross Bnd (J_1) at some point y_i. Let a_2 be a point of accumulation for the y_i on Bnd (J_1).



There must be a geodesic segment $\underline{a_2 a_1}$ from a_2 to a_1.

By considerations as presented in the proof of Corollary 3.9 this segment must be future directed causal. Indeed, it must be a null geodesic segment. For if it were timelike we would have $a_2 \in I^-(a_1)$. Hence there would be some point in $I^+[S] \cap I^-(a_1)$ and so $a_1 \in I^+[S]$ which is impossible.

Now $a_2 \in \mathcal{A}(I^+[S])$ and so we can repeat the same argument with respect to it. We can find a closed local diamond $J_2 = J^+_{O_2}(b_2) \cap J^-_{O_2}(c_2)$

containing a_2 such that $J_2 \cap S = \phi$. We can find an infinite set of points z_i in the interior of J_2 converging to a_2, all of which are in $I^+[S]$. So for each i there will be a point $\bar{s}_i \in S$ and a future directed timelike curve σ_i from $\bar{s}_i$ to z_i. Each σ_i will have to cross Bnd (J_2) at some point w_i and these points will have a point of accumulation a_3 in Bnd (J_2). Precisely as before we may conclude that there must be a null geodesic segment from a_3 to a_2. Further this segment must be extension of that from a_2 to a_1. Otherwise, by Corollary 3.8 we would have $a_3 << a_1$ and hence $a_1 \in I^+[S]$.

Continuing this way we generate a past directed null geodesic from a every point of which is in $\mathcal{G}(I^+[S])$. It never reaches S since then we would have $a \in J^+(s)$. Further it must be past endless. If b were a past endpoint it would follow that $b \in \mathcal{G}(I^+[S]) - J^+[S]$. Then we could redo the same argument within some local closed diamond containing b and extend the null geodesic a bit further - qed.

* * * * * * * * * * * * * *

Rather than proving more of these basic facts about causal structure, we turn now to the hierarchy of causality conditions.

The first condition is that of <u>chronology</u>, i.e. for all points a, $a \not< a$. Equivalently, this is the condition that there be no closed future directed timelike curves. If such a curve existed then under the standard interpretation it would be possible for a particle or signal, traveling strictly less

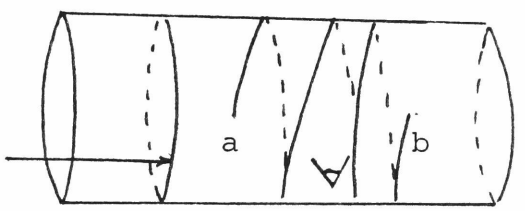
than the maximal possible velocity, to take a cosmic trip beginning and ending at the same point. Because of worries about the possibility of the particle "undoing what is already done" (shooting its particle grandfather before puberty), violation of chronology is often said to involve some sort of causal anomaly. Hence the term "causality condition." The stronger conditions will assert that there are no almost closed timelike or causal curves in several different senses of "almost."

Right off one has an easy result to the effect that every compact space-time exhibits chronology violation.

3.11 Theorem: If (M,g) is a space-time and if M is compact, there is a point a in M for which $a<<a$.

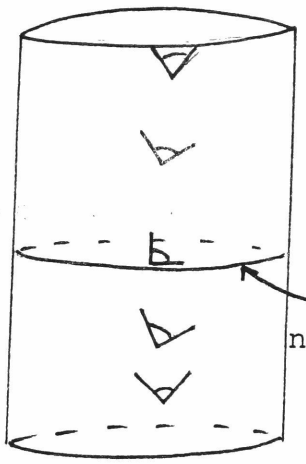
Proof: For every point b in M there is a point a such that $b<<a$, i.e. $b\in I^-(a)$. It follows that $\{I^-(a) : a\in M\}$ is an open cover of M. Suppose $I^-(a_1),\ldots, I^-(a_n)$ is a finite subcover. Now a_1 must be contained in some $I^-(a_{k_1})$, i.e. $a_1<<a_{k_1}$. Similarly a_{k_1} must be contained in some $I^-(a_{k_2})$, i.e. $a_{k_1}<<a_{k_2}$. Continuing in this way we generate a sequence $a_1<<a_{k_1}<<a_{k_2}<<\ldots.<<a_{k_n}<<a_{k_{n+1}}$. Since the k_i's come from $\{1,\ldots,n\}$, there will have to be two that are equal; say $k_l = k_{l'}$. Then taking $a = a_{k_i}$ we have $a<<a$ as claimed. qed.

The simplest example of a space-time in which chronology fails is Minkowski space-time "rolled up along a temporal axis." If $\{x_i\}$ are standard coordinates and $k>0$ we identify all points (x_0+nk, x_1, x_2, x_3), for integral n, with (x_0, x_1, x_2, x_3). Now the curves σ in this space-time defined by $\sigma(t) = (t, k_1, k_2, k_3)$ for fixed k_1, k_2, k_3 will all be closed, future-directed and timelike. Indeed they are geodesics. This space-time has the further property that for <u>all</u> points a and b we have $a \ll b$ (see figure). When this condition $[\forall a \forall b (a \ll b)]$ obtains we say that the space-time is <u>degenerate</u> in its causal structure.



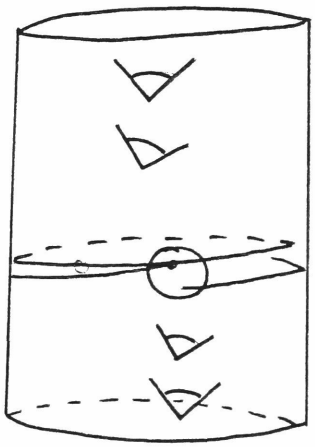
Chronology can be satisfied and yet there still be "causality violation" in the sense of there being a closed future-directed causal curve. The condition of <u>causality</u>, $\forall a \forall b [a<b \,\&\, b<a \rightarrow a=b]$, further excludes this possibility. Now if causality violation occurs in a space-time where chronology obtains, all closed causal curves must be closed null geodesics; this follows from Corollary 3.8. Consider, for example, the following two-dimensional space-time. Start with the (t,x) plane with metric $ds^2 = (\cosh t - 1)^2 (dt^2 - dx^2) + dtdx$. [Recall that $\cosh t = 1/2(e^t + e^{-t})$]. Along the line $t=0$, the metric reduces to the form $ds^2 = dtdx$ and its associated null cones are horizontal, pointing in the direction of increasing x. But as $|t| \rightarrow \infty$, the cones "tip to the left" asymptotically approaching

the upright position they have in Minkowski space-time. Now form a standing vertical cylinder by identifying all points $(t, x+kn)$, for some particular $k>o$, with the point (t,x) - see figure. The space-time is not causal as it admits a single closed null geodesic - the t=o equator. But it does not admit any closed timelike curves.



Chronology but not causality obtains

Suppose one now excises a single point from the t=o equator in the space-time. This renders it "causal." But there are still "almost closed" causal curves in the following sense. Given any point on the equator and any small open set containing that point, there will be a future directed causal curve starting at the point which leaves and then reenters the open set after circumnavigating the space-time; similarly there will be future directed causal curves which start in the open set, leave, and then reenter, reaching the point on the equator after the circumnavigational trip - see figure. This sort of behavior is ruled out in the next causality



causality but not past or future distinguishability obtains

condition.

> 3.12 Definition: A space-time (M,g) is future (past) distinguishing if for all points a and all open sets U containing a, there is an open subset $O \subseteq U$ containing a such that no future (past) directed causal curve from a which leaves O ever reenters.

The move to a subset here is necessary. If one only required that a have some open neighborhood O having the desired "no reentry" property, then the condition could be vacuously satisfied by taking O=M.

There are a number of equivalent ways in which future (past) distinguishability can be formulated. Several are given in the next theorem. Condition (ii) explains terminology.

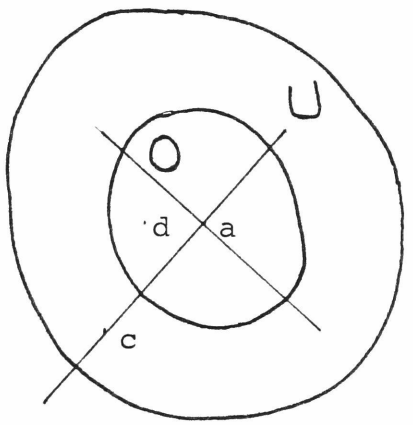
> 3.13 Theorem: In any space-time (M,g), future (past) distinguishability is equivalent to each of the following:
>
> (i) For all points a and all open sets U containing a, there is an open subset $O \subseteq U$ containing a such that for all b in O we have: $a<b \rightarrow a<b(O)$ $[b<a \rightarrow b<a(O)]$.
>
> (ii) $\forall a \forall b\ (I^+(a)=I^+(b) \rightarrow a=b)$
> $[\forall a \forall b\ (I^-(a)=I^-(b) \rightarrow a=b)]$
>
> (iii) $\forall a \forall b\ (a<b \& b<<I^+(a) \rightarrow a=b)$
> $[\forall a \forall b\ (a<b \& I^-(a)<<b \rightarrow a=b)]$

Proof: We prove the equivalences with future distinguishability. Past distinguishability is treated symmetrically.

First, suppose that (M,g) is future-distinguishing. We show that (i) is satisfied. Given any a and any open U containing a, let $O \subseteq U$ be as in the definition. Now suppose b is in O and $a<b$. If it were not the case that $a<b(O)$,

then any future directed causal curve from a to b would have to leave and then reenter O. This is impossible. So $a<b(O)$.

Next (i)$\Rightarrow$(ii). Suppose $a\neq b$, but $I^+(a)=I^+(b)$. Let U be a convex neighborhood of a not containing b. Since $I^+(b)\subseteq I^+(a)$ we know that $a\in\mathcal{C}\ell[I^+(b)]$ and hence, by Corollary 3.10, either $b<a$ or there is a past directed null geodesic from a which falls in $\mathcal{C}\ell[I^+(b)]$. In either case there is a point c in U where $c<a(U)$ and the causal geodesic from c to a in U falls in $\mathcal{C}\ell[I^+(b)]$. Now let O be <u>any</u> open subset of U containing a. We can certainly find a point d in O where $c<<d(U)$ but not $a\not<d(O)$. But $c\in\mathcal{C}\ell[I^+(b)]$ and $c<<d$ together imply $b<<d$. Since $I^+(b)\subseteq I^+(a)$ we may conclude that $a<<d$ even though not $a<d(O)$. This violates (i).

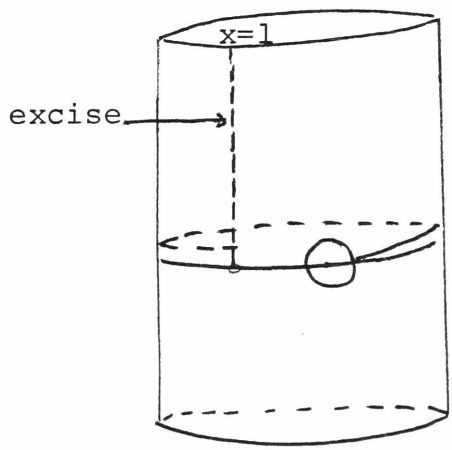


Third, (ii)$\Rightarrow$(iii). Suppose $a<b$ and $b<<I^+(a)$. These assumptions imply, respectively that $I^+(b)\subseteq I^+(a)$ and $I^+(a)\subseteq I^+(b)$. So $I^+(a) = I^+(b)$. Assuming (ii) we may conclude that $a = b$.

Finally, (iii)$\Rightarrow$future distinguishability. Suppose the latter fails for some point b and some open neighborhood U of b. Let I_i be a nested sequence of local open diamonds in U converging to b, i.e. for all i, $I_{i+1}\subseteq I_i\subseteq U$ and

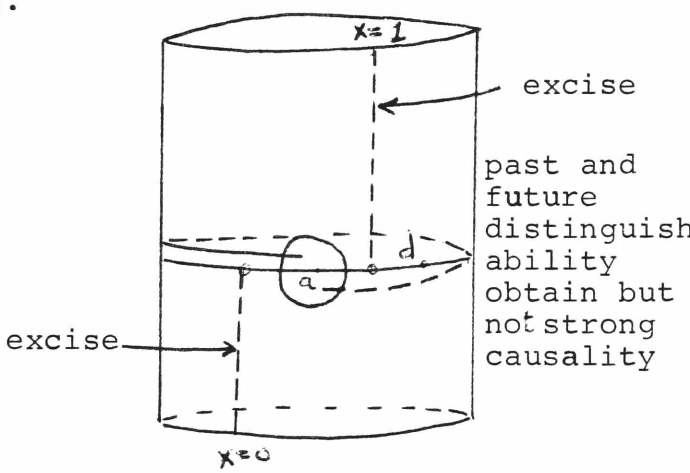
$\cap\{I_i\} = \{b\}$. For each i there will be a future directed causal curve γ_i from b which leaves and then reenters I_i. Each of these will, upon returning, hit $Bnd(J_o)$ - where J_o is the local closed diamond corresponding to I_o. If γ_i hits $Bnd(J_o)$ in a point a_i, the a_i will have a point of accumulation a in $Bnd(J_o)$. There will be a local geodesic segment from a to b. This segment, further, will have to be future directed causal. [The argument from the continuity of exp is as in Corollary 3.9.] So $a<b$. But now let c be any point in $I^+(a)$. Since $I^-(c)$ is open and contains a, it must also contain a_i for some i. Therefore $b<<a_i<<c$. Thus we have distinct points a and b where $a<b$, but $b<<I^+(a)$. This violates (iii). qed

The condition that a space-time be past and future distinguishing (PFD) will be of special interest later. It is strictly stronger than either past or future distinguishing alone. For example, the following two dimensional space-time is future, but not past distinguishing. Take the cylinder with metric $ds^2=(\cosh t-1)^2(dt^2-dx^2)+dt\,dx$, as described above. Now excise the closed set $\{(t,x):t\geq o \;\&\; x=1\}$ - see figure. It is still true that all points in the equator have as their common chronological past all points below the equator. But the space-



time is future distinguishing. Causal curves from points on the t=o equator which could previously circumnavigate the cylinder and return to arbitrarily small neighborhoods of the original point are now blocked by the cut.

If we cut a second slice from the example, say the set $\{(t,x): t \leq o \ \& \ x=o\}$, then we arrive at a space-time which is past and future distinguishing. But here there still remain "almost closed" causal curves of a sort - see figure. Given any point a on the t=o equator "between" the cuts and any small open set O containing the point there will be future directed causal curves which leave from O and eventually reenter. "Strong causality" excludes this possibility.



3.14 Definition: A space-time (M,g) is strongly causal if for all points a in M and open sets U containing a there is an open set $O \subseteq U$ containing a such that no future directed causal curve which leaves O ever reenters.

There are several different equivalent formulations; one in particular is important because it establishes a sufficient condition for the definability of the space-time topology from causal relations. Recall that in Minkowski space-time the sets of form $I^+(b) \cap I^-(c)$ form the basis for a topology

called the Alexandrov topology. It is easy to check that in the general case as well these sets form the basis for a topology. [If $I_1=I^+(b_1)\cap I^-(c_1)$ and $I_2=I^+(b_2)\cap I^-(c_2)$, and if $a\in I_1\cap I_2$, we can find points c_3 and b_3 such that $c_3<<c_2$, $c_3<<c_1$, $a<<c_3$ and $b_1<<b_3$, $b_2<<b_3$, $b_3<<a$. Therefore $a\in I^+(c_3)\cap I^-(b_3)\subseteq I_1\cap I_2$. So every intersection of basic Alexandrov sets is the union of basic Alexandrov sets.] Let the name carry over.

In Minkowski space-time the Alexandrov topology $\mathcal{T}_A$ is equal to the manifold topology $\mathcal{T}_M$. But in arbitrary space-times this is certainly not so. In rolled up Minkowski space-time above, where for all points a and b a<<b, the Alexandrov topology collapses to the indiscrete topology. And in the previous example of a space-time which was PFD but not strongly causal (SC), $\mathcal{T}_A$ was strictly coarser than the manifold topology (though it was T_0). Every Alexandrov neighborhood centered at the point a will include the point d. It is true, however, that if a space-time is strongly causal then its associated Alexandrov topology is equal to the manifold topology. Indeed, the two conditions are equivalent. This and other equivalences are established in the next theorem.

3.15: Theorem: If (M,g) is a space-time, the condition that it be strongly causal is equivalent to each of the following:

(i) For every point a and every open set U containing a, there is an open subset $O\subseteq U$ containing a such that for all b,c in O, $b<c\rightarrow b<c(O)$

(ii) $\mathcal{T}_A = \mathcal{T}_M$

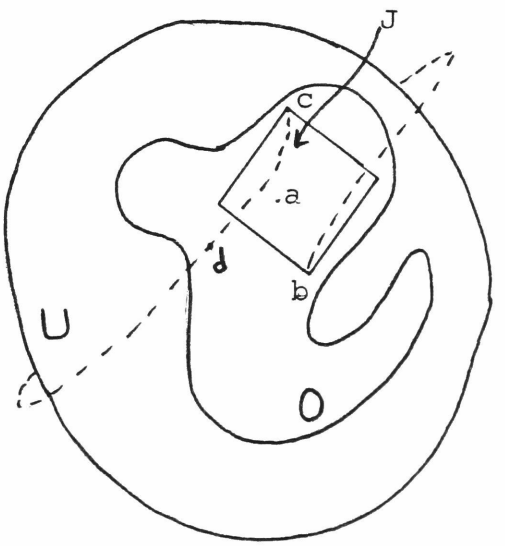
(iii) $\mathcal{T}_A$ is Hausdorff (i.e. T_2)

(iv) $\forall a,d\ [a<d\ \&\ I^-(d)<<I^+(a)\rightarrow a=d]$

Proof: The proofs here are very much like those in the previous theorem.

First, strong causality→(i). Given any a and any open set U containing a, let $O\subseteq U$ be as in the definition of strong causality. Suppose b,c are in O and b<c. It must be the case that b<c(O), since otherwise the future causal curve from b to c would have to leave and reenter O.

Second, (i)→(ii). Certainly $\mathcal{T}_A\subseteq\mathcal{T}_M$ is always true. So suppose $\mathcal{T}_M\not\subseteq\mathcal{T}_A$. Then there must exist some $\mathcal{T}_M$-open set U containing a point a such that no basic Alexandrov neighborhood of a is contained in U. We may assume that U is a convex neighborhood (by passing to a subset). Let O now be any open subset of U containing a. We can find points b,c in O such that $b<<a<<c(O)$ and $J=J^+_U(b)\cap J^-_U(c)\subseteq O$ is a local closed convex neighborhood. Since $I^+(b)\cap I^-(c)\not\subseteq U$, there must be a future directed timelike curve from b to c which leaves and reenters U, and hence J. Let d be a point on this curve in O just prior to its reentry in J - see figure. One has then that b<<d but not b<d(U), and hence not b<d(O). This violates (i). That (ii)→(iii) is immediate.



Next, (iii)→(iv). Suppose there exist distinct points a and d such that

$a<d$ and $I^-(d)<<I^+(a)$. Let $I_1=I^+(b_1)\cap I^-(c_1)$ and $I_2=I^+(b_2)\cap I^-(c_2)$ be any two Alexandrov neighborhoods of a and d respectively. We show that $I_1\cap I_2\neq\phi$ and hence that $\mathfrak{T}_A$ is not T_2. We have $b_1<<d$ and $b_2<<d$. Hence there must be a b_3 where $b_1<<b_3$, $b_2<<b_3$, and $b_3<<d$. It follows that $b_3\varepsilon I_2$ and, since $I^-(d)<<I^+(a)$, $b_3\varepsilon I_1$.

Finally, (iv) $\Rightarrow$ strong causality. Suppose failure of strong causality is witnessed by a point a and an open set U containing a. We can find a nested sequence of local open diamonds $\{I_i\}\subseteq U$ converging to {a}. For each there will be a future directed time-like curve γ_i which leaves and reenters U. Suppose each γ_i leaves J_o (for the first time) at a point d_i in $Bnd(J_o)$. Then the d_i will have a point of accumulation d in $Bnd(J_o)$. We may as well assume that the d_i converge to d (by passing to a subsequence). Similarly suppose each γ_i reenters J_o (for the last time) at a point c_i in $Bnd(J_o)$ and that the c_i converge to c in $Bnd(J_o)$. The usual continuity considerations guarantee that $c<a<d(U)$. Now consider any points $x\varepsilon I^-(d)$ and $y\varepsilon I^+(a)$. Eventually all d_i will be in $I^+(x)$ and eventually all c_i will be in $I^-(y)$. So there will be some i for which $x<<d_i<<c_i<<y$. Thus $I^-(d)<<I^+(a)$. This contradicts (iv). qed

The space-time in figure a is strongly causal, but given any small neighborhoods O_1 and O_2 of a_1 and a_2 respectively, there is

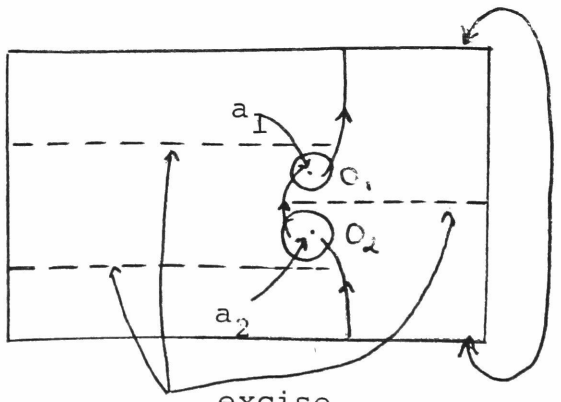


Fig. a

identify

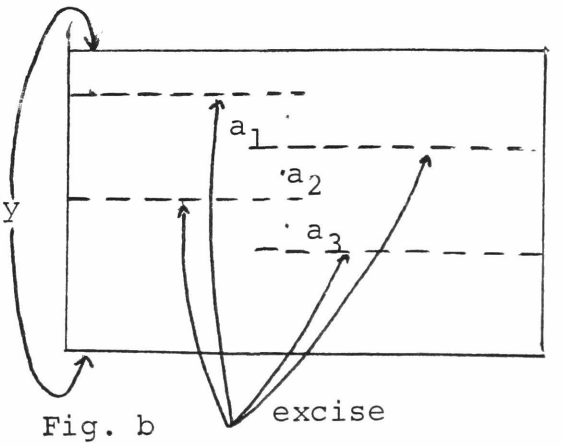


Fig. b

a future directed causal curve from O_1 which enters O_2 and one from O_2 which enters O_1. We may say that this is a violation of 2nd-order causality. Correspondingly we may say that the space-time in fig. b violates 3rd-order causality, but not 2nd-order causality. Generalizing this idea we come up with the following definition (which is a variation of the original due to Carter [1]).

> 3.16 Definition: If (M,g) is a space-time, it is nth order causal ($n \geq 2$) if for n arbitrary distinct points $a_1, \ldots, a_n$ with arbitrary open subsets $U_1, \ldots, U_n$ containing them respectively, we can find corresponding open subsets O_i with $a_i \in O_i \subseteq U_i$ such that there do not exist n future directed causal curves σ_i where for $i=1, \ldots, n-1$ σ_i leaves O_i and enters O_{i+1} and where σ_n leaves O_n and enters O_1.

Clearly by cutting more lines as aligned in the previous two figures we can, for any n, generate a space-time which is nth but not n+1st-order causal. Notice too that in a natural way strong causality can be construed as 1st-order causality. [Carter's scheme works somewhat differently.] We can also formulate the condition of nth-order causality using only causal relations.

> 3.17 Theorem: A space-time is nth order causal iff there do not exist n distinct points $a_1, \ldots, a_n$ in M such that for $i=1, \ldots, n-1$ $I^-(a_i) \ll I^+(a_{i+1})$ and $I^-(a_n) \ll I^+(a_1)$.

We skip the proof which only iterates a construction from the previous proof.

The strongest causality condition that has been studied is "stable causality." Consider again the space-times above exhibiting 2nd and 3rd order causality. It is clear that the slightest

"widening" of the cones whatsoever will allow a future directed timelike curve to scoot past the barriers and close back on itself. The space-time, then, admits "almost closed" causal curves with respect to metrics arbitrarily close to the original. "Closeness of metric" can be made precise in the following way. If (M,g) is a space-time, let $\mathcal{L}(M)$ be the set of all smooth, non-degenerate Lorentz metrics on M. Given g and $\bar{g}$ in $\mathcal{L}(M)$ define $g<\bar{g}$ to hold if for all points a in M and non-zero vectors $X\varepsilon M_a$, $g_a(X,X)\geq 0 \rightarrow \bar{g}_a(X,X)>0$. Intuitively, $g<\bar{g}$ just means that the associated cones of $\bar{g}$ are everywhere wider than those of g. One way of defining stable causality, then, is this:

3.18 Definition: A space-time (M,g) is stably causal if there is a $\bar{g}$ in $\mathcal{L}(M)$ where $g<\bar{g}$ and where $(M,\bar{g})$ is causal.

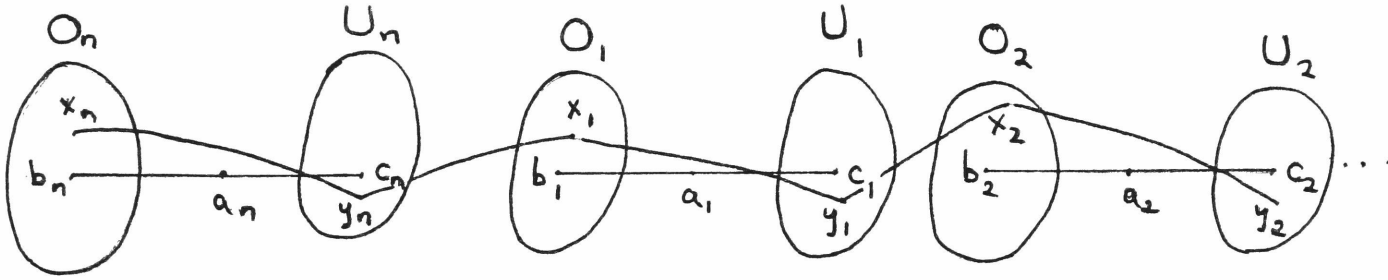
Clearly if $g<\bar{g}$ and $(M,\bar{g})$ is causal, then $(M,\bar{\bar{g}})$ is causal for all $\bar{\bar{g}}$ in $\mathcal{L}(M)$ where $g<\bar{\bar{g}}<\bar{g}$.

The examples above indicate that, for every n, there can be a space-time which is nth order causal but still not stably causal. It is easy to check, however, that stable causality implies nth order causality for every n.

3.19 Theorem: If a space-time (M,g) is stably causal then it is nth order causal for every $n\geq 2$.

Proof: Suppose distinct points $a_1, \ldots, a_n$ witness the failure of nth order causality. Then with the standard construction used before we can find disjoint local open diamonds I_i containing a_i, and points b_i and c_i in I_i, such that $b_i<a_i<c_i(I_i)$ for all i, and such that $I^-(c_i)<<I^+(b_{i+1})$ for $i=1, \ldots, n-1$ with $I^-(c_n)<<I^+(b_1)$. Now let $\bar{g}$ be a metric

in $\mathcal{L}(M)$ where $g<\bar{g}$ and where $(M,\bar{g})$ is causal. Then for each i we have $b_i<<c_i$ with respect to $\bar{g}$ - if previously $b_i \to c_i$, then the null geodesic (with respect to g) that connected them is now a timelike curve. Working with $\bar{g}$ we can find neighborhoods O_i of b_i and U_i of c_i such that for all i, $O_i << U_i$. If we take



x_i in $I^+(b_i) \cap O_i$ and y_i in $I^-(c_i) \cap U_i$ then we have $x_i << y_1 << x_2 << y_2 << \ldots << y_n << x_1$. This contradicts the assumption that $(M,\bar{g})$ is causal. qed

There remains the question whether a space-time that is nth order causal <u>for every n</u> is necessarily stably causal. I do not know if this has been settled but would guess that the claim is false.

Now we prove a theorem of Hawking which gives an interesting equivalent formulation of stable causality.

3.20 <u>Definition</u>: If (M,g) is a space-time, a smooth map $t: M \to R$ is called a <u>universal time function</u> on (M,g) if

for all distinct points a and b in M, $a<b$ implies $t(a)<t(b)$.

Hawking's Theorem asserts that a space-time is stably causal if and only if it admits a universal time function. The proof we give will fill in some of the detail that is omitted in Hawking and Ellis ([8], pg 198-201). First several Lemmas that we need are noted.

3.21 Lemma: If (M,g) is a space-time and $t: M \to R$ is a smooth function on M, then t is a universal time function on (M,g) iff ∇t is everywhere future directed timelike.

Proof: Consider an arbitrary point a in M and an arbitrary differentiable curve $\gamma : I \to M$ through a. We can choose coordinates x_i in some neighborhood of a such that at a, the $\frac{\partial}{\partial x_i}$ are orthogonal and $\frac{\partial}{\partial x_o} = \nabla t$. From the latter we have $\frac{\partial t}{\partial x_i} = \delta_{io}$. Now consider the function $t(s) = t \circ \gamma(s)$ defined over I. It follows from our choice of coordinates that at a:

$$||\nabla t|| \frac{dt}{ds} = ||\nabla t|| \frac{\partial t}{\partial x_i} \frac{d\gamma_i}{ds} = ||\nabla t|| \frac{d\gamma_o}{ds} = \nabla t \cdot \frac{d\gamma}{ds}$$

where $||\nabla t|| = g(\nabla t, \nabla t)$.

Now first suppose that ∇t is a universal time function on (M,g) but that at some point a ∇t is not future directed timelike. Suppose - first that ∇t is spacelike. Then we can find a future directed causal curve γ through a such that at a $\nabla t \cdot \frac{d\gamma}{ds} > o$. But we also have that $\frac{dt}{ds} \geq o$ at a since t is increasing along γ. So $o < \nabla t \cdot \frac{d\gamma}{ds} = ||\nabla t|| \frac{dt}{ds} \leq o$ which is impossible. Similarly, if ∇t is null we derive a contradiction by considering any future directed causal curve γ for which $\nabla t \cdot \frac{d\gamma}{ds} \neq 0$. And if ∇t is timelike, past-directed, we consider any future directed causal curve for which $\nabla t \cdot \frac{d\gamma}{ds} < 0$.

Conversely, suppose ∇t is future directed timelike at a and let γ be any future directed causal curve through a. Since $\nabla t \cdot \frac{d\gamma}{ds} > o$ it follows that $\frac{dt}{ds} > o$. Since a and γ are arbitrary here, t must increase along all future directed causal curves. Hence t must be a universal time function on (M,g). qed

The definition of $g < \bar{g}$ for metrics $g, \bar{g}$ in $\mathcal{L}(M)$ makes perfectly good sense for smooth, pseudo-Riemannian metrics on M even if they are not non-degenerate and Lorentzian. From now we use the relation in the extended sense.

3.22 Lemma: If (M,g) is a spacetime with $g < \bar{g}$ for $\bar{g}$ in $\mathcal{L}(M)$, then for any smooth metric $\bar{\bar{g}}$ on M if $g < \bar{\bar{g}} < \bar{g}$ it follows that $\bar{\bar{g}}$ is in $\mathcal{L}(M)$.

Proof: Suppose first that $\bar{\bar{g}}$ were degenerate at some point a. Then there would be a non-zero vector X in M_a such that $\bar{\bar{g}}_a(X,Y) = o$ for all Y in M_a. It would follow that $\bar{\bar{g}}_a(X,X) = o$, and hence both $g_a(X,X) < o$ and $\bar{g}_a(X,X) > o$. Let Y be some vector which is timelike with respect to g_a (and hence with respect to $\bar{\bar{g}}_a$ and $\bar{g}_a$). Certainly X and Y must be linearly independent. Then for all $r,s \in \mathbb{R}$ we would have $\bar{\bar{g}}_a(sX+rY, sX+rY) = r^2 \bar{\bar{g}}_a(Y,Y) \geq o$. Hence $\bar{g}_a(sX+rY, sX+rY) > o$ for all $r,s \in \mathbb{R}$ even though X,Y are both timelike with respect to $\bar{g}_a$. This is impossible since $\bar{g}_a$ is non-degenerate Lorentzian.

So $\bar{\bar{g}}$ must be everywhere non-degenerate. At any point a, then, the components of $\bar{\bar{g}}_a$ can be put in the form $\mathrm{diag}(+1, \pm 1, \ldots, \pm 1)$. If there were a second positive entry there would have to be two linearly independent vectors X, Y, timelike with respect to $\bar{\bar{g}}$ (and hence $\bar{g}$), such that all vectors $rX+sY$ were timelike with respect to $\bar{\bar{g}}$ (and hence $\bar{g}$). Again this is impossible. qed

The final lemma we need asserts the existence of a volume measure on all space-times. We state only the weak version that is actually used.

3.23 Lemma: If (M,g) is a space-time and $\mathcal{T}_M$ is the set of open subsets of M, there is a function $\mu: \mathcal{T}_M \to R$ such that:

(i) $\mu(M) = 1$
(ii) $O \neq \phi \to \mu(O) > 0$
(iii) $O \subset O' \to \mu(O) \leq \mu(O')$
(iv) $O \cap O' = \phi \to \mu(O \cup O') = \mu(O) + \mu(O')$
(v) For all a in M and all $\varepsilon > 0$ there is an open set $\mathcal{U}$ containing a where $\mu(\mathcal{U}) < \varepsilon$.

Proof: Every differentiable manifold which admits a C^1 affine connection must be paracompact (Geroch [ ]). So there must exist a countable atlas (U_i, ϕ_i) on M where the U_i are locally finite, i.e. for every point a in M there is an open set O containing a which intersects only finitely many of the U_i. Let ρ_i be the standard Lebesque measure on $\phi_i[U_i] \subseteq \mathbb{R}^n$, normalized so that $\rho_i(\phi_i[U_i]) = 1$. Now for each open O we define:

$$\mu_i(O) = \rho_i(\phi_i[O \cap U_i]) \text{ and}$$
$$\mu(O) = \sum 2^{-i} \mu_i(O).$$

Then: $\mu(M) = \sum 2^{-i} \mu_i(M) = \sum 2^{-i} \rho_i(\phi_i[U_i] = \sum 2^{-i} = 1.$

The other requirements on μ are checked similarly. In each case the property is inherited from the ρ_i. qed

Now we can prove Hawking's result.

3.24 Theorem: A space-time (M,g) admits a universal time function iff it is stably causal.

Proof: First assume (M,g) admits a universal time function t. Consider the smooth, pseudo-Riemannian metric on M defined by $\bar{g} = g + dt \otimes dt$. We claim $\bar{g} \in \mathcal{L}(M)$, $g < \bar{g}$, and $(M,\bar{g})$ is causal.

To check the first two claims we pick a convenient coordinate system at an arbitrary point a. We can find $\{x_i\}$ in some neighborhood of a such that at a the $\frac{\partial}{\partial x_i}$ are orthogonal (with respect to g_a), $\frac{\partial}{\partial x_o} = \nabla t$, and for $i = 1, \ldots, n-1$ $g_a(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_i}) = -1$. In these coordinates at a, $\bar{g}$ assumes the form $\bar{g}_a = (||\nabla t|| + 1)dx_o^2 - dx_1^2 - \ldots - dx_{n-1}^2$ where $||\nabla t|| = g(\nabla t, \nabla t) > o$. This insures $\bar{g} \in \mathcal{L}(M)$ and $g < \bar{g}$.

Now suppose γ: $[0,1] \to M$ is a closed, future directed causal curve with respect to $\bar{g}$. Consider $t(s) = to\gamma(s)$. Since $t(o) = t(1)$, the derivative $\frac{dt}{ds}$ must vanish at some point s_o. At $\gamma(s_o)$, by the equation derived in the proof of Lemma 3.21, we then have $g(\nabla t, \frac{d\gamma}{ds}) = o$ and hence that $\frac{d\gamma}{ds}$ is spacelike with respect to g. But this leads to a contradiction. At $\gamma(s_o)$ we must have:

$$\bar{g}(\frac{d\gamma}{ds}, \frac{d\gamma}{ds}) = g(\frac{d\gamma}{ds}, \frac{d\gamma}{ds}) + dt \otimes dt(\frac{d\gamma}{ds}, \frac{d\gamma}{ds}) = g(\frac{d\gamma}{ds}, \frac{d\gamma}{ds}) + (\frac{dt}{ds})^2 < o$$

even though γ is causal with respect to $\bar{g}$.

Conversely, assume (M,g) is stably causal. We shall construct a universal time function on (M,g). Suppose $\bar{g}$ in $\mathcal{L}(M)$ is such that $g < \bar{g}$ and $(M,\bar{g})$ is causal. Consider the metrics $g_k = (3-k)g + k\bar{g}$ defined for $k \in (o,3)$. Since $3g < g_k < 3\bar{g}$ it follows by Lemma 3.22 that g_k is in $\mathcal{L}(M)$ for each k. Let μ be a volume measure on (M,g) as guaranteed by Lemma 3.23. For each point a and $k \in (o,3)$ we define a "past volume": $V(a,k) = \mu[I^-(a,g_k)]$, where $I^-(a,g_k)$ is the set $I^-(a)$ with respect to the metric g_k. For given k,

$V(a,k)$ is a real valued function on M. We claim now that it increases along future directed causal curves. Suppose $a<a'$ for distinct points a and a'. (M,g_k) is itself stably causal and hence past distinguishing. So there must exist an open set O containing a' where $I^-(a,g_k)\cap O = \phi$. If $O' = I^-(a',g_k)\cap O$, then $O'\neq\phi$ and $I^-(a,g_k)\cup O'\subseteq I^-(a',g_k)$. So $V(a,k)<V(a,k)+\mu(O')$ $= \mu[I^-(a,g_k)\cup O']\leq V(a',k)$.

If $V(a,k)$ were continuous we would be almost done. We would have a "continuous universal time function" and it would only remain to smooth it. But it need not be continuous for any k because, as previous examples have shown, space-times can be rather jagged in their excisions. In figure a, for instance, the points p_i converge to p but their respective past volumes (in any g_k) do not approach the past volume of p. All of the p_i, no matter how close, can "see around the obstruction." Hawking's idea is to define a new function t which averages the past volumes over a range of k:

$$t(a) = \int_1^2 V(a,k)dk.$$

The function is well defined since for a given a, $V(a,k)$ is monotone increasing on the interval [1,2] and is, hence, integrable. Certainly t increases along future directed causal curves since $V(a,k)$ does so for each k. We now show that it is continuous.

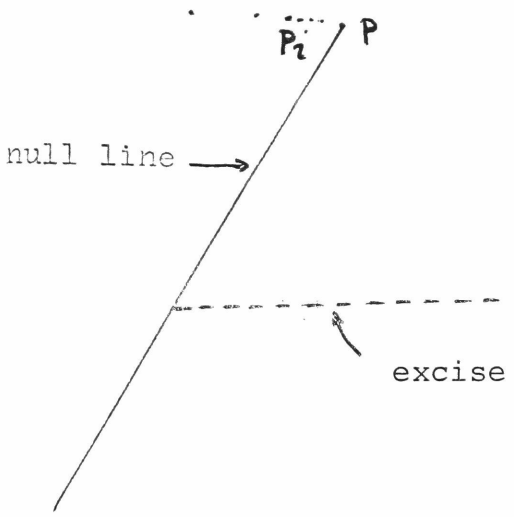


Fig a

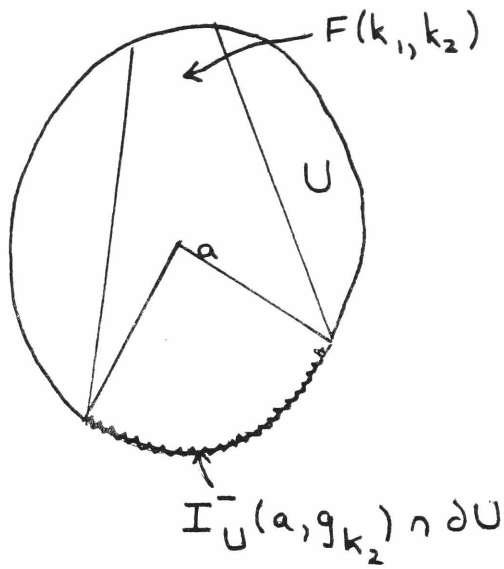


Fig b

For given a in M and $\varepsilon>0$ we must find an open set O containing a such that for all b in O: $t(a)-\varepsilon\leq t(b)\leq t(a)+\varepsilon$. Let U be an open neighborhood of a with compact closure such that $\mu(U)<\frac{\varepsilon}{2}$. Compactness guarantees that for all b in U, any inextendible curve starting at b must eventually intersect ∂U, the boundary of U. Now for $k_1,k_2\in[0,3]$ with $k_1<k_2$ consider the set $I_U^-(a,g_{k_2})\cap\partial U$.

Let $F(k_1,k_2)$ consist of those points in U such that every past directed, past inextendible g_{k_1}-timelike curve from the point hits $I_U^-(a,g_{k_2})\cap\partial U$ - see figure b. We note that a is in the interior of $F(k_1,k_2)$ which we write as $\mathring{F}(k_1,k_2)$. [If a were on the boundary of $F(k_1,k_2)$ we could find a sequence of points a_i converging to a and points $c_i\in\partial U-I_U^-(a,g_{k_2})$ where $c_i<<a_i(U)$ with respect to g_{k_1}. If c were a point of accumulation of the c_i it would follow that $c<a(U)$ relative to g_{k_1}. Hence by $k_1<k_2$ we would have $c<<a(U)$

relative to g_{k_2}. This runs afoul of $c_i \notin I_U^-(a,g_{k_2})$.] We also note that if $k_1 \leq \bar{k}_1 < \bar{k}_2 \leq k_2$ then $F(\bar{k}_1,\bar{k}_2) \subseteq F(k_1,k_2)$.

Now let n be an integer greater than $\frac{2}{\varepsilon}$. We can partition the interval $[1, 2+\frac{1}{2n}]$ into 2n+1 subintervals: $[1, 1+\frac{1}{2n}]$, $[1+\frac{1}{2n}, 1+\frac{2}{2n}]$, ..., $[1+\frac{2n-1}{2n}, 2]$, $[2, \frac{2n+1}{2n}]$. Corresponding to each interval $[1+\frac{i}{2n}, 1+\frac{i+1}{2n}]$ where $0 \leq i \leq 2n$ we have the set $\overset{\circ}{F}(1+\frac{i}{2n}, 1+\frac{i+1}{2n})$. Let O be the intersection of all of these $\overset{\circ}{F}$ sets. Then O is an open set containing a. Furthermore, for any $k \in [1,2]$ we can find an i where $0 \leq i \leq 2n$ such that $k \leq 1+\frac{i}{2n} < 1+\frac{i+1}{2n} \leq k+1/n$. Hence $O \subseteq \overset{\circ}{F}(k,k+1/n)$ for each $k \in [1,2]$. Let b be any point in O. If for some $k \in [1,2]$, $d \in I^-(b,g_k)-U$ then by the definition of $F(k,k+\frac{1}{n})$ it follows that $d \in I^-(a,g_{k+1/n})-U$. Thus for any $k \in [1,2]$ and any point b in O we have:

$$I^-(b,g_k) \subseteq I^-(a,g_{k+1/n}) \cup U \text{ and hence}$$

$$V(b,k) \leq V(a,k+1/n)+\mu(U) \leq V(a,k+\tfrac{\varepsilon}{2})+\mu(U).$$

Thus: $$t(b) = \int_1^2 V(b,k)\,dk \leq \int_{1+\epsilon/2}^{2+\epsilon/2} V(a,k)\,dk + \frac{\varepsilon}{2}$$

$$\leq t(a) + \int_2^{2+\frac{\varepsilon}{2}} V(a,k)\,dk + \frac{\varepsilon}{2} \leq t(a) + \varepsilon$$

Thus for every b in O we have $t(b) \leq t(a)+\varepsilon$. With a parallel argument we can find an O' such that for all b in O', $t(a)-\varepsilon \leq t(b)$. It only remains to take their intersection to establish that t is continuous. The argument that it can be smoothed is more involved

than the smoothing argument used to show that differentiable causal curves can be "rounded off" to smooth causal curves. Hawking gives a reference to Seifert's dissertation [17], but does not himself give the proof. Neither do we. qed

We summarize now the causality conditions mentioned in order of decreasing strength:

stable causality $\equiv$ existence of universal time function
$\downarrow$
$\forall n$ (nth order causality)
$\downarrow$
n+1st order causality
$\downarrow$
nth order causality ($n \geq 2$)
$\downarrow$
strong causality $\equiv$ $\mathcal{J}_M = \mathcal{J}_A$
$\downarrow$
past and future distinguishability
$\downarrow$
past distinguishability (future distinguishability)
$\downarrow$
causality
$\downarrow$
chronology

All arrows are strict except, possibly, for the first. Geroch has suggested the problem of proving, relative to some general notion of "causality condition," that stable causality is the strongest possible causality condition.

C. Determination of Spatio-Temporal Structure from Causal Structure

In the chapter on Minkowski space-time several senses were discussed in which topological, linear, and metrical structure were determined by causal structure. These senses were explicit definability, categorical axiomatizability, and what was called

"implicit definability." In this section we investigate to what extent the results discussed carry over to the class of general relativistic space-times. We also verify the claim made in section A that space-time manifold structure is uniquely determined by the class of causal curves. Only asymmetric causal structure will be considered in this section. The results will be generalized to the symmetric case in section D.

First, explicit definability. In Minkowski space-time, as we showed, it is possible to give explicit first order "causal definitions" for vector parallelism, metrical congruence, and "openness" of sets. From Theorem 3.15 we know that at least in strongly causal space-times the manifold topology can be defined via the Alexandrov topology (as in chapter II, pg. 60). The question arises whether some causally definable topology, more complex than the Alexandrov topology, might not be equal to the manifold topology under conditions weaker than strong causality. But one cannot ask after the possibility of defining vector parallelism or metrical congruence from causal relations. General space-times have no vector space structure, nor do they carry a notion of length between arbitrary points. The latter point is sometimes overlooked.

In Riemannian manifolds the distance between two points is taken to be the greatest lower bound of the lengths of (piecewise) differentiable curves connecting the two points. No such definition is applicable to the Lorentzian case. It will always be possible to connect any two points with a (piecewise) differentiable curve

whose "length" is zero, and with one whose (negative) length is smaller than any specified negative integer. One can try to define a distance between causally related points, taking it to be the least upper bound of the lengths of all causal curves connecting the points. When such a bound does exist it will be realized as the length of a causal geodesic between the points. (See Hawking & Ellis [8], pg. 213) And in the special case of Minkowski space-time this l.u.b. definition coincides with the standard definition of length for causally related points. But in general the l.u.b. will not exist. Even where it does exist for all causally related points, no corresponding definition of length is available for points which are not causally related. In any space-time it will always be possible to connect any two points by a spacelike curve with (negative) length less than any specified negative integer.

Locally, within a convex neighborhood, the l.u.b. definition of length for causally related points is always available; and the l.u.b. is realized by the unique local causal geodesic connecting the points. But again, no symmetric glb definition of length is available for space-like related points. In addition to the spacelike geodesic connecting them there will also be "shorter," flat, spacelike curves which wrap "horizontally" around the inner cone m-times, for any m. These, fully contained in the convex neighborhood, will realize arbitrarily small negative lengths. Thus in an arbitrary space-time there is no linear structure and no congruence relation which one can ask to have determined by causal relations.

Second, categorical axiomatizability. There is no problem here

as to how to formulate the problem. It carries over directly from the special case of Minkowski space-time. Given a set of causal relations $\mathcal{R}$ and a formal language with relation symbols for the elements of $\mathcal{R}$, let the $\mathcal{R}$-theory of a space-time be the class of sentences in the formal language true under interpretation in that space-time. We may say that a space-time is categorized by its $\mathcal{R}$-theory (relative to some specification of a language) if its $\mathcal{R}$-theory is categorical. We know from chapter II that any space-time conformally isometric to Minkowski space-time is categorized by, for example, its $<$-theory in the language of second order quantification theory. We might investigate the question whether any other space-time is similarly categorized. The answer is probably no. But proving this would be non-trivial and would turn more on technical points in logic than on features of space-time structure. More interesting are questions concerning the determination of spatio-temporal structure from causal relations in the sense of implicit definability. The requirements of implicit definability are hard enough to satisfy, too hard to be satisfied in all space-times. Rather than looking for yet stronger senses in which levels of spatio-temporal structure may be determined by causal structure, it is appropriate to look for a weaker sense which just might apply to all space-times.

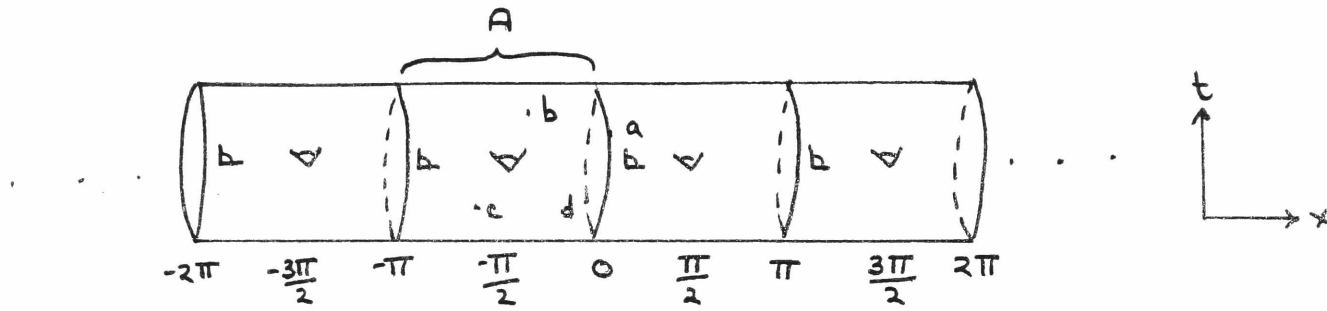
The notion of implicit definability introduced in connection with Minkowski space-time will not serve in the general case. Recall we said that a topology (resp. congruence relation) is

implicitly definable from causal relations $\mathcal{R}$ if all $\mathcal{R}$-automorphisms of Minkowski space-time are homeomorphisms with respect to the topology (resp. isometries with respect to the congruence relation). But, in general, a space-time will not admit any causal automorphisms (other than the identity); in such cases the criterion would render all topologies implicitly definable.

There is, however, a natural reformulation. Space-times as we have described them are mathematical models carrying several different levels of structure. The question we want to ask is whether it is possible to construe the causal structure as primitive and then uniquely recover the other levels from it. Given a space-time (M,g) and some class of causal relations $\mathcal{R}$, we say that its topological (resp. differential, conformal, projective, pseudo-Riemannian) structure is <u>implicitly definable from the relations</u> $\mathcal{R}$ if for all space-times (M',g') and all bijections $\phi:M\to M'$, if ϕ is an $\mathcal{R}$-isomorphism, then ϕ must be a homeomorphism (resp. smooth diffeomorphism, conformal isometry, projective isometry, isometry). We want to determine which levels of structure are implicitly definable from which (asymmetric) relations under which conditions.

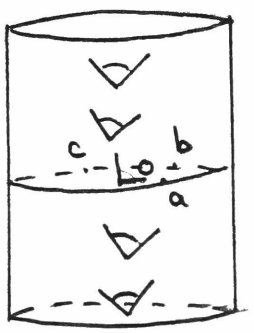
In Minkowski space-time the three asymmetric causal relations $<$, $<<$, and $\to$ are interdefinable and one can equally well work with any one of them as primitive. This is not the case in general. No one of the relations is definable from any of the others. A few simple two dimensional examples will establish this. First, consider

the following space-time with metric $ds^2=(\sin^2 x)(dt^2-dx^2)+\cos^2 x\,dt\,dx$ whose associated null cones as in the figure. They are horizontal



at points where $x = n\pi$ and assume standard Minkowski form at points where $x = \left(\frac{2n+1}{2}\right)\pi$. Let each region between a pair of $x = n\pi$ "rings" be called an A region. Every point in one such is chronologically prior to every other point in it. Suppose we map each A region according to any old bijection onto the ring forming the boundary to its immediate right. So, for example, the point b might be mapped to a, c to d, and so forth. The resulting bijection of the space-time certainly does not preserve $<<$ or $\rightarrow$ (e.g. $c<<b$ but $d\rightarrow a$). But it does preserve $<$. Certainly, then, $<<$ and $\rightarrow$ are not definable in general from $<$.

Next consider again the vertical punctured space-time which exhibited the fact that causality implies neither past nor future distinguishability. Let a,b,c be points on the equator with $c<a<b$ but not $b<<c$. If we map this space-time onto itself leaving all points fixed except b and c, and interchanging



them, we preserve $<<$ but not $<$ or $\rightarrow$. Hence neither of the latter two is definable in general from $<<$.

Finally we show that $<$ and $<<$ are not generally definable from $\rightarrow$. Consider again the first example. We map the space onto itself leaving fixed all points with x coordinate $<-\pi$ or $\geq\pi$, but interchanging the rings at $-\pi$ and 0, and interchanging the A sections $-\pi<x<0$ and $0<x<\pi$. More precisely we map:

$$(t,x) \longrightarrow \begin{cases} (t,x) & \text{if } x<-\pi \text{ or } x\geq\pi \\ (t,x+1) & \text{if } -\pi\leq x<0 \\ (t,x-1) & \text{if } 0\leq x<\pi \end{cases}$$

It is easy to check that $\rightarrow$ but neither $<$ nor $<<$ is preserved under the mapping.

Keeping track of all three relations and the minimal conditions for their relative interdefinability is a tedious business. Rather than doing so we shall work with $<$ as primitive and leave it at that. In the presence of future distinguishability (or past distinguishability) one has the following equivalences which establish the definability of $<<$ and $\rightarrow$ from $<$:

$$a<<b \leftrightarrow a<b \;\&\; \exists\, c,d_1,d_2\,[c\neq b \,\&\, a<d_1<c<b \;\&\; a<d_2<c \;\&\; \neg\,(d_1 \kappa d_2)]$$

$$a\rightarrow b \leftrightarrow a<b \;\&\; \neg\,(a<<b).$$

This is easy to check. But, interestingly, it does not seem possible to give a simple first order definition of $<$ and $\rightarrow$ from $<<$, or of $<$ and $<<$ from $\rightarrow$, not even in the presence of stable causality. Non-first order definitions are available in the presence of past and future distinguishability but they are messy.

We turn now to the questions about implicit definability of

spatio-temporal structure. First, we know that given any space-times (M,g) and (M',g') and any conformal isometry f: M→M', f preserves $<$, $<<$, and $\rightarrow$. So in no space-time is projective or full pseudo-Riemannian metrical structure implicitly definable from any causal relations. The basic positive result that one does have is due to Hawking. It asserts that in all strongly causal space-times, topological, differential, and conformal structure are implicitly definable from $<$.

3.25 Theorem: If (M,g) and (M',g') are strongly causal space-times and if f: M→M' is an $<$-isomorphism then f is a smooth conformal isometry.

Proof: First f must be a homeomorphism. Since f and f^{-1} preserve $<<$ they must preserve the Alexandrov topology, hence by Theorem 3.15 they must preserve the manifold topology. Next, f and f^{-1} must preserve (images of) null geodesics. [Suppose γ is a null geodesic through a in M but $f[\gamma]$ is not a null geodesic through f(a). Then given any open set O containing a we can find points b,c on $\gamma \cap O$, below and above a, so that $b \rightarrow c(O)$ but $f(b) << f(c)(f[O])$ Hence $b << c$. This means that there must be a future directed causal curve which leaves and then reenters O. Since O is arbitrary this runs into contradiction with the assumption that (M,g) is strongly causal. So f must preserve null geodesics and, similarly, so must f^{-1}.]

The non-trivial part of the proof is Hawking's argument [7] that f must be a smooth diffeomorphism. The basic idea is to use families of null geodesics to construct coordinate patches. One shows that since null geodesics themselves are preserved the resulting coordinates are mapped nicely by f and f^{-1}. The argument is rather

technical and we skip it here. Of course once it is established that f is a smooth diffeomorphism then, as discussed in section A, it follows automatically that f is a conformal isometry. qed

The theorem has as an immediate consequence the following reductive lemma.

3.26 Corollary: If (M,g) and (M',g') are space-times and $f:M\to M'$ is a homeomorphism such that for all open O in M and all $a,b\in O$: $a<b(O)\leftrightarrow f(a)<f(b)(f[O])$, then f is a smooth conformal isometry.

Proof: Given any point $a\in M$ let O be some convex neighborhood of a. Now considered as a space-time in its own right $(O,g|_O)$ is strongly causal. [For suppose there exist points a,b in O such that $a<b(O)$ and for all $d\in I^-_O(b)$, $e\in I^+_O(a)$ we have $d<<e(O)$. Then by the usual continuity arguments applying to convex neighborhoods it follows that $b<a(O)$. Hence a=b.] Therefore $(f[O],g'_{|f[O]})$ must be strongly causal. So $f|_O: O\to f[O]$ is an $<$-isomorphism between strongly causal space-times and so, by the theorem, must be a smooth conformal isometry. Since being a smooth conformal isometry is a local condition on f the result follows. qed

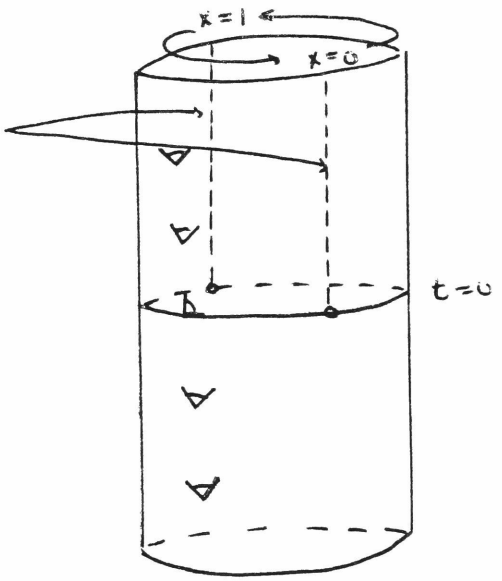
[We note that both the theorem and its corollary can be recast on the assumption that f and f^{-1} preserve $<<$, or preserve $\to$.]

A natural question to ask now is whether the hypothesis of strong causality can be weakened or eliminated. That it cannot be weakened to future distinguishability (or past distinguishability) is demonstrated by the following example. It is given in two-dimensions here, but higher dimensional analogues are readily available. One starts with the cylindrical space-time that has been used repeatedly

-- the (t,x) plane with metric $ds^2 = (\cosh t-1)^2(dt^2-dx^2) + dtdx$, where points (t,o) and say points (t,2n) are identified for all n. One excises two closed half lines: $\{(t,x): x=o \ \& \ t \geq o\}$ and $\{(t,x): x=1 \ \& \ t \geq o\}$ rendering the space-time future (but not past) distinguishing (see figure). Now let ϕ be a bijection of the space-time onto itself defined by:

$$\phi:(t,x) \to \begin{cases} (t,x) & \text{if } t<o \\ (t,x+1) & \text{if } t \geq o \end{cases}$$

ϕ leaves the lower open half of the space-time fixed but reverses the position of the two upper slabs. It preserves $<$, $\ll$, and $\to$ but is certainly discontinuous along the t=o axis. A symmetric example would serve for the past distinguishing case.



This leaves the question whether the hypothesis in the theorem can be weakened to past and future distinguishability. We show in this section that the answer is yes. First we reformulate the problem somewhat. Recall that a <u>continuous</u> curve is <u>causal</u> if for all points a and b on the curve and all open sets O containing the portion of the curve from a to b, there is a (differentiable) causal curve in O connecting the two points.

3.27 Lemma: If (M,g) and (M',g') are past and future distinguishing space-times, and if $f: M \to M'$ is a $<$-isomorphism, then f and f^{-1} must take future directed continuous causal curves to future directed continuous causal curves.

Proof: Suppose first that two sequences $\{b_i\}$ and $\{c_i\}$ converge to a causally from below and above respectively, i.e. for all i, $b_i < b_{i+1} < \ldots < a < \ldots < c_{i+1} < c_i$. We claim that the $f(b_i)$ and $f(c_i)$ must converge to $f(a)$. Suppose the $f(c_i)$ do not converge to $f(a)$. Let $U \subseteq M'$ be an open set containing $f(a)$ such that for all i_o there is an $i \geq i_o$ where $f(c_i)$ is not in U. By future distinguishability there must be an open subset $U' \subseteq U$ containing $f(a)$ such that no future directed causal curve from $f(a)$ which leaves U' ever reenters. Now let $f(d)$ be a point in U' where $f(a) << f(d)(U')$. Then of course $a << d$ and there must be an open set $O \subseteq I^-(d)$ containing a. For some i_o, c_i is in O for all $i \geq i_o$. Hence $f(c_i) << f(d)$ for all $i \geq i_o$. But for some $i \geq i_o$ $f(c_i)$ is not in U'. It follows that for this i, the future directed causal curve from $f(a)$ to $f(c_i)$ to $f(d)$ must leave and reenter U' which is a contradiction. Symmetrically we argue from past distinguishability to show that the $f(b_i)$ must converge to $f(a)$.

It follows that if $\gamma: I \to M$ is a future directed continuous causal curve, then $f \circ \gamma: I \to M'$ is a continuous curve in M'. Suppose now that $f(a)$, $f(b)$ are on $f \circ \gamma$ in M' with $f(a) < f(b)$. Let σ be the closed subsegment of the curve connecting them and let U be any open set containing σ. We must show $f(a) < f(b)(U)$. First there must be some point $f(d)$ on σ between $f(a)$ and $f(b)$

such that $f(a)<f(d)(U)$. [For let $U' \subseteq U$ be a subset of U containing $f(a)$ having the property that no future directed causal curve from $f(a)$ which leaves U' ever reenters. Consider any point $f(d) \in \sigma \cap U'$. Since $f(a)<f(d)$ we must have $f(a)<f(d)(U')$ and hence $f(a)<f(d)(U)$.] Let $f(d_o)$ be the lub of all points $f(d)$ on σ where $f(a)<f(d)(U)$. Let $U'' \subseteq U$ be any open subset of U containing $f(d_o)$ such that no past or future directed causal curve from $f(d_o)$ which leaves U'' ever reenters. We can find a point $f(d)$ on $\sigma \cap U''$ between $f(a)$ and $f(d_o)$ where $f(a)<f(d)(U)$ -- by our assumption about $f(d_o)$. Since $f(d)<f(d_o)$ we have $f(d)<f(d_o)(U'')$. This, with $f(a)<f(d)(U)$, entails $f(a)<f(d_o)(U)$. Now if $f(d_o) \neq f(b)$ we could repeat the earlier argument and find a point $f(d)$ on $\sigma \cap U''$ between $f(d_o)$ and $f(b)$ such that $f(d_o)<f(d)$. As before this would entail $f(d_o)<f(d)(U'')$ and hence $f(a)<f(d)(U)$. This would violate our assumption about $f(d_o)$. So $f(b) = f(d_o)$ and $f(a)<f(b)(U)$. qed

The theorem we are going to prove is:

3.28 Theorem: If (M,g) and (M',g') are space-times and if $f: M \to M'$ is a bijection such that f and f^{-1} preserve future directed continuous causal curves, then f must be a homeomorphism.

It is clear that this, together with Lemma 3.27 and corollary 3.26, will imply our claim.

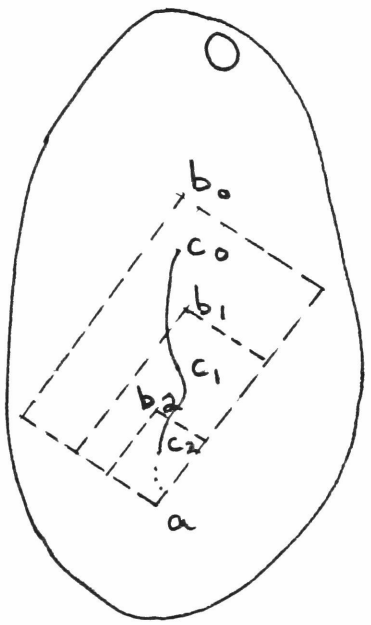
3.29 Corollary: If (M,g) and (M',g') are past and future distinguishing space-times and if $f: M \to M'$ is an $<$-isomorphism, then f is a smooth, conformal isometry.

We note that the Corollary remains true if the hypothesis that f is an $<$-isomorphism is replaced by the condition that f is an $<<$-isomorphism (or $\to$-isomorphism).

The proof of Theorem 3.28 is rather cumbersome and will be broken down into steps. In what follows we assume that (M,g), (M',g') together with a bijection $f: M \to M'$ satisfy the hypothesis of Theorem 3.28. Images under f will be denoted with primes, i.e. for all $x \in M$, $x' = f(x) \in M'$.

A Lemma: Given any point $a \in M$, and any convex neighborhood O of a, there is a point b in O where $a \ll b$ (O) such that f is continuous over the set $I^+_O(a) \cap I^-_O(b)$.

Proof: Let $\mathcal{U}$ be a convex neighborhood of a' in M. We show first that there is a $b \in I^+_O(a)$ such that $I^+_O(a) \cap I^-_O(b)$ is mapped into $\mathcal{U}$. Suppose not. Take b_o to be any point in $I^+_O(a)$. Then there is a point c_o in $I^+_O(a) \cap I^-_O(b_o)$ where c_o' is not in $\mathcal{U}$. Now let b_1 be a point in $I^+_O(a)$ where $I^+_O(a) \cap I^-_O(b_1) \subsetneq I^-_O(c_o)$.



Then there must be a point c_1 in $I^+_O(a) \cap I^-_O(b_1)$ where $c_1 \ll c_o$ (O) and c_1' is not in $\mathcal{U}$. Next let b_2 be a point in $I^+_O(a)$ where $I^+_O(a) \cap I^-_O(b_2) \subsetneq I^-_O(c_1)$. Then we can find a c_2 in $I^+_O(a) \cap I^-_O(b_2)$ where $c_2 \ll c_1$ (O) but c_2' is not in $\mathcal{U}$. Continuing in this way we generate a sequence $\{c_i\}$ in $I^+_O(a)$ converging chronologically to a such that for all i, c_i' is not in $\mathcal{U}$. For each i, let γ_i be a future directed timelike curve from c_{i+1} to c_i within O. Linking these segments together and adjoining the point a we have a future directed continuous

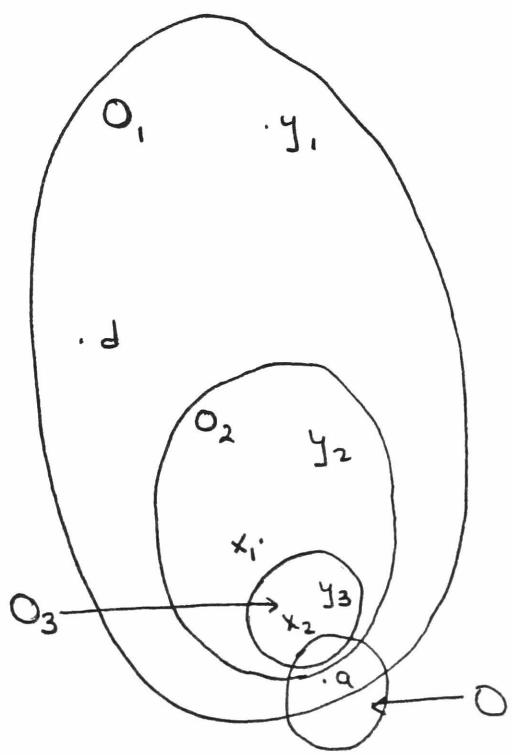
causal curve Γ from a to c_o threading all the c_i. By our construction no initial segment of Γ falls within $\bar{U}$. But this leads to a contradiction. The image of Γ must be a future directed continuous causal curve through a' and $\bar{U}$ contains an initial segment of every one of these.

So there must be a point b in $I^+_O(a)$ such that $I^+_O(a) \cap I^-_O(b)$ is mapped into U. But this immediately entails that f is continuous over $I^+_O(a) \cap I^-_O(b)$. Let x be any point in $I^+_O(a) \cap I^-_O(b)$. If U_1 is any open set containing x' in M' we can find points c,d in O where $a<<c<<x<<d<<b(O)$ such that $c',d' \in U$ and $J^+_U(c') \cap J^-_U(d') \subseteq U_1 \cap U$. But certainly $I^+_O(c) \cap I^-_O(d)$ is mapped into $J^+_U(c') \cap J^-_U(d')$. So f must be continuous at x. qed

Let $\mathcal{D}$ consist of those points in M at which f is discontinuous.

B. Lemma: If $\mathcal{D} \neq \phi$ then there is an open set O such that $\mathcal{D} \cap O \neq \phi$ and $\mathcal{D} \cap O$ is achronal in O [i.e. if $x,y \in \mathcal{D} \cap O$ then not $x<<y(O)$].

Proof: Suppose no such O exists. Let d be any point in $\mathcal{D}$ and let O_1 be any open set containing d. By passing to a subset if necessary we may as well assume that O_1 has compact closure. By our assumption we can find points x_1,y_1 in $O_1 \cap \mathcal{D}$ such that $x_1<<y_1(O_1)$. Now let O_2 be an open set containing x_1 where $O_2 \subseteq I^-_{O_1}(y_1)$ - see figure. Repeating the argument with respect to O_2 we can find points x_2,y_2 in $O_2 \cap \mathcal{D}$ where



$x_2 << y_2(O_2)$. Since $O_2 \subseteq I^-_{O_1}(y_1)$ it follows of course that $y_2 << y_1(O_1)$. Continuing in this way we generate a sequence of points $\{y_i\}$ in O_1 where $y_{i+1} << y_i(O_1)$ for each i. The sequence must have a point of accumulation, say a, in $\mathcal{C}\ell(O_1)$. Now let O be any convex neighborhood of a. Eventually all y_i must fall in $I^+_O(a)$. But this brings us into contradiction with Lemma A. Given any point b in $I^+_O(a)$, the set $I^+_O(a) \cap I^-_O(b)$ will contain points y_i at which f is discontinuous. qed

C. Lemma: f is continuous at a in M iff f^{-1} is continuous at a' in M'.

Proof: Certainly $M-\mathfrak{D}$ is open in M. So f restricted to $M-\mathfrak{D}$ is a continuous map from an open subset of M into M'. It follows by the classical "invariance of domain" theorem of Brouwer that $f[M-\mathfrak{D}]$ is open and f restricted to $M-\mathfrak{D}$ is a homeomorphism. So, if $a \in M-\mathfrak{D}$, it follows that f^{-1} is continuous at a'. The converse is symmetric. qed

D. Lemma: Suppose O is as in Lemma B and in addition O is geodesically convex. Then if d is in $\mathfrak{E} \cap O$ there is a null geodesic Γ through d where $O \cap \Gamma \subseteq \mathfrak{E}$. Further, if $\bar{\Gamma}$ is any other null geodesic through d, $\bar{\Gamma} \cap O = \{d\}$.

Proof: Since $\mathfrak{E} \cap O$ is achronal in O the second claim follows from the first. By the previous lemma we know that f^{-1} is discontinuous at d'. Therefore we can find a sequence (x'_i) converging to d' and a convex open set $\bar{O}$ containing d but none of the x_i. We may as

well assume that $\mathcal{U}[\bar{O}] \subset O$ and $\mathcal{U}[\bar{O}]$ is compact. We can also find two sequences $\{b'_i\}$ and $\{c'_i\}$ converging chronologically to d', respectively, from below and above such that for all i, $b'_i << x'_i << c'_i$ (locally). Say γ'_i is a local future directed timelike curve from b'_i to x'_i to c'_i.

Now the $\{b_i\}$ and $\{c_i\}$ will eventually fall in $\bar{O}$ converging to d. For each i, γ_i (- the preimage of γ'_i -) is a future directed continuous causal curve which leaves and then reenters $\bar{O}$. Let d_1 be an accumulation point on $\partial\bar{O} \subset O$ of the γ_i as they (first) leave $\bar{O}$. We have $d < d_1$ (O) by the usual continuity argument. We also have $d_1 \varepsilon \mathcal{E}$ - for we can find points $x_i \varepsilon \gamma_i$ converging to d_1 even though the x'_i must converge to $d' \neq d_1'$. So by the achronality of $\mathcal{E} \cap O$ in O we have $d \rightarrow d_1$ (O). Similarly if d_2 is a point of accumulation on $\partial\bar{O} \subset O$ of the γ_i as they last enter $\bar{O}$, we must have $d_2 \rightarrow d$ and $d_2 \varepsilon \mathcal{D}$. Hence $d_2 \rightarrow d_1$. Furthermore the argument used to establish $d_1 \varepsilon \mathcal{D}$ will now serve to show that the entire null geodesic segment from d_2 to d_1 must fall in $\mathcal{D}$.

Let Γ arise by extending this segment as far as possible in $\mathcal{D} \cap O$. We claim that Γ must reach ∂O on both ends. If not Γ would have a terminal point $\bar{d}$ in O - since $\mathcal{D}$ is closed. We could repeat the entire argument from above with respect to $\bar{d}$ and conclude that there must be a null geodesic segment $\bar{\Gamma}$ through $\bar{d}$ falling in $\mathcal{D} \cap O$. If $\bar{\Gamma}$ did not extend Γ we would have a violation of the achronality of $\mathcal{D} \cap O$ in O. So, contrary to assumption, $\bar{d}$ could not be a terminal point of Γ. qed

The next lemma records a geometrically intuitive fact about the behavior of null geodesics in any space-time. Given a null geodesic Δ we say that a sequence of null geodesics $\{\Delta_i\}$ converges to Δ if for all $x \varepsilon \Delta$ and all open sets $\mathcal{U}$ containing x, eventually all Δ_i intersect $\mathcal{U}$.

E. Lemma: Let d be any point in M, $\{d_i\}$ a sequence converging to d, and Δ any null geodesic through d. Then there is a subsequence $\{\bar{d}_i\}$ of the $\{d_i\}$ and a sequence of null geodesics $\{\Delta_i\}$ where $\bar{d}_i \varepsilon \Delta_i$ for each i and the Δ_i converge to Δ.

Proof: The argument arises from repeated invocation of continuity considerations. We first work locally and then extend outwards. Let N be a convex coordinate neighborhood of d and let e be some point on $\Delta \cap N$ other than d where, say, $e \to d$. Further let $\mathcal{U} \subseteq N$ be any open set containing e. We claim there is an i and a corresponding point e_i in $\mathcal{U}$ where $e_i \to d_i(N)$. One way to see this is the following. Let ρ be a short timelike curve through e. Let e_1 and e_2 be points on $\rho \cap N$, respectively below and above e. Then $e_1 << d(N)$ and $e_2 \sigma d(N)$. Hence there is a d_i where $e_1 << d_i(N)$ and $e_2 \sigma d_i(N)$. Unless there is a point on ρ which is null prior to d_i relative to N, it will be possible to disconnect σ into two non-empty, disjoint, open subsets -- those points timelike and those points spacelike related to d_i.

Now let $\mathcal{U}_i$ be a nested sequence of open sets converging to e. For each i we can find an $e_i \varepsilon \mathcal{U}_i$ and a $\bar{d}_i$ (- here we are moving to a subsequence -) such that $e_i \to \bar{d}_i(0)$. Let Δ_i for each i be the maximal extension of the null geodesic segment connecting e_i and $\bar{d}_i$.

We must verify that the Δ_i converge to Δ. Let b be any point on $\Delta \cap N$ other than d. Without loss of generality we may assume that $d \to b(N)$. We can find an open subset $\bar{N} \subset N$ containing d where $\mathcal{Q}(\bar{N})$ is compact and $\mathcal{Q}(\bar{N})$ only intersects Δ at two points, one b, the other some point "below" d on Δ. [This follows simply from the fact that N is homeomorphic to an open ball in $\mathbb{R}^n$.] Now every Δ_i must leave $\bar{N}$ and the departure points will accumulate at some c on $\mathrm{Bnd}(\bar{N})$. By continuity considerations again we must have $d \to c(N)$ and $e \to c(N)$. Therefore c must fall on Δ and, as things were set up, this is only possible if $c = b$.

Thus we know that the Δ_i converge to Δ inside N. Let $\mathring{\Delta}$ be the subset of Δ consisting of points on Δ which are <u>not</u> convergence points of the Δ_i. Clearly $\mathring{\Delta}$ is open. So if $\mathring{\Delta}$ is non-empty we can find a point x on $\Delta - \mathring{\Delta}$ which is a terminal point of a segment in $\Delta - \mathring{\Delta}$. Let V be a convex neighborhood of x. We can certainly find a point y in $\Delta \cap V$ other than x where $y \in \Delta - \mathring{\Delta}$. So repeating the argument from the previous paragraph, with x and y replacing d and e, we have that all of $\Delta \cap V$ must fall in $\Delta - \mathring{\Delta}$. Thus the assumption that $\mathring{\Delta}$ is non-empty leads to a contradiction; hence the Δ_i converge to Δ everywhere on Δ. qed

Our final lemma gives the heart of the argument.

> F. <u>Lemma</u>: Suppose O is a convex neighborhood satisfying the conditions of Lemma B (and hence Lemma D). Suppose $d \in O \cap \mathcal{D}$ and Γ is the unique null geodesic through d where $O \cap \Gamma \subset \mathcal{D}$. Then for every open set U in M' containing d', Γ' (the image of $O \cap \Gamma$) is <u>not</u> achronal in U.

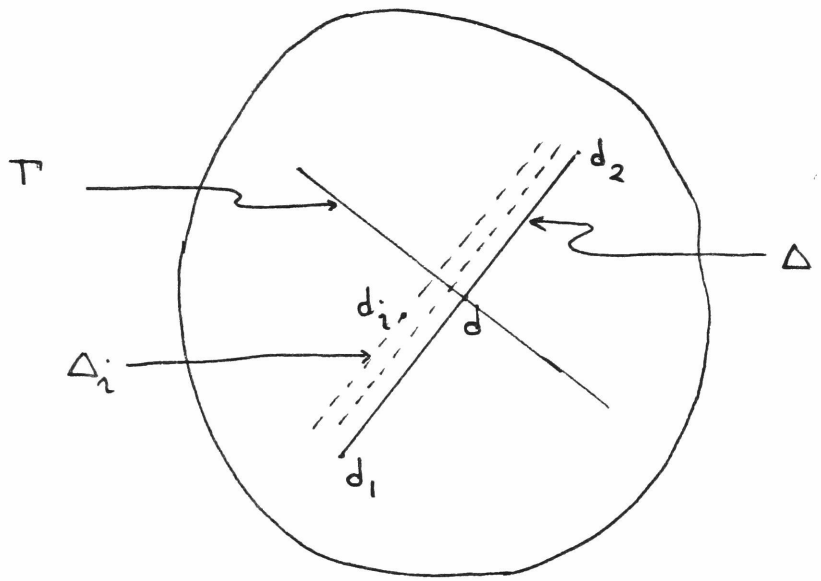
Before proving this lemma we verify that Theorem 3.28 follows from it. The argument is just a rerun <u>in</u> <u>M'</u> of Lemma B. Suppose $\mathcal{D} \neq \phi$. Then

we can find a convex neighborhood O in M satisfying the conditions of Lemma B (and hence Lemma D). Let $d \in \mathcal{D} \cap O$ and let U_1 be any convex neighborhood of d' in M' with compact closure. By hypothesis we can find points c_1, d_1 on $\Gamma \cap O$ where $d'_1 \ll c'_1(U_1)$. If $d'_1 \in U_2 \subseteq I^-_{U_1}(c'_1)$ then since $d_1 \in \mathcal{D} \cap O$ we can find points c_2, d_2 on $\Gamma \cap O$ where $d_2' \ll c_2'(U_2)$ and hence $c_2' \ll c_1'(U_1)$. Continuing in this way we generate a sequence of points c_i' in U_1, where $c_i \in \Gamma \cap O \subseteq \mathcal{D}$ and $c'_{i+1} \ll c_i'(U_1)$. These c_i' will accumulate at some c' in $U[U_1]$. The c_i' will converge chronologically to c from above. Since f^{-1} must be discontinuous at each c_i' we run into contradiction with the M'-symmetric version of Lemma A. Hence $\mathcal{D} = \phi$ and f must be a homeomorphism

Finally we prove Lemma F. Let O,d,Γ be as in the hypothesis and let U be any open set containing d'. Since $d \in \mathcal{D}$ we can find a sequence (d_i) converging to d whose image does not converge to d'. By passing to a subset of U and a subsequence of the $\{d_i\}$ we may as well assume that $d_i' \notin \bar{U}$ for all i. We also may as well take U to be convex with compact closure. Let Δ be any null geodesic through d other than Γ. We know that $\Delta \cap O \cap \mathcal{D} = \{d\}$. For each i we can find a null geodesic Δ_i through d_i where the Δ_i converge to Δ- here we are again moving to a subsequence and relabeling. Moving to a subsequence one final time we may assume that all Δ_i have the property that $\Delta_i \cap O \cap \mathcal{D}$ contains at most one point. This follows from Lemma D and the fact that $O-\mathcal{D}$ is an open set containing every point on Δ other than d, and hence must eventually intersect every Δ_i.

It will facilitate the argument to truncate Δ and the Δ_i. We can

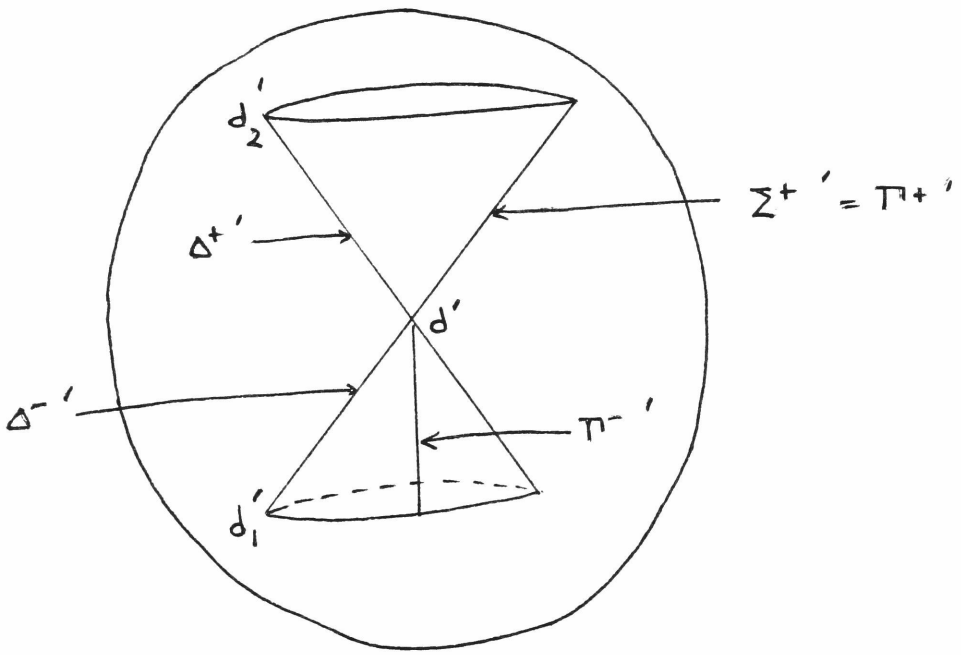
find points d_1 and d_2 on $\Delta \cap O$ where $d_1 \to d \to d_2 (O)$ and where the closed segment of Δ' (the image of Δ under f) between d'_1 and d'_2 falls in $\mathcal{U}$. By Δ we shall now understand the segment of Δ between d_1 and d_2. Correspondingly we truncate each Δ_i so that the resulting closed segments fall within O but still contain d_i. We can do so in such a way that every point on Δ is still an accumulation point of the Δ_i, but no point outside of (now shortened) Δ is one - see figure.



Notice now that Δ' is a "singly jointed" null geodesic in $\mathcal{U}$ with joint at d'. This is so because restricted to $O-\mathcal{L}$, f is a homeomorphism. [One can invoke Corollary 3.26 here, or easily check directly that homeomorphisms preserving continuous causal curves in both directions must preserve null geodesics.] Similarly the Δ'_i are singly jointed null geodesics in M'.

Consider further the situation in $\mathcal{U}$. The initial and terminal points of the Δ'_i converge, respectively, to d'_1 and d'_2 in U. By passing to a subsequence again if necessary we may

therefore assume that all the Δ_i' begin and end in U. But each must also leave U since it contains d_i'. This rapidly leads to the desired conclusion. Let Δ^- and Δ^+, Γ^- and Γ^+ be the respective upper and lower halves of Δ and Γ in O - d being the midpoint. Primes will denote their images in U. Initially, we know, the Δ_i' must converge to Δ^-'. Suppose now infinitely many of the Δ'_i leave U before they reach their "joint point." Then they will converge to the null geodesic extension of Δ^-' in U - see figure. The argument



is as in the proof of Lemma E. Let this upper extension be Σ^+'. Now Σ^+' cannot be an extension of Δ^+'. For if it were, since f^{-1} is continuous at d_2', the Δ_i on O would accumulate at points "beyond d_2" on Δ^+. We purposely truncated Δ so as to avoid this possibility. Furthermore, f^{-1} must be discontinuous along Σ^+', at least for an initial segment. Otherwise we would again be able to find accumulation points for the Δ_i off of Δ. So Σ^+, the preimage of Σ^+', must

be a future directed continuous causal curve from d where $\Sigma^+ \cap O \subseteq \mathcal{L}$. By the achronality of $\mathcal{L} \cap O$ in O it follows that $\Sigma^+ \cap O = \Gamma^+$. Hence $\Sigma^{+'}$ and $\Gamma^{+'}$ will overlap. Meanwhile f^{-1} must be discontinuous along $\Gamma^{-'}$. Since $\Gamma^{-'} \neq \Delta^{-'}$ points on $\Gamma^{-'}$ and $\Gamma^{+'}$ will be chronologically related to one another.

So our claim follows on the assumption that infinitely many of the Δ_i' leave U before they reach their joint point. A parallel argument establishes the claim if infinitely many of the Δ_i enter U (for the last time) after they reach their joint point. Clearly one assumption can be false only if the other is true. So lemma F, and with it Theorem 3.28, is proven.

We turn now to the claim from section A. As Theorem 3.28 is formulated, f and f^{-1} are assumed to preserve continuous causal curves. What if the hypothesis is changed to the assumption that they preserve (differentiable) causal curves in the sense explained in section A? The difference boils down to Lemma A since, as one can check, all the other steps in the argument are insensitive to the difference between types of causal curves.

Central to the proof of that Lemma is the argument that if $\{x_i\}$ is a sequence of locally chronologically related points in M converging to x from above, then f(x) is a point of accumulation of the $f(x_i)$. It is, in a sense, a fraction of the argument that f is continuous. Thereafter it remains to show that f(x) is a point of accumulation of the $f(x_i)$ even if the x_i converge haphazardly to x.

The idea is to generate a continuous causal curve from x which threads all (or at least infinitely many) of the x_i in reverse order so that the x_i converge to x on the curve. The desired conclusion then certainly follows since f preserves continuous causal curves. To prove the theorem while working with images of differentiable causal curves it suffices to show that one can interpolate a (differentiable) causal curve through x and infinitely many of the x_i in reverse order so that the x_i again converge to x on the curve. This one can certainly do though we skip the details of the argument.

It follows from this version of the theorem and Corollary 3.26 that:

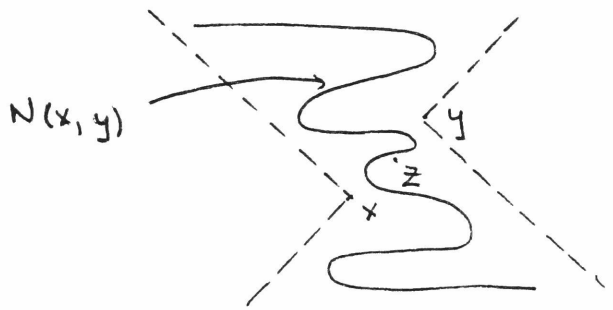
> 3.30 Theorem: If (M,g) and(M',g') are space-times and $f: M \to M'$ is a bijection where f and f^{-1} preserve (images of) future directed differentiable causal curves, then f is a smooth conformal isometry.

If in addition f and f^{-1} preserve (images of) timelike geodesics then f_*g and g' must agree up to a constant factor. This establishes the earlier claim.

Before turning to consider symmetric causal relations in the next section we pick up one small loose end. We know that the manifold topology is implicitly definable from $<$ in the presence of past and future distinguishability ($n \geq 2$) and explicitly definable from $<$ in the presence of strong causality ($n \geq 2$). There remains the question whether explicit definability is also possible under the

weaker hypothesis. This seems most unlikely for the case $n \geq 3$. The apparent dependence of our proof on the "invariance of domain" theorem speaks against the possibility. But in the special case $n = 2$ a contrived but simple definition is available. Given two points x,y in a space-time where $x \to y$ define:

$$N(x,y) = I^+(x) \cup I^+(y) \cup \{z: \ x \neq z \neq y \ \& \ x \to z \to y\}$$



It is easy to check that in dimension two $N(x,y)$ must be open and, in the presence of past and future distinguishability, the sets $I^+(a)$, $I^-(b)$, and $N(c,d)$ where $c \to d$ form a subbase for the manifold topology. This can be parlayed into an explicit first order definition of "O is open" from $<$. It will have the form of the definitions in section F of chapter II.

IV. BIBLIOGRAPHY

BIBLIOGRAPHY

1. Carter, B. "Causal Structure in Space-time," J. General Relativity and Gravitation, 1: 349-91 (1971).
2. Clarke, C.J.S. "On the Geodesic Completeness of Causal Space-times," Proc. Camb. Phil. Soc., 69: 319-24 (1971).
3. Domotor, Z. "Causal Models and Space-Time Geometries," Synthese, 24: 5-57 (1972).
4. Earman, J. "Notes on the Causal Theory of Time," Synthese, 24: 74-86 (1972).
5. Einstein, A. "Autobiographical Notes," in Albert Einstein: Philosopher-Scientist, P.A. Schilpp ed. Evanston: The Library of Living Philosophers, 1949.
6. Grünbaum, A. Philosophical Problems of Space and Time, New York: Alfred A. Knopf, 1963.
7. Hawking, S.W. "Singularities and the Geometry of Space-time," Adams Prize Essay (unpublished).
8. Hawking, S.W. and Ellis, G.F.R. The Large Scale Structure of Space-Time. Cambridge: Cambridge University Press, 1973.
9. Hicks, N.J. Notes on Differential Geometry. Princeton: Van Nostrand, 1965.
10. Kronheimer, E.H. and Penrose, R. "On the Structure of Causal Spaces," Proc. Camb. Phil Soc., 63: 481-501 (1967).
11. Lacey, H.M. "The Causal Theory of Time, a Critique of Grunbaum's Version," Philosophy of Science, 35: 322-54 (1968).
12. Latzer, R.W. "Nondirected Light Signals and the Structure of Time," Synthese, 24: 236-80 (1972).
13. Penrose, R. Techniques of Differential Topology in Relativity. Philadelphia: Society for Industrial and Applied Mathematics, 1972.
14. Reichenbach, H. Axiomatic der relativistischen Raum-Zeit-Lehre. Braunschweig: Vieweg, 1924.

15. Reichenbach, H. The Direction of Time. Los Angeles: University of California Press, 1958.
16. Robb, A.A. A Theory of Time and Space. Cambridge: Cambridge University Press, 1914 (2d edition, 1936).
17. Seifert, H.J. "Kausal Lorentzraüme," Doctoral Thesis, Hamburg University, 1968.
18. Sklar, L. Space, Time and Spacetime. Los Angeles: University of California Press, 1974.
19. van Fraassen, Bas C. An Introduction to the Philosophy of Time and Space. New York: Random House, 1970.
20. van Fraassen, Bas C. "Earman on the Causal Theory of Time," Synthese, 24: 87-95 (1972).
21. Winnie, J. "The Causal Theory of Minkowski Spacetime," forthcoming in Minnesota Studies in the Philosophy of Science, vol. 8.
22. Zeeman, E.C. "Causality Implies the Lorentz Group," J. Math. Phys., 5: 490-3 (1964).
23. Zeeman, E.C. "The Topology of Minkowski Space," Topology, 6: 161-70 (1967).

End