SECTION 2

Causality and Chronology

2.1. DEFINITION. A *trip* is a curve which is piecewise a future-oriented timelike geodesic. A trip *from* x *to* y is a trip with past endpoint x and future endpoint y. We write $x \ll y$ (read x *chronologically precedes* y) if and only if there exists a trip from x to y. Thus, the relation $x \ll y$ states the existence of points $x_0, x_1, \cdots, x_n$ with $n \geqq 1$, a timelike geodesic called a *segment* having past endpoint x_{i-1} and future endpoint x_i, for each $i = 1, \cdots, n$, where we set $x_0 = x$, $x_n = y$. Note that since the curves defined here are required to contain all their endpoints, the situation depicted in Fig. 9 (a "bad trip") in which the segments accumulate at a point p cannot occur.[1]

FIG. 9. *A "bad trip" has an infinite number of "joints" accumulating at p*

2.2. *Remark.* We shall see in 2.23 that timelike curves could equally well have been used in place of trips to define $\ll$, which would perhaps have been more physical, but trips turn out to be easier to handle mathematically. Compare [18].

Observe that in the above we could always choose $n = 1$ for $x \ll y$ in Minkowski space. On the other hand, space-times exist for which it is necessary for n to be allowed to be indefinitely large. An example (a "mutilated Minkowski space") is given in Fig. 10. A less artificial example, which shows that we need to allow $n \geqq 2$, is afforded by the anti-deSitter space (Fig. 7; see also Fig. 11).

2.3. DEFINITION. A *causal trip* is defined in the same way as a trip except that causal geodesics, *possibly degenerate*, replace the timelike geodesics of 2.1. We write $x \prec y$ (read x *causally precedes* y) if and only if there is a causal trip from x to y. See [18].

2.4. *Remark.* Note that $x \prec x$ for all $x \in M$, since degenerate causal geodesics are allowed. On the other hand, $x \ll x$ signifies the existence of a *closed trip* in M, that is, a trip whose past and future endpoints are identical. (Minkowski space, for example, possesses no closed trips.) A closed nondegenerate causal trip is signified by the existence of a pair of distinct points x, y such that $x \prec y$ and $y \prec x$.

[1] A trip with infinitely many segments is allowable of course, provided it is future- or past-endless.

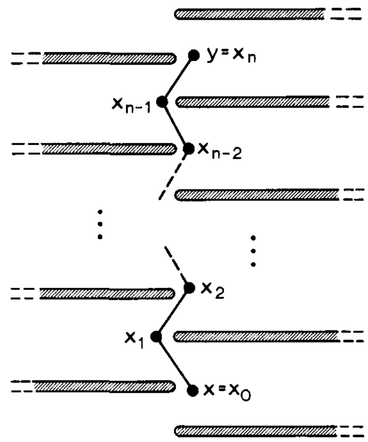

FIG. 10. *From Minkowski 2-space the half-lines $t = k$, $(-1)^k x \geqq 0$ are removed. To express the relation $x \ll y$, trips with arbitrarily large numbers of segments are required*

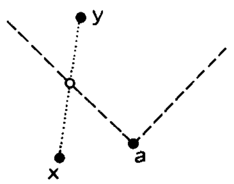

FIG. 11. *The space-time M is Minkowski space with one point removed. The set $J^+(a)$ is not closed since the null geodesic beyond the removed point, which extends that from a, is not part of $J^+(a)$, whereas it is part of $\partial J^+(a)$. (Small open circles in diagrams always denote removed points.)*

2.5. PROPOSITION.

$$a \ll b \quad \textit{implies} \quad a \prec b;$$

$$a \ll b, \quad b \ll c \quad \textit{implies} \quad a \ll c;$$

$$a \prec b, \quad b \prec c \quad \textit{implies} \quad a \prec c.$$

2.6. DEFINITION. The set $I^+(x) = \{y \in M | x \ll y\}$ is called the *chronological* (or open) *future* of x; $I^-(x) = \{y \in M | y \ll x\}$ is the *chronological past* of x; $J^+(x) = \{y \in M | x \prec y\}$ is the *causal future* of x; $J^-(x) = \{y \in M | y \prec x\}$ is the *causal past* of x. The chronological or causal future of a *set* $S \subset M$ is defined by $I^+[S] = \bigcup_{x \in S} I^+(x)$, $J^+[S] = \bigcup_{x \in S} J^+(x)$, respectively, and similarly for the pasts of S. (In general there will be a self-evident "duality" obtained by interchanging past and future in any result. The dual version of result will not normally be stated explicitly in what follows.) The slight abuse of notation $I^+[\gamma]$, etc., where γ is a trip, etc., will also be used.

2.7. *Remark.* In Minkowski space with the usual coordinates (t, x, y, z), if $a = (0, 0, 0, 0)$, then $I^+(a) = \{(t, x, y, z) | t > (x^2 + y^2 + z^2)^{1/2}\}$. Also $J^+(a)$ is the same but with "$\geqq$" replacing "$>$." Here $I^+(a)$ is an open set and $J^+(a)$ a closed set. In fact, every chronological future is open (cf. 2.9) but not all causal futures are closed. As an example of this, obtain the causal future $J^+(a)$ in Fig. 11.

2.8. PROPOSITION. *$I^+(a)$ is open for any $a \in M$.*

Proof. Let $x \in I^+(a)$; then there is a trip γ from a to x. Let $N \ni x$ be a simple region and let y be a point in N, other than x, on the terminal segment of γ. Now the vector $\exp_y^{-1}(x)$ is timelike and future-pointing (being a tangent to the terminal segment at y), and so belongs to the open set Q of timelike future-pointing vectors in $\exp_y^{-1}[N]$. Since $\exp_y$ is a homeomorphism in this neighborhood, it follows that $\exp_y Q$ is an open set in M (containing x) which lies in $I^+(y)$ and therefore in $I^+(a)$ (by 2.5), thus proving the result.

2.9. COROLLARY. *$I^+[S]$ is open, for any $S \subset M$.*

2.10. PROPOSITION. *$x \in I^+(y)$ if and only if $y \in I^-(x)$; $x \in J^+(y)$ if and only if $y \in J^-(x)$.*

2.11. PROPOSITION. *$I^+[S] = I^+[\bar{S}]$.*

Proof. If $y \gg x$, $x \in \bar{S}$, then $y \gg z$, $z \in S$ since $I^-(y)$ is open.

2.12. PROPOSITION. *$I^+[S] = I^+[I^+[S]] \subset J^+[S] = J^+[J^+[S]]$.*

Proof. This follows from 2.5, from the fact that $a \ll b$ implies the existence of c with $a \ll c \ll b$ and from the corresponding statement for $a \prec b$.

2.13. DEFINITION. Let N be a simple region and define [36], [19] the *world-function* $\Phi : N \times N \to \mathbb{R}$ by $\Phi(x, y) = g(\exp_x^{-1}(y), \exp_x^{-1}(y))$; in other words, $\Phi(x, y)$ is the squared length of the geodesic xy. Clearly $\Phi(x, y) = \Phi(y, x)$ and is positive, negative or zero according as xy is timelike, spacelike or null.

2.14. PROPOSITION. *$\Phi(x, y)$ is a continuous function of (x, y) in $N \times N$.*

Proof. See 1.11, [36], [19].

2.15. LEMMA. *The point $p \in N$ being kept fixed, the hypersurfaces $H_{p,K} = \{x | \Phi(p, x) = K\}$ are smooth in N (except at $x = p$) and are spacelike, timelike or null according as the constant K is positive, negative or zero. Furthermore, the geodesic px is normal to $H_{p,K}$ at x.*

Proof. The smoothness follows from the fact that $\exp_p$ is well-behaved in N, the equation of $H_{p,K}$ in Minkowski normal coordinates being $t^2 - x^2 - y^2 - z^2 = K$, which is smooth (except at the origin, when $K = 0$). A smooth hypersurface is said to be spacelike, timelike, or null according as its normal vectors are timelike, spacelike, or null. Let q be a point of $H_{p,K}$ and V a tangent vector to $H_{p,K}$ at q. Allowing q to vary on $H_{p,K}$ along a curve with tangent vector V, so that pq describes a 1-parameter system of a. p. geodesics of squared length K, we see that V belongs to a Jacobi field vanishing at p. Hence, by 1.16, V must be orthogonal, at q, to the direction of pq. The result follows.

2.16. LEMMA. *Let N be a simple region. Suppose $a, b, c \in \bar{N}$ are such that ab and bc are both future-causal, having distinct directions at b if both are null, or suppose a timelike curve or trip γ exists in $\bar{N}$ from a to c. Then ac is future-timelike.*

Proof. Consider $\Phi(x) = \Phi(a, x)$, as x varies from a to c along $\beta = ab \cup bc$ or along γ. As x proceeds in a future-causal direction defined by the vector T, the rate of change of Φ is measured by $T^i\nabla_i\Phi\,(= T(\Phi) = d\Phi(T) = g(g^{-1}\,d\Phi, T)) = g_{ij}T^i\nabla^j\Phi$. This, by 2.15, is nonnegative whenever ax is future-causal ($\nabla^i\Phi$, or $g^{-1}\,d\Phi$ being normal to $\Phi = \text{const.}$, i.e., to $H_{p,\Phi}$) and strictly positive unless ax

is null and T tangent to ax. (The scalar product of two future-causal vectors is nonnegative, being zero only if both are null and proportional.) Hence $\Phi(c) = \Phi(a, c) > 0$ and ac must be *future*-timelike, since $\exp_a^{-1} x$ never leaves the future component of the timelike vectors at a.

2.17. *Remark.* The proof of the lemma in 2.16 is based on 2.15 for causal $ax\,(K \geqq 0)$. Alternatively, the argument could equally well have been given using the result only for *null* $ax\,(K = 0)$. Essentially we require only the fact that the light cone $H_{a,0}$, being a null hypersurface (except at a) cannot be crossed from the inside to the outside by β or γ.

It is of some interest to note that the lemma in 2.16 is false for a $\tilde{\nabla}$ with torsion (but with $\tilde{\nabla} g = 0$ still holding). This is illustrated in Fig. 12. The light cone with respect to $\tilde{\nabla}$ is a *timelike* surface, being generated by null curves which are geodesics with respect to $\tilde{\nabla}$, but curl into the inside of the light cone with respect to ∇. Thus, β or γ can escape from inside to outside the light cone.

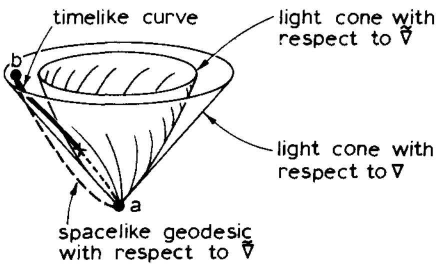

FIG. 12. *If we replace the Riemannian connection ∇ by another connection $\tilde{\nabla}$ which still preserves the metric ($\tilde{\nabla} g = 0$) but which possesses torsion, then 2.16 becomes untrue. We have a timelike curve connecting a to b, but the geodesic ab (according to $\tilde{\nabla}$) is spacelike*

2.18. PROPOSITION.

$$a \ll b, \quad b \prec c \quad \textit{implies} \quad a \ll c;$$

$$a \prec b, \quad b \ll c \quad \textit{implies} \quad a \ll c.$$

Proof. Without loss of generality, suppose $a \ll b$ and $b \prec c$. Let α be a trip from a to b and γ a causal trip from b to c. Then γ (being compact)[2] can be covered by a finite number of simple regions $N_1, \cdots, N_r$. (It is clear that we can assume that γ has no closed-loop parts, since redundant portions can be deleted.) Set $x_0 = b \in N_{i_0}$, say. Let x_1 be the future endpoint of the connected component of $\gamma \cap \bar{N}_{i_0}$ from x_0. Choose $y_1 \in N_{i_0}$ on the final segment of α, with $y_1 \neq x_0$ (see Fig. 13). Then by the lemma in 2.16, $y_1 x_1$ is future-timelike. Now, either $x_1 = c$, in which case the result is established, or $x_1 \notin N_{i_0}$, whence $x_1 \in N_{i_1}$, say. In the latter case, let x_2 be the future endpoint of the connected component of $\gamma \cap \bar{N}_{i_1}$ from x_1 and choose $y_2 \in N_{i_1}$ on $y_1 x_1$ with $y_2 \neq x_1$. Then either $x_2 = c$, in which case we are finished, or we can repeat the argument. The process must eventually terminate, since there are a finite number of connected components of the $\gamma \cap \bar{N}_i$.

[2] Cf. 1.13.

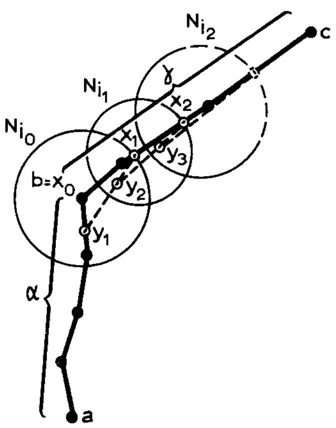

FIG. 13. *Construction for 2.18, to show that the trip α from a to b together with the causal trip γ from b to c can be replaced by a single trip from a to c*

2.19. PROPOSITION. *If α is a null geodesic from a to b, and β is a null geodesic from b to c, then either $a \ll c$ or else $\alpha \cup \beta$ constitutes a single null geodesic from a to c.*

Proof. If $\alpha \cup \beta$ fails to constitute a single geodesic, this is because the future direction of α at b does not agree with that of β at b (a "joint"). By 2.16, if x on α and y on β are sufficiently close to (but distinct from) b, then there is a timelike geodesic from x to y. Thus $a \prec x \ll y \prec c$, whence $a \ll c$ by 2.18.

2.20. PROPOSITION. *If $a \prec b$ but $a \not\ll b$, then there is a null geodesic from a to b.*

Proof. Let γ be a causal trip from a to b. If γ contains a timelike segment, then repeated application of 2.18 yields $a \ll b$. If all segments of γ are null, then repeated application of 2.19 yields $a \ll b$ unless γ is a null geodesic.

2.21. *Remark.* The relation: "$a \prec b$ but $a \not\ll b$"; is sometimes written $a \to b$ (or $a \nearrow b$) and is termed *horismos* [18], but I shall not concern myself with it explicitly here. The concepts of $\prec$, $\ll$ and $\to$ can refer to sets M more general than space-times, e.g., to a *causal space* (see Kronheimer and Penrose [18]), defined by relations $\prec$, $\ll$ on a set M subject to 2.5 and 2.18, and, in addition, to the requirements that $a \ll a$ hold for no a and that $a \prec b$, $b \prec a$ hold for no distinct pair a, b (stating the exclusion of "closed trips" or "closed causal trips").

2.22. *Remark.* The converse of 2.20 is false. (In the example illustrated in Fig. 14, there is a null geodesic from a to b, but $a \ll b$.) Observe, also, that we can have two distinct null geodesics from a to c and *not* have $a \ll c$ (cf. Fig. 14), but it is a consequence of 2.19 that any point x on the continuation of either geodesic beyond c must satisfy $a \ll x$.

2.23. PROPOSITION. *$a \ll b$ if and only if there is a timelike curve γ from a to b.*

Proof. Suppose γ exists. Cover γ with a finite number of simple regions N_i. Let $x_0 = a \in N_{i_0}$ and let x_1 be the future endpoint of the connected component of $\gamma \cap \bar{N}_{i_0}$, from x_0. Then by 2.16, x_0x_1 is future-timelike. Either $x_1 = b$, in which case $a \ll b$ as required, or else $x_1 \notin N_{i_0}$ so $x_1 \in N_{i_1}$, say. Let x_2 be the future endpoint of the connected component of $\gamma \cap \bar{N}_{i_1}$, from x_1. Then x_1x_2 is future-

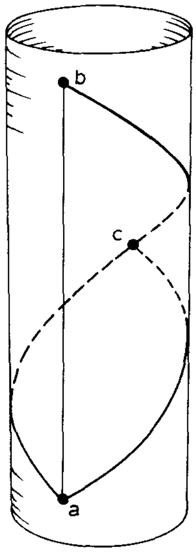

FIG. 14. *A two-dimensional "Einstein Universe" constructed by identifying* $(-1, t)$ *with* $(1, t)$ *in the strip* $-1 \leqq t \leqq 1$ *of Minkowski 2-space. Here a is* $(0, 0)$, *b is* $(0, 2)$ *and c is* $(1, 1)$. *We have* $a \ll b$ *even though b lies on a null geodesic through a*

timelike. Either $x_2 = b$, whence $a \ll b$, or else $x_2 \notin N_{i_1}$ so $x_2 \in N_{i_2}$ and the argument can be repeated. This terminates since there are a finite number of connected components of the $\gamma \cap \bar{N}_i$.

Conversely suppose $a \ll b$ and let α be a trip from a to b. I shall show that the "joints" of α can be smoothed so as to yield a timelike curve. Let μ and λ be consecutive segments of α. Let q be a point which is the future endpoint of the timelike geodesic λ and the past endpoint of the timelike geodesic μ. Consider $\exp_q^{-1}$ in some simple region $N \ni q$ and choose standard Minkowski coordinates (t, x, y, z) in $\underset{q}{T}$ so that the points of $\exp_q^{-1} \mu$ and $\exp_q^{-1} \lambda$ have coordinates of the form $(\tau, \tau \tan \chi, 0, 0)$ and $(-\tau, \tau \tan \chi, 0, 0)$, respectively, where τ varies over nonnegative values and where χ is fixed and satisfies $0 \leqq \chi < \pi/4$. Choosing $\tau_0 > 0$, connect $(-\tau_0, \tau_0 \tan \chi, 0, 0)$ to $(\tau_0, \tau_0 \tan \chi, 0, 0)$ by a C^∞ curve η in $\underset{q}{T}$ which joins on to $\exp_q^{-1} \lambda$ and $\exp_q^{-1} \mu$ smoothly (C^∞) and which is everywhere timelike according to the Minkowski metric $(dt^2 - dx^2 - dy^2 - dz^2)$ in $\underset{q}{T}$. For example, we could take η to be given by

$$R \cos\left(\frac{\theta\pi}{\pi - 2\chi}\right) = \exp\left(R^2 \sin^2\left(\frac{\theta\pi}{\pi - 2\chi}\right) - 1\right)^{-1},$$

where $t = \tau_0 R \sin \theta$, $x = \tau_0 R \cos \theta$ and $|R| \leqq 1$, $|\theta| \leqq \pi/2$. Measuring "angles" according to a "standard Euclidean metric" $dt^2 + dx^2 + dy^2 + dz^2$, we see that the slope of η is bounded away from the null cone in $\underset{q}{T}$ by an angle $\varepsilon\,(>0)$, say, where ε depends on χ but need not depend on τ_0. By choosing a small enough neighborhood of q in M we can ensure that the "error" in the slopes of the images of the null cones in M under $\exp_q^{-1}$ is less than ε. Hence, choosing τ_0 small enough, we ensure that $\exp_q \eta$ is timelike in M, thus achieving the required smoothing of the "joint" in $\lambda \cup \mu$.

2.24. *Remark.* Although 2.23 has some intrinsic interest in showing that trips and timelike curves are equivalent for defining the relation $\ll$, it will not in fact be required for any of the later results. All arguments can be carried out directly in terms of trips without any mention of smooth timelike curves.[3] On the other hand, the systematic use of timelike curves would be a little more awkward to handle since "smoothing arguments" would be required at various places (cf. 2.18 for example).

There is a similar result to 2.23 for causal trips (trivially, since by 2.20 and 2.23 a null geodesic or a timelike curve connects any two points for which $a \prec b$). However, I shall not restrict myself just to *smooth* causal curves here, since the role of a causal curve will be as a *limit* of timelike curves (or trips). A limit of a sequence of smooth curves need not be smooth. Let us therefore make the following definition which admits, under the term "causal curve," all such appropriate limits (cf. [21], [22], [6]).

2.25. Definition. A curve γ is a *causal curve* if and only if for all $a, b \in \gamma$ and for every open set Q containing the portion[4] of γ from a to b, there is a causal trip from a to b (or from b to a) lying entirely in Q.

2.26. *Remark.* Although a causal curve γ need not be smooth, there is a restriction on its "degree of wildness" imposed by the fact that it satisfies a Lipschitzian type of condition. As a consequence, γ must possess a tangent almost everywhere (remark due to R. P. Geroch), even though examples can be concocted in which γ fails to have a tangent at a set of points dense on γ.

[3] Except, strictly speaking, that given for 8.8.

[4] If the reader is concerned about a slight illogicality here, in the confusion of two notions of "curve," he may care to rephrase the statement (i.e., "the portion of γ from a to b" refers to the equivalence class of paths under parameter change, whereas to be contained in Q it must be a point set). This kind of looseness of terminology is also to be found in many other places in these notes.