What is a Singularity in General Relativity?*

ROBERT GEROCH[†,‡]

Palmer Physical Laboratory, Princeton University, Princeton, New Jersey 08540

The general covariance of relativity theory creates serious difficulties in formulating a suitable definition of a singularity in this theory. We review and, by means of an example, add to these difficulties. We examine the arguments which lead from one's intuitive picture of a singularity as "some quantity's becoming infinite" to the notion of geodesic completeness. Even within the framework of geodesic completeness there is still a great variety of definitions to choose from. It is claimed that none of these definitions is entirely satisfactory.

I. INTRODUCTION

The task of defining a singularity in general relativity is one of the most interesting—and perhaps also one of the most difficult—facets of the problem of singularities. Our intuitive idea of what a singularity should be in Einstein's theory comes from the relatively well-understood infinities which arise in other classical field theories, e.g., electrodynamics and hydrodynamics. Unfortunately, general relativity differs from these theories in one important respect: whereas in other field theories one has a background (Minkowskian) metric to which the field quantities can be referred, in general relativity the "background metric" is the very field whose singularities one wishes to describe. In view of the faulty analogy with which we must work, it is not suprising that (a) there is no widely accepted definition of a singularity in general relativity, and (b) each of the proposed definitions is subject to some inadequacy.

Our goal is to point out the serious difficulties involved in an attempt to define the term "singularity" in general relativity. In Section II we relate the line of argument which leads to geodesic completeness as the basic concept in all definitions of a singularity. These arguments are generally well known, but have not, as far as I know, been collected together and published before. In Section III we examine possible definitions based on geodesic completeness.

We shall not be concerned here with so-called "coordinate singularities". This term refers to a spacetime which has been expressed in an improper coordinate system. Thus, for example:

1. The Schwarzschild solution has a coordinate singularity at $r = 2m$ because Schwarzschild originally chose coordinates for his solution which are not applicable on this surface.

2. The solar system has a coordinate singularity on a line extending from the center of the sun to a point in the constellation Draco since planetary orbits are sometimes calculated in a spherical polar coordinate system with this line as axis.

The presence or absence of a coordinate singularity is not a property of the spacetime itself, but rather of the physicist who has chosen the coordinates by which the spacetime is described.

II. DIALOGUE ON SINGULARITIES

Sagredo: It should not be too difficult, Salviati, to discover a suitable definition of a singularity by analogy with electrodynamics. We know that there are solutions of Maxwell's equations in which the electromagnetic field becomes infinite, that is, undefined, at some points. It is this feature that characterizes electromagnetic singularities. By direct analogy, we may define a spacetime in general relativity as having a singularity if it contains points at which the metric tensor is not defined.

Salviati: But in general relativity the metric tensor may reside on whatever manifold we wish. In particular, given a spacetime which, according to the definition above, has a singularity, we simply remove from the manifold those regions in which the metric is undefined. The resulting spacetime has a metric defined everywhere, and so, by your very definition, must be nonsingular. Thus, by merely cutting the "real singularities" from the spacetime, one would be forced to admit the Schwarzschild and Friedmann solutions among the "nonsingular" spacetimes. Surely we should not wish to classify these two solutions as without singularity.

Sagredo: Let us then define a spacetime as having a singularity if either the metric is undefined somewhere, or if any regions have been removed from the spacetime.

Salviati: In electrodynamics, it is clear what is meant by "no regions have been removed from the spacetime". The Minkowski background metric allows us

to define the entire spacetime without reference to an electromagnetic field which might later be placed on it. But the situation is very different in general relativity. Here, we have no "background metric" which specifies the spacetime before the field of interest is set down. There is no obvious intrinsic way to tell whether or not "regions have been removed from the spacetime." That is, our definition of a singularity in general relativity must contain within it a technique to determine whether or not "real singularities" have been concealed by merely cutting them from the spacetime manifold.

Sagredo: I cannot help but feel, Salviati, that you have raised the essential point about the definition of a singularity. Perhaps a further analogy with electrodynamics will help us again. In the static coulomb solution, one would know of the presence of the singularity at the spatial origin even if the origin were removed from the spacetime, for the electric field becomes infinite as we approach the singularity. One might therefore define a spacetime in general relativity as having a singularity if the metric $g_{\alpha\beta}$ becomes infinite anywhere.

Salviati: But the $g_{\alpha\beta}$, being the components of a tensor, can be made as large or as small as we desire by a proper choice of the coordinate system. For example, the space:

$$ds^2 = -(1/t)^2\, dt^2 + dx^2 + dy^2 + dz^2,$$

defined for $t > 0$, would appear to have a singularity at $t = 0$, for the "metric becomes infinite" as we approach the "surface ($t = 0$) which has been removed from the spacetime." Yet this example is, of course, Minkowski space, which we should not like to describe as having a singularity.

Sagredo: Your point is well taken. In order to detect the presence of a singularity, we must inquire into the behavior of scalars rather than the components of a tensor. Let us consider, then, the physical components[1] of the Riemann tensor. In the first place, it is the Riemann tensor that determines many observable effects, for example, the deviation of geodesics. Secondly, the physical components of a tensor are scalars, and so the problem of coordinate transformations you raised a moment ago will be avoided. Let us call a spacetime nonsingular if the physical components of the Riemann tensor remain finite.

Salviati: Unfortunately, Sagredo, even the most reasonable spacetime can be made to appear singular by an appropriate choice of the frame in which the

---

[1] The physical components of any tensor field are defined as the components of that field along any orthonormal tetrad. For example, if $\xi^\alpha_{(i)}$, $i = 1, 2, 3, 4$, are any four orthonormal vector fields, the physical components of a tensor $A_{\alpha\beta\gamma}$ in the frame determined by the $\xi^\alpha_{(i)}$ are the sixty-four scalar fields $A_{(i)(j)(k)} \equiv A_{\alpha\beta\gamma}\xi^\alpha_{(i)}\xi^\beta_{(j)}\xi^\gamma_{(k)}$.

physical components are to be described. In particular, in any spacetime which is not of constant curvature, one can always choose a frame in which the physical components of the Riemann tensor become infinite on approaching any point.

Sagredo: It appears that we are left with only the scalar invariants to define singularities. Fortunately, an infinite number of scalar invariants can be derived from any given metric tensor (*1*). These might be interpreted as describing, in some sense, the curvature of the spacetime. In particular, if a scalar invariant becomes infinite, then this corresponds to infinite curvature. Let us say, then, that a spacetime has a singularity if any of the scalar invariants become infinite.

Salviati: Of course, each scalar invariant is a function of the point of the manifold. We must therefore ask of your definition: "scalar invariant approaches infinity as the manifold point approaches what?" One would like to answer: "as the point on the manifold approaches a singularity which has been cut out of the manifold". But as yet we have no prescription for determining which curves approach "singularities", and which curves trail off harmlessly to "infinity". Of course, should a scalar invariant become infinite along a curve of the latter type, we should not wish to call this circumstance a singularity. Our next task, it appears, is to make a distinction between the two types of curves.

Sagredo: One possible way to make the distinction you mention would be as follows: a curve which approaches "infinity" might be expected to have infinite total length, while a curve approaching a "singularity removed from the spacetime" would have a finite total length.

Salviati: Were the spacetime metric positive-definite, I would be inclined to agree with you. But since we deal with an indefinite metric, any curve in space-time may be approximated as closely as desired by a curve of arbitrarily small total length. The total length of a curve is of no help in deciding whether or not that curve should be construed as approaching "infinity".

Sagredo: There is another concept we may use, even with an indefinite metric, to replace the notion of "distance." We know that in Minkowski space each geodesic has the property that an affine parameter on that geodesic attains arbitrarily large values. This fact expresses the idea that each geodesic in Minkowski space goes off to infinity. On the other hand, if we remove some region R from Minkowski space, those geodesics which formerly entered R will now have only a finite affine length in the new space. Thus, affine length seems the perfect candidate to replace the notion of distance when the metric is indefinite. We may say that a spacetime has a singularity if there is a geodesic of finite total affine length along which a scalar invariant becomes infinite.

Salviati: I think, Sagredo, that you have raised a most important point. It does seem that incomplete geodesics are a signal that we have "removed a region from the spacetime." However, I am afraid that I'm now no longer clear as to what the scalar invariants have to do with singularities. I recall that Kundt (2) has pointed out that in a spacetime with indefinite metric there are two types of scalar invariants, which we may call type I and type II. Type I scalar invariants are formed by taking outer products of the Riemann tensor and its derivatives, and then contracting indices. Each type II invariant, on the other hand, is the "ratio" between two tensors (obtained from the Riemann tensor) which differ only by a factor. Now both types of invariants together are required to characterize the curvature properties of the spacetime. (For example, in the plane wave solutions all the type I invariants vanish, yet the Riemann tensor is not zero.) However, a type II invariant may become infinite in a region of spacetime in which the metric is defined and differentiable. In short, with an indefinite metric, we cannot find a collection of scalar invariants which both characterize the curvature of the spacetime and remain finite when the metric is regular.

Sagredo: Let us revise our definition, then, in the following way: a spacetime is singularity-free if it is geodesically complete. With this definition we would, of course, include within the class of singular spaces a number that result from simply removing some region from an otherwise regular space. (For example, Minkowski space with the origin removed would be singular.) But perhaps an over-inclusive definition is not too high a price to pay for the satisfaction of knowing that all spacetimes from which "real singularities" have been removed would be included in the class of singular spacetimes.

Salviati: Very well.

III. DEFINITIONS BASED ON GEODESIC COMPLETENESS

We now consider the formulation of a definition of a singularity in terms of geodesic completeness.

It is convenient to define a *half-geodesic* as a geodesic curve which has one endpoint and which has been extended as far as possible in some direction from that endpoint. A spacetime is called *timelike* (respectively, *null, spacelike*) *complete* if an affine parameter on every timelike (respectively, null, spacelike) half-geodesic assumes arbitrarily large values.

Let M be a spacetime which is complete in all three senses above, and let C be a closed subset of M. Then the spacetime M–C is timelike, null, and spacelike incomplete, for there are geodesics of each type that pass from M–C into C. One might hope, therefore, that a spacetime complete in any of the three senses is complete also in the other two.

Unfortunately, the three types of completeness are not equivalent (*3*). We illustrate this point by an example. Consider two-dimensional Minkowski space with metric $\eta_{\alpha\beta}$, written in the usual x, t coordinates. Define a new metric $g_{\alpha\beta} = \varphi\eta_{\alpha\beta}$ where the positive scalar field φ has the following three properties (Fig. 1):

1. $\varphi = 1$ outside of the region between the vertical lines $x = +1$ and $x = -1$;

2. φ is symmetric about the line $x = 0$, that is, $\varphi(t, x) = \varphi(t, -x)$;

3.   On the t-axis, φ goes to zero sufficiently quickly as $t \to \infty$ (say, as t^{-6}).

By condition 2, the t-axis is a timelike geodesic. By condition 3, that geodesic has finite proper (and therefore affine) length. The metric $g_{\alpha\beta}$ is therefore timelike incomplete. However, every spacelike or null geodesic which enters the region between the lines $x = +1$ and $x = -1$ must eventually leave and remain outside of this region. Therefore, by condition 1, $g_{\alpha\beta}$ is null and spacelike complete.

Slight modifications of this example, together with the examples of Kundt (*3*) provide spacetimes which fall into each of the following five categories:
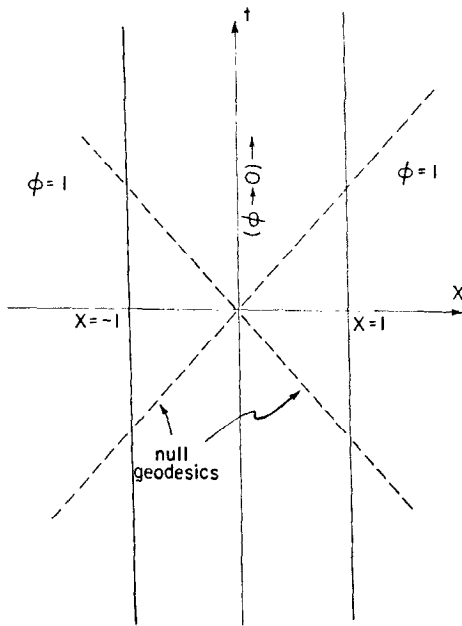

FIG. 1.   An example of a 2-dimensional spacetime which is timelike incomplete, but spacelike and null complete.

1.   timelike complete, spacelike and null incomplete;

2.   spacelike complete, timelike and null incomplete;

3.   null complete, timelike and spacelike incomplete;

4.   timelike and null complete, spacelike incomplete;

5.   spacelike and null complete, timelike incomplete.

No example is known of a spacetime which is spacelike and timelike complete but null incomplete. On the other hand, no proof has been given, as far as I know, of the nonexistence of such a spacetime.

We originally introduced geodesic completeness because this concept appeared to give a precise statement of

PROPERTY 1. In a nonsingular spacetime, one should like to be sure that "no regions have been deleted from the spacetime manifold".

We now find that timelike, spacelike, and null completeness are not equivalent.[2] Property 1 offers no guide as to which type of completeness should be selected in formulating the definition of a singularity. However, in a timelike incomplete spacetime, there are freely falling observers whose total proper time is finite. This fact suggests that we break the symmetry between the three types of completeness by introducing

PROPERTY 2. In a nonsingular spacetime, observers who follow "reasonable" (in some sense) world lines should have an infinite total proper time.

(We could not have required that *all* timelike world lines have infinite total length, for this property does not obtain in any spacetime.) Properties 1 and 2 have been stated in vague terms because they express vague concepts. Our task is to combine the two properties, using geodesic completeness, into a suitable definition of a singularity.

First of all, there is a serious problem involved in the use of geodesic completeness to describe property 1. It has been emphasized by Misner (4) that there are compact, geodesically incomplete, spacetimes. Now, it is not possible that a compact spacetime has resulted from the removal of a region from a larger spacetime, for no compact 4-manifold can be a proper submanifold of another connected (Hausforff) 4-manifold. Thus, there are spacetimes which are geodesically incomplete, but which satisfy a liberal interpretation of property 1. Geodesic completeness does not appear to describe property 1. This dilemma has led Shepley (5) and Misner (4) to propose the following definition.

_____

[2] The distinction between the three types of completeness is important in practice. For example, the Reissner-Nördstrom solution is timelike complete but neither spacelike nor null complete. Shall one say that the Reissner–Nördstrom solution has a singularity?

DEFINITION 1. A spacetime is nonsingular if every half-geodesic is either complete or else is contained in a compact set.

According to Definition 1, every compact spacetime is nonsingular. Perhaps the principal objection to Definition 1 is that it represents but one of several possibilities. Consider, for example,

DEFINITION 2. A spacetime is nonsingular if every half-geodesic $\gamma(\lambda)$, where λ is an affine parameter, is either complete or, if incomplete (say, with $0 \leqslant \lambda < \lambda_0$), has the property that for some compact set C and for every $\lambda' < \lambda_0$ there is a $\lambda \in (\lambda', \lambda_0)$ with $\gamma(\lambda) \in C$.

The two definitions are illustrated in Fig. 2. Every spacetime having a singularity according to Definition 2 also has a singularity according to Definition 1, but the converse is, presumably, not true. A number of other possible definitions could be offered. On what basis was Definition 1 selected from among the alternatives?
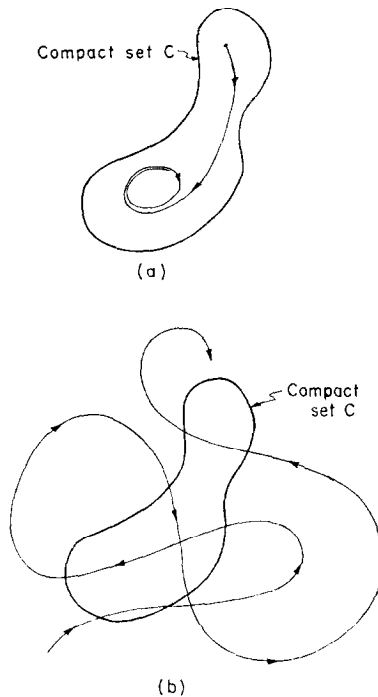

FIG. 2. Two possible definitions of a singularity. Definition 1 (a): Incomplete half-geodesics which are contained in a compact set are counted as complete. Definition 2 (b): Incomplete half-geodesics which continually reenter a compact set are counted as complete.

In view of the problems involved in expressing property 1 in terms of geodesic completeness, one is led to rely more heavily on property 2. In fact, property 2 cannot adequately be described by geodesic completeness either. In the Appendix we give an example of a spacetime which is geodesically complete in all three senses, but which contains a timelike curve γ with bounded acceleration and finite total proper length. The properties of γ imply that one could construct a space ship with a rocket of finite thrust (the acceleration of γ is bounded), using only a finite amount of fuel (the proper length of γ is finite), which would traverse the world line γ. An individual inside the rocket ship, however, has only a finite amount of time: after that he is no longer represented by a point on the spacetime manifold. However, the spacetime is geodesically complete: there is no possibility of extending the spacetime to include more points for our doomed observer to occupy. This example, though geodesically complete, does not satisfy even a very moderate interpretation of property 2.

Geodesic completeness, which at first appeared to be ideally suited to formulating a definition of a singularity, turns out to lead to several difficult problems. It is not clear what is the best direction to turn to find a suitable definition of a singularity.[3]

Finally, we remark that geodesic incompleteness (various combinations of the three types) is commonly used as a definition of a singularity because with such a definition one can show that large classes of solutions of Einstein's equations are singular (7).

APPENDIX

We give an example of a spacetime which is spacelike, timelike, and null geodesically complete, but which contains a timelike curve of bounded acceleration and finite total length.

Consider first the universal covering space of two-dimensional anti-de Sitter space (8). The metric may be written in the form:

$$ds^2 = -(1 + x^2)\, dt^2 + (1 + x^2)^{-1}\, dx^2.$$

Several timelike and null geodesics in this space are illustrated in Fig. 3. Note that each timelike geodesic which intersects the t axis at the point $t = t_0$ intersects that axis again at $t = t_0 + \pi$. In our example, we shall exploit this "focusing effect" on the timelike geodesics.

We now define a two-dimensional spacetime which depends on two continuous positive numbers, b and c, and on a positive integer n. This "$(b, c; n)$-space" is illustrated in Fig. 4. The metric in each of the four regions of Fig. 4 is given as follows.
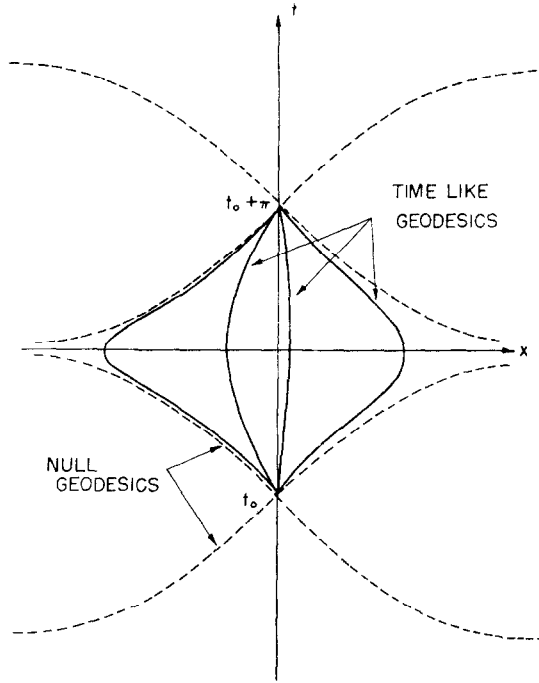
FIG. 3. Typical geodesics in anti-de Sitter space.

Region A:

$$ds^2 = \left(\frac{1}{2}\right)^{2n} [-(x^2 + 1) \, dt^2 + (x^2 + 1)^{-1} \, dx^2].$$

That is, the metric of region A is the metric of anti-de Sitter space multiplied by the constant factor $\left(\frac{1}{2}\right)^{2n}$.

Region B:

$$ds^2 = f\left(\frac{t}{b} - 3\right)\left(\frac{1}{2}\right)^{2n} [-(x^2 + 1) \, dt^2 + (x^2 + 1)^{-1} \, dx^2]$$

$$+ \left[1 - f\left(\frac{t}{b} - 3\right)\right]\left(\frac{1}{2}\right)^{2n} [-dt^2 + dx^2],$$

where we have defined the C^∞ function (Fig. 5):

$$f(z) = \begin{cases} 0, & z \leqslant 0 \\ \dfrac{\int_0^z \exp[-(y^{-2} + (y-1)^{-2})]\,dy}{\int_0^1 \exp[-(y^{-2} + (y-1)^{-2})]\,dy}, & 0 < z < 1, \\ 1, & z \geqslant 1. \end{cases}$$

We see that the metric in B joins smoothly with that in A across the line $t = 4b$, and approaches that of Minkowski space (multiplied by the constant factor $(\tfrac{1}{2})^{2n}$) as $t \to 3b$.

Region C: The metric in region C is to be conformal with Minkowski 2-space, the C^∞ conformal factor φ having the properties:

1. $\varphi(x, t) = \varphi(-x, t) = \varphi(x, -t)$;
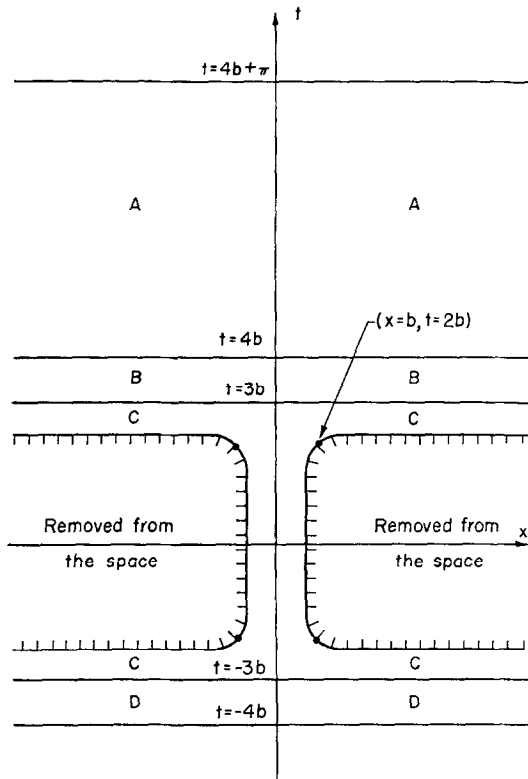

FIG. 4. $(b, c; n)$-space with its four regions, A, B, C, and D, shown.

2.   on the lines $t = 3b$ and $t = -3b$, $\varphi = (\frac{1}{2})^{2n}$ and all the derivitives of φ vanish;

3. $\varphi = (\frac{1}{2})^{2n}$ on the t-axis;

4. $\varphi \to \infty$ sufficiently quickly as we approach the two areas cut out of region C so that all geodesics which approach these areas do so in an infinite affine length.

By condition 2, the metric in region C joins smoothly with that in region B across the line $t = 3b$.

Region D:

$$ds^2 = f\left(-\frac{t}{b} - 3\right)\left(\frac{1}{2}\right)^{2n-2} [-((x - c)^2 + 1)\, dt^2 + ((x - c)^2 + 1)^{-1}\, dx^2]$$

$$+ \left[1 - f\left(-\frac{t}{b} - 3\right)\right]\left(\frac{1}{2}\right)^{2n} [-dt^2 + dx^2].$$

As $t \to -3b$, the metric in D approaches that of Minkowski space (multiplied by the constant factor $(\frac{1}{2})^{2n}$), thus joining smoothly with region C. As $t \to -4b$, the metric in D becomes that of anti-de Sitter space, multiplied by the factor $(\frac{1}{2})^{2n-2}$ and shifted by an amount c in the x-direction.
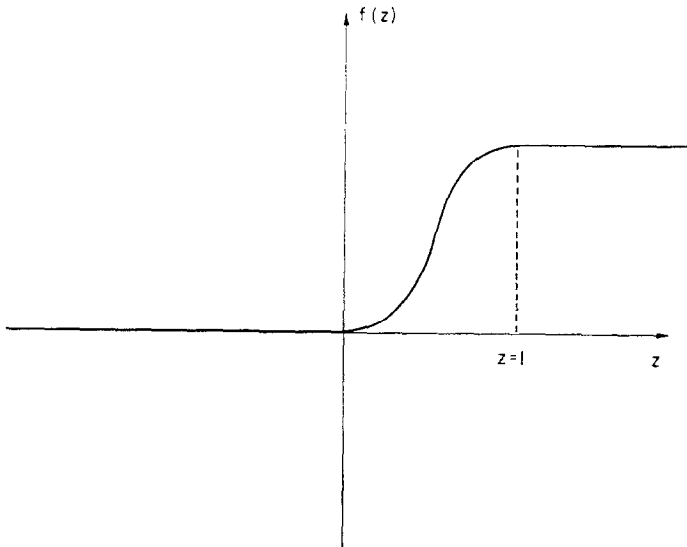

FIG. 5.   The C^∞ function $f(z)$.

Our $(b, c; n)$-space has been constructed so as to have the following three properties:

P_1. Every timelike geodesic which passes through region D either fails to reach the line $t = 4b$, or else crosses that line at a value of x in the interval $|x| < \tan 4b$. In the latter case, our geodesic strikes the line $t = 4b + \pi$ also in the interval $|x| < \tan 4b$.

P_2. If $c < 1$, then every timelike geodesic which intersects the line $t = -4b$ with a value of x less than $-4b$ fails to reach region B.

P_3. Consider the C^∞ curve γ consisting of two parts:

γ_1: $x = 0$ for $-4b \leqslant t \leqslant 4b$.

γ_2: $x = (\tan 4b + 4b) f((t - 4b)/\pi)$ for $4b < t < 4b + \pi$.

The curve γ_1 has a total length less than $9b$, and, if $c < 0.5$, an acceleration everywhere less than its acceleration at $t = -4b$. The curve γ_2 has total length less than $\pi(\frac{1}{2})^n$, and an acceleration everywhere less than $2^n a(b)$, where $a(b)$ is some strictly increasing continuous function of b which vanishes when $b = 0$.

We are now prepared to construct our example. Set $b_1 = 0.01$, $c_1 = 0$. Choose any sequence of positive numbers b_n, $n = 2, 3,...$, having the properties

P_4. $a(b_{n+1}) \leqslant \frac{1}{2}a(b_n)$

P_5. $b_{n+1} \leqslant \frac{1}{2}b_n$.

Set $c_n = \tan(4b_{n-1}) + 4b_{n-1}$ for $n = 2, 3,...$. Consider the following sequence of 2-dimensional spacetimes:

M_0:   The lower half ($t < 0$) of anti-de Sitter space.

M_n: $(b_n, c_n; n)$-space, $n = 1, 2,...$.

We connect all of these spaces together in the following way. Identify the lower edge of M_n with the upper edge of M_{n-1} according to the rule: the point $(x, t = -4b_n)$ of M_n is identified with the point $(x + c_n, t = 4b_{n-1} + \pi)$ of M_{n-1}. In the case of M_0, the point $(x, t = -4b_1)$ of M_1 is identified with the point $(x, t = 0)$ of M_0. We thus obtain a C^∞ spacetime M as shown in Fig. 6.

We note that M has the following two properties:

1. M is geodesically complete. We have chosen the conformal factor in each region C so that any geodesic which approaches one of the "cut-out" areas of Fig. 6 has infinite affine length. Anti-de Sitter space is geodesically complete. A geodesic may be incomplete, therefore, only if it can manage to pass from one M_n to the next, taking advantage of the decreasing conformal factor in each space to escape having infinite affine length. No spacelike or null geodesic can
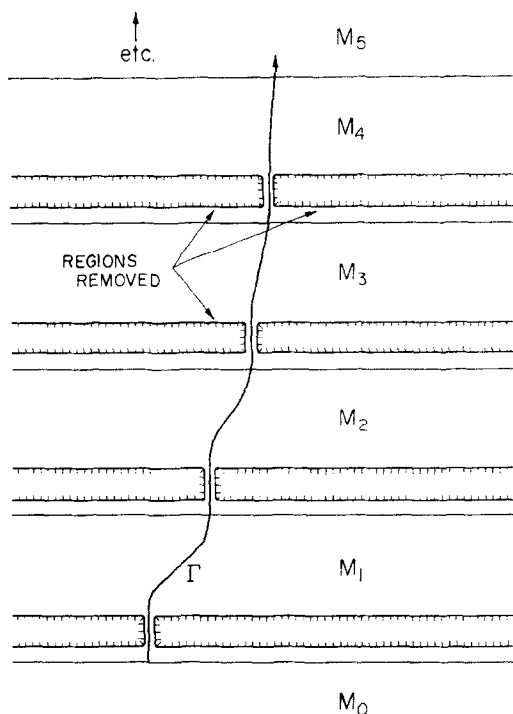
FIG. 6. The geodesically complete two-dimensional spacetime M. The timelike curve Γ has finite length and bounded acceleration.

pass through the narrow channel between the "cut out" areas of each region C (Fig. 4). Therefore, M is spacelike and null complete. Any timelike geodesic which passes through one channel must, according to property P_1, intersect the line $t = 4b_n + \pi$ in that M_n in the interval $|x| < \tan 4b_n$. But according to property P_2 and the definition of the c_n in terms of the b_n, this geodesic will not be unable to pass through the next channel. That is, no timelike geodesic in M is able to pass through more than one channel. We conclude that M is also timelike complete.

2. M contains a timelike curve of bounded acceleration and finite total length. Draw the curve γ in each of the spaces M_n ($n \geqslant 1$). These curves form a C^∞ curve Γ in M (Fig. 6). The total length of Γ may be written, according to property P_3 :

$$\text{length } \Gamma \leqslant \sum_{n=1}^{\infty} \left[9b_n + \pi \left(\frac{1}{2}\right)^n \right].$$

Because of property P_5, this length is finite. Further, the maximal acceleration

of the curve $\varGamma$ in the space M_n is $2^n a(b_n)$ (property P_3). Because of property P_4, this acceleration is bounded.

Our example may now be made into a four-dimensional spacetime by crossing M with the Euclidean plane.

Note that M is diffeomorphic to R^2. Thus, the entire spacetime M could be expressed in terms of a single coordinate patch with coordinates x, t. Aside from the presence in M of an inextendable timelike curve of bounded acceleration and finite total length, the geodesically complete spacetime M has no particularly pathological properties.

RECEIVED: July 18, 1967

REFERENCES

1. C. N. HASKINS, *Trans. Am. Math. Soc.* 3, 71 (1902).
2. W. KUNDT, *in* "Recent Developments in General Relativity," p. 307. Pergamon Press, New York, 1962; W. KUNDT, *in* "Les Theories Relativistes de la Gravitation," pp. 155. Centre National de la Recherche Scientifique, Paris, 1962.
3. W. KUNDT, *Z. Physik 172*, 488 (1963).
4. C. W. MISNER, *J. Math. Phys.* 4, 924 (1963).
5. L. SHEPLEY, Unpublished Doctoral dissertation, Department of Physics, Princeton University (1965).
6. A. AVEZ, *Compt. Rend. Acad. Sci. Paris* 240, 485 (1955).
7. R. PENROSE, *Phys. Rev. Letters* 14, 57 (1965); S. W. HAWKING, *ibid.* 17, 444 (1966); *Proc. Roy. Soc. (London)* 294, 511 (1966); *ibid.* 295A, 490 (1966); R. GEROCH, *Phys. Rev. Letters* 17, 445 (1966).
8. E. CALABI AND L. MARKUS, *Ann. Math.* 75, 63 (1962).