Eur. Phys. J. H
DOI: 10.1140/epjh/e2014-50013-6

THE EUROPEAN PHYSICAL JOURNAL H

Historical Document

Singularities and the geometry of spacetime*

Stephen Hawking

Gonville and Caius College, Cambridge, UK

Received 17 February 2014 / Received in final form 23 June 2014
Published online 10 November 2014

Abstract. The aim of this essay is to investigate certain aspects of the geometry of the spacetime manifold in the General Theory of Relativity with particular reference to the occurrence of singularities in cosmological solutions and their relation with other global properties. Section 2 gives a brief outline of Riemannian geometry. In Section 3, the General Theory of Relativity is presented in the form of two postulates and two requirements which are common to it and to the Special Theory of Relativity, and a third requirement, the Einstein field equations, which distinguish it from the Special Theory. There does not seem to be any alternative set of field equations which would not have some undeseriable features. Some exact solutions are described. In Section 4, the physical significance of curvature is investigated using the deviation equation for timelike and null curves. The Riemann tensor is decomposed into the Ricci tensor which represents the gravitational effect at a point of matter at that point and the Welyl tensor which represents the effect at a point of gravitational radiation and matter at other points. The two tensors are related by the Bianchi identities which are presented in a form analogous to the Maxwell equations. Some lemmas are given for the occurrence of conjugate points on timelike and null geodesics and their relation with the variation of timelike and null curves is established. Section 5 is concerned with properties of causal relations between points of spacetime. It is shown that these could be used to determine physically the manifold structure of spacetime if the strong causality assumption held. The concepts of a null horizon and a partial Cauchy surface are introduced and are used to prove a number of lemmas relating to the existence of a timelike curve of maximum length between two sets. In Section 6, the definition of a singularity of spacetime is given in terms of geodesic incompleteness. The various energy assumptions needed to prove the occurrence of singularities are discussed and then a number of theorems are presented which prove the occurrence

* The manuscript submitted to the adjudicators of the Adams Prize was typewritten with mathematical symbols and formulae inserted by hand. The handmade figures of the original are reproduced in the transcription. The essay is here published for the first time. See the accompanying paper by George Ellis [Ellis 2014] for more background information concerning the science history and content of this essay.

of singularities in most cosmological solutions. A procedure is given which could be used to describe and classify the singularites and their expected nature is discussed. Sections 2 and 3 are reviews of standard work. In Section 4, the deviation equation is standard but the matrix method used to analyse it is the author's own as is the decomposition given of the Bianchi identities (this was also obtained independently by Trümper). Variation of curves and conjugate points are standard in a positive-definite metric but this seems to be the first full account for timelike and null curves in a Lorentz metric. Except where otherwise indicated in the text, Sections 5 and 6 are the work of the author who, however, apologises if through ignorance or inadvertance he has failed to make acknowledgements where due. Some of this work has been described in [Hawking S.W. 1965b. Occurrence of singularities in open universes. *Phys. Rev. Lett.* **15** : 689–690 ; Hawking S.W. and G.F.R. Ellis. 1965c. Singularities in homogeneous world models. *Phys. Rev. Lett.* **17** : 246–247 ; Hawking S.W. 1966a. Singularities in the universe. *Phys. Rev. Lett.* **17** : 444–445 ; Hawking S.W. 1966c. The occurrence of singularities in cosmology. *Proc. Roy. Soc. Lond. A* **294** : 511–521]. Undoubtedly, the most important results are the theorems in Section 6 on the occurrence of singularities. These seem to imply either that the General Theory of Relativity breaks down or that there could be particles whose histories did not exist before (or after) a certain time. The author's own opinion is that the theory probably does break down, but only when quantum gravitational effects become important. This would not be expected to happen until the radius of curvature of spacetime became about 10^{-14} cm.

1 Preface

By comparison with the study of positive-definite metrics, that of Lorentz metrics has largely been neglected by pure mathematicians. The reasons for this seem to be, first, that many of the techniques used for positive-definite metrics fail when applied to Lorentz metrics, and second, that there is a feeling that such metrics are less natural and of not such interest. However, there is a Lorentz metric of great interest to physicists: namely, the metric of spacetime in the General Theory of Relativity. Thus a considerable amount of work has been done on the local properties of this metric. However, so far there has been little investigation of global properties.

This essay is intended as a small contribution to such an investigation. The principal tools employed are the variation of curves (developed in Sect. 4) and the concept of a null horizon (introduced in Sect. 5). These could probably be used for a number of global problems, but the one to which they are applied, namely singularities, seems to be that with the greatest physical interest.

While I hope that this essay contains no major errors, I am not so optimistic as to expect that there are no minor ones and I would ask the reader's indulgence for these. For various reasons, it was necessary to use a duplicating process and to put in the equations on the stencils by hand. This may result in them being not very legible in some places.

I am deeply indebted to Roger Penrose whose work introduced me to the problem of singularities in spacetime. I would like to thank Brandon Carter, George Ellis, and Denis Sciama, with whom I had many fruitful discussions, and Jill Powell, who did the typing. Above all I am grateful to my wife, without whose encouragement and help this essay would not have been written.

13 December 1966 Stephen Hawking

2 An outline of Riemannian geometry

2.1 Manifolds

Essentially, a manifold is a generalisation of Euclidean space. Let $\mathbb{R}^n$ denote Euclidean space of n dimensions, that is, the set of all n-tuples $(u^1, u^2, \ldots, u^n)$ with the usual topology. A map ϕ of an open set $O \subset \mathbb{R}^n$ to an open set $\bar{O} \subset \mathbb{R}^m$ is said to be of class C^r if the coordinates $(\bar{u}^1, \bar{u}^2, \ldots, \bar{u}^m)$ of $\phi(p)$ in $\bar{O}$ are r times continuously differentiable functions of the coordinates $(u^1, u^2, \ldots, u^n)$ of p in O. A map ϕ from a set $P \subset \mathbb{R}^n$ to a set $\bar{P} \subset \mathbb{R}^m$ is said to be C^r if ϕ is the restriction to P and $\bar{P}$ of a C^r map from an open set O containing P to an open set $\bar{O}$ containing $\bar{P}$.

Let $\mathbb{R}^{n+}$ denote the region of $\mathbb{R}^n$ for which $u^1 \geq 0$. Then an n dimensional C^r manifold (with boundary) is defined as a set M and an atlas (or differential structure) $\{U_\alpha, \phi_\alpha\}$, where U_α are subsets of M with $\cup_\alpha U_\alpha = M$ and each ϕ_α is a bijection (one-to-one correspondence) of the corresponding U_α to an open subset of $\mathbb{R}^n$ or of $\mathbb{R}^{n+}$ such that, if $U_\alpha \cap U_\beta$ is nonempty, then

$$\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \longrightarrow \phi_\alpha(U_\alpha \cap U_\beta)$$

is a C^r map of an open subset of $\mathbb{R}^n$ to an open subset of $\mathbb{R}^n$ or of an open subset of $\mathbb{R}^{n+}$ to an open subset of $\mathbb{R}^{n+}$. Another atlas is said to be compatible with the given atlas if their union is a C^r atlas for M. The atlas consisting of all the atlases compatible with the given atlas is called the complete atlas of the manifold. The topology of the manifold is defined to be that given by the basis consisting of all the subsets of the complete atlas. That is to say, it is the coarsest topology in which all these subsets are open.

The boundary of M, denoted by ∂M, is defined to be the set of all points of M whose image under a bijection ϕ_α lies on the boundary of $\mathbb{R}^{n+}$ in $\mathbb{R}^n$. Clearly, ∂M is an $(n-1)$ dimensional C^r manifold whose boundary is empty.

Let M be a C^r manifold with atlas $\{U_\alpha, \phi_\alpha\}$ and N a C^s manifold with atlas $\{V_\alpha, \psi_\alpha\}$. Then a map $\mu : M \to N$ is said to be $C^t (t \leq r, s)$, if for every nonempty $\mu(U_\alpha) \cap V_\beta$,

$$\psi_\beta \circ \mu \circ \phi_\alpha^{-1} : \phi_\alpha\big(U_\alpha \cap \mu^{-1}(V_\beta)\big) \longrightarrow \psi_\beta\big(\mu(U_\alpha) \cap V_\beta\big)$$

is a C^t map. In particular, a C^t map $f : M \to \mathbb{R}^1$ is called a C^t function on M. The set of n functions $u^1(p), u^2(p), \ldots, u^n(p)$ on U_α defined as the coordinates of $\phi_\alpha(p)$ in $\mathbb{R}^n$, are called local coordinates in U_α. It is not necessarily possible to find one set of local coordinates which covers M, as the example of the two-sphere demonstrates.

A manifold is said to be orientable if there is an atlas $\{V_\beta, \psi_\beta\}$ of the complete atlas such that, in every nonempty $V_\alpha \cap V_\beta$, the Jacobian $|\partial u^i / \partial \bar{u}^j|$ is positive, where $u^1, u^2, \ldots, u^n$ and $\bar{u}^1, \bar{u}^2, \ldots, \bar{u}^n$ are the local coordinates in V_α and V_β, respectively.

An atlas $\{V_\alpha, \psi_\alpha\}$ is said to be locally finite if every point $p \in M$ has an open neighbourhood which intersects only a finite number of the sets V_α. A manifold is said to be paracompact if it satisfies the Hausdorff separation axiom and if for every atlas $\{U_\alpha, \phi_\alpha\}$ there is a locally finite atlas $\{V_\beta, \psi_\beta\}$ with each V_β contained in some U_α. A Hausdorff manifold with a countable basis is paracompact and any paracompact manifold is normal [Hocking 1963, p. 78]. For a locally finite atlas $\{V_\beta, \psi_\beta\}$ of a paracompact manifold, one can find a partition of unity. This is a set of C^r functions $\{f_\beta\}$ such that [Kobayashi 1963, p. 273]:

1. $0 \leq f_\beta \leq 1$.
2. The support of f_β, i.e., the closure of the set $\{p \in M : f_\beta(p) \neq 0\}$, is contained in the corresponding V_β.

3. $\sum_\beta f_\beta(p) = 1$ for all $p \in M$.

Unless otherwise stated, all manifolds considered will be paracompact, at least C^4, and without boundary.

2.2 Tensors

A C^k curve $\lambda(t)$ in M is defined as a C^k map of a closed interval $[a, b]$ of $\mathbb{R}^1$ into M, that is to say, the restriction to $[a, b]$ of a C^k map of an open interval containing $[a, b]$. The tangent vector to $\lambda(t)$ at a point $p = \lambda(t_0)$ is defined as the map

$$\left(\frac{\partial}{\partial t}\right)_\lambda\Big|_p : \mathcal{F}(p) \longrightarrow \mathbb{R}^1,$$

where $\mathcal{F}(p)$ is the algebra of C^1 functions defined in an open neighbourhood of p. In other words, if $f \in \mathcal{F}(p)$, then $(\partial/\partial t)_\lambda f$ is the derivative of f in the direction of $\lambda(t)$.

Let $u^1, u^2, \ldots, u^n$ be local coordinates in a neighbourhood of p. Then

$$\left(\frac{\partial}{\partial t}\right)_\lambda\Big|_p = \sum_j \left.\frac{\mathrm{d}u^j\big(\lambda(t)\big)}{\mathrm{d}t}\right|_p \left.\frac{\partial}{\partial u^j}\right|_p.$$

Thus every tangent vector at p can be expressed as a linear combination of the coordinate derivatives

$$\left.\frac{\partial}{\partial u^1}\right|_p, \ldots, \left.\frac{\partial}{\partial u^n}\right|_p.$$

Conversely, given a linear combination

$$\sum_j v^j \left.\frac{\partial}{\partial u^j}\right|_p,$$

consider the curve defined by

$$u^j = u^j(p) + v^j t,$$

for t in some interval $[-\varepsilon, \varepsilon]$. Then the tangent vector to this curve at p is

$$\sum_j v^j \left.\frac{\partial}{\partial u^j}\right|_p.$$

Thus the tangent vectors at p form a vector space spanned by

$$\left.\frac{\partial}{\partial u^1}\right|_p, \ldots, \left.\frac{\partial}{\partial u^n}\right|_p.$$

To show that these are linearly independent, suppose

$$\sum_j v^j \left.\frac{\partial}{\partial u^j}\right|_p = 0.$$

Then applying this to u^k, one obtains

$$0 = \sum_j v^j \left.\frac{\partial u^k}{\partial u^j}\right|_p = v^k.$$

The space of all tangent vectors at p will be denoted $T_p(M)$ or simply T_p. Any vector $V \in T_p$ can be represented as

$$V = \sum_j V^j \left.\frac{\partial}{\partial u^j}\right|_p ,$$

where $V^j = Vu^j$ are the components of V with respect to the coordinate basis $\partial/\partial u^1\big|_p, \ldots, \partial/\partial u^n\big|_p$.

A form (one-form, covariant vector) at p is defined to be a linear map of T_p to $\mathbb{R}^1$. In other words, it is an element of T_p^*, the vector space dual to T_p. If $E_1, E_2, \ldots, E_n$ are a basis for T_p, there is a dual basis $E^1, E^2, \ldots, E^n$ for T_p^* such that $E^j(E_i) = \delta_{ij}$. Then a form $A \in T_p^*$ can be expressed as $\sum_j A_j E^j$, where $A_j = A(E_j)$ are called the components of the form in the basis dual to $E_1, \ldots, E_n$. For a function $f \in \mathcal{F}(p)$, the form $\mathrm{d}f$ defined by $\mathrm{d}f(X) = Xf$, for any $X \in T_p$, is called the differential of f at p. Then $\mathrm{d}u^1, \ldots, \mathrm{d}u^n$ form a basis of T_p^* dual to the coordinate basis $\partial/\partial u^1, \ldots, \partial/\partial u^n$ of T_p.

If P and Q are vector spaces over $\mathbb{R}^1$ with duals P^* and Q^*, the tensor product $P \otimes Q$ is defined as the space of all bilinear maps of $P^* \times Q^*$ to $\mathbb{R}^1$. If $p \in P$ and $q \in Q$, $p \otimes q$ denotes that element of $P \otimes Q$ which maps $(r, s) \in P^* \times Q^*$ to $\big[p(r)\big]\big[q(s)\big]$. If $p_1, \ldots, p_n$ and $q_1, \ldots, q_m$ are bases for P and Q, respectively, then $p_i \otimes q_j$ $(i = 1, \ldots, n,\ j = 1, \ldots, m)$ will be a basis for $P \otimes Q$.

The tensor space $T_s^r(p)$, of contravariant order r and covariant order s, at p is defined to be the tensor product

$$\underbrace{T_p \otimes \ldots \otimes T_p}_{r \text{ times}} \otimes \underbrace{T_p^* \otimes \ldots \otimes T_p^*}_{s \text{ times}},$$

whence $T_0^1(p) = T_p$ and $T_1^0(p) = T_p^*$. An element K of $T_s^r(p)$ will be called a tensor of type (r, s) at p. It will be a multilinear map of $T_p^* \times \ldots \times T_p^* \times T_p \times \ldots \times T_p$ to $\mathbb{R}^1$ and may be denoted as $K(A^1, \ldots, A^r, X_1, \ldots, X_s)$, where $A^1, \ldots, A^r \in T_p^*$ and $X_1, \ldots, X_s \in T_p$. In terms of the dual bases $E^1, \ldots, E^n$ and $E_1, \ldots, E_n$ of T_p^* and T_p, respectively, it can be expressed as

$$K = \sum K^{a\ldots d}{}_{i\ldots k} E_a \otimes \ldots \otimes E_d \otimes E^i \otimes \ldots \otimes E^k,$$

where the numbers $K^{a\ldots d}{}_{i\ldots k}$ are called the components of K with respect to the bases. Relations between tensors may be written either in terms of the tensors themselves considered as multilinear maps or in terms of their components. We shall be flexible in passing from one notation to the other.

The operation of contraction on a given contravariant and a given covariant position is the linear map $T_s^r(p) \to T_{s-1}^{r-1}(p)$ given by

$$K^{a\ldots b\ldots d}{}_{i\ldots j\ldots k} \quad \longmapsto \quad \sum_q K^{a\ldots q\ldots d}{}_{i\ldots q\ldots k},$$

or, using the dummy suffix notation, simply $K^{a\ldots q\ldots d}{}_{i\ldots q\ldots k}$.

The symmetrised (resp. antisymmetrised) part of a tensor on a given set of q contravariant positions is defined to be the tensor whose components are $1/q!$ times the sum (alternating sum) of components with all permutations of the indices. This will be denoted by placing round (square) brackets around the indices. Thus,

$$K^{(ab)} = \frac{1}{2!}\left(K^{ab} + K^{ba}\right), \qquad K^{[ab]} = \frac{1}{2!}\left(K^{ab} - K^{ba}\right).$$

Similarly, symmetrisation and antisymmetrisation may be defined on covariant positions. A tensor of type $(0, q)$ which equals its antisymmetric part on all q positions is called a q form. If A and B are p and q forms, respectively, we can define their wedge product $A \wedge B = (-1)^{pq} B \wedge A$ by

$$(A \wedge B)_{ab\ldots def\ldots h} = A_{[ab\ldots d} B_{ef\ldots h]}.$$

With this product, the forms constitute a Grassmann algebra. The forms

$$\mathrm{d}u^a \wedge \mathrm{d}u^b \wedge \ldots \wedge \mathrm{d}u^d$$

are a basis for p forms:

$$A = A_{ab\ldots d}\, \mathrm{d}u^a \wedge \mathrm{d}u^b \wedge \ldots \wedge \mathrm{d}u^d.$$

A C^k tensor field K of type (r, s) on a set G is an assignment of an element of $T^r_s(p)$ for every $p \in G$ such that the components of K with respect to a set of local coordinates in an open neighbourhood of every point p are C^k functions of the coordinates.

2.3 Maps of manifolds

Let ϕ be a differentiable map of an n dimensional manifold M to an $\bar{n}$ dimensional manifold $\overline{M}$. Then if f is a function on $\overline{M}$, $\phi^+ f$ is defined to be the functional on M whose value at a point $p \in M$ is that of f at $\phi(p)$. Thus ϕ^+ is a linear map of functions on $\overline{M}$ to functions on M. Let $\lambda(t)$ be a curve through $p \in M$. Then $\phi\big(\lambda(t)\big)$ will be a curve in $\overline{M}$ through $\phi(p)$. The tangent vector to this curve at $\phi(p)$ will be called $\phi_+ \left((\partial/\partial t)_\lambda\right)$. Then ϕ_+ is a linear map of $T_p(M)$ to $T_{\phi(p)}(\overline{M})$. It is easy to see that $X(\phi^+ f) = (\phi_+ X) f$, for all $X \in T_p(M)$. Similarly, the linear map

$$\phi^+ : T^*_{\phi(p)}(\overline{M}) \longrightarrow T^*_p(M)$$

can be defined by

$$\big(\phi^+ A\big)(X) = A(\phi_+ X), \quad A \in T^*_{\phi(p)}(\overline{M}).$$

Thus the maps ϕ^+ and ϕ_+ can be regarded respectively as maps of covariant tensor fields from $\overline{M}$ to M and contravariant tensor fields from M to $\overline{M}$.

The map ϕ is said to be of rank r at p if the dimension of $\phi_+\big(T_p(M)\big)$ is r. It is said to be injective at p if $r = n$ and surjective if $r = \bar{n}$. it is said to be a diffeomorphism if $\phi^{-1} : \overline{M} \to M$ is a differentiable map. It follows from the inverse function theorem that, if $n = \bar{n}$ and ϕ is injective at p, there is an open neighbourhood U of p such that $\phi : U \to \phi(U)$ is a diffeomorphism.

The map ϕ is said to be an immersion if ϕ is injective at every point $p \in M$. An immersion ϕ is said to be an imbedding if ϕ is a homeomorphism onto its image in the induced topology. M, or $\phi(M)$, is then said to be an imbedded submanifold, or simply a submanifold.

2.4 Differentiation

The exterior differential operator d acting on a function, i.e., a 0-form field, is defined as in Section 2.2. It is defined acting on an r-form field $A = A_{ab\ldots d}\mathrm{d}u^a \wedge \mathrm{d}u^b \wedge \ldots \wedge \mathrm{d}u^d$ by

$$\mathrm{d}A = \mathrm{d}A_{ab\ldots d} \wedge \mathrm{d}u^a \wedge \mathrm{d}u^b \wedge \ldots \wedge \mathrm{d}u^d.$$

The following properties may easily be verified:

1. d maps r-form fields linearly to $(r+1)$-form fields.
2. If A is an r-form, then $\mathrm{d}(A \wedge B) = \mathrm{d}A \wedge B + (-1)^r A \wedge \mathrm{d}B$.

3. $\mathrm{d}^2 A = 0$.
4. If $\phi : M \to \overline{M}$ is a C^2 differentiable map and A is a form field on $\overline{M}$, then

$$\mathrm{d}\left(\phi^+ A\right) = \phi^+(\mathrm{d}A).$$

Let M be a compact, orientable, n dimensional manifold with boundary and let $\{f_\alpha\}$ be a partition of unity for a finite orientated atlas $\{U_\alpha, \phi_\alpha\}$. Then if A is an n-form field on M, the integral of A over M is defined as

$$\int_M A = \sum_\alpha \int_{\phi_\alpha(U_\alpha)} f_\alpha A_{12\ldots n} \mathrm{d}u^1 \mathrm{d}u^2 \ldots \mathrm{d}u^n,$$

where $A_{12\ldots n}$ are the components of A in the local coordinate neighbourhood U_α and the integrals on the right are ordinary multiple integrals over open sets $\phi_\alpha(U_\alpha)$ of $\mathbb{R}^n$. It may be verified that $\int_M A$ is independent of the atlas chosen and that if $\phi : \overline{M} \to M$ is a differentiable map, then

$$\int_{\psi^{-1}(M)} \psi^+ A = \int_M A.$$

If B is an $(n-1)$-form field on M, the generalised Stokes equation can be expressed as

$$\int_{\partial M} B = \int_M \mathrm{d}B.$$

This may be verified from the definition of the integral given above.

2.4.1 Lie derivative

Let X be a C^1 vector field on M. Then by the fundamental theorem of differential equations, through each point of M there is a unique curve (called the integral curve of X) whose tangent vector is X. For every $p \in M$, there is an open neighbourhood U and an $\varepsilon > 0$ such that there is a family of differentiable maps $\phi_t : U \to M$ $(|t| < \varepsilon)$ which are diffeomorphisms, $U \to \phi_t(U)$ and which are defined by taking each point of U a parameter distance t along the integral curves of X. If K is a tensor field of type (r, s) on U, the isomorphism $(\phi_t^{-1})_+ : T_{\phi_t^{-1}(p)} \to T_p$ induces an isomorphism $\bar{\phi}_t : T_s^r\left(\phi_t^{-1}(p)\right) \to T_s^r(p)$.

The field $\bar{\phi}_t(K)$ is said to be 'dragged along' by the diffeomorphism ϕ_t. Then the Lie derivative of K with respect to X is defined to be the derivative with respect to t of this dragged along field, that is

$$\mathcal{L}_X K\big|_p = \lim_{t\to 0} \frac{1}{t}\left[K\big|_p - \bar{\phi}_t(K)\big|_p\right].$$

By its definition, $\mathcal{L}_X K$ will also be a tensor field of type (r, s). In terms of components with respect to a coordinate basis:

$$(\mathcal{L}_X K)^{ab\ldots d}{}_{ef\ldots h} = \frac{\partial K^{ab\ldots d}{}_{ef\ldots h}}{\partial u^i} X^i - K^{ib\ldots d}{}_{ef\ldots h} \frac{\partial X^a}{\partial u^i} - \ldots + K^{ab\ldots d}{}_{if\ldots h} \frac{\partial X^i}{\partial u^e} + \ldots$$

In particular, $\mathcal{L}_X f = Xf$, where f is a function. If Y is a vector field, $\mathcal{L}_X Y$ is sometimes written as $[Y, X] = -[X, Y]$.

2.4.2 Covariant derivative

A connection at a point $p \in M$ is a rule which assigns to each C^1 vector field Y in a neighbourhood of p a tensor ∇Y of type $(1,1)$ at p called the covariant derivative of Y, such that:

1. ∇Y is linear in Y.
2. $\nabla(fY) = \mathrm{d}f \otimes Y + f\nabla Y$.

A tensor of type $(1,1)$ is a bilinear map $T_p^* \times T_p \to \mathbb{R}^1$. It can also be regarded as a linear map $T_p \to T_p$. Thus if $X \in T_p$, we denote the action of ∇Y on X by $\nabla_X Y$, the covariant derivative of Y in the direction of X.

A C^k connection on M is a rule which assigns a connection at each point on M such that if Y is a C^{k+1} vector field on M, then ∇Y is a C^k tensor field. In terms of local coordinates $u^1, \ldots, u^n$ on a neighbourhood U, the connection is determined by n^3 C^k functions on U such that

$$\nabla \frac{\partial}{\partial u^j} = \Gamma^k_{ij} \frac{\partial}{\partial u^k} \otimes \mathrm{d}u^i.$$

Then by rules (1) and (2) above,

$$\nabla Y = Y^i{}_{;j} \frac{\partial}{\partial u^i} \otimes \mathrm{d}u^j,$$

where $Y^i{}_{;j}$ are the coordinate components of the covariant derivative of Y and are given by

$$Y^i{}_{;j} = \frac{\partial Y^i}{\partial u^j} + \Gamma^i_{jk} Y^k.$$

The definition of covariant derivative can be extended to any C^1 tensor field by the rules:

1. If K is a tensor field of type (r,s), then ∇K is a tensor field of type $(r, s+1)$.
2. $\nabla(K \otimes L) = \nabla K \otimes L + K \otimes \nabla L$.
3. ∇ commutes with contraction.
4. $\nabla f = \mathrm{d}f$, where f is a function.

These give that the components of ∇K are

$$K^{ab\ldots d}{}_{ef\ldots h;i} = \frac{\partial K^{ab\ldots d}{}_{ef\ldots h}}{\partial u^i} + \Gamma^a_{ij} K^{jb\ldots d}{}_{ef\ldots h} + \ldots - \Gamma^j_{ie} K^{ab\ldots d}{}_{jf\ldots h} - \ldots$$

If K is a C^1 tensor field along a curve $\lambda(t)$, we may define $\mathrm{D}K/\partial t$, the covariant derivative of K along $\lambda(t)$, as $\nabla_{\partial/\partial t}\overline{K}$, where $\overline{K}$ is any C^1 tensor field extending K in an open neighbourhood of λ. It is not difficult to see that $\mathrm{D}K/\partial t$ is independent of the extension $\overline{K}$. K is said to be parallelly transported along λ if $\mathrm{D}K/\partial t = 0$.

If ∇Y and $\hat{\nabla} Y$ are covariant derivatives obtained from two different connections, then

$$\nabla Y - \widehat{\nabla} Y = \left(\Gamma^i_{jk} - \widehat{\Gamma}^i_{jk}\right) Y^k \frac{\partial}{\partial u^i} \otimes \mathrm{d}u^j$$

will be a tensor. Thus $\Gamma^i_{jk} - \widehat{\Gamma}^i_{jk}$ will be the components of a tensor. Similarly, $\Gamma^i_{jk} - \Gamma^i_{kj}$ will be the components of a tensor called the torsion tensor of the connection. We shall deal only with connections that are torsion free (symmetric).

The exterior and Lie derivatives may be expressed in terms of the covariant derivative. Thus,

$$\mathrm{d}A = A_{ab\ldots d;e} \mathrm{d}u^a \wedge \mathrm{d}u^b \wedge \ldots \wedge \mathrm{d}u^d \wedge \mathrm{d}u^e$$

and

$$[X,Y] = \left(X^a{}_{;b}Y^b - Y^a{}_{;b}X^b\right)\frac{\partial}{\partial u^a}.$$

However, by their construction, they are independent of the connection.

The curvature (or Riemann) tensor of a connection is a measure of the extent to which the scond covariant derivative $\nabla_{\partial/\partial u^i}(\nabla_{\partial/\partial u^j}Z)$ is not symmetric in i and j. Given C^2 vector fields X, Y, and Z, define a new vector field by

$$R(X,Y)Z = \nabla_X\nabla_Y Z - \nabla_Y\nabla_X Z + \nabla_{[X,Y]}Z.$$

It is easy to verify that the value of $R(X,Y)Z$ at a point $p \in M$ depends only on the values of X, Y, and Z at p, and not on their values at nearby points, and that $R(X,Y)Z$ is trilinear in X, Y, and Z. In other words, R is a tensor. In component form, one has

$$z^a{}_{;bc} - z^a{}_{;cb} = R^a{}_{dcb}Z^d,$$

where

$$R^a{}_{bcd} = \mathrm{d}u^a\left(R(\partial/\partial u^c, \partial/\partial u^d, \partial/\partial u^b)\right)$$

are the coordinate components of the Riemann tensor and are related to the Γ^a_{bc} by

$$R^a{}_{bcd} = \frac{\partial\Gamma^a_{db}}{\partial u^c} - \frac{\partial\Gamma^a_{cb}}{\partial u^d} + \Gamma^e_{db}\Gamma^a_{ce} - \Gamma^e_{cb}\Gamma^a_{de}.$$

As $\Gamma^a_{[bc]} = 0$, the Riemann tensor has the symmetry $R^a{}_{[bcd]} = 0$ and satisfies the Bianchi identity $R^a{}_{b[cd;e]} = 0$.

The Ricci tensor is defined to be the contraction of the Riemann tensor

$$R(X,Y) = \mathrm{d}u^a\left(R(\partial/\partial u^a, X)Y\right).$$

In components, $R_{bd} = R^a{}_{bad}$.

2.5 The metric

A metric at $p \in M$ is a scalar product on T_p. Thus it can be represented by a symmetric tensor g of type (0,2) with coordinate components $g_{ab} = g(\partial/\partial u^a, \partial/\partial u^b)$. Then

$$g = g_{ab}\mathrm{d}u^a \otimes \mathrm{d}u^b.$$

In fact, the tensor product sign is normally omitted and g is denoted by $\mathrm{d}s^2$. The metric is said to be nondegenerate if there is no nonzero $X \in T_p$ such that $g(X,Y) = 0$ for all $Y \in T_p$. In terms of components, the metric is nondegenerate if and only if the matrix (g_{ab}) of the components is nonsingular. By a metric, we shall in future always mean a nondegenerate metric. For such a metric, one can define a symmetric contravariant metric tensor with components g^{ab}, such that

$$g^{ab}g_{bc} = \delta^a_c.$$

These tensors can be used to give an isomorphism between covariant and contravariant tensors (in other words, to raise and lower indices). For example, if X^a are the components of a contravariant vector, then X_a will be the components of a covariant vector, where

$$X_a = g_{ab}X^b, \qquad X^a = g^{ab}X_b.$$

A C^k metric on M is a C^k tensor field g. The signature of g at p is the number of positive eigenvalues of the matrix (g_{ab}) minus the number of negative ones. As g is

nondegenerate, the signature will be constant on M. A metric whose signature is n is called a positive definite metric and one whose signature is $2-n$ is called a Lorentz metric. Any paracompact C^r manifold admits a C^{r-1} positive-definite metric. This may be shown as follows. Let $\{f_\alpha\}$ be a partition of unity for a locally finite atlas $\{U_\alpha, \phi_\alpha\}$. Then we may define $g(X,Y)$ by

$$g(X,Y) = \sum f_\alpha \big\langle (\phi_\alpha)_+ X, (\phi_\alpha)_+ Y \big\rangle,$$

where $\langle\ ,\ \rangle$ is the natural scalar product in Euclidean space $\mathbb{R}^n$.

However, a C^r paracompact manifold admits a C^{r-1} Lorentz metric if and only if it admits a C^{r-1} line-element field $(X,-X)$ (by this is meant a non-vanishing C^{r-1} vector field X which is determined up to a sign). This may be seen as follows. Let $\overline{g}$ be a C^{r-1} positive-definite metric. Then we may define a Lorentz metric g by

$$g(Y,Z) = -\overline{g}(Y,Z) + 2\left[\overline{g}(X,X)\right]^{-1}\overline{g}(X,Y)\overline{g}(X,Z).$$

Conversely, if g is a Lorentz metric, consider the equation $g_{ab}X^b = \lambda \overline{g}_{ab}X^b$. This will have one positive and $n-1$ negative eigenvalues λ. Thus the eigenvector X corresponding to the positive eigenvalue will be determined up to a sign and a normalising factor. It may be normalised by $X^aX^bg_{ab} = 1$. In fact, any noncompact manifold admits a line-element field, while a compact manifold does if and only if its Euler invariant is zero.

Given a metric g, there is a unique symmetric, i.e., torsion-free, connection for which the covariant derivative of g is zero. It is easy to verify that the components of this connection are

$$\Gamma^a_{bc} = \frac{1}{2}g^{ad}\left(\frac{\partial g_{db}}{\partial u^c} + \frac{\partial g_{dc}}{\partial u^b} - \frac{\partial g_{bc}}{\partial u^d}\right).$$

Henceforth, we shall deal only with the connection defined by the metric. The Riemann tensor then satisfies the additional identity $R_{(ab)cd} = 0$ (where one index has been lowered by g_{ab}), as well as the identities $R_{ab(cd)} = 0$ (from the definition of R_{abcd}) and $R_{a[bcd]} = 0$ (from the fact that $\Gamma^a_{[bc]} = 0$). These identities imply that there are

$$\frac{1}{12}n^2(n+1)(n-1)$$

algebraically independent components R_{abcd}. When $n > 2$, $n(n+1)/2$ of them can be represented by the components R_{ab} of the Ricci tensor (since $R_{[ab]} = 0$). When $n > 3$, the remaining

$$\frac{1}{12}n(n+1)(n+2)(n-3)$$

components can be represented by the components C_{abcd} of the Weyl or conformal tensor defined by

$$R_{abcd} = C_{abcd} - \frac{2}{n-2}\Big(g_{a[d}R_{c]b} + g_{b[c}R_{d]a}\Big) - \frac{2}{(n-2)(n-1)}Rg_{a[c}g_{d]b},$$

where $R = g^{ab}R_{ab}$ is the curvature scalar. The Weyl tensor satisfies the identities

$$C_{abcd} = C_{[ab][cd]}, \qquad C_{a[bcd]} = 0, \qquad C^a{}_{bad} = 0.$$

Two metrics $\tilde{g}$ and g are said to be conformal if $\tilde{g} = \Omega^2 g$ for some suitably defined nonzero differentiable function Ω. Then the connections they define are related by

$$\tilde{\Gamma}^a_{bc} = \Gamma^a_{bc} + \Omega^{-1}\left(\delta^a_b\frac{\partial\Omega}{\partial u^c} + \delta^a_c\frac{\partial\Omega}{\partial u^b} - g_{bc}g^{ad}\frac{\partial\Omega}{\partial u^d}\right),$$

and their Weyl tensors by

$$\tilde{C}^a{}_{bcd} = C^a{}_{bcd}.$$

Thus the Weyl tensor is a conformal invariant.

A curve $\lambda(t)$ is said to be a geodesic curve if

$$\frac{\mathrm{D}}{\partial t}\left(\frac{\partial}{\partial t}\right)_\lambda \quad \text{is parallel to} \quad \left(\frac{\partial}{\partial t}\right)_\lambda.$$

For such a curve, one can find an affine parameter v such that

$$\frac{\mathrm{D}}{\partial v}\left(\frac{\partial}{\partial v}\right)_\lambda = 0.$$

The curve λ with the parameter v is called a geodesic. The affine parameter of a geodesic curve is determined up to a constant multiplying and a constant additive factor. If the connection is C^1, then given any $X \in T_p$, there is a unique geodesic $\lambda(v)$ with $\lambda(0) = p$ and

$$\left(\frac{\partial}{\partial v}\right)_\lambda\Bigg|_p = X.$$

These geodesics give a differentiable map $\exp : T_p \to M$ which takes $X \in T_p$ to $\lambda(1)$. This map will be defined for some open neighbourhood of the origin of T_p. M is said to be geodesically complete if it is defined on T_p for all $p \in M$. As the exponential map is injective at M, there will be an open neighbourhood V of p such that it is a diffeomorphism: $\exp^{-1}(V) \to V$. Let $E_1, E_2, \ldots, E_n$ be an orthonormal basis for T_p, i.e., $g(E_i, E_j) = \pm\delta_{ij}$. Then we may define local coordinates $u_1, u^2, \ldots, u^n$ on V by setting $u^i(q)$ equal to the components with respect to $E_1, E_2, \ldots, E_n$ of $\exp^{-1}(q)$. At p, we then have $\partial/\partial u^i = E_i$ and $\Gamma^i_{jk} = 0$. The neighbourhood V may be chosen so that it is convex, that is, any two points of it may be joined by a unique geodesic in V [Kobayashi 1963, p. 149]. The neighbourhood V will be called a normal coordinate neighbourhood.

The Kronecker tensor of order p ($p \leq n$) has components

$$\delta^{ab\ldots d}_{ef\ldots h} = p!\,\delta^a_{[e}\delta^b_f \ldots \delta^d_{h]}$$

and has zero covariant derivative. The components $\delta^{12\ldots n}_{ef\ldots h}$ of the Kronecker tensor of order n have the correct antisymmetry to be components of an n-form. However, if $\bar{u}^1, \bar{u}^2, \ldots, \bar{u}^n$ are another set of local coordinates, then

$$\bar{\delta}^{12\ldots n}_{ef\ldots h} = \left|\frac{\partial \bar{u}^i}{\partial u^d}\right| \frac{\partial u^r}{\partial \bar{u}^e}\frac{\partial u^s}{\partial \bar{u}^f} \ldots \frac{\partial u^v}{\partial \bar{u}^h}\delta^{12\ldots n}_{rs\ldots v}.$$

This is not the correct transformation law for the components of an n-form since it includes the Jacobian $|\partial \bar{u}^i/\partial u^d|$. However,

$$\eta_{ef\ldots h} = \sqrt{g}\delta^{12\ldots n}_{ef\ldots h}, \qquad f = \det|g_{ab}|$$

will be the components of an n-form called the canonical form, which has zero covariant derivative. using this one may define the Hodge star operation which maps p-forms linearly to $(n-p)$-forms:

$$({}^*A)_{ab\ldots d} = p!\eta_{ab\ldots def\ldots h}A_{ij\ldots l}g^{ei}g^{fj}\ldots g^{hl}.$$

Then

$${}^{**}A = (-1)^{(n-p)p+s-n}A,$$

where $2s$ is the signature of the metric. This operation enables one to integrate p-forms over $n-p$ submanifolds. One may also define the operator δ by

$$\delta = (-1)^{(n-p)p+s-p} {}^*d^*.$$

This may be regarded as a generalisation of the divergence operator. For a one-form,

$$\delta A = A_{a;b} g^{ab},$$

and Green's formula may be expressed as

$$(-1)^{(n-p)p+s-n} \int_M {}^*\delta A = \int_M \mathrm{d}{}^*A = \int_{\partial M} {}^*A.$$

3 General relativity

3.1 Special relativity

The special theory of relativity was proposed by Einstein in 1905. Since then its predictions have been extensively tested and found to agree well with experiment, provided that gravitational effects are neglected. In order to include gravitation, Einstein formulated the general theory of relativity in 1916. This theory includes all the experimentally tested features of special relativity and also predicts results for gravitational fields very similar to those of the well tried Newtonian theory.

We shall present the general theory in the form of two postulates which describe the mathematical model to be used, but which do not have any physical content until they are supplemented by three requirements which relate the mathematical structure to physically observable quantities. The two postulates and the first two requirements are statements of the special theory of relativity and as such are well tested. The third requirement which distinguishes the general from the special theory is not so well established experimentally. Nevertheless, we shall see that it would be difficult to think of any alternative requirement which did not have some undesirable features.

Postulate (a). Space and time are represented together as a four-dimensional, connected, paracompact manifold M of class at least C^4.

A manifold corresponds naturally to our intuitive ideas about the continuity of space and time. However, it is of interest to note that theories have been proposed in which spacetime has a discrete structure [e.g., Hill 1955, Coxeter 1950]. Nevertheless, even in these theories a manifold is a good approximation for regions larger than about 10^{-14} cm.

On the question of what class of differentiability we should assume for the spacetime manifold M, we shall adopt the view later that the differential structure can only be physically determined by use of the metric. Thus unless we assumed a C^∞ metric, and there does not seem to be any physical reason for doing so, we could not physically determine which C^∞ atlas was *the* C^∞ atlas. Lichnerowicz [Lichnerowicz 1955] has suggested that one might assume that M was only of class C^2 with a C^2, piecewise C^4 differential structure. This is defined as follows: an atlas $\{V_\alpha, \psi_\alpha\}$ of the complete C^2 atlas of M will be said to be piecewise C^4 if, for every non-empty $V_\alpha \cap V_\beta$,

$$\psi_\alpha \psi_\beta^{-1} : \phi_\beta(V_\alpha \cap V_\beta) \longrightarrow \phi_\alpha(V_\alpha \cap V_\beta)$$

is a piecewise C^4 map. The set of all piecewise C^4 atlases compatible with the given atlas constitutes the complete C^2, piecewise C^4 atlas of M. A function on M is said

to be C^2 piecewise C^4 if it is a C^2, piecewise C^4 function of the local coordinates of a piecewise C^4 atlas. Similarly, a tensor field on M is said to be C^1 piecewise C^3 if its components with respect to the local coordinates of some piecewise C^4 atlas are C^1, piecewise C^3 functions of the coordinates.

The assumption of this piecewise C^4 structure is mathematically convenient for the construction of certain exact solutions since it allows certain discontinuities in the curvature, etc. All such solutions, however, can be arbitrarily closely approximated by C^4 and even C^∞ solutions which would probably be more physically realistic anyway. All the results that will be obtained in this paper could be proved on the assumption of a piecewise C^4 structure instead of a C^4 structure. However, to perform all the calculations on this basis would be very tedious and would not really produce anything new. We will therefore assume a C^4 structure for M.

Postulate (b). On M, there is a C^3 (Lichnerowicz: C^1, piecewise C^3) Lorentz metric $g(X, Y)$ of signature -2 and components

$$g_{ab} = g\left(\partial/\partial u^a, \partial/\partial u^b\right).$$

We saw in Section 2 that the existence of such a metric implied that, if M was compact, its Euler invariant must be zero. However, it will be shown in Section 5 that there is good reason to suppose that M is not compact.

The metric enables the non-zero vectors at a point $p \in M$ to be divided into three classes: a non-zero vector $X \in T_p$ is said to be timelike, null, or spacelike, according to whether $g(X, X)$ is greater than, equal to, or less than zero, respectively. The physical significance of the metric comes from the following requirements which relate it to physically observable quantities.

Requirement 1. Local causality. Let U be a normal coordinate neighbourhood of a point p. Then events at another point $q \in U$ cannot be causally related to events at p by effects confined to U unless q can be reached from p by a C^1 curve whose tangent vector is everywhere timelike or null (we shall call such a curve non-spacelike). That is to say, it is impossible to send a signal which does not cross the boundary of U from one point to the other if they cannot be joined by a non-spacelike curve.

Taking u^1, u^2, u^3, and u^4 to be normal coordinates in U about p with $\partial/\partial u^4$ timelike, ($g_{11} = g_{22} = g_{33} = -g_{44} = -1$, other g_{ab} zero at p), it is easy to see that the points that may be causally related to p in U are those whose coordinates satisfy

$$\left(u^4\right)^2 - \left(u^3\right)^2 - \left(u^2\right)^2 - \left(u^1\right)^2 \geq 0.$$

The boundary of these points in U is formed by the points whose coordinates satisfy the equality above. This surface, called the null cone of p, is generated by the null geodesics through p, that is, geodesics with null tangent vectors. Thus observation of causal relationships on M enables us to determine N_p, the space of null vectors at p. But once N_p is known, the metric at p may be determined up to a conformal factor. This may be shown as follows. Let $X \in T_p$ be a timelike vector and $Y \in T_p$ a spacelike vector. Consider the equation

$$g(X + \lambda Y, X + \lambda Y) = g(X, X) + 2\lambda g(X, Y) + \lambda^2 g(Y, Y) = 0.$$

Since $g(X, X) > 0$ and $g(Y, Y) < 0$, this will have real roots λ_1 and λ_2. If N_p is known, λ_1 and λ_2 may be determined. But

$$\lambda_1 \lambda_2 = \frac{g(X, X)}{g(Y, Y)},$$

and thus the ratio of the magnitudes of a timelike and a spacelike vector may be found. Then if W and Z are any two vectors at p,

$$g(W,Z) = \frac{1}{2}\left[g(W+Z,W+Z) - g(W,W) - g(Z,Z)\right].$$

Each of the magnitudes on the right may be compared with the magnitudes of either X or Y, and so $g(W,Z)/g(X,X)$ may be found. Thus observation of the causal structure of M allows us to measure the metric up to a conformal factor. In practice, this measurement is most conveniently carried out using the fact that light travels approximately on null geodesics. This, however, is a consequence of the particular equations the electromagnetic theory obeys, not of the theory of relativity itself. Causality will be considered further in Section 5, where it will be shown that, with certain assumptions, it may be used to define the topological and differential structure of M.

Requirement 2. Covariance. The equations which govern the behaviour of the physical fields on M should all be expressible as relations between tensors on M with all derivatives with respect to position being covariant derivatives with the symmetric connection defined by the metric. In particular, if T^{ab} are the components of the total energy-momentum tensor of all the physical fields (except the gravitational fields), the equations of local conservation of energy and momentum are expressed as

$$T^{ab}{}_{;b} = 0.$$

If the metric is flat, we may introduce in some region U coordinates u^1, u^2, u^3, u^4 for which the only non-zero components of the metric are

$$g_{11} = g_{22} = g_{33} = -g_{44} = -1,$$

and all the components Γ^a_{bc} of the connection are zero. It is then easy to see that the metric has zero Lie derivative with respect to the following vector fields:

$$\underset{\alpha}{L} = \frac{\partial}{\partial u^\alpha}, \quad \alpha = 1,2,3,4,$$

$$\underset{\gamma\delta}{M} = u^\gamma \frac{\partial}{\partial u^\delta} - u^\delta \frac{\partial}{\partial u^\gamma}, \quad \underset{\gamma 4}{M} = -\underset{4\gamma}{M} = -u^\gamma \frac{\partial}{\partial u^4} - u^4 \frac{\partial}{\partial u^\gamma}, \quad \gamma,\delta = 1,2,3.$$

That is, the ten vector fields $\underset{\alpha}{L}$ and $\underset{\gamma\delta}{M}$ are Killing vectors. They generate a ten parameter Lie group of isometries known as the inhomogeneous Lorentz group. We may use the $\underset{\alpha}{L}$ to define one-forms $\underset{\alpha}{P}$ whose components are

$$\underset{\alpha}{P}_a = g_{ab} g_{cd} T^{bc} \underset{\alpha}{L}{}^d.$$

We may think of $\underset{4}{P}$ as representing the flow of energy and $\underset{1}{P}$, $\underset{2}{P}$, and $\underset{3}{P}$ the flow of the three components of momentum. We have

$$\delta \underset{\alpha}{P} = \underset{\alpha}{P}_{a;e} g^{ae} = T^{bc}{}_{;b} \underset{\alpha}{L}_c + T^{bc} \underset{\alpha}{L}_{c;b}.$$

The first term is zero by the conservation equations and the second term vanishes because $\underset{\alpha}{L}_{(c;b)}$ is zero as $\underset{\alpha}{L}$ is a Killing vector. Thus if D is some compact region contained in U with boundary ∂D, we have

$$\int_{\partial D}^{*} \underset{\alpha}{P} = \int_D {}^{*}\delta \underset{\alpha}{P} = 0.$$

This means that the total flux of energy and each component of momentum over a closed surface is zero. This is the integral form of the law of conservation of energy and momentum. Similar integrals may be defined using the Killing vectors $\underset{\gamma\delta}{M}$. They represent the conservation of angular momentum.

If the metric is not flat, there will not in general be any Killing vectors and so the above integral conservation laws will not hold. However, in a neighbourhood U of a point p, we may introduce normal coordinates u^1, u^2, u^3, and u^4 in which, at p, the non-zero components of the metric are $g_{11} = g_{22} = g_{33} = -g_{44} = -1$ and the components Γ^a_{bc} of the connection are zero. We may take a neighbourhood D of p in which they differ from their values at p by an arbitrarily small amount. Then $\underset{\alpha}{L}_{(a;b)}$ and $\underset{\gamma\delta}{M}_{(a;b)}$ will not exactly vanish in D, but will differ from zero by an arbitrarily small amount. Thus $\int_{\partial D} {}^*\underset{\alpha}{P}$ will still be zero in the first approximation. That is to say, we still have approximate conservation of energy-momentum in a small region. This local conservation enables us to prove that a small isolated body moves approximately on a geodesic in M.

We think of the body as being represented by a thin timelike tube C in M, outside which T_{ab} is zero. We take a timelike curve $\gamma(s)$, parametrised by path length s, in C as representing the motion of the body. We proceed as follows. The Fermi derivative of a vector field X along a timeline curve $\gamma(s)$ with unit tangent vector V, so $g(V,V) = 1$, is defined by

$$\frac{\mathrm{D_F}X}{\partial s} = \frac{\mathrm{D}X}{\partial s} + Vg\left(X, \frac{\mathrm{D}V}{\partial s}\right) - \frac{\mathrm{D}V}{\partial s}g(X,V).$$

A vector field X is said to be Fermi propagated along $\gamma(s)$ if

$$\frac{\mathrm{D_F}X}{\partial s} = 0.$$

The product $g(X,V)$ will be constant along $\gamma(s)$. Let E_1, E_2, E_3, E_4 be an orthonormal basis of vectors at some point $p = \gamma(0)$ with $E_4 = V$. They may be Fermi propagated along $\gamma(s)$ to give an orthonormal basis at all points of $\gamma(s)$. Let H_q be the subspace of T_q, the tangent space at $q = \gamma(s)$, which is spanned by the vectors E_1, E_2, E_3. Then in a neighbourhood of q, $\exp_q(H_q)$ will be a three-surface orthogonal to $\gamma(s)$. We define the Fermi coordinates u^1, u^2, u^3, u^4 of a point r in $\exp_q(H_q)$ by

$$u^i = -g\big(E_i, \exp_q^{-1}(r)\big), \quad i = 1,2,3, \qquad u^4 = s.$$

It is easy to show that, in these coordinates, the only non-zero components of the connection are

$$\Gamma^4_{a4} = \dot{V}_a, \qquad \Gamma^a_{44} = \dot{V}^a,$$

where $\dot{V} = \mathrm{D}V/\partial s$ is the acceleration of the curve $\gamma(s)$. If we define the vectors $\underset{\alpha}{L} = \partial/\partial u^\alpha$ as before, then $\underset{\alpha}{L}{}^a{}_{;b} = \Gamma^a_{\alpha b}$. We put $\underset{\alpha}{P}_a = T_{ab}\underset{\alpha}{L}{}^b$ and take the region D to be that bounded by the two surfaces $\Sigma(u^4 = s)$ and $\Sigma'(u^4 = s')$ and a tube C' lying outside C. Then,

$$\int_{\Sigma'} {}^*\underset{\alpha}{P} - \int_{\Sigma} {}^*\underset{\alpha}{P} = \int_D {}^*\left(T_{ab}\Gamma^a_{\alpha c}g^{bc}\right).$$

Taking Σ' very close to Σ, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}\int_{\Sigma(s)} {}^*\underset{\alpha}{P} = \int_{\Sigma(s)} {}^*\left(T_{ab}\Gamma^a_{\alpha c}g^{bc}\mathrm{d}u^4\right).$$

Multiplying $\underset{\alpha}{P}$ by u^i, $i = 1, 2, 3$, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}\int_{\Sigma(s)} {}^*\left(\underset{\alpha}{P}u^i\right) = \int_{\Sigma(s)} {}^*\left[\left(T_{ab}\underset{\alpha}{L}{}^a g^{ib} + T_{ab}\Gamma^a_{\alpha c}g^{bc}u^i\right)\mathrm{d}u^4\right].$$

We may regard integrals of the form $\int_\Sigma {}^*\left(\underset{\alpha}{P}u^i\right)$ as representing dipole and higher moments of the matter distribution. As we make the body arbitrarily small, we shall assume that these may be neglected in comparison with the first order moments. We may also approximate the components of the connection by their values on $\gamma(s)$. We then get

$$\int_\Sigma {}^*\left(T_{ab}\underset{\alpha}{L}{}^a g^{ib}\mathrm{d}u^4\right) = 0, \qquad \dot{V} = 0,$$

$$\int_\Sigma {}^*(\underset{\alpha}{P}) = 0, \qquad \alpha = 1, 2, 3, \qquad \int_\Sigma {}^*\underset{4}{P} = \text{const.}$$

Thus a sufficiently small isolated body moves on a geodesic independent of its internal constitution. This may be thought of as corresponding to Galileo's principle that all bodies fall at the same rate. In Newtonian terms, one would say that the inertial mass (the m that appears in $F = ma$) and the passive gravitational mass (the mass acted on by a gravitational field) are equal for small bodies.

Requirement (1) enabled us to measure the metric physically up to a conformal factor. Requirement (2) enables us to determine this factor. For it demands that the rate at which physical processes occur in a small isolated system should depend only on the metric and the past history of the system. However, there are found to be systems (such as electronic states of atoms) whose rates are independent of their past history. The fact that there are large numbers of similar such systems enables us to compare the conformal factors at different points and so completely determine the metric in terms of some physical standard of time.

Before proceeding to the third requirement, which distinguishes the general from the special theory of relativity, we shall consider in the next section how the equations and the energy-momentum tensor of the physical fields may be derived.

3.2 Lagrangian formulation

It may be possible to obtain the equations of the physical fields from a Lagrangian L which is a function of the fields ψ^A, ψ^B, and so on (which may be scalar functions, tensors, or spinors), their covariant derivatives, and the metric. One requires that the action integral

$$I = \int_D {}^*L$$

be stationary under small variations of the fields $\psi^A, \psi^B, \ldots$

This may be stated more precisely as follows. We regard the fields as cross-sections of some tensor bundles $P^A, P^B, \ldots$ over M. Then we may define a variation of ψ^A as a C^2 map

$$\alpha : (-\varepsilon, \varepsilon) \times M \longrightarrow P^A,$$

for some $\varepsilon > 0$ such that

1. $\alpha(0, q) = \psi^A(q), \quad q \in M,$
2. $\alpha(u, r) = \psi^A(r), \quad r \in M\backslash D, \;\; u \in (-\varepsilon, \varepsilon).$

We shall denote by $\Delta\psi^A$ the variation vector

$$\Delta\psi^A = \pi\left(\left(\frac{\partial}{\partial u}\right)\alpha\Big|_{u=0}\right).$$

Under this variation α, the derivative of the action will be

$$\frac{\partial I}{\partial u}\Big|_{u=0} = \int_D \left[\frac{\partial^* L}{\partial\psi^A}\Delta\psi^A + \frac{\partial^* L}{\partial\psi^A{}_{;a}}\Delta(\psi^A{}_{;a})\right],$$

where $\psi^A{}_{;a}$ are the components of the covariant derivative of ψ^A. But

$$\Delta\left(\psi^A{}_{;a}\right) = \left(\Delta\psi^A\right)_{;a}.$$

Thus the second term may be expressed as

$$\int_D \left[\left(\frac{\partial^* L}{\partial\psi^A{}_{;a}}\Delta\psi^A\right)_{;a} - \left(\frac{\partial^* L}{\partial\psi^A{}_{;a}}\right)_{;a}\Delta\psi^A\right].$$

The first term above can be transformed into an integral over the boundary of D which vanishes as $\Delta\psi^A$ is zero there. Thus, in order that $\partial I/\partial u$ should be zero for all variations α, we require

$$\boxed{\frac{\partial^* L}{\partial\psi^A} - \left(\frac{\partial^* L}{\partial\psi^A{}_{;a}}\right)_{;a} = 0}.$$

These are the equations of the fields. They automatically satisfy requirement (2).

We may define the energy-momentum tensor from the Lagrangian by considering the change in the action induced by a small change in the metric. Suppose we have a variation α which leaves the fields $\psi^A, \psi^B, \ldots$, unchanged but which alters the components g_{ab} of the metric, then

$$\frac{\partial I}{\partial u}\Big|_{u=0} = \int_D \left[\sum_A \frac{\partial^* L}{\partial\psi^A{}_{;a}}\Delta(\psi^A{}_{;a}) + \frac{\partial^* L}{\partial g_{ab}}\Delta g^{ab}\right].$$

In this case, $\Delta(\psi^A{}_{;a})$ will not necessarily be zero even though $\Delta\psi^A$ is, because the variation in the metric will induce a variation in Γ^a_{bc}, the components of the connection. Since the difference between two connections transforms as a tensor, $\Delta\Gamma^a_{bc}$ may be regarded as the components of a tensor. They are related to the variation in the components of the metric by

$$\Delta\Gamma^a_{bc} = \frac{1}{2}g^{ad}\left[(\Delta g_{db})_{;c} + (\Delta g_{dc})_{;b} - (\Delta g_{bc})_{;d}\right].$$

Using this relation, $\Delta(\psi^A{}_{;a})$ may be expressed in terms of $(\Delta g_{bc})_{;d}$, and the usual integration by parts employed to give an integrand involving Δg_{ab} only. Thus we may write $\partial I/\partial u$ as

$$\int {}^* \left(T^{ab}\Delta g_{ab}\right),$$

where T^{ab} are the components of a symmetric tensor which we call the energy-momentum tensor of the fields. This satisfies the conservation equations as a consequence of the equations of the fields $\psi^A, \psi^B, \ldots$ For suppose we had a diffeomorphism

$\phi : M \to M$ which was the identity everywhere except in the interior of D. Then by the invariance of integrals under differential maps,

$$I = \int_D {}^*L = \int_{\phi^{-1}(D)} \phi^+({}^*L) = \int_D \phi^+({}^*L).$$

Thus

$$\int_D \left[{}^*L - \phi^+({}^*L) \right] = 0.$$

If the diffeomorphism ϕ is generated by a vector field X (non-zero only in the interior of D), it follows that

$$\int_D \mathcal{L}_X({}^*L) = 0.$$

But

$$\int_D \mathcal{L}_X({}^*L) = \int_D \left\{ \sum_A \left[\frac{\partial^* L}{\partial \psi^A} - \left(\frac{\partial^* L}{\partial \psi^A{}_{;a}} \right)_{;a} \right] \mathcal{L}_X \psi^A + {}^* \left(T^{ab} \mathcal{L}_X g_{ab} \right) \right\}.$$

The first term vanishes as a consequence of the equations of the fields. In the second term,

$$\mathcal{L}_X g_{ab} = 2X_{(a;b)}.$$

Thus

$$0 = \int_D {}^* \left(T^{ab} X_{a;b} \right) = \int_D {}^* \left[\left(T^{ab} X_a \right)_{;b} - T^{ab}{}_{;b} X_a \right].$$

The first term may be transformed into an integral over the boundary of D which vanishes as X is zero there. Since the second term must be zero for arbitrary X, it follows that

$$T^{ab}{}_{;b} = 0.$$

We shall give two examples which illustrate methods by which the equations of the fields and the energy-momentum tensor can be derived from a Lagrangian. First, the electromagnetic field (without sources). This is described by a covariant vector A called the potential. The electromagnetic field tensor F is defined as $\mathrm{d}A$. In component form, one has $F_{ab} = A_{[a;b]}$. The Lagrangian is taken to be

$$-\frac{1}{2} F_{ab} F^{ab}.$$

Then requiring the action to be stationary, we have

$$F^{ab}{}_{;b} = 0.$$

Since the field tensor F is the exterior derivative of A, it follows that $\mathrm{d}F = 0$. In component form this is $F_{[ab;c]} = 0$. This and the above equation are known as the Maxwell equations for the source-free field. Varying the metric, we have

$$\left. \frac{\partial I}{\partial u} \right|_{u=0} = - \int_D \left(\eta F_{ab} F_{cd} g^{bd} \Delta g^{ac} + \frac{1}{2} \Delta \eta F_{ab} F_{cd} g^{ac} g^{bd} \right),$$

where

$$\eta_{abcd} = \sqrt{-g} \delta^{1234}_{abcd}$$

is the canonical 4-form. So

$$\Delta\eta_{abcd} = -\frac{1}{2}g_{ef}\Delta g^{ef}\eta_{abcd}.$$

But

$$\Delta g^{ef} = -g^{ea}g^{fb}\Delta g_{ab}.$$

Thus

$$T^{ab} = F^{ac}F^{b}{}_{c} - \frac{1}{4}g^{ab}F^{cd}F_{cd}.$$

One may note the following interesting properties of this energy-momentum tensor: it has zero contraction, i.e., $T^{a}{}_{a} = 0$, and if W is any timelike vector, then

$$T_{ab}W^{a}W^{b} > 0.$$

This may be interpreted as meaning that the energy density is positive to any observer. The significance of this will become apparent in Section 6.

The second example of a Lagrangian will be that for an isentropic perfect fluid. The technique here is rather different. We shall consider the fluid in a region D of M to be described by a function ρ called the density and a congruence of timelike curves which are called flow lines. By a congruence of curves we mean a diffeomorphism

$$\gamma : [a, b] \times N \longrightarrow D,$$

where $[a, b]$ is some closed interval of $\mathbb{R}$ and N is some three-dimensional manifold. The curves are said to be timelike if their tangent vector

$$W = \left(\frac{\partial}{\partial t}\right)_{\gamma}, \quad t \in [a, b],$$

is timelike everywhere. We will define the tangent vector V as $g(W, W)^{-1/2}W$ and the fluid current vector J as ρV. This is required to be conserved, that is, $J^{a}{}_{;a} = 0$. We also introduce the elastic potential π, which is some function of the density ρ. This may be thought of as the potential energy per unit density ρ or as the specific enthalpy. The Lagrangian L is taken to be $2\rho(1 + \pi)$ and the action I is required to be stationary under variation of the flow lines. A variation α of the flow lines is a differentiable map

$$\alpha : (-\varepsilon, \varepsilon) \times [a, b] \times N \longrightarrow D,$$

such that

$$\alpha\big(0, [a, b], N\big) = \gamma\big([a, b], N\big).$$

It then follows that $\Delta W = \mathcal{L}_K W$, where the vector $K = (\partial/\partial u)_{\alpha}$, $u \in (-\varepsilon, \varepsilon)$ may be thought of as representing the displacement under the variation of a point on the flow line. We have

$$\Delta V^{a} = V^{a}{}_{;b}K^{b} - K^{a}{}_{;b}V^{b} - V^{a}V_{c}K^{c}{}_{;b}V^{b},$$

and using the fact that $(\Delta J^{a})_{;a} = 0$,

$$\Delta\rho = \big(\rho K^{b}\big)_{;b} - \rho K_{b;c}V^{b}V^{c}.$$

Substituting this into the action integral,

$$\left.\frac{\partial I}{\partial u}\right|_{u=0} = 2\int_{D}{}^{*}\left\{\left[(\rho K^{b})_{;b} - \rho K_{b;c}V^{b}V^{c}\right]\left[1 + \frac{\mathrm{d}(\rho\pi)}{\mathrm{d}\rho}\right]\right\}.$$

Integrating by parts, we have

$$\left.\frac{\partial I}{\partial u}\right|_{u=0} = 2\int_D {}^* \left(\left\{ \rho \left[1 + \frac{\mathrm{d}(\rho\pi)}{\mathrm{d}\rho} \right] \dot{V}^a - \rho \left[\frac{\mathrm{d}(\rho\pi)}{\mathrm{d}\rho} \right]_{;c} (g^{ac} - V^a V^c) \right\} K_a \right).$$

Thus

$$(\mu + \rho)\dot{V}^a = p_{;b} \left(g^{ab} - V^a V^b \right),$$

where $\mu = \rho(1+\pi)$ is called the energy density and $p = \rho^2 \mathrm{d}\pi/\mathrm{d}\rho$ is called the pressure.

To obtain the energy-momentum tensor, one varies the metric. The calculations may be simplified by noting that the conservation of the current may be expressed as

$$J^a{}_{;a} = \frac{1}{\sqrt{-g}} \frac{\partial}{\partial u^a} \left(\sqrt{-g} J^a \right) = 0.$$

Thus $\sqrt{-g}J^a$ is unchanged when the metric is varied. We have

$$\rho^2 = \frac{1}{-g} \left(\sqrt{-g} J^a \sqrt{-g} J^b \right) g_{ab},$$

so

$$2\rho\Delta\rho = \left(J^a J^b - J^c J_c g^{ab} \right) \Delta g_{ab}$$

and

$$\begin{aligned} T^{ab} &= \left[\rho(1+\pi) + \rho^2 \frac{\mathrm{d}\pi}{\mathrm{d}\rho} \right] V^a V^b - \rho^2 \frac{\mathrm{d}\pi}{\mathrm{d}\rho} g^{ab} \\ &= (\mu + \rho) V^a V^b - p g^{ab}. \end{aligned}$$

We shall call a perfect fluid any matter whose energy-momentum tensor is of the above form, whether or not it is derived from a Lagrangian. From the conservation of energy-momentum, we have

$$\begin{aligned} \mu_{;a} V^a + (\mu + p) V^a{}_{;a} &= 0, \\ (\mu + p)\dot{V}^a - (g^{ab} - V^a V^b) p_{;b} &= 0. \end{aligned}$$

These are the same as the equations derived from a Lagrangian. We shall call a perfect fluid isentropic if the pressure p is a function of the energy density only. In this case we can introduce a conserved density ρ and a potential π and derive the equations and the energy-momentum tensor from a Lagrangian.

3.3 General relativity

We still have to specify the metric that was introduced in Section 3.1. In the special theory of relativity the metric is fixed by imposing the requirement that its Riemann tensor vanishes. However, we cannot adopt this requirement if we wish to include gravitational effects. For as we have seen the equations of energy-momentum conservation imply that a small isolated body moves on a geodesic. But if the metric were flat there would not be a geodesic going round the Sun the way our planet does. Thus we must relate the curvature of the metric to the matter distribution in such a way as to describe the observed gravitational effects.

In choosing this relationship, we will be guided by the four following principles:

- First, by requirement (2), it must be expressible as a relation between tensors only.

- Second it should not involve derivatives higher than the second of any field or the metric, as all our experience has led us to believe that we need to specify only the initial values of fields and their first derivatives to specify the future development of a system.
- Third, it should incorporate the Newtonian principle that active gravitational mass (the mass producing a gravitational field) and passive gravitational mass (the mass acted on by a gravitational field) are equal. If this were not so we could couple together two bodies with the same passive but different active gravitational mass and obtain a combination that would continuously accelerate, which would be contrary to experience. We saw above that passive gravitational mass and inertial mass are both described in the theory of relativity by the energy-momentum tensor. Thus this principle requires that the source of the curvature should be the energy-momentum tensor only.
- Lastly, of course, the relationship must predict results for weak fields similar to those of Newtonian theory.

Principles one, two, and three imply that the relationship must be of the form $K_{ab} = T_{ab}$, where K_{ab} is a symmetric tensor which depends on the curvature and the metric tensors only. Since T_{ab} satisfies the equations $T^{ab}{}_{;b} = 0$, we see that K^{ab} must satisfy $K^{ab}{}_{;b} = 0$. The analogy with Newtonian theory suggests that K^{ab} ought to be linear in second derivatives of the metric tensor, as this corresponds to the Newtonian potential. In this case it can be shown that the only possible tensor satisfying all these requirements is

$$K^{ab} = \gamma \left(R^{ab} - \frac{1}{2} g^{ab} R + \lambda g^{ab} \right),$$

where R^{ab} is the Ricci tensor, R is the curvature scalar, and γ and λ are constants.

Even if one did not require that K^{ab} be linear in second derivatives of the metric tensor, there would not seem to be any other tensors satisfying the other conditions, though the author knows no proof of this. This point is important since the results which will be obtained in Section 6 seem to show that singularities will occur if the equations above hold. Thus one might ask whether there was any alternative relation which would not lead to singularities. However, there do not appear to be any having all the properties that seem desirable.

It may be, of course, that we will have to abandon one or more of those properties. Thus we might allow third or higher derivatives of the metric to appear in the relations. However, experience with the radiation damping force in electrodynamics suggests that, when we have higher derivatives, we get unphysical 'runaway' solutions. Thus we might be no better off. On the other hand one might abandon the equations of conservation of energy-momentum. However, this would mean that T^{ab} could no longer be derived from a Lagrangian. Indeed it would be difficult to see what meaning energy and momentum could have if they were not conserved in some sense.

There might be a further difficulty as follows. The ten equations

$$\gamma \left(R^{ab} - \frac{1}{2} g^{ab} R + \lambda g^{ab} \right) = T^{ab}$$

are not all independent since there are the four identities

$$\gamma \left(R^{ab} - \frac{1}{2} g^{ab} R + \lambda g^{ab} \right)_{;b} = 0 = T^{ab}{}_{;b}.$$

They are however sufficient since the ten components g_{ab} cannot be completely determined by the equations as there must always be four degrees of freedom to make

coordinate transformations. If, however, one had a relation between T^{ab} and a tensor K^{ab} which did not satisfy any identities, the system would be overdetermined unless T^{ab} satisfied only six independent equations as a consequence of the equations of the fields. These would have to be chosen so as to be compatible with the tensor and this might be difficult to arrange. For example, empty space would not in general be compatible.

Requirement 3. Field equations. The Einstein equations hold:

$$\gamma\left(R^{ab}-\frac{1}{2}g^{ab}R+\lambda g^{ab}\right)=T^{ab}.$$

These equations may be derived by requiring that the action

$$I=\int_D{}^*\big[\gamma(R-2\lambda)+L\big]$$

be stationary under variation of g_{ab}. For,

$$\Delta\big[{}^*(R-2\lambda)\big]=\Delta\big[\eta(R-2\lambda)\big],$$

where η is the canonical 4-form, whence

$$\Delta\left[{}^*(R-2\lambda)\right]=(R-2\lambda)\Delta\eta+\eta\left(R_{ab}\Delta g^{ab}+g^{ab}\Delta R_{ab}\right).$$

The last term can be written as

$$\begin{aligned}\eta g^{ab}\Delta R_{ab}&=\eta g^{ab}\left[(\Delta\Gamma^c_{ab})_{;c}-(\Delta\Gamma^c_{ac})_{;b}\right]\\&=\eta\left(\Delta\Gamma^c_{ab}g^{ab}-\Delta\Gamma^d_{ad}g^{ac}\right)_{;c}.\end{aligned}$$

Thus it may be transformed into an integral over the boundary ∂D, which vanishes as $\Delta\Gamma^a_{bc}$ vanishes on the boundary. The remaining terms then give the Einstein equations.

One might ask whether varying an action derived from some other scalar combination of the metric and curvature tensors might not give an alternative set of equations. However, the curvature scalar is the only such scalar linear in second derivatives of the metric tensor. This allows one to transform away a surface integral and be left with an equation involving only second derivatives of the metric. If one tried any other scalar such as $R_{ab}R^{ab}$ or $R_{abcd}R^{abcd}$, one would obtain an equation involving fourth derivatives of the metric.

It remains, of course, to show that the Einstein equations lead to results similar to Newtonian theory for weak, almost static fields. This may be done as follows. Suppose the metric is static. By this we mean that there exists a timelike Killing vector W, with $W_{(a;b)}=0$, which is proportional to the gradient of a scalar t ($W_{a;b}W_c\eta^{abcd}=0$). We define the unit timelike vector V as $f^{-1}W$, where $f^2=W^aW_a$. Then $V^a{}_{;b}=\dot{V}^aV_b$, where $\dot{V}^a=V^a{}_{;b}V^b=f^{-1}f_{;b}g^{ab}$ represents the acceleration of the integral curves of V. Then

$$\begin{aligned}\dot{V}^a{}_{;a}&=V^a{}_{;ba}V^b+V^a{}_{;b}V^b{}_{;a}\\&=R_{cb}V^cV^b.\end{aligned}$$

The integral curves of V define the static frame of reference. That is to say, a particle travelling on one of these curves would appear to remain at rest. Thus a particle released from rest and following a geodesic would appear to have an initial acceleration

of $-\dot{V}$ with respect to the static frame. If f differs only slightly from a constant f_0, we may put $f = f_0(1+\phi)$, where ϕ is small. Then $\dot{V}_a \approx \phi_{;a}$. Thus the initial acceleration of a freely moving particle released from rest is minus the gradient of ϕ. This suggests that we should regard ϕ as the quantity analogous to the Newtonian potential.

We have

$$\phi_{;ab}g^{ab} \approx R_{ab}V^aV^b = 2\lambda + \gamma^{-1}\left(T_{ab}V^aV^b - \frac{1}{2}T^a{}_a\right).$$

Suppose for simplicity that the energy-momentum tensor is that of a perfect fluid:

$$T^{ab} = (\mu + p)V^aV^b - pg^{ab},$$

where μ is the energy density and p the pressure of the fluid. Then

$$\phi_{;ab}g^{ab} \approx 2\lambda + \frac{1}{2}\gamma^{-1}(\mu + 3p).$$

The term on the left is the Laplacian of ϕ with respect to the induced metric in the surface $t = \text{const.}$. If the metric is almost flat, this will correspond to the Newtonian Laplacian of the potential. On the right, the pressure p is normally small compared to the density μ and may be neglected. Thus we obtain approximate agreement with Newtonian theory if λ is small or zero and $\gamma^{-1}/2$ is equal to $4\pi G$, where G is the Newtonian gravitational constant. By choosing units of mass appropriately, we can arrange for γ to equal one. As λ must anyway be small, we shall generally take it to be zero, though we shall bear in mind the possibility of other values.

3.4 Some exact solutions

By an exact solution we mean a manifold and metric which comply with postulates (a) and (b) and which satisfy the Einstein equations for some specified form of the energy-momentum tensor. However, to make the calculations humanly possible, it is necessary to assume that the metric has a high degree of symmetry and that the energy-momentum tensor has some very simple form which, at best, can be regarded as only an approximation to that of matter in the universe, and which may fail to be even that under extreme conditions. Thus exact solutions tend to be unrealistic in two ways. Nevertheless, they are of interest because they may be reasonable approximations to certain regions of spacetime and because they give examples of certain global properties that spacetime could have. However, one should be cautious about assuming that more realistic solutions would necessarily also have these properties.

The first and simplest example for empty space (zero energy-momentum tensor) is Minkowski space. This is just the manifold $\mathbb{R}^4$ with a flat Lorentz metric. Taking u^1, u^2, u^3, u^4 as coordinates, the metric may be expressed as

$$\mathrm{d}s^2 = \left(\mathrm{d}u^4\right)^2 - \left(\mathrm{d}u^3\right)^2 - \left(\mathrm{d}u^2\right)^2 - \left(\mathrm{d}u^1\right)^2.$$

These coordinates cover the entire manifold. Another form is obtained using coordinates t, r, θ, ϕ, related to u^1, u^2, u^3, u^4 by

$$u^4 = t, \quad u^3 = r\cos\theta, \quad u^2 = r\sin\theta\cos\phi, \quad u^1 = r\sin\theta\sin\phi,$$

where $-\infty < t < \infty$, $0 \le r < \infty$, $0 \le \theta \le \pi$, $0 \le \phi < 2\pi$. Then the metric takes the form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - \mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

Note that the metric is apparently singular for $r = 0$ and for $\sin\theta = 0$. This is because t, r, θ, ϕ are not admissible coordinates at these points. This is easily recognisable in this case, though it is not always so simple to tell that an apparent singularity in the metric is just due to a bad choice of coordinates.

Another, very interesting, representation of Minkowski space has been given by Penrose [Penrose 1965a]. We introduce new coordinates v and w defined as

$$2v = t + r, \qquad 2w = t - r,$$

and the metric becomes

$$\mathrm{d}s^2 = \mathrm{d}v\mathrm{d}w - (v-w)^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

The 3-surfaces $v = \text{const.}$ or $w = \text{const.}$ are null, by which we mean that

$$g^{ab}v_{;a}v_{;b} = 0, \qquad g^{ab}w_{;a}w_{;b} = 0.$$

We now introduce new 'null' coordinates p and q defined by

$$\tan p = v, \quad \tan q = w, \quad -\frac{\pi}{2} < p < \frac{\pi}{2}, \quad -\frac{\pi}{2} < q < \frac{\pi}{2}, \quad 0 \le p - q.$$

Then the metric is expressed as

$$\mathrm{d}s^2 = \sec^2 p \sec^2 q\left[\mathrm{d}p\mathrm{d}q - \sin^2(p-q)\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right)\right].$$

Thus the physical metric g is conformal to the metric $\tilde{g}$ given by

$$\mathrm{d}\tilde{s}^2 = \mathrm{d}p\mathrm{d}q - \sin^2(p-q)\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

This can be put in a more recognisable form by introducing new coordinates t^1 and r^1 defined by

$$t^1 = p + q, \qquad r^1 = p - q.$$

Then

$$\mathrm{d}s^2 = \left(\mathrm{d}t^1\right)^2 - \left(\mathrm{d}r^1\right)^2 - \sin^2 r^1\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

This metric has only been given on the manifold defined by

$$-\pi < t^1 - r^1 < \pi, \quad -\pi < t^1 + r^1 < \pi, \quad r^1 > 0.$$

It can, however, be analytically extended to the manifold $\mathbb{R}^1 \times \mathbb{S}^3$, where $-\infty < t^1 < \infty$ and r^1, θ, ϕ are regarded as coordinates on $\mathbb{S}^3$. The apparent singularities in the metric at $r^1 = 0$ and $r^1 = \pi$ are similar to the singularity at the origin of polar coordinates. They could be removed by transforming to local coordinates in some regions of those points.

This shows that Minkowski space is conformal to the interior of the region of $\mathbb{R}^1 \times \mathbb{S}^3$ shown in Figure 1. We may think of the boundary of this region as representing the conformal structure of infinity of Minkowski space. We see that it consists of the null surfaces $p = \pi/2$ (labelled $\mathcal{J}^+$) and $q = -\pi/2$ (labelled $\mathcal{J}^-$) and the points $p = \pi/2, q = \pi/2$ (labelled $\mathcal{I}^+$), $p = \pi/2, q = -\pi/2$ (labelled $\mathcal{I}^0$), and $p = -\pi/2, q = -\pi/2$ (labelled $\mathcal{I}^-$). The image of any timelike geodesic in Minkowski space will approach $\mathcal{I}^+$ and $\mathcal{I}^-$, while that of a null geodesic will approach $\mathcal{J}^+$ and $\mathcal{J}^-$, and both ends of the image of a spacelike geodesic will approach $\mathcal{I}^0$. Thus we may think of these as representing timelike, null, and spacelike infinity, respectively.

This representation is a bit difficult to grasp as it is hard to visualise objects in four dimensions or to draw diagrams of them. However, it can be simplified using the

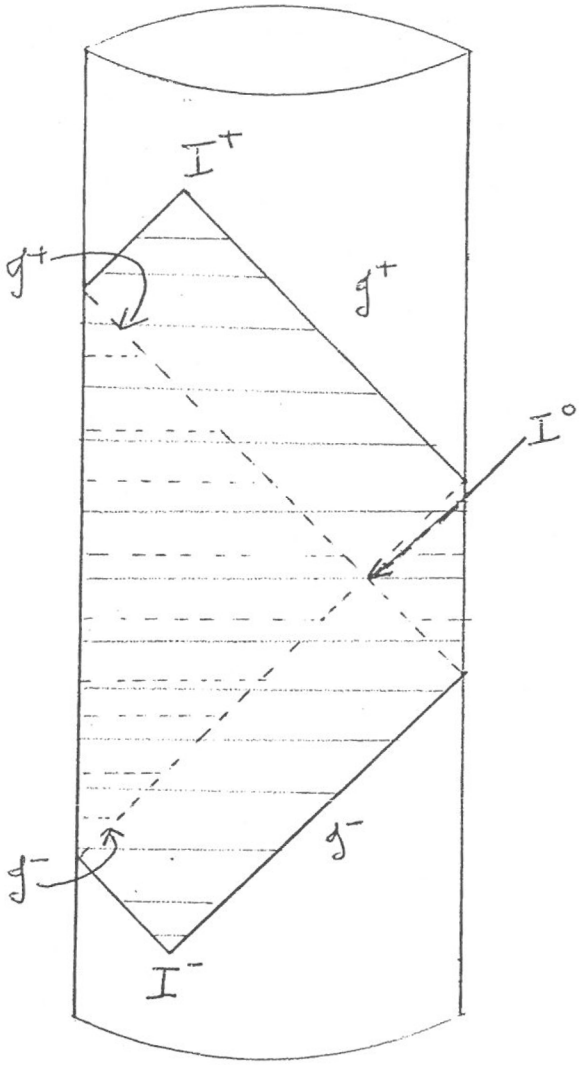

Fig. 1. The cylinder represents $\mathbb{R}^1 \times \mathbb{S}^3$ where two spatial dimensions have been suppressed. The *shaded region* is the part conformal to Minkowski space.

fact that Minkowski space, as all the other solutions to be described, has spherical symmetry. Thus it is sufficient to consider only the geometry of the $t - r$ plane and to regard each point of this plane as representing an $\mathbb{S}^2$ whose radius is the value of r at the point. As any metric on a two-dimensional manifold is locally conformal to a flat metric, this geometry can be represented by a diagram in which null geodesics run at $\pm 45°$ to the vertical. Such a representation will be called a Penrose diagram. That for Minkowski space is shown in Figure 2. We shall adopt the convention that boundaries representing infinity will be denoted by single lines, those representing the origin of polar coordinates by a dotted line, and those representing irremovable singularities of the metric by double lines.

The Schwarzschild solution which will be described next represents the spherically symmetric gravitational field outside some massive body. All the experiments which have been carried out to test differences between the general theory of relativity and

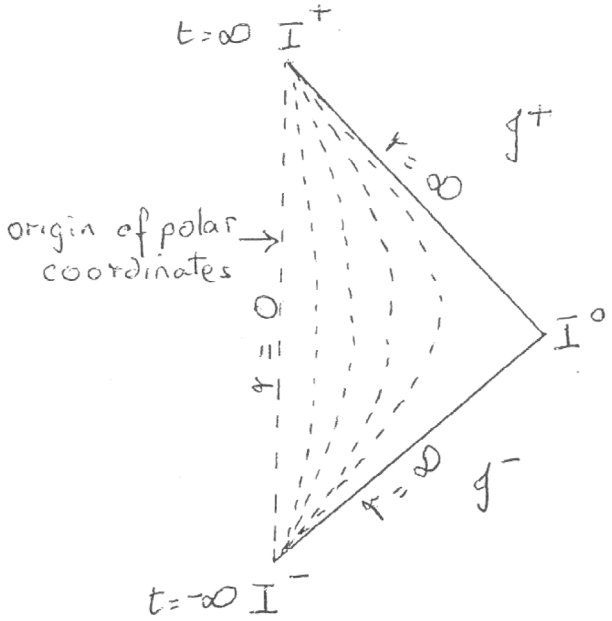

Fig. 2. The Penrose diagram of the $t-r$ plane of Minkowski space. The *dotted lines* represent curves of constant r.

Newtonian theory are based on predictions by this solution. The metric has the form:

$$\mathrm{d}s^2 = \left(1 - \frac{2m}{r}\right)\mathrm{d}t^2 - \left(1 - \frac{2m}{r}\right)^{-1}\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

It can be seen that this is static, i.e., $\partial/\partial t$ is a Killing vector. Comparison with Newtonian theory shows that m should be regarded as the gravitational mass, as measured from infinity, of the body producing the field.

Normally one would regard the above metric as being the solution only outside some spherical body, that is, for r greater than some value, and that inside the body the metric would have a different form. However, it is interesting to see what happens when the metric is thought of as being an empty space solution for all values of $r > 0$. There is an apparent singularity in the metric when $r = 2m$. However, this is simply due to a bad choice of coordinates. We can introduce a new advanced time coordinate v defined by

$$v = t + r + 2m\log(r - 2m).$$

The metric then takes the form

$$\mathrm{d}s^2 = \left(1 - \frac{2m}{r}\right)\mathrm{d}v^2 - 2\mathrm{d}r\mathrm{d}v - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right).$$

This has only been given on the manifold for which $r > 2m$, but clearly it may be analytically extended to give a nonsingular metric on the manifold for which $r > 0$. Similarly, we could extend the original solution another way by introducing a retarded time coordinate

$$w = t - \left[r + 2m\log(r - 2m)\right].$$

Using a combination of such extensions, one may obtain the maximal analytic extension, which was found by Kruskal [Kruskal 1960]. The Penrose diagram of its $t - r$

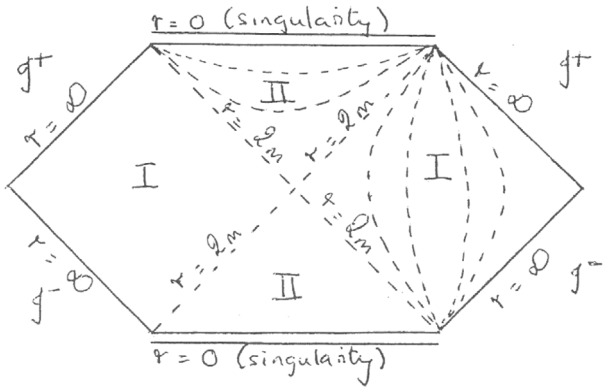

Fig. 3. The $t - r$ plane of the Schwarzschild solution.

plane is shown in Figure 3. This has some very interesting features. One sees that there are two exterior regions where $r > 2m$ (labelled I). As r tends to infinity the metric tends to that of Minkowski space and the boundary at infinity is null, as for Minkowski space. The two exterior regions are joined together by two interior regions where $r < 2m$ (labelled II). There are two singularities corresponding to $r = 0$: one in the past and one in the future. As one approaches them, the scalar $R_{abcd}R^{abcd}$ tends to infinity. Thus they are true singularities of the metric and cannot be removed by choosing different coordinates. The surface $r = 2m$ is null and is called the Schwarzschild surface. It has the property that any matter crossing it inevitably hits the singularity at $r = 0$.

The Reissner-Nordström solution represents the field outside a spherically symmetric body carrying an electric charge. The metric is rather similar to that of the Schwarzschild solution:

$$\mathrm{d}s^2 = \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)\mathrm{d}t^2 - \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)^{-1}\mathrm{d}r^2 - r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right),$$

where m represents the gravitational mass and e the charge of the body. The above metric may be regarded as the solution outside some body. However, as in the case of the Schwarzschild solution, it is interesting to see what happens if we regard it as the solution for all values of r. If $e^2 > m^2$, the metric is nonsingular everywhere except at $r = 0$, where there is an irremovable singularity. This may be thought of as representing a point charge which produces the field. If $e^2 \leq m^2$, the metric has apparent singularities at $r = r_+$ and $r = r_-$, where $r_\pm = m \pm \sqrt{m^2 - e^2}$. As in the Schwarzschild case, these may be removed by introducing suitable coordinates and extending the manifold. The maximal analytic extension has been obtained by Graves and Brill [Graves 1960] for the case $e^2 < m^2$ and by Carter [Carter 1966] for $e^2 = m^2$. The Penrose diagrams of their $t - r$ planes are shown in Figure 4. One sees that there is now an infinite series of exterior regions where $r > r_+$ (labelled I), joined together by intermediate regions where $r_- < r < r_+$ (II) and interior regions where $r < r_-$ (III). There are still irremovable singularities where $r = 0$.

In the earliest cosmologies, man, as lord of creation, placed himself firmly at the centre. However, since the time of Copernicus, we have been successively demoted to a medium-sized planet going round a medium-sized star on the outer edge of a fairly average galaxy which is itself simply part of the local group of galaxies. Indeed, we are now so modest that we would claim that our position was in no way specially distinguished. We shall call this the *Copernican principle*, after Bondi [Bondi 1952].

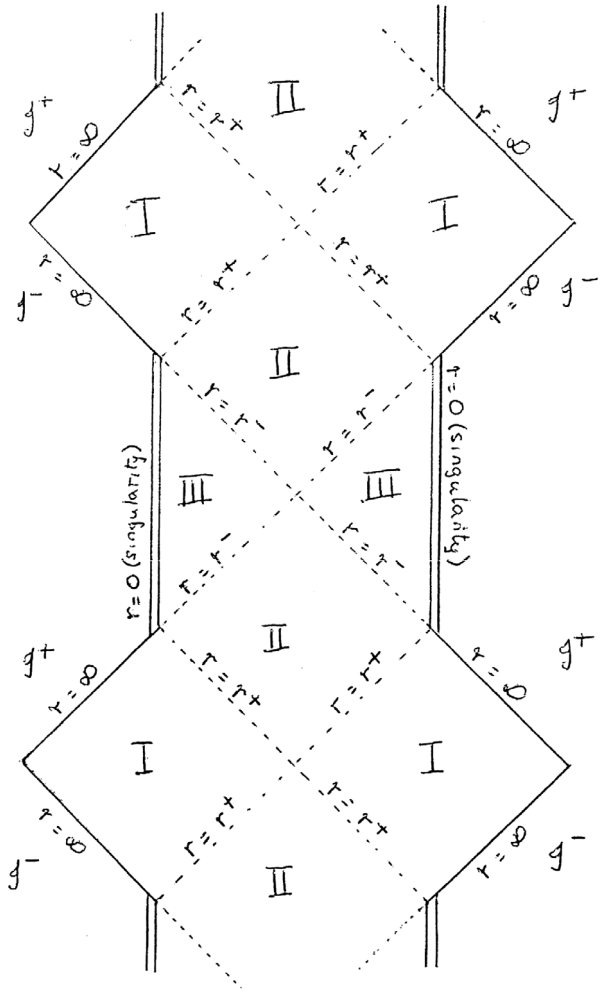

Fig. 4. The $t - r$ plane of the Reissner-Nordström solution for $e^2 < m^2$.

This would seem to rule out the metric of spacetime being asymptotically flat as in the Schwarzschild and Reissner-Nordström solutions. For such a metric, we would have to be near the centre. This is not to say that such metrics cannot be reasonable approximations in the vicinity of some massive body, but they could not be taken to represent the whole of spacetime.

The Copernican principle as stated is somewhat vague. However, it would seem reasonable to interpret it as implying that the universe is approximately spatially homogeneous. By spatially homogeneous, we mean that there is a three-parameter Lie group of isometries which acts freely on M and whose surfaces of transitivity are spacelike 3-surfaces. In other words any point on one of these surfaces would be equivalent to any other point on the same surface. Of course, the universe is not exactly spatially homogeneous. There are local irregularities like stars and galaxies. Nevertheless it might seem reasonable to suppose that the universe was spatially homogeneous on a large enough scale. It is difficult to test homogeneity directly by observation because of the lack of any simple way of measuring the separation from us of distant objects. However, observations seem to indicate that the universe is approximately

spherically symmetric about us. Unless we assume that we occupy a special position in the universe, we must conclude that the universe will be approximately spherically symmetric about every point.

As has been shown by Walker [Walker 1944], exact spherical symmetry about every point would imply that the universe was spatially homogeneous and admitted a six-parameter Lie group of isometries whose surfaces of transitivity are spacelike 3-surfaces of constant curvature. The metric would have the Robertson-Walker or Friedmann form

$$\mathrm{d}s^2 = \mathrm{d}t^2 - S^2(t)\left[\frac{\mathrm{d}r^2}{1-Kr^2} + r^2\left(\mathrm{d}\theta^2 + \sin^2\theta\mathrm{d}\phi^2\right)\right],$$

where the quantity K is minus one, zero, or one according to whether the 3-surfaces $t = \text{const.}$ have negative, zero, or positive constant curvature, respectively. They are diffeomorphic to $\mathbb{R}^3$ in the first and second cases and to $\mathbb{S}^3$ in the third. In this case, the above coordinates are admissible over only half the surface, but one could use a combination of such coordinate neighbourhoods to cover the whole surface. Of course, one could identify suitable points in these surfaces. It would be possible to do this even for the negative curvature case in such a way that the resultant surface was compact. However, such a compact surface of constant negative curvature would have no continuous group of isometries [Yano 1953]. Thus there would seem little point in making such an identification, as the original reason for considering this class of solutions was that they had a six-parameter group of isometries. In fact, the only identification which would not reduce the dimension of the isometry group would be to identify antipodal points on $\mathbb{S}^3$ in the case of a surface of positive constant curvature.

The symmetry of these Robertson-Walker solutions requires that the energy-momentum tensor have the form of that of a perfect fluid whose density μ and pressure p are functions of the coordinate t only and whose flow times are the curves r, θ, ϕ constant. This fluid should be thought of as a smeared out approximation to the matter in the universe. The function $S(t)$ represents the separation of neighbouring flow lines. By the Einstein equations, we have

$$3\frac{\dot{S}^2 + K}{S^2} + \lambda = \mu, \qquad \frac{2S\ddot{S} + \dot{S}^2 + K}{S^2} + \lambda = -p,$$

where a dot indicates differentiation with respect to t. For completeness, we have included the possibility of λ being nonzero.

It would be reasonable to assume that μ is positive and that p is non-negative. Then if λ is zero, it can be seen that S could not be constant. In other words, the universe would be either expanding or contracting. In fact, observations of other galaxies indicate that they are moving away from us, hence that the universe is expanding at the present time.

Eliminating K between the first and second equations, we obtain

$$\dot{\mu} = -3(\mu + p)\frac{\dot{S}}{S}, \qquad \frac{1}{2}(\mu + 3p) - \lambda = -\frac{2\ddot{S}S}{S^2}.$$

The first equation could have been obtained directly from conservation of energy-momentum. It shows that the density decreases as the universe expands, as one would expect. From the second equation, one can see that S must have been zero a finite time ago if $\mu + 3p - 2\lambda$ is positive. This would be a real singularity of the metric as the density and hence some components of the Ricci tensor would be infinite there. This singularity is the most striking feature of the Robertson-Walker solutions. It

would imply that the universe (or at least that part of which we can have physical knowledge) had a beginning a finite time ago. However, it must be emphasised that this result depended on assuming exact spatial homgeneity and spherical symmetry. While spatial homogenity and spherical symmetry may be reasonable approximations on a large enough scale at the present time they certainly do not hold locally. One might think that, as one traced the evolution of the universe back in time, the local irregularities would grow and could prevent the occurrence of singularity, causing the universe to 'bounce' instead. Whether this could happen or whether physically realistic solutions with inhomogeneities would contain singularities is a question of great importance for cosmology and constitutes the principal problem dealt with in this essay. In Section 6, it will be shown that singularities are inevitable in solutions which satisfy certain reasonable global conditions and in which the energy-momentum tensor satisfies a reasonable inequality.

If the relation between p and μ is specified, the Einstein equation can be solved to give S as a function of the time t. In fact, the pressure is very small at the present epoch. If we take it and λ to be zero, we have the first integrals

$$\mu = \frac{M}{S^3}, \qquad 3M\dot{S}^2 - \frac{M^2}{S} = E,$$

where M and E are constants. The first equation expresses the conservation of mass when the pressure is zero. In the second equation, we may think of E as representing the kinetic plus the potential energy of the matter. If E is negative, S will increase to some maximum value and then decrease to zero. if E is positive or zero, S will increase indefinitely. E is related to the quantity K by $K = E/3M$. If we introduce a function $\tau(t)$ given by

$$\frac{\mathrm{d}\tau}{\mathrm{d}t} = \frac{1}{S},$$

we have the three cases:

$$\begin{aligned}
&K = -1, \quad S = \frac{E}{18}(\cosh\tau - 1), \quad t = \frac{E}{18}(\sinh\tau - \tau),\\
&K = 0 \ \text{ (Einstein-de Sitter solution)}, \quad S = t^{2/3},\\
&K = +1, \quad S = -\frac{E}{18}(1 - \cos\tau), \quad t = -\frac{E}{18}(\tau - \sin\tau).
\end{aligned}$$

The Penrose diagrams of their $t - r$ planes are shown in Figure 5.

We shall also describe three other spaces which are interesting for their global properties but which probably do not correspond to anything occurring in nature. The first two are the Lorentz spaces of constant curvature. For these,

$$R_{abcd} = F(g_{ac}g_{bd} - g_{ad}g_{bc}),$$

where F is a constant. Then,

$$R_{ab} - \frac{1}{2}g_{ab}R = -3Fg_{ab}.$$

They may be regarded as solutions for $\lambda = 3F$ and no matter, or for $\lambda = 0$ and a perfect fluid with density $-3F$ and pressure $3F$. However, the latter would seem unreasonable physically as it would necessitate either negative density or negative pressure. The constant curvature space for F zero is of course Minkowski space. That for F negative is known as de Sitter space. It has the topology $\mathbb{R}^1 \times \mathbb{S}^3$ (the possible

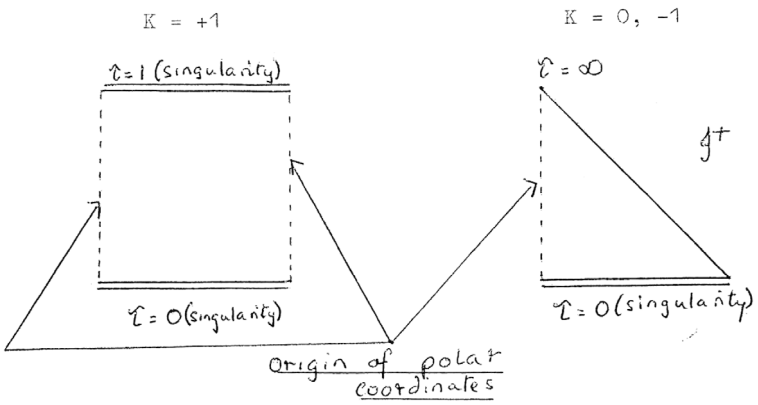

Fig. 5. The $t - r$ planes of the Robertson-Walker solutions for $p = 0 = \lambda$.

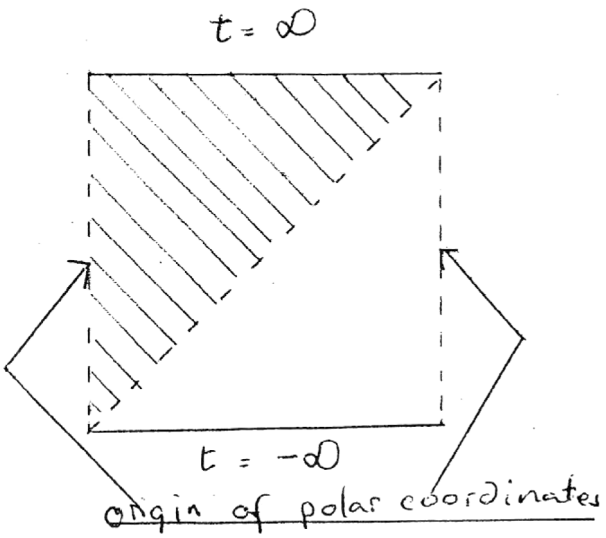

Fig. 6. The $t - r$ plane of de Sitter space. The *shaded region* is the part representing the steady-state universe.

identifications have been studied by Calabi and Markus [Calabi 1962]). It can be represented in the Robertson-Walker form with $K = 1$ and

$$S = \frac{1}{\sqrt{-F}} \cosh \sqrt{-F}t.$$

Although it is geodesically complete, there are pairs of points which cannot be joined by any geodesic. This is in contrast to the case for a geodesically complete positive definite metric, where there is at least one geodesic between each pair of points.

Another form of the metric of de Sitter space is obtained by taking $K = 0$ and $S = e^{\sqrt{-F}t}$ (see Fig. 6). This metric covers only half of de Sitter space and so is geodesically incomplete in the past. It was proposed as a model of the steady-state universe by Bondi and Gold [Bondi 1948] and by Hoyle [Hoyle 1948]. In this theory the flow lines of the matter are taken to be the curves r, θ, ϕ constant. As the universe expands and the matter moves further apart it is assumed that more matter is continuously created to maintain the density at a constant value. Bondi and Gold did not seek to provide

field equations for this theory, but Hoyle and Narlikar [Hoyle 1964a] have pointed out that the ordinary Einstein equations (with $\lambda = 0$) can be satisfied if, in addition to the ordinary matter, one introduces a scalar field of negative energy density. This C field would also be responsible for the continuous creation of matter.

The steady-state theory has the advantage of making simple and definite predictions. However, these do not seem to be in agreement with current observations, so the theory has largely (and regretfully) been abandoned.

The space of constant curvature for which F is positive is called the anti-de Sitter space. It has the topology $\mathbb{R}^3 \times \mathbb{S}^1$ and so has a covering space with topology $\mathbb{R}^4$. The metric can be represented in the Robertson-Walker form with $K = -1$ and

$$S = \frac{1}{\sqrt{F}} \cos \sqrt{F} t.$$

However, this only covers part of the space and has apparent singularities at

$$t = \pm \frac{\pi}{2\sqrt{F}}.$$

These can be removed by taking a different set of coordinates t', r', θ, ϕ, for which the metric has the form

$$\mathrm{d}s^2 = \frac{1}{\sqrt{3F}} \cosh^2 \sqrt{3F} r' \, \mathrm{d}t'^2 - \mathrm{d}r'^2 - \sinh^2 r' (\mathrm{d}\theta^2 + \sin^2 \theta \mathrm{d}\phi^2).$$

This covers the whole space. The Penrose diagram of its $t' - r'$ plane is shown in Figure 7. The timelike geodesics which are orthogonal to a surface $t' = \text{const.}$ intersect at points to the past and future of the surface. As these are the curves r, θ, ϕ constant in the Robertson-Walker coordinates, one can see why the apparent singularities arise. One can also see that there are pairs of points which can be joined by a timelike curve but not be a timelike geodesic.

The last example to be described is a flat two-dimensional space due to Misner [Misner 1965]. This is diffeomorphic to $\mathbb{R}^1 \times \mathbb{S}^1$. In other words, it is a cylinder. The coordinates are t $(-\infty < t < \infty)$ and ϕ $(0 \leq \phi \leq 2\pi)$, where $\phi = 0$ and $\phi = 2\pi$ are identified. The metric is

$$\mathrm{d}s^2 = 2\mathrm{d}\phi \, \mathrm{d}t + t \, \mathrm{d}\phi^2.$$

If we regarded this metric as being given only on the portion of the cylinder for which $t > 0$, we could obviously extend it analytically to get a non-degenerate metric on the entire manifold $\mathbb{R}^1 \times \mathbb{S}^1$ described by the coordinates $-\infty < t < \infty$ and $0 \leq \phi \leq 2\pi$. This extension would be maximal in the sense that it could not be regarded as an open subspace of a larger two-dimensional manifold with an analytic non-degenerate metric. However, the region $t > 0$ has another inequivalent maximal analytic extension. If for $t > 0$ we define new coordinates $\tilde{t}, \tilde{\phi}$ by

$$\tilde{t} = t, \qquad \tilde{\phi} = \phi + 2 \log t,$$

the metric takes the form

$$\mathrm{d}s^2 = -2\mathrm{d}\tilde{\phi} \, \mathrm{d}\tilde{t} + \tilde{t} \, \mathrm{d}\tilde{\phi}^2.$$

This can obviously be analytically extended to the manifold $\mathbb{R}^1 \times \mathbb{S}^1$ defined by $-\infty < \tilde{t} < \infty$ and $0 \leq \tilde{\phi} \leq 2\pi$. However, this maximal extension is inequivalent as $\tilde{\phi}$ is not a continuous function of the coordinates t, ϕ at $t = 0$.

This example has a property which will be of interest later and which is related to the above behaviour. Consider the extension defined by the coordinates t, ϕ. There

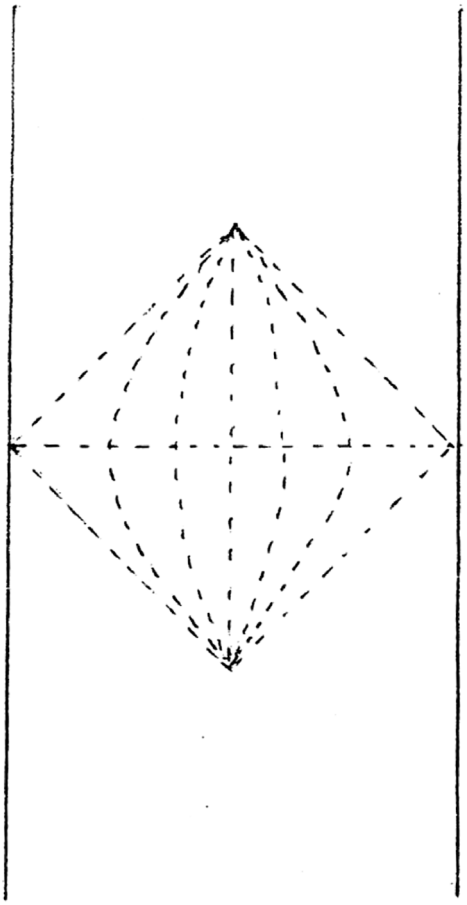

Fig. 7. The $t' - r'$ plane of anti-de Sitter space. The diagram shows the timelike geodesics orthogonal to a surface $t' = \text{const.}$. They intersect at points to the past and future of the surface.

are geodesics which enter the compact region between $t = 1$ and $t = 0$, do not leave it again, and are not extendible to arbitrary values of the affine parameter. We shall call this kind of incompleteness in a compact region Misner incompleteness. It cannot occur in positive-definite metrics.

Actually, the incomplete geodesics in the t, ϕ extension could be completed if we made the $\tilde{t}, \tilde{\phi}$ extension instead. However, there are other geodesics which can be continued to arbitrary values of the affine parameter in the t, ϕ extension, but which cannot be so continued in the $\tilde{t}, \tilde{\phi}$ extension. As each extension is maximal, we cannot complete all the geodesics.

4 The physical significance of curvature

4.1 Timelike curves

In Section 3, we saw that if the metric was static there was a relation between the magnitude of the timelike Killing vector and the Newtonian potential. We were able to tell whether a body was in a gravitational field by whether, if released from rest,

it would accelerate with respect to the static frame defined by the Killing vector. However, in general, spacetime will not have any Killing vectors. Thus we will not have any special frame against which to measure the acceleration of a body. The best we can do is to take two bodies close together and measure their relative acceleration. This will enable us to measure the gradient of the field. If we think of the metric as being analogous to the Newtonian potential, the gradient of the Newtonian field would correspond to the second derivatives of the metric. These are described by the Riemann tensor. Thus one would expect that the relative acceleration of two neighbouring bodies would be related to some components of the Riemann tensor.

In order to investigate this relation more precisely, we shall examine the behaviour of a congruence of timelike curves with timelike unit tangent vector v, with $g(v,v)=1$. These curves could represent the paths of small test particles. In this case, they would be geodesics. Or they might represent the flow lines of a fluid. If this was a perfect fluid, the energy-momentum tensor would be

$$T^{ab} = \mu V^a V^b - p h^{ab},$$

where μ is the energy density, p the pressure, and $h^{ab} = g^{ab} - V^a V^b$ is the metric in the subspace H_q of T_q orthogonal to V. In this case, the conservation equations $T^{ab}{}_{;b} = 0$ would give

$$\dot{\mu} + (\mu - p) V^a{}_{;a} = 0, \qquad (\mu + p)\dot{V}^a - p_{;b} h^{ab} = 0, \tag{1}$$

where $\dot{\mu} = \mu_{;a} V^a$ is the rate of change of the energy density with time on a flow line and $\dot{V}^a = V^a{}_{;b} V^b$ is the acceleration of the flow line with respect to a geodesic curve at each point. We see that $\dot{V}^a$ is given by the gradient of the pressure, as one would expect.

Suppose $\lambda(t)$ is a curve with tangent vector $z = (\partial/\partial t)_\lambda$. Then we may construct a family $\lambda(t,s)$ of curves by moving each point of the curve $\lambda(t)$ a distance s along the integral curves of V. If we now define Z as $(\partial/\partial t)_\lambda(t,s)$, it follows that $\mathcal{L}_V Z = 0$. We may interpret Z as representing the separation of points equal distances along two neighbouring integral curves of V. We have

$$\frac{\mathrm{D}}{\partial s} Z^a = V^a{}_{;b} Z^b. \tag{2}$$

If Z is initially orthogonal to V, it will not remain orthogonal unless V is geodesic ($\dot{V}^a = 0$). We shall define $\perp Z$ as the part of Z orthogonal to V, where $\perp$ indicates projection by $h^a{}_b$, i.e.,

$$\perp Z^a = h^a{}_b Z^b.$$

We may think of $\perp Z$ as representing the separation in the 3-surface orthogonal to V of two neighbouring curves. It obeys

$$\perp \frac{\mathrm{D}}{\partial s} (\perp Z^a) = V^a{}_{;b} \perp Z^b. \tag{3}$$

This gives the rate of separation in the 3-surface orthogonal to V of two neighbouring curves. Operating with $\mathrm{D}/\partial s$,

$$\frac{\mathrm{D}}{\partial s}\left[\perp \frac{\mathrm{D}}{\partial s}(\perp Z^a)\right] = -R^a{}_{bcd} \perp Z^c V^b V^d + \dot{V}^a{}_{;c} \perp Z^c - \dot{V}^a \dot{V}^b \perp Z_b. \tag{4}$$

This equation, which is known as the deviation or Jacobi equation, gives the relative acceleration of two neighbouring curves. We see that this depends only on the Riemann

tensor if the curves are geodesic. One could think of the term $R^a{}_{bcd}V^bV^d$ as a sort of tidal force, causing neighbouring bodies to accelerate relative to each other.

In order to investigate further the significance of this equation, we shall introduce dual orthonormal bases E_1, E_2, E_3, E_4 and E^1, E^2, E^3, E^4 of T_q and T_q^*, respectively, at some point q on an integral curve $\gamma(s)$ of V with $E_4 = V$ and Fermi propagate them along the curve to obtain orthonormal bases at all points of the curve. Then a tesnor field K of type (r, s) along the curve can be expressed in terms of its components with respect to these bases:

$$K = \overset{ab\ldots d}{\underset{ij\ldots l}{K}} \underset{a}{E} \otimes \underset{b}{E} \otimes \ldots \otimes \underset{d}{E} \otimes \overset{i}{E} \otimes \overset{j}{E} \otimes \ldots \otimes \overset{l}{E},$$

where we have placed the indices above and below the letter K to distinguish them from indices with respect to a coordinate basis. The indices may be raised and lowered using the components $\overset{ab}{g}$ and $\underset{ab}{g}$ of the metric tensor with respect to the bases. As the bases are chosen to be orthonormal, the nonzero components will be

$$\underset{11}{g} = \overset{11}{g} = \underset{22}{g} = \overset{22}{g} = \underset{33}{g} = \overset{33}{g} = -\underset{44}{g} = -\overset{44}{g} = -1.$$

We shall be interested in tensors in the subspaces H_q and H_q^* of T_q and T_q^* orthogonal to V. These are spanned by the bases E_1, E_2, E_3 and E^1, E^2, E^3. We will denote the indices of components of such tensors by Greek letters α, β, and so on, which will take the values 1, 2, 3 only. Using this convention, we may express the vector $\perp Z$ as

$$\perp Z = \overset{\alpha}{Z}\underset{\alpha}{E},$$

where

$$\frac{\mathrm{d}}{\mathrm{d}s}Z^\alpha = \overset{\alpha}{V}{}_{;\beta}\overset{\beta}{Z}$$

and $\overset{\alpha}{V}{}_{;\beta}$ are the components with respect to the bases E_1, E_2, E_3 and E^1, E^2, E^3 of the covariant derivative of V. Since the components $\overset{\alpha}{Z}$ obey a linear differential equation, they can be expressed in terms of their values at some point q by

$$\overset{\alpha}{Z}(s) = \overset{\alpha}{\underset{\beta}{A}}(s)\ \overset{\beta}{Z}\big|_q,$$

where $\overset{\alpha}{\underset{\beta}{A}}(s)$ is a 3×3 matrix which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}s}\overset{\alpha}{\underset{\beta}{A}}(s) = \overset{\alpha}{V}{}_{;\gamma}\overset{\gamma}{\underset{\beta}{A}}$$

and is the unit matrix at q.

For convenience, we shall adopt matrix notation for products. Thus the above equations will be written as

$$\underline{Z}(s) = \underline{A}(s) \cdot \underline{Z}(q), \qquad \frac{\mathrm{d}}{\mathrm{d}s}\underline{A} = \underline{V} \cdot \underline{A},$$

where

$$(\underline{V})_{\alpha\beta} = \underset{\alpha;\beta}{V}.$$

Note that, because of the signature of the metric,

$$(\underline{V}\cdot\underline{A})_{\alpha\beta} = -\sum_{\gamma} \underset{\alpha;\gamma}{V}\ \underset{\gamma\beta}{A}.$$

The matrix $\underline{A}$ can be expressed as

$$\underline{A} = \underline{O}\cdot\underline{S},$$

where $\underline{O}$ is an orthogonal matrix and $\underline{S}$ is a symmetric matrix. These will both be the unit matrix at q. The matrix $\underline{O}$ may be thought of as representing the rotation that neighbouring curves have undergone with repsect to the Fermi propagated basis. The matrix $\underline{S}$ represents the separation of these curves from $\gamma(s)$. The determinant of $\underline{S}$, which equals the determinant of $\underline{A}$, may be thought of as representing the 3-volume of the element of the surface orthogonal to $\gamma(s)$ marked out by neighbouring curves.

At q, where A is the unit matrix,

$$\frac{\mathrm{d}}{\mathrm{d}s}(\underline{O})\cdot\underline{S} + \underline{O}\cdot\frac{\mathrm{d}}{\mathrm{d}s}\underline{S} = \underline{V}\cdot\underline{O}\cdot\underline{S} = \underline{V}.$$

But

$$\underline{O}\cdot\underline{O}^{\mathrm{T}} = \underline{I},$$

where $\underline{I}$ is the identity matrix. So

$$\frac{\mathrm{d}}{\mathrm{d}s}(\underline{O})\cdot\underline{O}^{\mathrm{T}} + \underline{O}\cdot\frac{\mathrm{d}}{\mathrm{d}s}(\underline{O})^{\mathrm{T}} = 0.$$

Thus $\mathrm{d}\underline{O}/\mathrm{d}s$ is antisymmetric at q. However, $\mathrm{d}\underline{S}/\mathrm{d}s$ is symmetric. Therefore, the rate of change of the rotation of neighbouring curves at q is given by the antisymmetric part of the matrix $\underline{V}$, while the rate of change of their separation is given by the symmetric part of $\underline{V}$, and the rate of change of volume is given by the trace of $\underline{V}$. Thus we may define the instantaneous vorticity as

$$\underline{\omega} = \frac{1}{2}\left(\underline{V} - \underline{V}^{\mathrm{T}}\right),$$

the instantaneous rate of separation as

$$\underline{\psi} = \frac{1}{2}\left(\underline{V} + \underline{V}^{\mathrm{T}}\right),$$

and the volume expansion as

$$\theta = \mathrm{Tr}\ (\underline{V}).$$

We may also define the shear as the trace-free part of $\underline{\psi}$, viz.,

$$\underline{\sigma} = \underline{\psi} - \frac{1}{3}\theta\underline{I}.$$

Returning for a moment to coordinate indices, we will define the vorticity and shear tensors and the expansion as follows:

$$\omega_{ab} = h_a{}^c h_b{}^d V_{[c;d]}, \quad \sigma_{ab} = h_a{}^c h_b{}^d V_{(c;d)} - \frac{1}{3}h_{ab}\theta, \quad \theta = V_{a;b}h^{ab} = V_{a;b}g^{ab}.$$

We may also define the vorticity vector as

$$\omega^a = \frac{1}{2}\eta^{abcd}V_b\omega_{cd} = \frac{1}{2}\eta^{abcd}V_b V_{c;d}.$$

The covariant derivative of the vector V can be expressed in terms of these quantities as

$$V_{a;b} = \omega_{ab} + \sigma_{ab} + \frac{1}{3}h_{ab}\theta + \dot{V}_a V_b. \tag{5}$$

This decomposition is directly analogous to that of the gradient of the fluid velocity vector in hydrodynamics.

To obtain equations for the propagation of the vorticity, shear, and expansion, we use the deviation equation (4), which can be expressed as

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\underline{A} = (-\underline{G} + \underline{F}) \cdot \underline{A}, \tag{6}$$

where

$$(\underline{G})_{\alpha\beta} = \underset{\alpha 4 \beta 4}{R}$$

is the 'tidal force', and

$$(\underline{F})_{\alpha\beta} = \underset{\alpha;\beta}{\dot{V}} - \underset{\alpha}{\dot{V}}\underset{\beta}{\dot{V}}$$

depends on the acceleration. By definition we have

$$\underline{\omega} = \frac{1}{2}\left[\frac{\mathrm{d}}{\mathrm{d}s}(\underline{A}) \cdot \underline{A}^{-1} - \left(\underline{A}^{\mathrm{T}}\right)^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}s}\underline{A}^{\mathrm{T}}\right],$$

$$\underline{\psi} = \frac{1}{2}\left[\frac{\mathrm{d}}{\mathrm{d}s}(\underline{A}) \cdot \underline{A}^{-1} + \left(\underline{A}^{\mathrm{T}}\right)^{-1} \cdot \frac{\mathrm{d}}{\mathrm{d}s}\underline{A}^{\mathrm{T}}\right],$$

and

$$\theta = \mathrm{Tr}\left(\underline{\psi}\right) = (\det \underline{A})^{-1} \frac{\mathrm{d}}{\mathrm{d}s}(\det \underline{A}).$$

Substituting in (5),

$$\frac{\mathrm{d}}{\mathrm{d}s}\underline{\omega} = -\underline{\omega} \cdot \underline{\psi} - \underline{\psi} \cdot \underline{\omega} + \frac{1}{2}\left(\underline{F} - \underline{F}^{\mathrm{T}}\right). \tag{7}$$

We see that the propagation of vorticity depends on the antisymmetric gradient of the acceleration but not on the 'tidal force'. If the curves are geodesics and the vorticity vanishes at one point on a curve, then it will vanish everywhere on that curve. Another form of the above equation is

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A}\right) = \frac{1}{2}\underline{A}^{\mathrm{T}} \cdot \left(\underline{F} - \underline{F}^{\mathrm{T}}\right) \cdot \underline{A}. \tag{8}$$

Thus $\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A}$ is constant if the curves are geodesics. If the curves are the flow lines of an isentropic fluid, a straightforward calculation gives

$$\frac{1}{2}\left(\underline{F} - \underline{F}^{\mathrm{T}}\right) = -\underline{\omega}\frac{\dot{p}}{\mu + p}.$$

Thus we have the interesting conservation law

$$p\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A} = \text{const.},$$

where

$$\log p = \int \frac{\mathrm{d}p}{\mu + p}.$$

For the rate of separation tensor, we have

$$\frac{\mathrm{d}}{\mathrm{d}s}\underline{\psi} = -\underline{G} - \underline{\omega}\cdot\underline{\omega} - \underline{\psi}\cdot\underline{\psi} + \frac{1}{2}\left(\underline{F} + \underline{F}^{\mathrm{T}}\right).$$

This contains the 'tidal force' term $\underline{G}$. Taking the trace,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\theta &= -\mathrm{Tr}(\underline{G}) + 2\omega^2 - 2\sigma^2 - \frac{1}{3}\theta^2 + \mathrm{Tr}(\underline{F}) \\ &= -R_{ab}V^aV^b + 2\omega^2 - 2\sigma^2 - \frac{1}{3}\theta^2 + \dot{V}^a{}_{;a},\end{aligned} \tag{9}$$

where

$$2\omega^2 = \mathrm{Tr}(\underline{\omega}\cdot\underline{\omega}^{\mathrm{T}}) \geq 0, \qquad 2\sigma^2 = \mathrm{Tr}(\underline{\sigma}\cdot\underline{\sigma}^{\mathrm{T}}) \geq 0.$$

This equation, which was discovered by Landau and independently by Raychaudhuri, will be of great importance later. From it, one sees that vorticity induces expansion, as might be expected by analogy with centrifugal force, while shear induces contraction.

The above equations enable one to calculate the propagation of the vorticity, shear, and expansion along the integral curves of V if one knows the Riemann tensor. We saw in Section 2 that the Riemann tensor could be decomposed into the Weyl tensor and terms involving the Ricci tensor:

$$R_{abcd} = C_{abcd} - g_{a[d}R_{c]b} - g_{b[c}R_{d]a} - \frac{1}{3}Rg_{a[c}g_{d]b}.$$

The Ricci tensor is given by the Einstein equations:

$$R_{ab} - \frac{1}{2}g_{ab}R = T_{ab}.$$

Thus we see that the Weyl tensor is that part of the curvature which is not determined locally by the matter distribution. However, it cannot be entirely arbitrary as the Riemann tensor must satisfy the Bianchi identities. These give

$$R_{ab[cd;e]} = 0, \qquad C^{abcd}{}_{;d} = J^{abc},$$

where [Kundt 1962]

$$J^{abc} = R^{c[a;b]} + \frac{1}{6}g^{c[b}R^{;a]}. \tag{10}$$

These equations are rather similar to Maxwell's equations in electrodynamics:

$$F^{ab}{}_{;b} = J^a,$$

where F^{ab} is the electromagnetic field tensor and J^a is the source current. Thus in a sense we could regard the Bianchi identities (10) as field equations for the Weyl tensor, giving that part of the curvature at a point that depends on the matter distribution at other points.

In electrodynamics, we may split the field tensor F^{ab} into electric and magnetic components with respect to a unit timelike vector field V:

$$E_a = -F_{ab}V^b, \qquad H_a = -\frac{1}{2}F^{bc}\eta_{bcad}V^d.$$

Then the equations $F^{ab}{}_{;b} = J^a$ may be written in terms of the electric and magnetic fields:

$$E^b{}_{;b} + E_b\dot{V}^b - 2H_b\omega^b = J^bV_b$$

and

$$-\eta_{abcd}V^bH^{c;d} - h_a{}^b\dot{E}_b + \omega_a{}^bE_b + \sigma_a{}^bE_b - \frac{2}{3}\theta E_a - \eta_{abcd}V^b\dot{V}^cH^d = h_{ab}J^b.$$

These correspond to the equations

$$\operatorname{div}\underline{E} = \rho, \qquad \operatorname{curl}\underline{H} + \dot{\underline{E}} = -\underline{J},$$

where $\rho = V_bJ^b$ is the charge density. The extra terms arise because the vector field V used to obtain the splitting into electric and magnetic components does not in general have vanishing covariant derivative. The remaining Maxwell's equations $F_{[ab;c]} = 0$ may be expressed in a similar manner as

$$H^b{}_{;b} + H_b\dot{V}^b + 2E_b\omega^b = 0$$

and

$$-\eta_{abcd}V^bE^{c;d} + h_a{}^b\dot{H}_b - \omega_a{}^bH_b - \sigma_a{}^bH_b + \frac{2}{3}\theta H_a - \eta_{abcd}V^b\dot{V}^cE^d = 0.$$

These correspond to the equations

$$\operatorname{div}\underline{H} = 0, \qquad -\operatorname{curl}\underline{E} + \dot{\underline{H}} = 0.$$

Following the analogy between the Maxwell equations and the Bianchi identities, we shall split the Weyl tensor into 'electric' and 'magnetic' components:

$$E_{ab} = -C_{acbd}V^cV^d, \quad E_{ab} = E_{(ab)}, \quad E^a{}_a = 0,$$
$$H_{ab} = \frac{1}{2}C_a{}^{pqr}\eta_{qrbs}V_pV^s, \quad H_{ab} = H_{(ab)}, \quad H^a{}_a = 0.$$

The quantities E_{ab} and H_{ab} each have five independent components. We call E_{ab} the 'electric' components since they contribute to the 'tidal force', inducing relative acceleration of neighbouring curves in a manner analogous to that in which the electric components of the electromagnetic field cause acceleration of a charged particle. Expressing the Bianchi indentities in terms of the fields E_{ab} and H_{ab}, we have [Trümper 1964; Hawking 1965a, 1966b]:

$$h_a{}^bE_{bc;d}h^{cd} - 3H_{ab}\omega^b + \eta_{abcd}V^b\sigma^c{}_eH^{de} = J_a{}^{cd}V_cV_d, \tag{11}$$

$$h_a{}^bH_{bc;d}h^{cd} + 3E_{ab}\omega^b - \eta_{abcd}V^b\sigma^c{}_eE^{de} = \frac{1}{2}\eta_{abcd}J^{cde}V^bV_e, \tag{12}$$

$$h_a{}^ch_b{}^d\dot{E}_{cd} + h_{(a}{}^f\eta_{b)cde}V^cH_f{}^{d;e} + E_{ab}\theta - E^c{}_{(a}\omega_{b)c} - 3E^c{}_{(a}\sigma_{b)c} + h_{ab}E_{cd}\sigma^{cd}$$
$$-2H^d{}_{(a}\eta_{b)cde}V^c\dot{V}^e = J^{cde}V_ch_{d(a}h_{b)e}, \tag{13}$$

$$h_a{}^ch_b{}^d\dot{H}_{cd} - h_{(a}{}^f\eta_{b)cde}V^cE_f{}^{d;e} + H_{ab}\theta - H^c{}_{(a}\omega_{b)c} - 3H^c{}_{(a}\sigma_{b)c} + h_{ab}H_{cd}\sigma^{cd}$$
$$-2E^d{}_{(a}\eta_{b)cde}V^c\dot{V}^e = -\frac{1}{2}h_{f(a}\eta_{b)cde}V^cJ^{def}. \tag{14}$$

If we assume that the energy-momentum tensor is that of a perfect fluid, then the source terms on the right are:

$$J_a{}^{bc}V_bV_c = -\frac{1}{3}h_a{}^b\mu_{;b}, \qquad \frac{1}{2}\eta_{abcd}J^{cde}V^bV_e = (\mu+p)\omega_a,$$

$$J^{cde}V_ch_{d(a}h_{b)e} = \frac{1}{2}(\mu+p)\sigma_{ab}, \qquad -\frac{1}{2}h_{f(a}\eta_{b)cde}V^cJ^{def} = 0.$$

It can be seen that equations (11)–(14) are of the form:

$$\operatorname{div}\underline{E} = \rho, \qquad \operatorname{div}\underline{H} = \tilde{\rho}, \qquad \operatorname{curl}\underline{H} + \dot{\underline{E}} = J, \qquad \operatorname{curl}\underline{E} - \dot{\underline{H}} = 0,$$

where the quantity ρ, analogous to the electric charge, is the gradient of the density, the quantity $\tilde{\rho}$, which would correspond to the magnetic charge if such existed, depends on the vorticity, and the quantity J, analogous to the electric current, depends on the shear. One may also notice that the term which would correspond to the magnetic current vanishes, though of course this would not necessarily be true for more general energy-momentum tensors.

Equations (7)–(9) and (11)–(14) do not form a complete set since, to evaluate the covariant derivatives, we would also have to know the components of the metric and the connection and to relate them to the Ricci and Weyl tensors. However, there are certain important cases in which we can avoid doing this. Suppose the metric differed only slightly from a metric that was conformally flat (Weyl tensor vanishes). Then E_{ab} and H_{ab} would be small quantities and to the first order we could perform all derivatives with the connection of the conformally flat metric. For example, if the metric differed only slightly from a flat metric, we could introduce a timelike vector field V whose covariant derivative was also a small quantity and, neglecting products of such quantities, we would have:

$$E_a{}^b{}_{;b} = -\frac{1}{3}h_a{}^b\mu_{;b}, \tag{15}$$

$$H_a{}^b{}_{;b} = (\mu+p)\omega_a, \tag{16}$$

$$\dot{E}_{ab} + H_{(a}{}^{d;e}\eta_{b)cde}V^c = \frac{1}{2}(\mu+p)\sigma_{ab}, \tag{17}$$

$$\dot{H}_{ab} - E_{(a}{}^{d;e}\eta_{b)cde}V^c = 0, \tag{18}$$

where the derivatives are performed with the connection of the flat metric. Another example would be a metric which differed only slightly from the metric of one of the Friedmann models discussed in the previous section, as these are all conformally flat. In such a metric, we may introduce a vector field V whose covariant derivative $V_{a;b}$ differs only slightly from $h_{ab}\theta_0/3$, where θ_0 is the value of the expansion in the undisturbed Friedmann model. Then, to first order, we get the same equations as above except that extra terms $E_{ab}\theta_0$ and $H_{ab}\theta_0$ appear on the left of equations (17) and (18) and the derivatives are performed with the connection of the undisturbed model.

In the vicinity of the Earth the metric is very nearly flat, so equations (15)–(18) should hold to very high accuracy. On a larger scale, it seems probable that the metric is similar to that of a Friedmann model so, with extra terms, equations (15)–(18) should be reasonably accurate on a cosmological scale too. They may be used to describe disturbances in the intergalactic medium and to investigate the propagation and absorption of gravitational radiation [Hawking 1966b].

4.2 Null curves

We may also consider the deviation equation for a congruence of null curves with tangent vector K, where $g(K,K)=0$. For simplicity, we shall assume that the curves are geodesic. They could be thought of as representing the paths of rays of light. There are two important differences between this case and that of the timelike curves considered in the previous section. First, we could normalise the tangent vector V to the timelike curves by requiring $g(V,V)=1$. In effect, this meant that we parametrised the curves by the arc length s. However, this is clearly impossible with null curves, as they have zero arc length. The best we can do is to choose an affine parameter v. Then the tangent vector K will obey

$$\frac{\mathrm{D}}{\partial v}K^a = K^a{}_{;b}K^b = 0.$$

However, we could multiply v by a function f which was constant along each curve. Then fv would be another affine parameter and the corresponding tangent vector would be $f^{-1}K$. Thus given the curves, the tangent vector is only really unique up to a constant factor along each curve. The second difference is that H_q, the subspace of T_q orthogonal to K, includes the vector K itself since $g(K,K)=0$. We will call the vector space H_q modulo the vector K, the screen space S_q. That is S_q is the space of equivalence clases of vectors of H_q which differ only by a multiple of K.

As before, we will introduce dual bases E_1, E_2, E_3, E_4 and E^1, E^2, E^3, E^4 of T_q and T_q^* at some point q on a curve $\gamma(v)$. However, we will not choose them to be orthonormal. We will take E_4 equal to K, E_3 to be some other null vector having unit product with E_4, i.e.,

$$g(E_3,E_3)=0, \qquad g(E_3,E_4)=1,$$

and E_1 and E_2 to be unit spacelike vectors, orthogonal to each other and to E_3 and E_4, i.e.,

$$g(E_1,E_1)=-1=g(E_2,E_2), \qquad g(E_1,E_2)=0=g(E_1,E_3)=g(E_1,E_4),$$

and so on. It can be seen that E_1, E_2, and E_4 constitute a basis for H_q, while E_1 and E_2 alone are a basis for S_q. We shall call a basis having the properties of E_1, E_2, E_3, E_4 above pseudo-orthonormal. By parallel transporting them along the geodesic $\gamma(v)$, we may obtain a pseudo-orthonormal basis at each point of $\gamma(v)$.

We shall use this basis to analyse the deviation equation for null geodesics. If Z is the vector representing the separation of corresponding points on neighbouring curves, we have, as before $\mathcal{L}_K Z = 0$, so

$$\frac{\mathrm{D}}{\partial v}Z^a = K^a{}_{;b}Z^b, \tag{19}$$

and

$$\frac{\mathrm{D}^2}{\partial v^2}Z^a = -R^a{}_{bcd}Z^cK^bK^d. \tag{20}$$

If we take Z to be orthogonal to K initially, it will remain orthogonal, because K is geodesic since

$$\frac{\mathrm{D}}{\partial v}(Z_aK^a) = \frac{1}{2}(K_aK^a)_{;b}Z^b = 0.$$

Then we can express Z as $\overset{\alpha}{Z}E_\alpha$, where $\alpha, \beta, \ldots$, now take the values 1, 2, and 4. We will have

$$\frac{\mathrm{d}}{\mathrm{d}v}\overset{\alpha}{Z} = \overset{\alpha}{K}{}_{;\beta}\overset{\beta}{Z}.$$

However, $\overset{\alpha}{K}_{;4} = 0$, since K is geodesic. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}v}\overset{m}{Z} = \overset{m}{K}_{;n}\overset{n}{Z},$$

where $m, n, \ldots$, take the values 1 and 2 only. But $\overset{m}{Z}$ are the components of Z in the screen space S_q with respect to the basis E_1 and E_2 of S_q. So we see that the propagation of the projection of Z onto S_q depends only on itself and not on the other components of Z.

As in the previous section, we may express $\overset{m}{Z}$ in terms of their values at some point q by

$$\overset{m}{Z}(v) = \tilde{A}^m{}_n(v)\overset{n}{Z}\big|_q,$$

where $\tilde{A}^m{}_n(v)$ is a 2×2 matrix which satisfies

$$\frac{\mathrm{d}}{\mathrm{d}v}\tilde{A}^m{}_n(v) = K^m{}_{;p}\tilde{A}^p{}_n(v),$$

and which is the unit matrix at q. Using matrix notation as before, we will express $\underline{\tilde{A}}$ as the product of an orthogonal 2×2 matrix $\underline{\tilde{O}}$ and a symmetric 2×2 matrix $\underline{\tilde{S}}$. The matrix $\underline{\tilde{O}}$ may be thought of as representing the rotation, measured in the screen space, which neighbouring geodesics have undergone with repsect to the parallel propagated basis E_1 and E_2, while $\underline{\tilde{S}}$ represents the separation measured in the screen space of neighbouring geodesics from the geodesic $\gamma(v)$. We have

$$\frac{\mathrm{d}}{\mathrm{d}v}(\underline{\tilde{O}} \cdot \underline{\tilde{S}}) = \underline{K} \cdot \underline{\tilde{O}} \cdot \underline{\tilde{S}},$$

where

$$\left(\underline{K}\right)_{mn} = \underset{m;n}{K}.$$

As before, we will call the antisymmetric part of the matrix $\underline{K}$ the vorticity $\underline{\tilde{\omega}}$, the symmetric part of $\underline{K}$ the rate of separation $\underline{\tilde{\psi}}$, and the trace of $\underline{K}$ the expansion $\tilde{\theta}$. We may also define the shear $\underline{\tilde{\sigma}}$ as the trace-free part of $\underline{\tilde{\psi}}$.

Using equations (19) and (20), we may obtain equations for the propagation of these quantities analogous to those of the previous section. We have

$$\frac{\mathrm{d}}{\mathrm{d}v}\overset{m}{Z} = \overset{m}{K}_{;n}\overset{n}{Z}, \qquad \frac{\mathrm{d}^2}{\mathrm{d}v^2}\overset{m}{Z} = -\underset{4n4}{\overset{m}{R}}\overset{n}{Z},$$

since

$$\underset{444}{\overset{m}{R}} = 0,$$

as the Riemann tensor is antisymmetric in the last two positions. Thus

$$\frac{\mathrm{d}}{\mathrm{d}v}\underline{\tilde{A}} = \underline{K} \cdot \underline{\tilde{A}}$$

and

$$\frac{\mathrm{d}^2}{\mathrm{d}v^2}\underline{\tilde{A}} = -\underline{\tilde{G}} \cdot \underline{\tilde{A}}, \tag{21}$$

where

$$\left(\underline{\tilde{G}}\right)_{mn} = \underset{m4n4}{R}.$$

Then we have, as before,

$$\frac{\mathrm{d}}{\mathrm{d}v}\underline{\tilde{\omega}} = -\underline{\tilde{\omega}}\cdot\underline{\tilde{\psi}} - \underline{\tilde{\psi}}\cdot\underline{\tilde{\omega}}, \tag{22}$$

$$\frac{\mathrm{d}}{\mathrm{d}v}\left(\underline{\tilde{A}}^{\mathrm{T}}\cdot\underline{\tilde{\omega}}\cdot\underline{\tilde{A}}\right) = 0, \tag{23}$$

$$\frac{\mathrm{d}}{\mathrm{d}v}\underline{\tilde{\psi}} = -\underline{\tilde{G}} - \underline{\tilde{\omega}}\cdot\underline{\tilde{\omega}} - \underline{\tilde{\psi}}\cdot\underline{\tilde{\psi}}, \tag{24}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}v}\tilde{\theta} &= -\mathrm{Tr}\big(\underline{\tilde{G}}\big) + 2\tilde{\omega}^2 - 2\tilde{\sigma}^2 - \frac{1}{2}\tilde{\theta}^2 \\ &= -R_{ab}K^aK^b + 2\tilde{\omega}^2 - 2\tilde{\sigma}^2 - \frac{1}{2}\tilde{\theta}^2.\end{aligned} \tag{25}$$

4.3 Conjugate points

In Section 4.1 we saw that the components of the vector Z which represented the separation between a curve $\gamma(s)$ and a neighbouring curve in a congruence of timelike geodesics satisfied the Jacobi equation

$$\frac{\mathrm{D}^2}{\partial s^2}\overset{\alpha}{Z} = -\underset{4\beta4}{\overset{\alpha}{R}}\overset{\beta}{Z} \qquad (\alpha,\beta = 1,2,3). \tag{26}$$

A solution of this equation will be called a Jacobi field along $\gamma(s)$. Since a solution may be specified by giving the values of $\overset{\alpha}{Z}$ and $\mathrm{d}\overset{\alpha}{Z}/\mathrm{d}s$ at some point on $\gamma(s)$, there will be six independent Jacobi fields along $\gamma(s)$. There will be three independent Jacobi fields which vanish at some point q of $\gamma(s)$. They may be expressed as

$$\overset{\alpha}{Z}(s) = A^{\alpha}{}_{\beta}(s)\left.\frac{\mathrm{d}}{\mathrm{d}s}\overset{\beta}{Z}\right|_q, \tag{27}$$

where

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}A^{\alpha}{}_{\beta}(s) = -\underset{4\gamma4}{\overset{\alpha}{R}}A^{\gamma}{}_{\beta}(s),$$

and $A^{\alpha}{}_{\beta}(s)$ is a 3×3 matrix which vanishes at q. These Jacobi fields may be thought of as representing the separation of neighbouring geodesics through q. Adopting matrix notation, we may define the vorticity, shear, and expansion of the Jacobi fields along $\gamma(s)$ which vanish at q:

$$\begin{aligned}\underline{\omega} &= \frac{1}{2}\left[\frac{\mathrm{d}}{\mathrm{d}s}\big(\underline{A}\big)\cdot\underline{A}^{-1} - \big(\underline{A}^{\mathrm{T}}\big)^{-1}\cdot\frac{\mathrm{d}}{\mathrm{d}s}\underline{A}^{\mathrm{T}}\right], \\ \underline{\sigma} &= \frac{1}{2}\left[\frac{\mathrm{d}}{\mathrm{d}s}\big(\underline{A}\big)\cdot\underline{A}^{-1} + \big(\underline{A}^{\mathrm{T}}\big)^{-1}\cdot\frac{\mathrm{d}}{\mathrm{d}s}\underline{A}^{\mathrm{T}}\right] - \frac{1}{3}\underline{I}\theta, \\ \theta &= (\det\underline{A})^{-1}\frac{\mathrm{d}}{\mathrm{d}s}(\det\underline{A}).\end{aligned}$$

These will obey the equations for geodesics derived in Section 4.1. In particular,

$$\underline{A}^{\mathrm{T}}\cdot\underline{\omega}\cdot\underline{A} = \frac{1}{2}\left[\underline{A}^{\mathrm{T}}\cdot\frac{\mathrm{d}}{\mathrm{d}s}\underline{A} - \frac{\mathrm{d}}{\mathrm{d}s}\big(\underline{A}^{\mathrm{T}}\big)\underline{A}\right]$$

will be constant along $\gamma(s)$. But it vanishes at q as $\underline{A}$ is zero. Therefore $\underline{\omega}$ will be zero wherever $\underline{A}$ is non-singular.

We shall say that a point p on $\gamma(s)$ is conjugate to q along $\gamma(s)$ if there is a Jacobi field along $\gamma(s)$, not identically zero, which vanishes at q and p. We may think of p as a point where neighbouring geodesics through q intersect. The Jacobi fields along $\gamma(s)$ which vanish at q are described by the matrix $\underline{A}$. Thus a point p is conjugate to q along $\gamma(s)$ if and only if $\underline{A}$ is singular at p. The expansion θ is defined as

$$\theta = (\det \underline{A})^{-1} \frac{\mathrm{d}}{\mathrm{d}s}(\det \underline{A}).$$

Since $\underline{A}$ obeys the equation

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2}\underline{A} = -\underline{G} \cdot \underline{A},$$

where $\underline{G}$ is finite, $\mathrm{d}(\det \underline{A})/\mathrm{d}s$ will be finite. Thus a point p will be conjugate to q along $\gamma(s)$ if θ becomes infinite there. The converse will also be true since

$$\theta = \frac{\mathrm{d}}{\mathrm{d}s} \log(\det \underline{A}),$$

and $\underline{A}$ can be singular only at isolated points or else it would be singular everywhere.

We shall take s to be zero at q. Near q for $s > 0$, the expansion θ of the matrix $\underline{A}$ will be positive. However, for greater values of s, θ may become negative. We have the following lemma:

Lemma 1. *If at some point $\gamma(s_1)$, $s_1 > 0$, the expansion θ has a negative value $\theta < 0$, and if $R_{ab}V^aV^b \geq 0$ everywhere, then there will be a point conjugate to q along $\gamma(s)$ between $\gamma(s_1)$ and $\gamma(s_1 + 3/ - \theta_1)$.*

The expansion θ of the matrix $\underline{A}$ obeys the Raychaudhuri equation derived in Section 4.1:

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -R_{ab}V^aV^b - 2\sigma^2 - \frac{1}{3}\theta^2, \tag{28}$$

where we have used the fact that the vorticity is zero. All the terms on the right-hand side are negative. Thus for $s > s_1$,

$$\theta \leq \frac{3}{s - (s_1 + 3/ - \theta_1)},$$

so θ will become infinite and there will be a point conjugate to q for some value of s between s_1 and $s_1 + 3/ - \theta_1$.

In other words, if $R_{ab}V^aV^b \geq 0$ and the neighbouring geodesics through q start converging on $\gamma(s)$, some neighbouring geodesic will actually intersect $\gamma(s)$. By the Einstein equations,

$$R_{ab}V^aV^b = T_{ab}V^aV^b - \frac{1}{2}T.$$

If we assume that the energy-momentum tensor is that of an electromagnetic field, the expression on the right above is always non-negative. If the energy-momentum tensor is that of a perfect fluid with energy density μ and pressure p, then the above expression will be non-negative if

$$\mu + p \geq 0, \qquad \mu + 3p \geq 0.$$

These would seem very reasonable requirements. This point will be discussed further in Section 6.

We shall also consider the congruence of timelike geodesics normal to a spacelike 3-surface H. By a spacelike 3-surface, we mean the imbedded three-dimensional submanifold of an open set D of M defined by $f = 0$, where f is a C^2 function on D and

$$g^{ab} f_{;a} f_{;b} > 0 \quad \text{where } f = 0.$$

We define Y, the unit normal to H, by

$$Y^a = \left(g^{bc} f_{;b} f_{;c}\right)^{-1/2} g^{ad} f_{;d},$$

and the second fundamental tensor X of H by

$$X_{ab} = h_a{}^c h_b{}^d Y_{c;d},$$

where $h_{ab} = g_{ab} - Y_a Y_b$ is called the first fundamental tensor (or induced metric tensor) of H. It follows from the definition that X is symmetric. The congruence of timelike geodesics normal to H will consist of the timelike geodesics whose unit tangent vector V equals the unit normal Y at H. Then we have

$$V_{a;b} = X_{ab} \text{ at } H.$$

The vector Z which represents the separation of a neighbouring geodesic normal to H from $\gamma(s)$, a normal geodesic to H, will obey the Jacobi equation (26). At a point q on $\gamma(s)$ at H, it will satisfy the initial condition

$$\frac{\mathrm{d}}{\mathrm{d}s} \overset{\alpha}{Z} = \underset{\beta}{\overset{\alpha}{X}} \overset{\beta}{Z}. \tag{29}$$

We shall express the Jacobi fields along $\gamma(s)$ which satisfy the above condition as

$$\underline{Z}(s) = \underline{A}(s) \cdot \underline{Z}\big|_q,$$

where

$$\frac{\mathrm{d}^2}{\mathrm{d}s^2} \underline{A} = -\underline{G} \cdot \underline{A}$$

and at q, $\underline{A}$ is the unit matrix and

$$\frac{\mathrm{d}}{\mathrm{d}s} \underline{A} = \underline{X}.$$

We shall say that a point p on $\gamma(s)$ is conjugate to H along $\gamma(s)$ if there is a Jacobi field along $\gamma(s)$ not identically zero, which satisfies the initial conditions (29) at q and vanishes at p. In other words, p is conjugate to H along $\gamma(s)$ if and only if $\underline{A}$ is singular at p. We may think of p as being a point where neighbouring geodesics normal to H intersect. As before, $\underline{A}$ will be singular where and only where the expansion θ becomes infinite. At q, the initial value of $\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A}$ will be zero as $\underline{X}$ is symmetric. Thus $\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A}$ will be zero everywere on $\gamma(s)$. The initial value of θ will be $\mathrm{Tr}(\underline{X})$.

Lemma 2. *If $R_{ab} V^a V^b \geq 0$ and $\mathrm{Tr}\,(\underline{X}) < 0$, there will be a point conjugate to H along $\gamma(s)$ within a distance $3/-\mathrm{Tr}\,(\underline{X})$ from H.*

This may be proved using the Raychaudhuri equation as for Lemma 1.

We shall call a solution of the equation

$$\frac{\mathrm{d}^2}{\mathrm{d}v^2} \overset{m}{Z} = -\underset{4n4}{\overset{m}{R}} \overset{n}{Z} \qquad (m, n = 1, 2), \tag{30}$$

along a null geodesic $\gamma(v)$ a Jacobi field along $\gamma(v)$. The components $\overset{m}{Z}$ should be thought of as the components with respect to the basis E_1 and E_2 of a vector in the screen space at each point. We shall say that p is the conjugate to q along the null geodesic $\gamma(v)$ if there is a Jacobi field along $\gamma(v)$ not identically zero, which vanishes at q and p. Representing the Jacobi fields along $\gamma(v)$ which vanish at q by the 2×2 matrix $\underline{\tilde{A}}$, so that

$$\underline{Z}(s) = \underline{\tilde{A}} \cdot \frac{\mathrm{d}}{\mathrm{d}s}\underline{Z}\Big|_q ,$$

we have as before

$$\underline{\tilde{A}}^{\mathrm{T}} \cdot \underline{\tilde{\omega}} \cdot \underline{\tilde{A}} = 0.$$

Also p will be conjugate to q along $\gamma(s)$ if and only if

$$\tilde{\theta} = \big(\det \underline{\tilde{A}}\big)^{-1} \frac{\mathrm{d}}{\mathrm{d}s}\big(\det \underline{\tilde{A}}\big)$$

becomes infinite at p. Analogous to Lemma 1, we have:

Lemma 3. *If $R_{ab}K^aK^b \geq 0$ everywhere and if at some point $\gamma(v_1)$, $v_1 > 0$, the expansion $\tilde{\theta}$ has the negative value $\tilde{\theta}_1 < 0$, then there will be a point conjugate to q along $\gamma(v)$ between $\gamma(v_1)$ and $\gamma(v_1 + 2/-\theta_1)$.*

The expansion $\tilde{\theta}$ of the matrix $\underline{\tilde{A}}$ obeys

$$\frac{\mathrm{d}}{\mathrm{d}v}\tilde{\theta} = -R_{ab}K^aK^b - 2\tilde{\sigma}^2 - \frac{1}{2}\tilde{\theta}^2. \tag{31}$$

The proof is as for Lemma 1.

By the Einstein equations

$$R_{ab}K^aK^b = T_{ab}K^aK^b.$$

It seems reasonable to assume that this is also non-negative.

Similarly, we may also consider the null geodesics normal to a spacelike 2-surface $\tilde{H}$. By a spacelike 2-surface in an open set D of M, we mean the imbedded two-dimensional submanifold of D defined by $f_1 = 0$, $f_2 = 0$, where f_1 and f_2 are C^2 functions on D such that when $f_1 = 0$, $f_2 = 0$, then $f_{1;a}$ and $f_{2;a}$ are non-vanishing and

$$(f_{1;a} + \mu f_{2;a})(f_{1;b} + \mu f_{2;b})g^{ab} = 0,$$

for some real value of μ. Then there will be two real values μ_1 and μ_2 which satisfy the above equation. We shall define $\underset{3}{\tilde{Y}}^a$ and $\underset{4}{\tilde{Y}}^a$ to be the two null vectors normal to $\tilde{H}$ proportional to $g^{ab}(f_{1;b} + \mu_1 f_{2;b})$ and $g^{ab}(f_{1;b} + \mu_2 f_{2;b})$, respectively, and normalised so that

$$\underset{3}{\tilde{Y}}^a \underset{4}{\tilde{Y}}^b g_{ab} = 1.$$

We may complete the pseudo-orthonormal basis by introducing two spacelike unit vectors $\underset{1}{\tilde{Y}}^a$ and $\underset{2}{\tilde{Y}}^a$ orthogonal to each other and to $\underset{3}{\tilde{Y}}^a$ and $\underset{4}{\tilde{Y}}^a$. Then we shall define the two null second fundamental tensors of $\tilde{H}$ as

$$\underset{1}{\tilde{X}}_{ab} = -\underset{3}{\tilde{Y}}_{c,d}\left(\underset{1}{\tilde{Y}}^c \underset{1}{\tilde{Y}}_a + \underset{2}{\tilde{Y}}^c \underset{2}{\tilde{Y}}_a\right)\left(\underset{1}{\tilde{Y}}^d \underset{1}{\tilde{Y}}_b + \underset{2}{\tilde{Y}}^d \underset{2}{\tilde{Y}}_b\right),$$

and so on. $\underset{1}{\tilde{X}}$ and $\underset{2}{\tilde{X}}$ are then symmetric.

There will be two families of null geodesics normal to $\tilde{H}$, correspond to the two null normals $\underset{3}{\tilde{Y}}$ and $\underset{4}{\tilde{Y}}$. Consider the family whose tangent vector K equals $\underset{4}{\tilde{Y}}$ at $\tilde{H}$. We may fix our pseudo-orthonormal basis E_1, E_2, E_3, E_4 by taking $E_1 = \underset{1}{\tilde{Y}}$, etc., at $\tilde{H}$ and parallel propagating along the null geodesics. The projection into the screen space of the vector Z representing the separation of neighbouring null geodesics from the null geodesic $\gamma(v)$ will satisfy equation (30) and the initial conditions

$$\frac{\mathrm{d}}{\mathrm{d}v}\underline{Z} = \underset{2}{\tilde{X}} \cdot \underline{Z} \tag{32}$$

at q on $\gamma(v)$ at $\tilde{H}$. As before, the vorticity of these fields will be zero. The initial value of the expansion $\tilde{\theta}$ will be $\mathrm{Tr}\big(\underset{2}{\tilde{X}}\big)$. Analogous to Lemma 2, we have

Lemma 4. *If $R_{ab}K^aK^b \geq 0$ everywhere and $Tr\big(\underset{2}{\tilde{X}}\big)$ is negative, there will be a point conjugate to $\tilde{H}$ along $\gamma(v)$, within an affine distance $2/-\mathrm{Tr}\big(\underset{2}{\tilde{X}}\big)$ from $\tilde{H}$.*

The proof is as for Lemmas 2 and 3.

The significance of conjugate points will be seen in the next section.

4.4 Variation of arc length

By a broken timelike curve $\gamma(t)$ from a point $q = \gamma(0)$ to $p = \gamma(t_p)$, we shall mean a connected, piecewise C^3 curve from q to p such that, at every regular point, the tangent vector $(\partial/\partial t)_\gamma$ is timelike and, at every singular point, the two tangent vectors $\partial/\partial t|_-$ and $\partial/\partial t|_+$ are timelike and satisfy

$$g\left(\left.\frac{\partial}{\partial t}\right|_-, \left.\frac{\partial}{\partial t}\right|_+\right) > 0.$$

That is, they point into the same half of the null cone. We define the length of such a curve as

$$L = \sum \int \sqrt{g(\partial/\partial t, \partial/\partial t)}\,\mathrm{d}t,$$

where the integrals are taken over the differentiable sections of the curve. If we choose the parameter to be the arc length s, then $g(\partial/\partial t, \partial/\partial t) = 1$ and we have

$$L = \sum \int \mathrm{d}s.$$

In a positive-definite metric, we may find the shortest curve between two points, but in an indefinite, Lorentz metric, there will not be any shortest curve, as any timelike curve can be deformed into a null curve of zero length. However, in certain cases there will be a longest timelike curve between two points or between a point and a spacelike 3-surface.

Consider a spacelike 3-surface H in a normal coordinate neighbourhood U. By the implicit function theorem, there will be an open neighbourhood W of H such that, in W, the map

$$\beta : H \times [0, \epsilon] \longrightarrow M$$

will be a diffeomorphism from some $\epsilon > 0$, where β is defined by taking a point of H a distance $v \in [0, \epsilon]$ along the geodesics normal to H. For constant v, the image

$\beta(H \times v)$ will be a 3-surface. It will be orthogonal to V, the unit tangent vector to the geodesics as a vector representing the separation of points equal distances along neighbouring geodesics will remain orthogonal to V. Let p be a point of $\beta(H \times [0, \epsilon])$ such that every timelike curve from p to H remains in $\beta\big(H \times (0, \epsilon)\big)$. There will be a geodesic normal to H through p. We shall show that this is the longest timelike curve in U from p to H. Let $\gamma(t)$ be such a curve. We may choose the parameter t to be equal to v. Then the tangent vector $(\partial/\partial t)_\gamma$ can be expressed as $V + Y$, where Y is some vector orthogonal to V. Hence,

$$g\big((\partial/\partial t)_\gamma, (\partial/\partial t)_\gamma\big) = g(V, V) + g(Y, Y) \leq 1,$$

the equality holding if and only if $(\partial/\partial t)_\gamma = V$. Thus the length of $\gamma(t)$ will be less than or equal to the value of v at p, the equality holding if and only if $\gamma(t)$ is a geodesic curve normal to H.

A similar construction may be used to show that, in a normal coordinate neighbourhood, a timelike geodesic curve is the longest broken timelike curve between two points. To investigate whether the broken timelike curve $\gamma(t)$ from q to p is the longest such curve from q to p, we shall consider the change in its length under a small variation. A variation α of $\gamma(t)$ is a map

$$\alpha : (-\epsilon, \epsilon) \times [0, t_p] \longrightarrow M$$

such that:

1. $\alpha(0, t) = \gamma(t)$.
2. There is a subdivision $0 = t_1 < t_2 < \ldots < t_n = t_p$ of $[0, t_p]$ such that α is C^3 on each $(-\epsilon, \epsilon) \times [t_i, t_{i+1}]$.
3. $\alpha(u, 0) = q$, $\alpha(u, t_p) = p$.
4. For each constant u, $\alpha(u, t)$ is a broken timelike curve.

The vector $(\partial/\partial u)_\alpha\big|_{u=0}$ will be called the variation vector Z. Conversely, given a continuous, piecewise C^2 vector field Z along $\gamma(t)$, vanishing at q and p, we may define a variation α for which Z will be the variation vector by

$$\alpha(u, t) = \exp_r \left(uZ\big|_r\right),$$

where $u \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$ and $r = \gamma(t)$.

Lemma 5. *The variation of the length from q to p under α is*

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = -\sum \int g\left(\frac{\partial}{\partial u}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\frac{\partial f}{\partial t}\frac{\partial}{\partial t}\right\}\right) \mathrm{d}t - \sum g\left(\frac{\partial}{\partial u}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right),$$

where $f^2 = g(\partial/\partial t, \partial/\partial t)$ is the squared magnitude of the tangent vector and

$$\left[f^{-1}\frac{\partial}{\partial t}\right]$$

is the discontinuity at one of the singular points of $\gamma(t)$.

We have

$$\begin{aligned}\left.\frac{\partial L}{\partial u}\right|_{u=0} &= \sum \frac{\partial}{\partial u}\int \sqrt{g(\partial/\partial t, \partial/\partial t)}\,\mathrm{d}t \\ &= \sum \int g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) f^{-1}\mathrm{d}t \\ &= \sum \int g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right) f^{-1}\mathrm{d}t \\ &= \sum \int \left\{\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\right) f^{-1} - g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right) f^{-1}\right\}\mathrm{d}t.\end{aligned}$$

Integrating the first term by parts, we have the required formula.

We may simplify the formula by choosing the parameter t to be the arc length s. Then $g(\partial/\partial t, \partial/\partial t) = 1$. We shall call $\partial/\partial s$ the unit tangent vector V. We have

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = -\sum \int g(Z, \dot{V})\mathrm{d}s - \sum g(Z, [V]),$$

where $\dot{V} = \mathrm{D}V/\partial s$ is the acceleration. From this we see that a necessary condition for $\gamma(t)$ to be the longest curve from q to p is that it should be an unbroken geodesic curve, as otherwise we could choose a variation which would yield a longer curve.

We may also consider a curve $\gamma(t)$ from a spacelike 3-surface to a point p. We define a variation α of this curve as before, except that we replace condition (3) above by

3′. $\alpha(u, 0)$ lies on H and $\alpha(u, t_p) = p$.

Thus at H the variation vector $Z = \partial/\partial u$ lies in H.

Lemma 6.

$$\left.\frac{\partial L}{\partial u}\right|_{u=0} = -\sum \int g(\dot{V}, Z)\mathrm{d}s - \sum g(Z, [V]) - g(Z, V)\big|_{s=0}.$$

The proof is as for Lemma 5. From this we see that a necessary condition for $\gamma(t)$ to be the longest curve from H to p is that it is an unbroken geodesic orthogonal to H.

We have seen that, under a variation α, the first derivative of the length of a geodesic curve is zero. To proceed further, we shall calculate the second derivative. We define a two-parameter variation α of a geodesic curve $\gamma(t)$ from q to p as a C^1 map

$$\alpha : (-\epsilon_1, \epsilon_1) \times (-\epsilon_2, \epsilon_2) \times [0, t_p] \longrightarrow M,$$

with properties as before and

$$\underset{1}{Z} = \left(\frac{\partial}{\partial u_1}\right)_\alpha \Bigg|_{\substack{u_1 = 0 \\ u_2 = 0}}, \qquad \underset{2}{Z} = \left(\frac{\partial}{\partial u_2}\right)_\alpha \Bigg|_{\substack{u_1 = 0 \\ u_2 = 0}},$$

as the two variation vectors. Conversely, given two continuous, piecewise C^2 vector fields $\underset{1}{Z}$ and $\underset{2}{Z}$ along $\gamma(t)$, we may define a variation for which they will be the variation vectors, by

$$\alpha(u_1, u_2, t) = \exp_r\left(u_1 \underset{1}{Z} + u_2 \underset{2}{Z}\right), \qquad r = \gamma(t).$$

Lemma 7. *Under the two-parameter variation of the geodesic curve $\gamma(t)$, the second derivative of the length will be*

$$\left.\frac{\partial^2 L}{\partial u_2 \partial u_1}\right|_{\substack{u_1=0\\u_2=0}} = -\sum \int g\left(\underset{1}{Z}, \left\{\frac{\mathrm{D}^2}{\partial s^2}\left(\underset{2}{Z} - Vg(V, \underset{2}{Z})\right) + R(V, \underset{2}{Z})V\right\}\right) \mathrm{d}s$$
$$-\sum g\left(\underset{1}{Z}, \left\{\frac{\mathrm{D}}{\partial s}\left(\underset{2}{Z} - Vg(V, \underset{2}{Z})\right)\right\}\right).$$

By Lemma 5, we have

$$\left.\frac{\partial L}{\partial u_1}\right|_{\substack{u_1=0\\u_2=0}} = -\sum \int g\left(\frac{\partial}{\partial u}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\frac{\partial f}{\partial t}\frac{\partial}{\partial t}\right\}\right) \mathrm{d}t$$
$$-\sum g\left(\frac{\partial}{\partial u}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right).$$

Therefore,

$$\left.\frac{\partial^2 L}{\partial u_2 \partial u_1}\right|_{\substack{u_1=0\\u_2=0}} = -\sum \int g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \left\{f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - f^{-2}\frac{\partial f}{\partial t}\frac{\partial}{\partial t}\right\}\right) \mathrm{d}t$$
$$+\sum \int g\left(\frac{\partial}{\partial u_1}, \left\{f^{-2}\frac{\partial f}{\partial u_2}\frac{\mathrm{D}}{\mathrm{d}t}\frac{\partial}{\partial t} - f^{-1}\frac{\mathrm{D}}{\partial u_2}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} - 2f^{-3}\frac{\partial f}{\partial u_2}\frac{\partial f}{\partial t}\frac{\partial}{\partial t}\right.\right.$$
$$\left.\left. + f^{-2}\frac{\partial^2 f}{\partial u_2 \partial t}\frac{\partial}{\partial t} + f^{-2}\frac{\partial f}{\partial t}\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}\right\}\right)\mathrm{d}t$$
$$-\sum g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \left[f^{-1}\frac{\partial}{\partial t}\right]\right) - \sum g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial u_2}\left[f^{-1}\frac{\partial}{\partial t}\right]\right).$$

The first and third terms vanish as $\gamma(t)$ is an unbroken geodesic curve. In the second term, we can write

$$\frac{\mathrm{D}}{\partial u_2}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} = R\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t} + \frac{\mathrm{D}}{\partial t}\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}$$

and

$$\frac{\partial^2 f}{\partial u_2 \partial t} = \frac{\partial}{\partial t}\left\{f^{-1}g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right\}$$
$$= \frac{\partial}{\partial t}\left\{f^{-1}\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u_2}, \frac{\partial}{\partial t}\right)\right) - f^{-1}g\left(\frac{\partial}{\partial u_2}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)\right\}.$$

In the fourth term,

$$\frac{\mathrm{D}}{\partial u_2}\left[f^{-1}\frac{\partial}{\partial t}\right] = \left[f^{-1}\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2} - f^{-3}g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}, \frac{\partial}{\partial t}\right)\frac{\partial}{\partial t}\right].$$

Then taking t to be the arc length s, we obtain the required result.

Although it is not immediately obvious from the appearance of the expression, we know from its definition that it is symmetric in the two variation vector fields $\underset{1}{Z}$

and $\underset{2}{Z}$. We see that it only depends on the projections of $\underset{1}{Z}$ and $\underset{2}{Z}$ into the space orthogonal to V. Thus we can confine our attention to variations α whose variation vectors are orthogonal to V. We shall define T_γ to be the (infinite-dimensional) vector space consisting of all continuous, piecewise C^2 vector fields along $\gamma(t)$ orthogonal to V and vanishing at q and p. Then $\partial^2 L/\partial u_2 \partial u_1$ will be a symmetric map of $T_\gamma \times T_\gamma$ to $\mathbb{R}^1$. We may think of it as a symmetric tensor on T_γ and write it as

$$L\left(\underset{1}{Z}, \underset{2}{Z}\right) = \left.\frac{\partial^2 L}{\partial u_2 \partial u_1}\right|_{\substack{u_1 = 0 \\ u_2 = 0}}, \qquad \underset{1}{Z}, \underset{2}{Z} \in T_\gamma.$$

We may also calculate the second derivative of the length from H to p of a geodesic curve $\gamma(t)$ normal to H. We proceed as before, except that one end point of $\gamma(t)$ is allowed to vary over H instead of being fixed.

Lemma 8. *The second derivative of the length of $\gamma(t)$ from H to p is*

$$\left.\frac{\partial^2 L}{\partial u_2 \partial u_1}\right|_{\substack{u_1 = 0 \\ u_2 = 0}} = -\sum \int g\left(\underset{1}{Z}, \left\{\frac{\mathrm{D}^2}{\partial s^2}\underset{2}{Z} + R\left(V, \underset{2}{Z}\right)V\right\}\right) \mathrm{d}s$$
$$- \sum g\left(\underset{1}{Z}, \left[\frac{\mathrm{D}}{\partial s}\underset{2}{Z}\right]\right) - \left. g\left(\underset{1}{Z}, \frac{\mathrm{D}}{\partial s}\underset{2}{Z}\right)\right|_H + X\left(\underset{1}{Z}, \underset{2}{Z}\right),$$

where $\underset{1}{Z}$ and $\underset{2}{Z}$ have been taken orthogonal to V and $X\left(\underset{1}{Z}, \underset{2}{Z}\right)$ is the second fundamental tensor of H. The first two terms are as for Lemma 7. The extra terms are

$$-\left.\frac{\mathrm{D}}{\partial u_2} g\left(\frac{\partial}{\partial u_1}, f^{-1}\frac{\partial}{\partial t}\right)\right|_H = -f^{-1}\left. g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\right|_H$$
$$+ f^{-3}\, g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\Big|_H$$
$$- f^{-1}\left. g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}\right)\right|_H.$$

The second term vanishes as $\partial/\partial u_1$ is orthogonal to $\partial/\partial t$. If we take t to be the arc length, then $\partial/\partial t$ will be equal to the unit normal Y at H. Since the end point of $\gamma(t)$ is restricted to varying over H, $\partial/\partial u_1$ will always be orthogonal to Y. Thus

$$g\left(\frac{\mathrm{D}}{\partial u_2}\frac{\partial}{\partial u_1}, Y\right) = \frac{\mathrm{D}}{\partial u_2} g\left(\frac{\partial}{\partial u_1}, Y\right) - g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial u_2} Y\right)$$
$$= X\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial u_2}\right).$$

We shall say that a timelike geodesic curve $\gamma(t)$ from q to p is maximal if $L\left(\underset{1}{Z}, \underset{2}{Z}\right)$ is negative-definite. In other words, $\gamma(t)$ is maximal if all small variations α yield a shorter curve from q to p. Similarly, we shall say that a timelike geodesic curve from H to p, normal to H, is maximal if all small variations yield a shorter curve from H to p.

Lemma 9. *A timelike geodesic curve $\gamma(t)$ from q to p is maximal if there is no point conjugate to q along $\gamma(t)$ in $[q, p]$.*

We wish to show that $L(Z,Z) < 0$ for any non-vanishing $Z \in T_\gamma$. Adopting matrix notation, $\underline{Z}$ will be zero at q and p, while the matrix $\underline{A}$ representing the Jacobi fields which vanish at q will be zero at q, but non-singular elsewhere in $[q,p]$. Thus we may express $\underline{Z}$ as

$$\underline{Z} = \underline{A} \cdot \underline{b}, \quad \text{where} \quad \underline{b} = \underline{A}^{-1} \cdot \underline{Z}.$$

Note that $\underline{b}$ will vanish at p. Then

$$\begin{aligned} L(\underline{Z},\underline{Z}) &= -\sum \int \underline{b}^{\mathrm{T}} \cdot \underline{A}^{\mathrm{T}} \cdot \left\{ \frac{\mathrm{d}^2}{\mathrm{d}s^2}(\underline{A} \cdot \underline{b}) + \underline{G} \cdot \underline{A} \cdot \underline{b} \right\} \mathrm{d}s - \sum \underline{b}^{\mathrm{T}} \cdot \underline{A}^{\mathrm{T}} \cdot \left[\frac{\mathrm{d}}{\mathrm{d}s}(\underline{A} \cdot \underline{b}) \right] \\ &= -\sum \int \underline{b}^{\mathrm{T}} \cdot \underline{A}^{\mathrm{T}} \cdot \left\{ 2\frac{\mathrm{d}}{\mathrm{d}s}\underline{A} \cdot \frac{\mathrm{d}}{\mathrm{d}s}\underline{b} + \underline{A} \cdot \frac{\mathrm{d}^2}{\mathrm{d}s^2}\underline{b} \right\} \mathrm{d}s - \sum \underline{b}^{\mathrm{T}} \cdot \underline{A}^{\mathrm{T}} \cdot \underline{A} \cdot \left[\frac{\mathrm{d}}{\mathrm{d}s}\underline{b} \right] \\ &= \sum \int \left\{ \frac{\mathrm{d}\underline{b}^{\mathrm{T}}}{\mathrm{d}s} \cdot \underline{A}^{\mathrm{T}} \cdot \underline{A} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} + \underline{b}^{\mathrm{T}} \cdot \left(\frac{\mathrm{d}\underline{A}^{\mathrm{T}}}{\mathrm{d}s} \cdot \underline{A} - \underline{A}^{\mathrm{T}} \cdot \frac{\mathrm{d}\underline{A}}{\mathrm{d}s} \right) \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} \right\} \mathrm{d}s. \end{aligned}$$

But

$$\frac{\mathrm{d}\underline{A}^{\mathrm{T}}}{\mathrm{d}s} \cdot \underline{A} - \underline{A}^{\mathrm{T}} \cdot \frac{\mathrm{d}\underline{A}}{\mathrm{d}s} = -\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A} = 0,$$

so

$$L(\underline{Z},\underline{Z}) = \sum \int \frac{\mathrm{d}\underline{b}^{\mathrm{T}}}{\mathrm{d}s} \cdot \underline{A}^{\mathrm{T}} \cdot \underline{A} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} \mathrm{d}s.$$

But

$$\frac{\mathrm{d}\underline{b}^{\mathrm{T}}}{\mathrm{d}s} \cdot \underline{A}^{\mathrm{T}} \cdot \underline{A} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} = \left(\underline{A} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} \right)^{\mathrm{T}} \cdot \left(\underline{A} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} \right) \leq 0,$$

because of the signature of the metric, and $\mathrm{d}\underline{b}/\mathrm{d}s$ must be nonzero somewhere if $\underline{Z}$ is not identically zero. Thus $L(\underline{Z},\underline{Z}) < 0$.

Lemma 10. *A timelike geodesic curve $\gamma(t)$ from q to p is not maximal if there is a point r between q and p conjugate to q along $\gamma(t)$.*

Let $\underline{b}$ be the Jacobi field along $\gamma(t)$ which vanishes at q and r. Let $\underline{c} \in T_\gamma$ be such that

$$\underline{c}^{\mathrm{T}} \cdot \frac{\mathrm{d}\underline{b}}{\mathrm{d}s} = 1,$$

at r. Extend $\underline{b}$ to p by putting it zero in rp. Let $\underline{Z}$ be

$$\underline{Z} = \varepsilon \underline{c} + \varepsilon^{-1} \underline{b},$$

where ε is some constant. Then,

$$\begin{aligned} L(\underline{Z},\underline{Z}) &= \varepsilon^2 L(\underline{c},\underline{c}) + 2L(\underline{c},\underline{b}) + 2\varepsilon^{-2} L(\underline{b},\underline{b}) \\ &= \varepsilon^2 L(\underline{c},\underline{c}) + 2. \end{aligned}$$

Thus by taking ε small enough, $L(\underline{Z},\underline{Z})$ may be made positive.

We may obtain similar results for the case of a timelike geodesic curve $\gamma(t)$ normal to H.

Lemma 11. *A timelike geodesic curve $\gamma(t)$ normal to H from H to p is maximal if there is no point conjugate to H in $[H,p]$.*

Lemma 12. *It is not maximal if there is a point r between H and p conjugate to H along $\gamma(t)$.*

The proofs are as for Lemmas 9 and 10.

We shall consider variations of a broken null curve $\gamma(t)$ from q to p. The definition of a broken null curve is the same as that for a broken timelike curve except that the tangent vector $(\partial/\partial t)_\gamma$ is required to be null everywhere. We shall be interested in the circumstance under which it is possible to find a variation α of $\gamma(t)$ which makes $g(\partial/\partial t, \partial/\partial t)$ positive everywhere or, in other words, yields a timelike curve from q to p. To decide these it would not be convenient to study the behaviour of L under a small variation since $\sqrt{g(\partial/\partial t, \partial/\partial t)}$ will not be differentiable when $g(\partial/\partial t, \partial/\partial t) = 0$. Instead we shall consider the variation in

$$\Lambda = \sum \int g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) \mathrm{d}t.$$

Clearly, a necessary but not sufficient condition that a variation α of $\gamma(t)$ should yield a timelike curve from q to p is that Λ should become positive.

Under a variation α,

$$\begin{aligned}\frac{\partial}{\partial u}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right) &= 2g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right) = 2g\left(\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right) \\ &= 2\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\right) - 2g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right).\end{aligned}$$

In order to obtain a timelike curve from q to p, we require this to be greater than or equal to zero everywhere on $\gamma(t)$. If $\gamma(t)$ is an unbroken geodesic curve, we may take t to be equal to an affine parameter v. Then

$$\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t} = 0 \quad \text{and} \quad \left.\frac{\partial \Lambda}{\partial u}\right|_{u=0} = 0.$$

We see also that the variation vector $\partial/\partial u|_{u=0}$ must be orthogonal to the tangent vector $\partial/\partial t$ everywhere on $\gamma(t)$, otherwise

$$\frac{\partial}{\partial t}\left(g\left(\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\right)$$

would be negative somewhere on $\gamma(t)$. On the other hand, if $\gamma(t)$ is not an unbroken geodesic curve, it is not difficult to see that we could define a variation α which would have

$$\frac{\partial}{\partial u}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right)$$

positive everywhere on $\gamma(t)$. Thus if q and p are joined by a null curve which is not an unbroken geodesic, they can also be joined by a timelike curve.

We shall now consider a two-parameter variation α of an unbroken null geodesic curve $\gamma(t)$ from q to p. The variation α will be defined as before except that, for the reason given above, we shall restrict ourselves to variations whose variation vectors

$$\begin{aligned}\frac{\partial^2}{\partial u_2 \partial u_1}\left(g\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\right) &= \frac{\partial^2}{\partial u_2 \partial t}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\right) - \frac{\partial}{\partial u_2}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial t}\right)\right) \\ &= \frac{\partial^2}{\partial u_2 \partial t}\left(g\left(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial t}\right)\right) \\ &\quad - g\left(\frac{\partial}{\partial u_1}, \left\{\frac{\mathrm{D}^2}{\partial t^2}\frac{\partial}{\partial u_2} + R\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t}\right\}\right)\end{aligned}$$

and

$$\left.\frac{\partial^2 \Lambda}{\partial u_2 \partial u_1}\right|_{\substack{u_1=0\\u_2=0}} = -\sum \int g\left(\frac{\partial}{\partial u_1}, \left\{\frac{\mathrm{D}^2}{\partial t^2}\frac{\partial}{\partial u_2} + R\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial u_2}\right)\frac{\partial}{\partial t}\right\}\right) \mathrm{d}t$$
$$-\sum g\left(\frac{\partial}{\partial u_1}, \left[\frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u_2}\right]\right).$$

This formula is very similar to that for the variation of the length of a timelike curve. It can be seen that the variation of Λ is zero for a variation vector proportional to the tangent vector $\partial/\partial t$, since $\partial/\partial t$ is null and

$$R\left(\frac{\partial}{\partial t}, \frac{\partial}{\partial t}\right)\frac{\partial}{\partial t} = 0,$$

as the Riemann tensor is antisymmetric. Such a variation would be equivalent to simply reparametrizing $\gamma(t)$. Thus if we want a variation which will give a timelike curve, we need consider only the projection of the variation vector into the screen space at each point of $\gamma(t)$. In other words, if we introduce a pseudo-orthonormal basis E_1, E_2, E_3, E_4 along $\gamma(t)$ with $E_4 = \partial/\partial t$, the variation of Λ will depend only on the components $\underset{m}{Z}$ of the variation vector. Then we have:

Lemma 13. *If there is no point in $[q,p]$ conjugate to q along $\gamma(t)$, then $\partial^2\Lambda/\partial u^2|_{u=0}$ will be negative for any variation α of $\gamma(t)$ whose variation vector $\partial/\partial u|_{u=0}$ is orthogonal to the tangent vector $\partial/\partial t$ on $\gamma(t)$ and is not everywhere zero or proportional to $\partial/\partial t$. In other words, if there is no point in $[q,p]$ conjugate to q, then there is no small variation of $\gamma(t)$ which gives a timelike curve from q to p.*

The proof is similar to that for Lemma 9, using instead the 2×2 matrix $\underline{\tilde{A}}$ of Section 4.2.

Lemma 14. *If there is a point r between q and p conjugate to q along $\gamma(t)$, then there will be a variation of $\gamma(t)$ which will give a timelike curve from q to p.*

The proof is a bit finicky, since we have to show that the tangent vector becomes timelike everywhere.

Let $\underline{Z}$ be the screen components of the Jacobi field along $\gamma(t)$ which vanishes at q and r. It obeys

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\underline{Z} = -\underline{\tilde{G}} \cdot \underline{Z},$$

where for convenience we have taken t to be equal to an affine parameter. We may write $\underline{Z}$ as

$$\underline{Z} = |Z|\underline{\hat{Z}},$$

where $\underline{\hat{Z}}$ is a unit vector. We shall take $|Z|$ and $\underline{\hat{Z}}$ to be continuous at r. Then,

$$\underline{\hat{Z}}\frac{\mathrm{d}^2}{\mathrm{d}t^2}|Z| + 2\frac{\mathrm{d}\underline{\hat{Z}}}{\mathrm{d}t}\frac{\mathrm{d}|Z|}{\mathrm{d}t} + |Z|\frac{\mathrm{d}^2\underline{\hat{Z}}}{\mathrm{d}t^2} = -\underline{\tilde{G}} \cdot \underline{\hat{Z}}|Z|.$$

Multiplying by $\underline{\hat{Z}}^{\mathrm{T}}$,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}|Z| + f|Z| = 0,$$

where

$$f = \underline{\hat{Z}}^{\mathrm{T}} \cdot \frac{\mathrm{d}^2\underline{\hat{Z}}}{\mathrm{d}t^2} + \underline{\hat{Z}}^{\mathrm{T}} \cdot \underline{\tilde{G}} \cdot \underline{\hat{Z}}.$$

Let p' be a point of $\gamma(t)$ between r and p such that $\underline{Z}$ is not zero in $(r, p']$. Let f' be the greatest lower bound of f in rp'. Let $a > 0$ be such that

$$a^2 + f' > 0 \quad \text{and} \quad b = -\left.\frac{|Z|}{\mathrm{e}^{at} - 1}\right|_{p'}.$$

Then the field $\underline{Z}'$ given by

$$\underline{Z}' = \left[b(\mathrm{e}^{at} - 1) + |Z|\right]\hat{\underline{Z}}$$

will vanish at q and p' and satisfy

$$\underline{Z}'^{\mathrm{T}} \cdot \left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\underline{Z}' + \tilde{\underline{G}} \cdot \underline{Z}'\right) > 0$$

in (q, p'). We shall choose a variation α' of $\gamma(t)$ from q to p' such that the screen components of its variation vector $\partial/\partial u|_{u=0}$ equal $\underline{Z}'$ and such that

$$g\left(\frac{\mathrm{D}}{\partial u}\frac{\partial}{\partial u}, \frac{\partial}{\partial t}\right)\Big|_{u=0} + g\left(\frac{\partial}{\partial u}, \frac{\mathrm{D}}{\partial t}\frac{\partial}{\partial u}\right)\Big|_{u=0} = \begin{cases} \varepsilon t, & 0 \le t \le \frac{1}{4}tp', \\ \varepsilon\left(\frac{1}{2}tp' - t\right) & \frac{1}{4}tp' < t < \frac{3}{4}tp', \\ \varepsilon(t - tp') & \frac{3}{4}tp' \le t \le tp', \end{cases}$$

where tp' is the value of t at p' and $\varepsilon > 0$ but less than the least value of

$$\underline{Z}'^{\mathrm{T}} \cdot \left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\underline{Z}' + \tilde{\underline{G}} \cdot \underline{Z}'\right)$$

in the range $tp'/4 \le t \le 3tp'/4$. Then the variation α will give a timelike curve from q to p'. If we join this curve to the section of $\gamma(t)$ from p' to p, we will have a non-spacelike curve from q to p which is not an unbroken null geodesic. Thus there will be a variation of this curve which gives a timelike curve from q to p.

Lemma 15. *If $\gamma(t)$ is an unbroken null geodesic curve normal to $\tilde{H}$ from $\tilde{H}$ to p, and if there is no point conjugate to $\tilde{H}$ along γ in $[\tilde{H}, p]$, then no small variation of $\gamma(t)$ can give a timelike curve from $\tilde{H}$ to p.*

Lemma 16. *If there is a point conjugate to $\tilde{H}$ along $\gamma(t)$ between $\tilde{H}$ and p, then there is a variation which gives a timelike curve from $\tilde{H}$ to p.*

The proofs are similar to those of the previous lemmas. These results will be of great importance in Section 6.

5 Horizons and causality

5.1 Conformal geometry

By the conformal geometry of spacetime, we shall mean the set of all Lorentz metrics $\tilde{g}(X, Y)$ on M conformal to the physical Lorentz metric $g(X, Y)$, i.e.,

$$\tilde{g}(X, Y) = \Omega^2 g(X, Y),$$

where $\Omega > 0$ is a C^3 function on M. Clearly, the division of vectors into timelike, spacelike, and null is a property of the conformal geometry. Thus all causal relations

on M are properties of the conformal geometry. Conversely, as we saw in Section 3, local causal relations determine the metric up to a conformal factor. All the properties of causal relations which will be considered in this section are thus also properties of the conformal geometry.

We saw in Section 2 that, if $\tilde{g} = \Omega^2 g$, then $\tilde{C}^a{}_{bcd} = C^a{}_{bcd}$, these being the components of the Weyl tensor. Thus the Weyl tensor is a property of the conformal geometry. On the other hand, we have

$$\tilde{R}_{ab} = R_{ab} - \Omega^{-2}\left(2\Omega\Omega_{;ab} + g_{ab}g^{cd}\Omega\Omega_{;cd} - 4\Omega_{;a}\Omega_{;b}\right),$$

so the Ricci tensor is not a property of the conformal geometry. This means that the metric $\tilde{g}$ will not in general satisfy the Einstein equations. If $\gamma(t)$ is a curve with tangent vector $K = (\partial/\partial t)_\gamma$, then

$$\frac{\tilde{\mathrm{D}}}{\partial t}K^a = \frac{\mathrm{D}}{\partial t}K^a + 2\Omega^{-1}\Omega_{;b}K^bK^a - \Omega^{-1}K^bK^cg_{bc}\Omega_{;d}g^{ad}.$$

If the curve is geodesic with respect to the metric g, we have

$$K^{[b}\frac{\mathrm{D}}{\partial t}K^{a]} = 0.$$

It can be seen that the curve will not in general also be geodesic with respect to the metric $\tilde{g}$ unless K is null. Thus being a null geodesic curve is a conformal property. The affine parameter on a null geodesic will in general be different in the metric $\tilde{g}$. Thus completeness (whether a geodesic can be extended to arbitrary values of the affine parameter) will not in general be a conformal property though we shall see later a special case in which it is. Since null geodesic curves are properties of the conformal geometry, the projection into the screen space of a Jacobi field along a null geodesic and the relation of being conjugate points along a null geodesic curve will also be conformal properties.

5.2 Time orientability

In our own neighbourhood, there is a clear division of non-spacelike vectors into future- and past-directed. However, it is not so obvious that we could extend this to give a continuous division at all points of M. This problem has been dealt with by Markus [Markus 1955]: as there is a Lorentz metric on M, we can find a continuous line-element field $(X, -X)$, where X is a timelike vector. The manifold $\tilde{M}$ is defined as the set of all pairs $(p, X|_p)$ and $(p, -X|_p)$ with the natural structure. Clearly, $\tilde{M}$ covers M twice. There are two possibilities: either $\tilde{M}$ is connected, in which case it is impossible to introduce a continuous division into future- and past-directed vectors in M, but one exists in $\tilde{M}$, or $\tilde{M}$ consists of two disconnected components, in which case one can find a continuous division in M. Physically it would seem very reasonable to expect M to be time orientable, but even if it was not, we could apply all the following theorems to the time orientable covering manifold $\tilde{M}$. In particular, if we can prove the occurrence of a singularity in $\tilde{M}$, this will imply the existence of one in M.

Similarly, we could ask whether it was possible to assign continuously right- and left-handed orientations of frames of spacelike vectors. As before, if it is not possible, there will be a doubly covering manifold in which it is possible. However, there is no such compelling physical reason for thinking that M is space orientable. If it were not, it would be possible globally to distinguish between neutrinos and anti-neutrinos, and it would raise the intriguing possibility that a right-handed space traveller might return home left-handed.

5.3 Horizons

As was stated in the last section, we shall assume either that M is time orientable or that we are considering the time orientable manifold $\tilde{M}$. We introduce the following notation which has been suggested by Kronheimer and Penrose [Kronheimer 1967] and Carter (private communication). A point p is said to causally precede q, denoted by $p < q$, if there is a non-spacelike curve from p to q whose tangent vector is future-directed or if $p = q$. The set of all points which causally precede q will be denoted by $(q >$ and the set of all points which q causally precedes will be denoted by $< q)$. The abbreviation $\langle p, q\rangle$ will be used for $< p) \cap (q >$. A point p will be said to chronologically precede q, denoted $p \ll q$, if there is a future-directed timelike curve from p to q. The sets $(q \gg, \ll q)$, and $\langle\langle p, q\rangle\rangle$ are defined similarly. As any timelike curve from q to r can be deformed into a timelike curve from q to some neighbourhood of r, these will be open sets.

A point p is said to precede q in the sense of the horismos, denoted by $p \prec\!\!\succ q$, if $p < q$ but not $p \ll q$. The sets $\prec\!\!\succ q)$, $(q \prec\!\!\succ$, and $\prec\!\!\succ p, q \prec\!\!\succ$ are defined similarly. The relations $<$, $\ll$, and $\prec\!\!\succ$ satisfy the following conditions:

1. $p \ll q \Longrightarrow p < q$.
2. $p < q$ and $q < r \Longrightarrow p < r$.
3. $p < q$ and $q \ll r$ or $p \ll q$ and $q < r \Longrightarrow p \ll r$.

The last property may be proved by considering a variation of the curve from p to q and r. We have as a direct result that, if $p < q < r$ and $p \prec\!\!\succ r$, then $p \prec\!\!\succ q \prec\!\!\succ r$. If λ was a non-spacelike curve from p to q which was not a null geodesic curve, we could find a varation of λ which gave a timelike curve from p to q. Thus if $p \prec\!\!\succ q$, q must lie on a future-directed null geodesic curve through p, though the converse is not necessarily true.

If N is a submanifold of M, we denote by $\overset{N}{<}$, $\overset{N}{\ll}$, and $\overset{N}{\prec\!\!\succ}$ the causal relations on N as a manifold with the metric induced from M, so that $p \overset{N}{<} q$ if and only if $p = q$ or there is a future-directed non-spacelike curve in N from p to q, and so on. These relations may not agree with the causal relations of M restricted to N. If u^1, u^2, u^3, u^4 are normal coordinates in a neighbourhood U of q, with $\partial/\partial u^4$ future-directed, it is easy to see that $\overset{U}{<} q)$ and $\overset{U}{\ll} q)$ consist of points whose coordinates satisfy

$$\left(u^4\right)^2 - \left(u^1\right)^2 - \left(u^2\right)^2 - \left(u^3\right)^2 \geq 0, \qquad u^4 \geq 0,$$

and the strict inequality, respectively. The boundary in U of these sets is formed by the points whose coordinates satisfy the above equality. They lie on the future-directed null geodesics through q.

To derive properties of the boundaries of more general sets, we shall use the following lemmas. We denote the boundary in M of a set S by $\dot{S} = \overline{S} \cap (\overline{M - S})$, where a bar denotes closure.

Lemma 17. *If a set N is such that $p \in N$ implies $(p \gg \subset N$, and if $M - N$ is not empty, then $\dot{N}$ is a closed, three-dimensional C^0 submanifold and there are no two points $q, r \in \dot{N}$ such that $q \ll r$.*

If $q \in \dot{N}$, any open neighbourhood of q intersects N and $M - N$. If $p \ll q$, then there is an open neighbourhood of q in $\ll p)$. Thus $(q \gg \subset N$. Similarly, $\ll q) \subset M - N$. If $r \gg q$, then there would be an open neighbourhood V of r such that $V \subset \ll q) \subset M - N$. Thus r could not belong to $\dot{N}$. We may introduce normal coordinates $u^1_\alpha, u^2_\alpha, u^3_\alpha, u^4_\alpha$ in an open neighbourhood U_α of q, with $\partial/\partial u^4_\alpha$ timelike, such that the curves $u^i_\alpha = \text{const.}$ $(i = 1, 2, 3)$ intersect $\ll q)$ and $(q \gg$. Then each

of these curves contains precisely one point of $\dot{N}$. The u^4_α coordinates of these points are continuous since no two points of $\dot{N}$ have timelike separation. Then the bijection $\phi_\alpha : \dot{N} \cap U_\alpha \to \mathbb{R}^3$ defined by $\phi_\alpha(q) = u^i(q)$ for $q \in \dot{N} \cap U_\alpha$ is a homeomorphism. Thus, $\{\dot{N} \cap U_\alpha, \phi_\alpha\}$ is a C^0 atlas for $\dot{N}$.

We shall call a boundary with the properties of $\dot{N}$ above a non-timelike horizon.

Lemma 18. *If N satisfies the conditions of the previous lemma and if for a point $q \in \dot{N}$ there is a normal coordinate neighbourhood U of q and a ball $B \subset U$ of constant coordinate radius about q which contains a sequence of points r_n converging on q such that there is a future-directed non-spacelike curve λ_n from each r_n which intersects $(\overline{\dot{B} \cap N})$, then $\dot{N}$ contains a future-directed null geodesic segment from q.*

The author is indebted to Dr R. Penrose for the following proof. Let p be a limit point of $\lambda_n \cap (\overline{\dot{B} \cap N})$. Then $p \in \dot{N}$ and so is not contained in $\ll q)$. Select a subsequence r'_n such that $\lambda'_n \cap (\overline{\dot{B} \cap N})$ converges to p. If V is any open neighbourhood of p, all r'_n for n large enough are contained in $(V \overset{U}{\gg}$. Thus $q \in (V \overset{U}{\gg}$ and $p \in \overline{\overset{U}{\ll} q})$. But

$$\overline{\overset{U}{\ll} q)} = \overset{U}{<} q).$$

Therefore, $q \rightarrowtail p$. Any neighbourhood of p must intersect $M - N$. Thus $p \in \dot{N}$. Moreover, if y is any point such that $q \rightarrowtail y \rightarrowtail p$, then $y \in \dot{N}$. This shows that the future-directed null geodesic segment from q to p lies in $\dot{N}$.

Corollary 1. *If there is a past-directed null geodesic segment from q in $\dot{N}$, this must be a continuation of the future-directed null geodesic from q, since if it were in any other direction, we would have points of $\dot{N}$ which were joined by a broken null geodesic curve and which could therefore be joined by a timelike curve. If there is more than one future-directed null geodesic segment from q in $\dot{N}$, there can be no past-directed such segment from q in $\dot{N}$.*

The above lemmas have duals in which future and past are interchanged. We shall regard such dual results as self-evident.

We shall say that a non-timelike horizon is null wherever the conditions laid down in Lemma 18 are satisfied. For example, consider the boundary of the chronological past of a point p. As $(p \gg$ satisfies the conditions of Lemma 17, its boundary will be a non-timelike horizon, provided that $M - (p \gg$ is not empty. It is easy to see that the conditions of Lemma 18 would be satisfied at each point of $(\dot{p} \gg$, except at the point p itself, if it lay on $(\dot{p} \gg$ (it need not: there might be a past-directed timelike curve from p which returned to p). Thus $(\dot{p} \gg$ would be null everywhere except at p. It would be generated by null geodesic segments which would have a past end point if they intersected another generating segment, but which could have a future end point only at p. There might be generating segments of $(\dot{p} \gg$ which did not pass through p. These could have no future end point. If there were a point r conjugate to p along a past-directed null geodesic λ from p, then all points on λ beyond r would be in $(p \gg$, by Lemma 14. Thus if some segment of λ lay in $(\dot{p} \gg$, it would have a past end point before or at r.

All points q such that $q \rightarrowtail p$ will lie on the boundary of $(p \gg$, since any point $r \ll q$ will be in $(p \gg$. However, such points may not comprise the whole of $(p \gg$. We shall say that a region S of M is causally simple if for every compact set V contained in S the intersections with S of $\ll \dot{V})$ and $(\dot{V} \gg$ consist entirely of points q such that $V \rightarrowtail q$ and $q \rightarrowtail V$, respectively. An equivalent statement is that every null geodesic

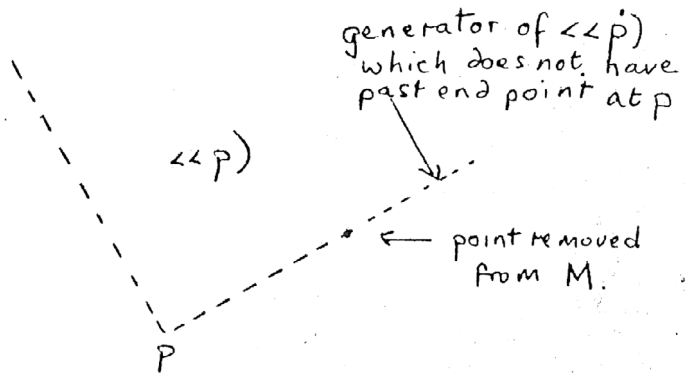

Fig. 8. An example of a space which is not causally simple.

generating segment of $\ll \dot{V}$) and ($\dot{V} \gg$ which intersects S has a past or future end point, respectively, at V.

The set ($p >$ consists of ($p \gg$ and ($p \rightarrow\!\!\!\!\succ$. Thus it has the same boundary as ($p \gg$. It can be seen that a region S is causally simple if and only if $S \cap (V >$ and $S \cap < V)$ are closed in S.

All the examples given in Section 3 are causally simple. Thus it would seem a fairly natural property. However, one can construct examples of spaces which are not causally simple (see Fig. 8).

Another example of a horizon would be the absolute event horizon of a point p [Rindler 1956]. This may be defined as the boundary of that region of spacetime in which no observer could ever learn about events at p. In other words, it would be the boundary of the set of points whose causal futures did not intersect the causal future of p, or equivalently, the boundary of the causal past of the causal future of p. It is easy to see that this set satisfies the condition of Lemma 17 and that the conditions of Lemma 18 are satisfied at each point of the boundary. Thus if the absolute event horizon of p exists, it will be generated by null geodesic segments which will have a past end point if they intersect another generating segment, but which can have no future end point.

It can be see from the Penrose diagram in Section 3 that there are no absolute event horizons in Minkowski space, as the boundary of its diagram is null. On the other hand, the diagram of de Sitter space has a spacelike boundary, so any point will have an absolute event horizon.

The dual of an event horizon is sometimes called a particle horizon [Penrose 1963, Rindler 1956]. The existence of such a horizon implies that an observer cannot see all the particles in the universe (see Fig. 9). Particle horizons occur in the Friedmann models filled with normal matter, as their diagrams all have spacelike boundaries at the initial singularity.

5.4 Causality

There was a young lady of Wight
Who travelled much faster than light.
She departed one day,
In a relative way,
And arrived on the previous night.

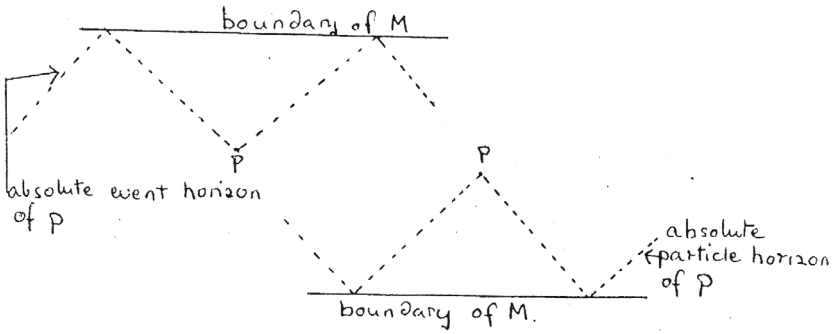

Fig. 9. Examples of event and particle horizons when the Penrose diagram of M has a spacelike boundary [Penrose 1963].

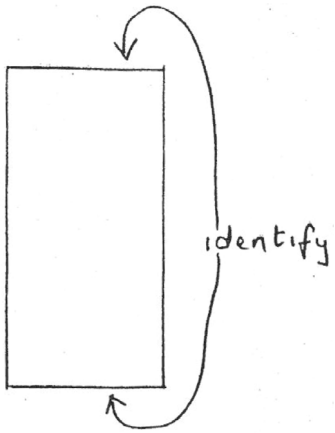

Fig. 10. Example where the Alexandroff topology is smaller than the manifold topology. In this case, the only open set is M itself.

There is a common type of science fiction story in which the hero travels into the past and accidentally kills one of his ancestors, or commits some other chronoclasmic act. The logical contradictions which arise from such stories would seem to be strong evidence for believing that such behaviour is impossible. Even if to travel into one's past involved going right round the universe, one would feel that this was a matter of principle and should not depend on the practical difficulties of constructing suitable spaceships. It would therefore seem reasonable to adopt:

> **The chronology assumption.** There are no closed timelike curves. Equivalent statements are that $\ll p) \cap (p \gg$ is empty or that there are no points q for which $q \ll q$.

Following Kronheimer and Penrose [Kronheimer 1967], we shall define the Alexandroff topology of M as that defined by the basis consisting of all sets of the form $\langle\langle p, q\rangle\rangle$, for $p, q \in M$. As $\langle\langle p, q\rangle\rangle$ is open in the manifold topology, the Alexandroff topology is not larger (finer) than the manifold topology. In fact, Figure 10 shows that it may be smaller. If we adopt the chronology assumption, M cannot be covered by a finite number of sets of the form $\langle\langle p, q\rangle\rangle$, as $\langle\langle p, q\rangle\rangle$ does not contain p. Thus M cannot be compact with respect to the Alexandroff topology or with respect to the manifold topology, as this is larger. This would seem to indicate that a compact manifold would

not be a physically realistic model for spacetime. In a sense this is a pity because there are a number of techniques such as harmonic analysis which can be applied effectively only to compact manifolds.

In order to get a deeper insight into the physical significance of chronology, it may be useful to consider how violations of it could occur. The following result is due to Carter. Let Q be the set of points of M at which chronology is violated. That is, from each point $q \in Q$, there is a timelike curve which returns to q. Then Q is the union of disjoint sets of the form $\langle\langle q_1, q_1\rangle\rangle$, $\langle\langle q_2, q_2\rangle\rangle$, and so on, where $q_1, q_2, \ldots \in Q$.

Let $q_1 \in Q$ and λ_1 be a timelike curve with past and future end points at q_1. Then $\langle\langle q_1, q_1\rangle\rangle$ is non-empty and is contained in Q, for if $p \in \langle\langle q_1, q_1\rangle\rangle$, then p can be joined to q by future- and past-directed timelike curves μ and $\overline{\mu}$, and $\mu \circ \lambda_1 \circ \overline{\mu}^{-1}$ is a future-directed timelike curve from p which returns to p. Thus $p \in Q$. Moreover, if $p \in \langle\langle q_1, q_1\rangle\rangle \cap \langle\langle q_2, q_2\rangle\rangle$, then $q_2 \in \langle\langle q_1, q_1\rangle\rangle$ and $\langle\langle q_1, q_1\rangle\rangle = \langle\langle q_2, q_2\rangle\rangle$. This shows that Q is the union of disjoint sets of the form $\langle\langle q_1, q_1\rangle\rangle$, etc.

We shall call the boundary of Q a chronology horizon. It is easy to see that the boundary of each disjoint component $\langle\langle q_1, q_1\rangle\rangle$ of Q consists of two parts: one of which forms part of the boundary of $\langle\langle q_1)$ and the other, part of the boundary of $(q_1\rangle\rangle$. As the boundary of $\langle\langle q_1)$ cannot contain q_1, it will be generated by null geodesic segments which have no past end point. Similarly, the boundary of $(q_1\rangle\rangle$ will be generated by null geodesic segments which have no future end points. Thus the chronology horizon, the boundary of $\langle\langle q_1, q_1\rangle\rangle$, will consist of two parts: the future (past) part being generated by null geodesic segments that have no future (past) end points. The two parts may or may not be connected.

Carter has pointed out that a chronology horizon occurs in the Kerr rotating solution [Boyer 1967]. This raises grave doubts as to whether it can be regarded as a physically realistic solution.

As well as making the chronology assumption, we could adopt:

The causality assumption. There are no closed non-spacelike curves. An equivalent statement is that $\langle p, p\rangle = p$ for any $p \in M$.

This is slightly stronger than the chronology assumption. Nevertheless, it would still seem a very reasonable requirement.

As before, we have the result that the set Q on which causality is violated will be the union of disjoint sets of the form $\langle q_1, q_1\rangle$, etc. Suppose that the chronology assumption held. Then each set $\langle q_1, q_1\rangle$ would consist of a single closed null geodesic curve, as any closed non-spacelike curve which was not a null geodesic curve could be deformed to give a closed timelike curve. We mentioned in Section 5.1 that completeness of a null geodesic curve was not in general a conformal property. However, it is in the case of a closed null geodesic curve. For the affine parameter $\tilde{v}$ in a metric $\tilde{g} = \Omega^2 \tilde{g}$ is related to the affine parameter v in the metric g by

$$\frac{\mathrm{d}\tilde{v}}{\mathrm{d}v} = \Omega^2.$$

The conformal factor Ω will have an upper and lower bound on the closed null geodesic curve. Thus arbitrary values of the parameter $\tilde{v}$ will be attained if and only if arbitrary values of v are. Let $v_1, v_2, \ldots,$ be successive values of v at q. The tangent vector at q, viz.,

$$\left.\frac{\partial}{\partial v}\right|_{v=v_1},$$

will be parallel to the tangent vector

$$\left.\frac{\partial}{\partial v}\right|_{v=v_2}.$$

Thus

$$\left.\frac{\partial}{\partial v}\right|_{v=v_1} = a^{-1}\left.\frac{\partial}{\partial v}\right|_{v=v_2} = a^{-2}\left.\frac{\partial}{\partial v}\right|_{v=v_3},$$

and so on, where a is some constant and

$$v_2 - v_1 = a(v_3 - v_2) = a^2(v_4 - v_3),$$

and so on. If $a < 1$, v will never attain the value of

$$v_2 - (v_2 - v_1)(1 - a)^{-1},$$

and if $a > 1$, v will not attain the value of

$$v_1 + (v_2 - v_1)\left(1 - \frac{1}{a}\right)^{-1}.$$

Thus the closed null geodesic curve is complete if and only if $a = 1$. The Misner space described in Section 3 contains closed null geodesic curves for which $a \neq 1$.

Consider for a moment the physical metric g. Suppose that the energy-momentum tensor satisfied the inequality $T_{ab}W^aW^b \geq 0$ for any non-spacelike vector W and that, in some region of a point q on a closed null geodesic $\lambda(v)$, the energy-momentum tensor satisfied the strict inequality $T_{ab}W^aW^b > 0$ (this could be interpreted physically as meaning that there was some matter at q). Let E_1, E_2, E_3, E_4 be a pseudo-orthonormal tetrad parallel transported along $\lambda(v)$ and let Z_m $(m = 1, 2)$ be variation vectors along $\lambda(v)$ defined by

$$Z_m = E_m \cos\frac{\pi v}{\ell},$$

where ℓ is some constant. Let Λ be

$$\Lambda = \int_{-\ell/2}^{\ell/2} g\left(\frac{\partial}{\partial v}, \frac{\partial}{\partial v}\right) \mathrm{d}v.$$

Then using the formula derived in Section 4, the variation in Λ induced by the variation vector Z_m is

$$\Lambda(Z_m, Z_m) = -2\int_{-\ell/2}^{\ell/2} g\left(Z_m, \left\{\frac{\mathrm{D}^2 Z_m}{\partial v^2} + R\left(\frac{\partial}{\partial v}, Z_m\right)\frac{\partial}{\partial v}\right\}\right) \mathrm{d}v.$$

But

$$\sum_m g\big(E_m, R(E_4, E_m)E_4\big) = -R(E_4, E_4).$$

Therefore,

$$\sum_m \Lambda(Z_m, Z_m) = -3\ell^{-1} + 2\int_{-\ell/2}^{\ell/2} R(E_4, E_4)\cos^2\frac{\pi v}{\ell}\mathrm{d}v.$$

But by the Einstein equations, $R(E_4, E_4) \geq 0$, and it is strictly positive in some neighbourhood of q. Thus by taking ℓ greater than some value $\overline{\ell}$, it follows that the quantity $\sum_m \Lambda(Z_m, Z_m)$ will be positive. This shows that there would be a point conjugate to $r_1 = \lambda(-\overline{\ell}/2)$ before $r_2 = \lambda(\overline{\ell}/2)$, and there would be a variation of $\lambda(v)$ between r_1 and r_2 which would give a timelike curve between r_1 and r_2. There are two possibilities: either the null geodesic $\lambda(v)$ is not complete and v does not

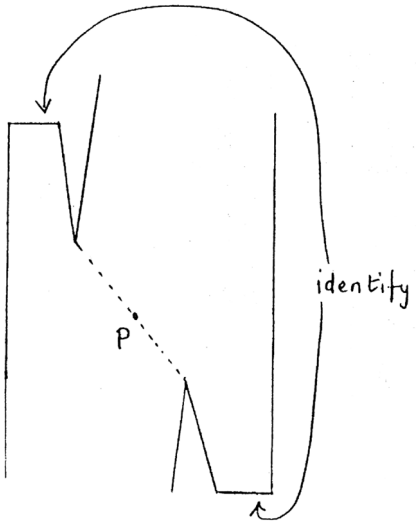

Fig. 11. Example of a space in which the causality but not the strong causality assumption holds. There is no small open neighbourhood of p that does not intersect some timelike curve more than once.

attain one of the values $-\overline{\ell}/2$ and $\overline{\ell}/2$, or there is also a violation of chronology. In general, if we assumed that every null geodesic encountered some matter somewhere along its length, then a violation of causality would imply either that there existed incomplete null geodesics or that there was also a violation of chronology. This would be an additional reason for regarding the causality assumption as essential on physical grounds.

As well as making the chronology and causality assumptions, it would seem reasonable to exclude the possibility of almost closed non-spacelike curves. In fact, Carter (private communication) has pointed out that there is an infinite hierarchy of such assumptions. However, for our purposes it will be sufficient to consider only what we shall call:

> **The strong causality assumption.** Every open neighbourhood of each point $p \in M$ contains an open neighbourhood D such that no non-spacelike curve intersects D more than once. An equivalent statement in terms of causal relations alone has been given by Kronheimer and Penrose [Kronheimer 1967]: if p_1 and p_2 are such that every $q_1 \ll p_1$ causally precedes every $q_2 \gg p_2$ and if $p_2 < p_1$, then $p_2 = p_1$.

The equivalent statement is proven as follows. Suppose $p_2 \neq p_1$. If V_1 and V_2 were any two open neighbourhoods of p_1 and p_2, respectively, there would be a past-directed non-spacelike curve from V_2 to V_1. If V_1 were taken sufficiently small, this curve could be extended to intersect V_2 again.

Figure 11 shows a pathological example where the causality assumption holds but not the strong causality assumption. However, it can be seen that, if the strong causality assumption did not hold, M would be on the verge of violating the chronology assumption. That is to say, the slightest disturbance of the metric (by quantum fluctuations, for instance) coud lead to the existence of closed timelike curves. It would not seem realistic to suppose that spacetime could judge as accurately as that to avoid a violation of chronology. Thus it would seem physically reasonable to assume strong causality.

Suppose that the causality assumption held but that strong causality was violated at some point p. Let U be a normal coordinate neighbourhood of p and $B \subset U$ a ball of constant coordinate radius aout p. Let $V_n \subset B$ be a series of open neighbourhoods of p such that, if W is any open neighbourhood of p, then W contains all V_n for n greater than some value. For each V_n, there will be a future-directed non-spacelike curve λ_n which leaves B and returns to V_n. Let r_1 be a limit point of where the λ_n intersect $\dot{B}$ for the first time. Then $r_1 \in < p)$, but not of $\ll p)$, since otherwise we could obtain a closed non-spacelike curve. Thus $p \rightarrow\!\!\!\!\succ r_1$. Choose a subsequence λ'_n which converges to r_1. Let r_2 be a limit point of where the λ'_n intersect $\dot{B}$ for the second time. Then $r_2 \rightarrow\!\!\!\!\succ p$. Moreover, r_2 cannot be in $(r_1 \gg$. Thus r_2, p, and r_1 lie on a null geodesic. Each point of the null geodesic segment $r_2 r_1$ will be a limit point of λ'_n. But then all points of the null geodesic must be limit points of λ'_n, since if q was the last point on the geodesic which was a limit point, a similar construction about q could be used to show that there were limit points beyond q. As the causality assumption holds, no two points on the null geodesic could be joined by a timelike curve.

If the energy-momentum tensor obeyed the inequality $T_{ab}W^aW^b \geq 0$ for any non-spacelike vector W, and if every null geodesic encountered some matter (somewhere where the strict inequality was satisfied), then the above result shows that a violation of strong causality would imply either that M was geodesically incomplete or that there was also a violation of chronology. This would be a further ground for thinking that the strong causality assumption is physically reasonable.

Let U be a normal coordinate neighbourhood of p and $D \subset U$ an open neighbourhood of p such that no non-spacelike curve intersects D more than once. Causal relations on D as a submanifold will coincide with the causal relations on M restricted to D (we shall call D a local causality neighbourhood). Thus the Alexandroff topology will agree with the manifold topology on D. If the strong causality assumption holds, M can be covered by the neighbourhoods D and the two topologies will be identical everywhere. This shows that the topological structure of M could be determined physically by observation of causal relationships. One would like to use these relations to determine the differential structure also, that is, to determine the admissible local coordinates. This is a non-trivial problem: in two dimensions, causality does not determine the differential structure, but in higher dimensions, it does. We first establish the following lemma.

Lemma 19. *If M and $\tilde{M}$ are n dimensional C^r manifolds $(n \geq 3, r \geq 3)$ with C^{r-1} Lorentz metrics g and $\tilde{g}$, such that the strong causality assumption holds on M, and if ϕ is a bijection $\phi : M \rightarrow \tilde{M}$ such that ϕ and ϕ^{-1} preserve causal relationships, then ϕ is a C^r diffeomorphism.*

As ϕ and ϕ^{-1} preserve causal relationships, they are continuous with respect to the Alexandroff topology on M and $\tilde{M}$. Since the strong causality assumption holds on M, it will also hold on $\tilde{M}$ and the Alexandroff topologies of M and $\tilde{M}$ will coincide with their manifold topologies. Let D be a local causality neighbourhood in M. Then $\phi(D)$ will be such a neighbourhood in $\tilde{M}$. Let $U \subset D$ and $\tilde{U} \subset \phi(D)$ be normal coordinate neighbourhoods of $p \in D$ and $\phi(p) \in \phi(D)$, respectively. A point $q \in U \cap \phi^{-1}(\tilde{U})$ can be reached from p by a future-directed null geodesic curve if and only if $p \rightarrow\!\!\!\!\succ q$. However, ϕ and ϕ^{-1} preserve causal relationships, so $p \rightarrow\!\!\!\!\succ q$ if and only if $\phi(q)$ can be reached from $\phi(p)$ by a future-directed null geodesic curve. Thus ϕ and ϕ^{-1} map null geodesic curves to null geodesic curves, though they may not map the parameter differentiably.

To prove that the parameter is mapped differentiably, we shall employ a construction used by Zeeman [Zeeman 1964]. Let $\lambda_1(t)$, $\lambda_2(t)$, $\lambda_3(t)$ be null geodesic curves in

U such that, for each point $q_2 \in \lambda_2$, there is a $q_1 \in \lambda_1$ with $q_1 \nrightarrow q_2$, for each point $q_3 \in \lambda_3$, there is a $q_2 \in \lambda_2$ with $q_2 \nrightarrow q_3$, and for some range of $q_1' \in \lambda_1$, there is a $q_3 \in \lambda_3$ with $q_1' \nrightarrow q_3$. [If U is sufficiently small, the components of the metric will differ by arbitrarily small amounts from their values in Minkowski space. Comparison with Minkowski space shows that such $\lambda_1(t)$, $\lambda_2(t)$, $\lambda_3(t)$ can be found.] The maps ψ and $\tilde{\psi}$ defined by $\psi(q_1') = q_1$ and $\tilde{\psi}\big(\phi(q_1')\big) = \phi(q_1)$ will be C^r diffeomorphisms. If M and $\tilde{M}$ were two-dimensional, ψ and $\tilde{\psi}$ would be the identity, but for higher dimensions, this is not necessarily the case. In fact, comparison with Minkowski space shows that each point $q_1 \in \lambda_1$ has an open neighbourhood V on λ_1 such that, if $p \in V$, then we could find λ_2 and λ_3 such that $\psi(p) = q_1$. Let $\tilde{t}$ be a C^r parameter on the null geodesic curve $\phi\big(\lambda_1(t)\big)$. Then ϕ maps t to $\tilde{t}$ continuously and monotonically. Thus $\phi(t)$ is differentiable almost everywhere by Lebesgue's theorem. Let E be the set of all values of t for which $\phi'(t)$ exists. We have

$$\phi\big(\psi(t)\big) = \tilde{\psi}\big(\phi(t)\big),$$

where ψ and $\tilde{\psi}$ are C^r functions. Then for $t \in E$,

$$\frac{\phi\big(\psi(t+\delta t)\big) - \phi\big(\psi(t)\big)}{\psi(t+\delta t) - \psi(t)} = \frac{\tilde{\psi}\Big\{\phi(t) + \big[\phi'(t) + \varepsilon\big]\delta t\Big\} - \tilde{\psi}\big(\phi(t)\big)}{\big[\psi'(t) + \eta\big]\delta t},$$

where $\varepsilon, \eta \to 0$ as $\delta t \to 0$.

Thus $\psi(E) \subset E$. But for any value of t, there is an open neighbourhood V such that if $t^1 \in V$, we can choose ψ so that $\psi(t^1) = t$. Thus $\phi(t)$ must be differentiable everywhere, and as

$$\phi'\big(\psi(t)\big)\psi'(t) = \tilde{\psi}'\big(\phi(t)\big)\phi'(t),$$

it follows that $\phi'(t)$ must be continuous. By repeating the above process, it may be shown that $\phi(t)$ is of class C^r. Thus ϕ maps a C^r parameter of a null geodesic curve to a C^r parameter. Comparison with Minkowski space shows that, in some neighbourhood $W \subset D$ of p, we can introduce four congruences of null geodesic curves such that any point $q \in W$ can be uniquely and differentiably described by the composition of a displacement a given parameter distance t_1 along a curve of the first congruence, followed by a displacement a given parameter distance t_2 along a curve of the second congruence, and so on. Then t_1, t_2, t_3, t_4 will be local coordinates of the complete C^r atlas of M. As ϕ maps null geodesic curves to null geodesic curves and maps their C^r parameters to C^r parameters, t_1, t_2, t_3, t_4 will also be local coordinates of the complete C^r atlas of M. Thus ϕ is a C^r diffeomorphism.

In Section 3, we saw how the metric could be measured physically using causal relations. However, this presupposed that we could recognise the differential structure of M, i.e., that we could tell which local coordinates would be admissible. This would normally be fairly obvious, but in the vicinity of a singularity, it might not be so simple. However, by Lemma 19, we may determine the differential structure of M by observation of causal relationships. For if V_α is a subset of M and ϕ_α is a bijection of V_α to an open set of $\mathbb{R}^4$ such that, for some C^r Lorentz metric on $\phi_\alpha(V_\alpha)$, both ϕ and ϕ^{-1} preserve causal relations, then (V_α, ϕ_α) belong to the complete C^r atlas of M. It can be seen that, if the metric on M is assumed to be only C^{r-1}, then it is possible to determine physically only the C^r differential structure of M. Thus, as was said in Section 3, there is no point in assuming a C^∞ structure for M unless we also assume that the metric is C^∞.

The strong causality assumption has another consequence which will be important in the next section. This is the non-existence of partially imprisoned non-spacelike lines. By a future-directed non-spacelike line we shall mean a broken non-spacelike

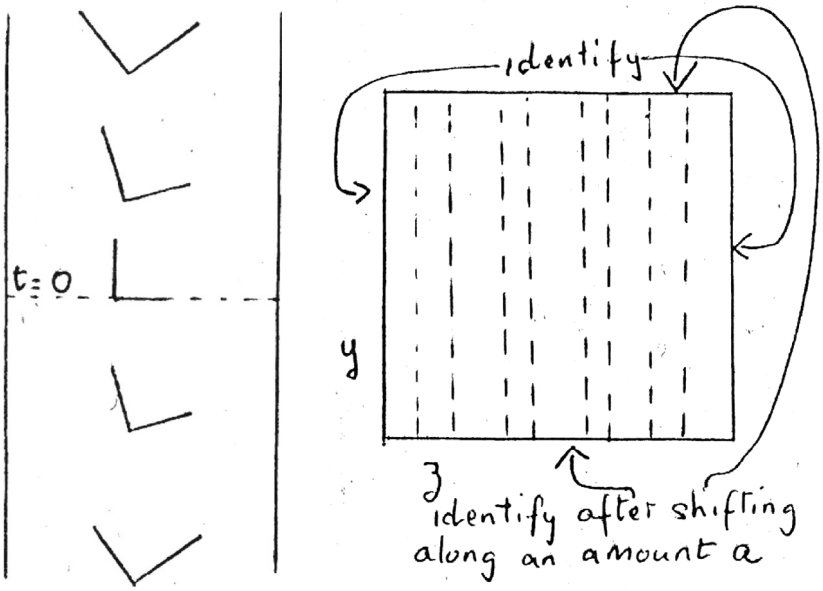

Fig. 12. An example with imprisoned non-spacelike lines but no closed non-spacelike curves, viz., the manifold $\mathbb{R}^1 \times \mathbb{S}^1 \times \mathbb{S}^1$ described by coordinates t, y, ξ, where (t, y, ξ) and $(t, y, \xi + 1)$ are identified and (t, y, ξ) and $(t, y + 1, \xi + a)$ are identified, with a an irrational number. The Lorentz metric may be taken as: $\mathrm{d}s^2 = (\cosh t - 1)^2(\mathrm{d}t^2 - \mathrm{d}y^2) + \mathrm{d}t\mathrm{d}y - \mathrm{d}\xi^2$.

curve which is inextendible in a future direction. In other words, it has no future end point. We shall say that a future-directed non-spacelike line λ is partially future-imprisoned if there is a compact set N such that for every $p \in \lambda$, there is a $q \in \lambda \cap N$ with $p < q$, or in other words, λ keeps on intersecting N. We shall say that λ is totally future-imprisoned if for some point $p \in \lambda$, we have $< p) \cap \lambda \subset N$. In other words, λ does not leave N in the future direction. Such a line must wind round and round inside N and one might feel that this was possible only if there existed closed non-spacelike curves. However, Carter has given an example, shown in Figure 12, which disproves this. Nevertheless, we have the following result.

Lemma 20. *If the strong causality assumption holds, there can be no partially (or totally) imprisoned non-spacelike lines.*

Any compact set N can be covered by a finite number of local causality neighbourhoods D. A future-directed non-spacelike line which intersects a neighbourhood D must leave it again and not re-enter. Thus there must be a point $p \in \lambda$ such that there is no point $q \in \lambda \cap N$ with $p < q$.

5.5 Cauchy surfaces

We shall call a closed (not necessarily compact), connected, spacelike 3-surface (without boundary) a slice, and a slice which does not intersect any non-spacelike curve more than once a partial Cauchy surface. We have the following result:

Lemma 21. *If there exists a slice H in M, then there is a covering manifold $\widehat{M}$ with projection $\pi : \widehat{M} \to M$ such that each connected component of $\pi^{-1}(H)$ is diffeomorphic to H and is a partial Cauchy surface in $\widehat{M}$. The latter may be identical to M.*

Let $\widehat{M}$ be the space of all pairs $(p, [\lambda])$, where $p \in M$ and $[\lambda]$ is an equivalence class of curves from p to H, homotopic modulo p and H. The topological and differential structure of $\widehat{M}$ is defined in a natural way as follows. Let $\{V_\alpha, \phi_\alpha\}$ be the complete atlas of M. Let p and q be in V_α and μ be a curve in V_α from q to p. The set $\widehat{V}_\alpha$ is defined as the union over V_α of all parts of the form $(q, [\lambda \circ \mu])$, where $\lambda \circ \mu$ is the juxtaposition of λ and μ. Then $\{\widehat{V}_\alpha, \phi_\alpha \pi\}$ is an atlas for $\widehat{M}$, where $\pi : \widehat{M} \to M$ is the covering map defined by

$$\pi(p, [\lambda]) = p.$$

The manifold $\widehat{M}$ has a natural metric induced from M. Let $\widehat{H}$ be the set $(p, [0])$, $p \in H$. Then $\widehat{H}$ is a connected component of $\pi^{-1}(H)$ and is diffeomorphic to H. As M and $\widehat{M}$ are time orientable, $\widehat{H}$ will be a two-sided surface. Suppose there were a non-spacelike curve γ in $\widehat{M}$ which intersected $\widehat{H}$ only at two distinct points r_1 and r_2. We could join r_2 to r_1 by a curve μ in $\widehat{H}$, and $\mu \circ \gamma$ would be a closed loop from r_1 which crossed $\widehat{H}$ only once. It could not be deformed to zero, since a continuous deformation could change the number of times it crossed $\widehat{H}$ by only an even number, as $\widehat{H}$ has no edges. Thus γ would be a curve from r_1 to $\widehat{H}$ which could not be deformed to a zero curve from r_1 to $\widehat{H}$. This is impossible, so no non-spacelike curve can intersect $\widehat{H}$ more than once.

The physical significance of a partial Cauchy surface H is that data on it determine events in some region J of M which we shall call the Cauchy development of H. We may define J as $J^+ \cup J^-$, where J^+ (J^-) is called the future (past) Cauchy development of H and is defined as the set of all points q such that every past- (future-) directed non-spacelike line through q intersects H. Clearly, points sufficiently near H will be in J. By the future (past) Cauchy horizon L^+ (L^-) of H we shall mean $\dot{J}^+ - H$ ($\dot{J}^- - H$). Using Lemmas 5.1 and 5.2, it can be seen that L^+ (L^-) is generated by null geodesic segments which have no past (future) end points. We shall say that H is a Cauchy surface if L^+ and L^- are empty, or equivalently, if $J = M$. A necessary and sufficient condition for H to be a Cauchy surface is that it intersect every null geodesic, for then, if its Cauchy horizons were non-empty, they would intersect H, which would be impossible. The existence of a Cauchy surface H would imply that M was diffeomorphic to $H \times \mathbb{R}^1$. For as we saw in Section 5.2, M admits a non-vanishing future-directed timelike vector field X. Each integral curve of X will intersect H and can be parametrised in a differentiable manner from $-\infty$ to ∞.

We shall deduce a number of properties of partial Cauchy surfaces which will be used in the theorems on singularities in Section 6. For a set S, we shall define $Q^+(S)$, resp. $Q^-(S)$, to be the set of all points q which have an open neighbourhood W such that $\overline{\langle S, W\rangle}$, resp. $\overline{\langle W, S\rangle}$, is compact or empty and such that the strong causality assumption holds at all points of $\overline{\langle S, W\rangle}$, resp. $\overline{\langle W, S\rangle}$. We shall define $C^+(S)$, resp. $C^-(S)$, the future, resp. past, compact region of S to be $\langle S) \cap Q^+(S)$, resp. $(S > \cap Q^-(S)$. If $q_1, q_2 \in C^+(S)$, then $\langle q_1, q_2\rangle \subset C^+(S)$. Moreover, as the strong causality assumption holds on the compact set $\overline{\langle q_1, q_2\rangle}$, every null geodesic segment generating the boundary of $\langle q_1)$, resp. $(q_2\rangle$, which intersects $\overline{\langle q_1, q_2\rangle}$ will have a past, resp. future, end point at q_1, resp. q_2. thus $\langle q_1, q_2\rangle$ will be closed.

Lemma 22. *If $Q^+(S)$ is non-empty, it will be a null horizon generated by null geodesic segments which have no past end point.*

The set N defined as $M - Q^+(S)$ is such that $q \in N$ implies $\ll q) \subset N$. We shall show that the conditions of Lemma 18 are satisfied for each point $q \in \dot{N}$. Let U be a normal coordinate neighbourhood of q and $B \subset U$ a ball of constant coordinate radius

about q. Let r_n be a sequence of points in $B \cap N$ converging to q, with $r_{n+1} \ll r_n$. Let V_n be a sequence of open neighbourhoods with $r_{n+1} \in V_{n+1} \subset (r_n \gg$. Suppose that, for some n, $\dot{B} \cap (\overline{V}_n \overset{U}{>}$ was contained in $Q^+(S)$. Then since $\dot{B} \cap (\overline{V}_n \overset{U}{>}$ is compact, it could be covered by a finite number of neighbourhoods W_i for which $\overline{\langle S, W_i \rangle}$ was compact. Thus $\overline{\langle S, V_n \rangle}$ would be contained in

$$\left[\bigcup_i \overline{\langle S, W_i \rangle} \right] \cup \left[\overline{B} \cap (\overline{V}_n \overset{U}{>} \right],$$

and so would be compact. Suppose that the strong causality assumption did not hold at some point $p \in \overline{B} \cap (\overline{V}_n \overset{U}{>}$. Then using the construction employed in the previous section, there would be a null geodesic γ (without end points) through p at every point of which the strong causality assumption would not hold. This would be impossible as γ would intersect $Q^+(S)$. Thus the strong causality assumption would hold on $\overline{\langle S, V_n \rangle}$. But this would imply that r_n was in $Q^+(S)$, which is not the case. Thus from each $\overline{V}_n$ and from each r_{n-1}, there must be a past-directed non-spacelike curve which intersects $\dot{B} \cap N$. By Lemma 18, $\dot{N} = \dot{Q}^+(S)$ will be a null horizon generated by null geodesic segments which have no past end points.

Corollary 1. *For a partial Cauchy surface H, $C^+(H) = J^+$, the future Cauchy development of H.*

Clearly, H is in $Q^+(H)$. The boundary of $Q^+(H)$ cannot intersect J^+ since it would then intersect H, which would be impossible. Thus $C^+(H)$ contains J^+. Through a point $q \in L^+$, the future Cauchy horizon of H, there would a past-directed null geodesic line λ generating L^+ which was contained in $\overline{\rangle H, q \rangle}$. Thus the strong causality assumption could not hold everywhere on $\overline{< H, q >}$ if it were compact. This shows that q would not be in $C^+(H)$. Therefore $C^+(H) = J^+$.

Corollary 2. *If S is a compact set contained in J, then $\langle S) \cap J$ will be contained in $C^+(S)$.*

By the previous corollary, a point $q \in J$ will have an open neighbourhood W such that the strong causality assumption holds on the set $\overline{\langle H, W \rangle}$, which is compact or empty. Similarly, S can be covered by a finite number of open neighbourhoods W_i such that the strong causality assumption holds on the sets $\overline{\langle W_i, H \rangle}$, which are compact. Thus $q \in Q^+(S)$, since $\overline{\langle S, W \rangle}$ will be contained in

$$\left[\bigcup_i \overline{\langle W_i, H \rangle} \right] \cup \overline{\langle H, W \rangle}.$$

Lemma 23. *$C^+(S)$ is causally simple.*

We have to show that, if V is a compact set in $C^+(S)$, then each null geodesic segment generating the boundary of $(V >$, resp. $< V)$, which intersects $C^+(S)$ has a future, resp. past, end point on V. As V is compact, it can be covered by a finite number of open neighbourhoods W_i for which $\overline{\langle S, W_i \rangle}$ is compact. Thus $\overline{\langle S, V \rangle}$ is compact. Suppose that there is a point $q \in C^+(S)$ on geodesic segments λ of $(\dot{V} >$. The future-directed segment λ could have a future end point only on V. It will enter $\overline{\langle S, V \rangle}$ and not leave it again. Thus as the strong causality assumption holds on $\overline{\langle S, V \rangle}$, λ must have a future end point on V. Similarly, if $q \in C^+(S)$ is a point on a generating segment of $< \dot{V})$, the segment will enter the compact set $\overline{\langle S, q \rangle}$, and so will have a past end point on V.

Corollary 1. *J is causally simple.*

A null geodesic generating segment of ($\dot{V} >$ which intersects J^- will either have a future end point on $V \cap J^-$, or it will intersect H and the compact set $\overline{\langle H, V\rangle}$ and so have a future end point on $V \cap J^+$.

For points q_1, q_2 with $q_2 \gg q_1$, $d(q_1, q_2)$ will be defined to be the least upper bound of the lengths of timelike curves from q_1 to q_2. It will be defined to be zero for q_2 not in $\langle\langle q_1)$. For sets S_1, S_2, $d(S_1, S_2)$ will be defined as the least upper bound of $d(q_1, q_2)$ for $q_1 \in S_1$, $q_2 \in S_2$. If $d(q_1, q_2)$ is finite, given any $\delta > 0$, we could find a timelike curve λ from q_1 to q_2 of length greater than $d(q_1, q_2) - \delta/2$. Then we could find an open neighbourhood V of q_2 such that λ could be deformed to give a timelike curve from q_1 to any point of V of length greater than $d(q_1, q_2) - \delta$. Thus $d(q_1, q_2)$, where finite, is lower semi-continuous in q_1 and q_2. It is not, however, necessarily continuous.

Lemma 24. *$d(S, q)$ is finite for $q \in C^+(S)$.*

Let N be the set of all points q for which $d(S, q)$ is infinite. By Lemmas 17 and 18, $\dot{N}$, if non-empty, will be generated by null geodesic segments which have no past end points. Suppose that there was a point $r \in C^+(S) \cap N$. Let W be an open neighbourhood of r such that $\overline{\langle S, W\rangle}$ was compact. There would be a point $p \in \overline{\langle S, W\rangle}$ for which there was no point $q \in S$ such that $q \ll p$, as otherwise there would be an infinite sequence of points $p_n \in \overline{\langle S, W\rangle}$ with $p_{n+1} \ll p_n$ such that $\{\ll p_n)\}$ would give an open covering of $\overline{\langle S, W\rangle}$ which contained no finite subcovering. Then $d(S, p) = 0$. This shows that there would be a generating segment λ of $\dot{N}$ which intersected $\langle S, r\rangle$. By Lemma 20, the past-directed null geodesic line λ would have to leave $\overline{\langle S, W\rangle}$. As λ would not leave the open set ($W \gg$, it would leave $\overline{\langle S)}$. That would be impossible as any open neighbourhood of a point on λ would contain a point q for which $d(S, q)$ was infinite and $d(S, q)$ is zero for q not in $\langle\langle S)$. Thus $d(S, q)$ is finite for $q \in C^+(S)$.

Lemma 25. *$d(q_1, q_2)$ is continuous in q_1 and q_2, and $d(S, q_2)$ is continuous in q_2, for q_1, q_2 restricted to $C^+(S)$.*

Suppose q_1 and q_2 lie in a local causality neighbourhood D. As D is convex, there will be a unique geodesic $\gamma(v)$ in D with $\gamma(0) = q_1$ and $\gamma(1) = q_2$. This geodesic will depend differentiably on q_1 and q_2, and the quantity

$$\Delta(q_1, q_2) = \int_0^1 g\left(\frac{\partial}{\partial v}, \frac{\partial}{\partial v}\right) \mathrm{d}v$$

will be a differentiable function on $D \times D$. If $\gamma(v)$ is timelike, i.e., $\Delta > 0$, it will be the longest timelike curve from q_1 to q_2. If $\gamma(v)$ is null or spacelike, there will be no timelike curve from q_1 to q_2. Thus $d(q_1, q_2)$ will be $\left[\Delta(q_1, q_2)\right]^{1/2}$ if $\Delta > 0$ and $q_2 \gg q_1$ and will be zero otherwise. It will therefore be a continuous function on $D \times D$.

Suppose that $d(q_1, q_2)$ had a discontinuity $\delta > 0$ in q_2 for $q_2 = r \in C^+(S)$. By this is meant that δ is the least upper bound of $\varepsilon > 0$ such that every open neighbourhood of r contains a point q_2 such that $d(q_1, q_2) > d(q_1, r) + \varepsilon$. Let D be a local causality neighbourhood of r and $B \subset D$ a ball of constant coordinate radius about r. As d is continuous on $D \times D$, we could find an open neighbourhood V of r such that

$$d\Big(V, (\dot{r} > \cap B\Big) < \frac{1}{2}\delta.$$

Let y_n be a sequence of points in V converging to r, such that

$$d(q_1, y_n) > d(q_1, r) + \delta\left(1 - \frac{1}{n+3}\right).$$

Then from each y_n, we could find a timelike curve λ_n to q_1 of length greater than

$$d(q_1, r) + \delta\left(1 - \frac{1}{n+2}\right).$$

Let ζ be a limit point of $\lambda_N \cap \dot{B}$. It would be in $\overline{\langle r\rangle\rangle} \cap \dot{B}$ but not in $\langle r\rangle\rangle \cap \dot{B}$, as otherwise there would be a curve λ_n which could be deformed to give a timelike curve from q_1 to r of length greater than $d(q_1, r)$. Thus ζ would lie on a past-directed null geodesic γ through r. Any open neighbourhood of ζ would contain a point q_2 such that $d(q_1, q_2)$ was arbitrarily close to $d(q_1, r) + \delta$. However, $d(q_1, \zeta)$ would be less than or equal to $d(q_1, r)$. [If not, there would be an open neighbourhood of ζ on which $d(q_1, q_2) > d(q_1, r)$ and so there would be a timelike curve from q_1 to r of length greater than $d(q_1, r)$.] Thus $d(q_1, q_2)$ would have a discontinuity in q_2 of not less than δ for $q_2 = \zeta$. In fact, this would be true for all points on the past-directed null geodesic line γ, as if p were the least bound of the points at which there was such a discontinuity. We could perform a similar construction about p and show that there were points on γ and beyond p at which there was the same discontinuity. If W was an open neighbourhood of r such that $\overline{\langle S, W\rangle}$ was compact, γ would leave $\overline{\langle S, W\rangle}$. As before, this would imply that γ left $\overline{\langle S\rangle}$, which would be impossible as $d(q_1, q_2)$ is zero and therefore, trivially, continuous for q_2 not in $< q_2$). Thus $d(q_1, q_2)$ is continuous in q_2. Similarly, $d(S, q_2)$ will be continuous in q_2.

Suppose that, at $q_1 \in C^+(S)$, $d(q_1, q_2)$ had a discontinuity $\delta > 0$ in q_1 restricted to $\langle S)$. By this is meant that δ is the least upper bound of $\varepsilon > 0$ such that the intersection of $\langle S)$ with every open neighbourhood of r contains a point q_1 for which $d(q_1, q_2)$ is greater than $d(r, q_2) + \varepsilon$. By a construction similar to the above, there would be a future-directed null geodesic line γ through r at each point of which $d(q_1, q_2)$ would have a discontinuity δ in q_1 for q_1 restricted to $\langle S)$. But γ would have to leave the compact set $\overline{\langle S, q_2\rangle}$. This would be impossible as there would then be a point of γ which had an open set which did not intersect $\langle S, q_2\rangle$. Thus $d(q_1, q_2)$ is continuous in q_1 and q_2 for q_1 and q_2 restricted to $C^+(S)$.

Corollary 1. *If S is a compact set in J, then $d(S, q)$ is continuous in q for $q \in J$.*

By the second corollary to Lemma 22, $\langle S)\cap J$ is in $C^+(S)$. By Lemma 23, J is causally simple. Therefore $\langle S) \cap J = \overline{\langle S)} \cap J$. But $d(S, q) = 0$, for q not in $\langle\langle S)$. Thus $d(S, q)$ is continuous in q for $q \in J$. In particular, $d(q_1, q_2)$ is continuous in q_1 and q_2 for $q_1, q_2 \in J$.

The example of anti-de Sitter space shows that $d(q_1, q_2)$ and $d(H, q_2)$ are not necessarily continuous if q_1 or q_2 are not in J. We shall use the continuity of $d(q_1, q_2)$ to prove the existence of a timelike curve from q_1 to q_2 of maximum length.

Lemma 26. *If $q_1, q_2 \in C^+(S)$ with $q_2 \gg q_1$, there is a timelike geodesic curve of length $d(q_1, q_2)$ from q_1 to q_2.*

Let $D \subset \langle q_2\rangle\rangle$ be a local causality neighbourhood of q_1 and $B \subset D$ a ball of constant coordinate radius about q_1. Let r be such that $d(q_1, q) + d(q, q_2)$ with $q \in \langle q_1) \cap \dot{B}$ is maximised for $q = r$. Let γ be the future-directed non-spacelike geodesic line from q_1 through r. The relation $d(q_1, q) + d(q, q_2) = d(q_1, q_2)$ will hold for all points q on γ between q_1 and r. Let p be the least bound of the points of γ for which the relation holds. Suppose that $d(q_1, p) < d(q_1, q_2)$. As $d(q, q_2) = 0$ for q not in $\langle q_2\rangle\rangle$, p would be in $\overline{\langle q_1, q_2\rangle}$. But $\overline{\langle q_1, q_2\rangle}$ is contained in $C^+(S)$. Thus, as d is continuous on $C^+(S)$, the relation would also hold at p and p would be in $\langle q_2\rangle\rangle$. Let $\tilde{D} \subset \langle q_2\rangle\rangle$ be a local causality neighbourhood of p and $\tilde{B} \subset \tilde{D}$ a ball of constant coordinate radius

about p. Let y_1 and y_2 be the two intersections of γ with $\dot{\dot{B}}$ and let $\tilde{r}$ be such that $d(p,q) + d(q,q_2)$ for $q \in < p) \cap \dot{\dot{B}}$ was maximised for $q = \tilde{r}$. Then

$$d(p,\tilde{r}) + d(\tilde{r},q_2) = d(p,q_2),$$

and $\tilde{r}$ would have to coincide with y_2, as otherwise

$$d(y_1,\tilde{r}) > d(y_1,p) + d(p,\tilde{r}),$$

and so

$$d(q_1,y_1) + d(y_1,\tilde{r}) + d(\tilde{r},q_2) > d(q_1,q_2).$$

Thus the relation

$$d(q_1,q) + d(q,q_2) = d(q_1,q_2)$$

holds for all points q for which $d(q_1,q) < d(q_1,q_2)$. As $\overline{\langle q_1,q_2\rangle}$ is compact, γ must leave $(q_2\rangle$ at some point ζ. As d is continuous on $\overline{\langle q_1,q_2\rangle}$, the relation will also hold at ζ. But $d(\zeta,q_2) = 0$. Thus $d(q_1,\zeta) = d(q_1,q_2)$. Suppose that ζ lay on the boundary of $(q_2\rangle$, not at q_2. As $C^+(S)$ is causally simple, ζ would lie on a null geodesic segment λ from q_2. Then the juxtaposition of γ, from q_1 to ζ, and λ would give a broken non-spacelike geodesic curve from q_1 to q_2 of length $d(q_1,q_2)$. There would be a variation of this curve which gave a longer timelike curve from q_1 to q_2. Thus ζ coincides with q_2 and γ is a timelike geodesic curve of length $d(q_1,q_2)$ from q_1 to q_2.

Corollary 1. *If q is in J^+, the future Cauchy development of H, there is a geodesic curve orthogonal to H of length $d(H,q)$ from H to q.*

By Lemmas 22 and 25, $d(p,q)$ for $p \in H$ will attain its maximum value for some p in the compact set $\overline{H \cap (q >}$. This timelike geodesic curve of length $d(H,q)$ from p to q must be orthogonal to H as otherwise there would be a variation which would give a longer timelike curve from H to q.

Corollary 2. *If S_1 and S_2 are compact sets in J with $\langle\langle S_1,S_2\rangle\rangle$ non-empty, there is a timelike geodesic curve from S_1 to S_2 of length $d(S_1,S_2)$.*

The quantity $d(p_1,S_2)$ for $p_1 \in S_1$ will attain its maximal value for some p_1. Then $d(p_1,p_2)$ for $p_2 \in S_2$ will attain its maximum value for some p_2. There will be a timelike geodesic curve from p_1 to p_2 of length $d(S_1,S_2)$.

6 Singularities

6.1 Incompleteness and inextendibilty

We must decide what we are going to mean by a singularity of spacetime. An obvious definition would seem to be that it is a point where the metric is singular (fails to be Lorentz or fails to be suitably differentiable). However, the trouble with this is that we could simply cut out the singular points and say that the remaining manifold represented all of spacetime. Indeed, it would seem undesirable to include the singular points in the definition of spacetime, as if we did, we would be introducing something into the theory which was not physically observable, namely the manifold structure and metric at those points. For as we saw in Section 5, the manifold structure can be physically determined only where the metric is non-singular. On the other hand, we want to omit only the singular points and not perfectly non-singular points as well. We shall say that the spacetime manifold M is extendible if it can be imbedded as an

open submanifold in a larger four-dimensional, connected, paracompact, C^4 manifold and that it is metrically extendible if there is a C^3 (Lichnerowicz: C^1, piecewise C^3) Lorentz metric on the larger manifold which coincides with the physical metric on M. The Schwarzschild solution in the original Schwarzschild coordinates provides a good example of a manifold which is metrically extendible.

In order to make sure that no non-singular points are left out of the definition of spacetime, we shall supplement postulates (a) and (b) of Section 3 by:

Postulate (c). The spacetime manifold M is metrically inextendible.

Although we have omitted the singular points from the definition of spacetime, we can still recognise the 'holes' left where they have been cut out by the existence of incomplete geodesics. Thus it would seem reasonable to make geodesic incompleteness the basis of our definition of singularities of spacetime. We can distinguish three kinds of incompleteness: that of spacelike, null, and timelike geodesics. They are not equivalent [Kundt 1963]. The first kind has no particular physical significance as spacelike geodesics are not important in the theory of relativity, but the second and third kinds have a most immediate significance: they imply that there could be particles or photons whose histories would not exist after (or before) a certain length of time or affine parameter, as measured by them. In other words, they would apparently be annihilated (or created). It is this feature which many people have found so objectionable and which has led them to suggest that, under realistic conditions, the general theory of relativity would not predict the occurrence of singularities.

Of course, the cutting out of a singular point is not the only way in which geodesic incompleteness could occur. The Misner space described in Section 3 contains future-directed timelike geodesics which are totally future-imprisoned in a compact set and which do not attain arbitrary length. However, this Misner incompleteness would seem equally objectionable. It would imply that particles could apparently be annihilated (or created) in a compact region. By Lemma 20 in Section 5, Misner incompleteness would also imply that the strong causality assumption was violated.

We shall therefore take timelike and null geodesic incompleteness as our definition of a singularity of spacetime. More precisely, we shall say that M is singularity-free if and only if it is timelike and null geodesically complete.

6.2 The energy assumption

In order to predict the occurrence of singularities, we have to make some assumption about the energy-momentum tensor, as otherwise any manifold and Lorentz metric could be regarded as a solution of the Einstein equations. On the other hand, we do not know the exact form of the energy-momentum tensor of the matter in the universe. Thus to be physically realistic, a theorem predicting the occurrence of singularities should depend only on some fairly general assumption about the nature of the energy-momentum tensor. The theorems that will be presented satisfy this requirement. In order of increasing strength, the assumptions that will be used are:

The weak energy assumption. The energy-momentum tensor obeys the inequality $T_{ab}W^aW^b \geq 0$ for any timelike vector W.

By continuity, this will then also be true for any null vector W. To get an idea of what this means, consider an observer whose worldline has unit tangent vector V. The energy density of matter as measured by him is $T_{ab}V^aV^b$. Thus the weak energy assumption is equivalent to assuming that the energy density is non-negative to every observer. This would seem very reasonable physically. Indeed, there would seem to be grave quantum mechanical difficulties associated with the existence of negative energy

density. For it seems that there would not be anything to prevent the creation of pairs consisting of a quantum of negative energy and a quantum of positive energy. Even if the cross-section for this pair production were very low, the infinite phase space available to the quanta would cause an infinite number of such pairs to be produced in any given region of spacetime.

In general, the energy-momentum tensor can be expressed in the form

$$T_{ab} = \mu V_a V_b + \sum_{i=1}^{3} p_i \underset{i}{Z}_a \underset{i}{Z}_b,$$

where V and $\underset{i}{Z}$ are eigenvectors of the energy-momentum tensor and form an orthonormal basis. In the exceptional case when one of the eigenvectors is null, it can be expressed as

$$T_{ab} = \mu K_a K_b + \overline{\mu}(K_a L_b + K_b L_a) + \sum_{i=1}^{2} p_i \underset{i}{Z}_a \underset{i}{Z}_b,$$

where K and L are null and $\underset{i}{Z}$ are two unit spacelike vectors orthogonal to K and L and to each other. The quantity μ will be the energy density to an observer whose worldline is tangent to V and p_i will be the pressures in the three spatial directions $\underset{i}{Z}$. The weak energy assumption will hold if

$$\mu + p_i \geq 0,$$

in the general case, and if

$$\mu \geq 0, \qquad \overline{\mu} \geq 0, \qquad p_i \geq 0,$$

in the exceptional case. This certainly holds for any known form of matter and also for all the projected equations of state for matter at supernuclear densities. In particular, it can be seen that it does not matter how great the pressure becomes: the only things that are ruled out are negative energy densities, or large negative pressures.

The energy assumption. The energy-momentum tensor obeys the inequality

$$T_{ab} W^a W^b \geq \frac{1}{2} W_a W^a T,$$

for any timelike vector W.

In terms of the decomposition given above, this holds if

$$\mu + p_i \geq 0, \quad \mu + \sum_i p_i \geq 0,$$

or in the exceptional case, if

$$\overline{\mu} \geq 0, \quad \mu \geq 0, \quad p_i \geq 0.$$

This assumption is slightly stronger but again seems very reasonable physically and would hold for any known form of matter. It rules out only negative energy densities and large negative pressures.

The strong energy assumption. The energy-momentum tensor obeys the inequality

$$T_{ab}W^aW^b > \frac{1}{2}W_aW^aT,$$

for any non-spacelike vector W.

This holds in the general case if

$$\mu + p_i > 0, \quad \mu + \sum_i p_i > 0,$$

but not in the exceptional case. It may be thought of as implying that there is some matter with nonzero rest mass present everywhere.

6.3 Theorems on singularities

In Section 3, we saw that there were singularities in any Robertson-Walker solution for which the strong energy assumption held. However, we could not conclude from this that there would necessarily be singularities in more realistic solutions with local irregularities. The exact spatial homogeneity and spherical symmetry of the Robertson-Walker solutions required that the matter was a perfect fluid whose flow lines formed a geodesic congruence with zero vorticity and shear, and whose density and pressure had zero spatial gradients, i.e., were constant in the 3-surfaces orthogonal to the flow lines. To see what could be the effects of acceleration, vorticity, and shear of the flow lines of a perfect fluid, consider the Raychaudhuri equation. This gives the rate of change of the expansion θ as

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -\frac{1}{3}\theta^2 - 2\sigma^2 - \frac{1}{2}(\mu + 3p) + 2\omega^2 + \dot{V}^a{}_{;a}.$$

This shows that shear induces contraction, as do the density and pressure if the strong energy assumption holds. However, vorticity induces expansion, as might be expected, while aceleration of the flow lines can induce either expansion or contraction depending on the sign of its divergence. The acceleration is given by

$$(\mu + p)\dot{V} = h^{ab}p_{;b}.$$

It will vanish if the spatial gradient of the pressure is zero. This explains why high pressure cannot prevent the occurrence of a singularity in the Robertson-Walker solutions: it is only pressure gradients that help.

If the vorticity and pressure were zero, the Raychaudhuri equation would give that θ would become infinite within a finite distance to the future or to the past along each geodesic flow line. This would imply a singularity as infinitesimally neighbouring flow lines would intersect and the density would become infinite. However, vorticity and pressure vanishing is clearly a very special case. If vorticity were present it might prevent the flow lines from converging in the two spatial dierections orthogonal to the axis of vorticity. One might think that there would be nothing to stop the flow lines converging along the axis of vorticity. However, the rate of change of the rate of separation tensor is

$$\dot{\psi}_{ab} = -C_{acbd}V^cV^d - \frac{1}{3}h_{ab}R_{cd}V^cV^d - \omega_a{}^c\omega_{cb} - \psi_a{}^c\psi_c{}^b + \dot{V}_{(a;b)} - \dot{V}_a\dot{V}_b.$$

This involves the 'electric' components E_{ab} of the Weyl tensor. These could be such as to induce an expansion in the direction of the axis of vorticity which would counteract

the contraction induced by the Ricci tensor term $h_{ab}(\mu + 3p)/6$. We saw in Section 4 that E_{ab} could be regarded as representing the gravitational effect at a point of matter at other points of spacetime. Thus a purely local proof of the occurrence of singularities does not seem possible because we could always suppose that a sufficiently large E_{ab} field existed locally, maybe as a gravitational wave propagated from another part of the universe.

It will therefore be necessary to make some global assumptions about the universe. Our guide as to what global assumptions would be reasonable will be the Copernican principle. However, it would be difficult to check such assumptions by observation as we can only see part of the universe.

An obvious generalisation of the Robertson-Walker solutions, which retained the Copernican principle, would be solutions which were spatially homogeneous but not spherically symmetric. Such solutions would permit the flow lines of the matter to have acceleration, vorticity, and shear. Although the universe seems approximately spherically symmetric at the present time, there might have been large anisotropies at an earlier epoch. The following result is an improved version of a theorem given by Hawking and Ellis [Ellis 1965c]:

Theorem 1. *M cannot be timelike geodesically complete if:*

1. *The strong energy assumption holds.*
2. *There is a slice, i.e., a spacelike 3-surface without boundary, H in which there are at least three vector fields K which are independent at each point and for which $\mathcal{L}_k h_{ab} = 0$ and $\mathcal{L}_k X_{ab} = 0$, where h_{ab} and X_{ab} are the first and second fundamental tensors of H (this means that M is homogeneous on H).*
3. *The evolution of the Cauchy development of a partial Cauchy surface is determined by the Cauchy data on the surface.*

By Lemma 21 of Section 5, M will have a covering manifold $\widehat{M}$ in which each connected component of the image of H will be a partial Cauchy surface. We shall assume that $\widehat{M}$ is timelike geodesically complete and show that this is inconsistent with conditions (1), (2), and (3). Let $\widehat{H}$ be a connected component of the image of H. By condition (2), the Cauchy data on $\widehat{H}$ is homogeneous. Therefore, by condition (3), the Cauchy evolution of any region of $\widehat{H}$ is equivalent to the Cauchy evolution of any other similar region of $\widehat{H}$ (there being no time bombs). This implies that the surfaces $s = \text{const.}$ are homogeneous, if they lie within the Cauchy development of $\widehat{H}$, where s is the distance measured along the geodesics normal to $\widehat{H}$. These surfaces $s = \text{const.}$ must lie either entirely within or entirely without the Cauchy development of $\widehat{H}$, as otherwise there would be equivalent regions of $\widehat{H}$ which had inequivalent Cauchy evolutions. As the Cauchy horizon of $\widehat{H}$ is null, this imples that, while the surfaces $s = \text{const.}$ remain spacelike 3-surfaces, they will lie within the Cauchy development of $\widehat{H}$. The geodesics orthogonal to $\widehat{H}$ will also be orthogonal to the surfaces $s = \text{const.}$, as a vector representing the separation of points equal distances along neighbouring geodesics will remain orthogonal to the geodesics if it is so initially.

As in Section 4, we can represent the separation of neighbouring geodesics orthogonal to $\widehat{H}$ by a matrix $\underline{A}$ which is the unit matrix on $\widehat{H}$. By homogeneity, it will be constant on the surfaces $s = \text{const.}$, while these lie in the Cauchy development of $\widehat{H}$. While $\underline{A}$ is non-degenerate, the map from $\widehat{H}$ to a surface $s = \text{const.}$ defined by the geodesics will be of rank 3, and so the surfaces will be spacelike 3-surfaces contained within the Cauchy development of $\widehat{H}$. The expansion

$$\theta = (\det \underline{A})^{-1} \frac{\mathrm{d}}{\mathrm{d}s} \det \underline{A}$$

obeys the Raychaudhuri equation

$$\frac{\mathrm{d}}{\mathrm{d}s}\theta = -\frac{1}{3}\theta^2 - 2\sigma^2 - R_{ab}V^aV^b,$$

where the vorticity is zero. By the Einstein equations and the strong energy assumption, $R_{ab}V^aV^b$ is positive. Thus θ will become infinite and $\underline{A}$ degenerate for some finite positive or negative value s_0 of s. As the map from $\widehat{H}$ to the surface $s = s_0$ can have rank at most 2, this surface will be at most two-dimensional. As the geodesics lie within the Cauchy development of $\widehat{H}$ for $|s| < |s_0|$, the surface $s = s_0$ will lie in the Cauchy development or on the Cauchy horizon of $\widehat{H}$, and its position will be uniquely determined by the Cauchy data on $\widehat{H}$. By the strong energy assumption, the energy-momentum tensor has a unique eigenvector everywhere. These eigenvectors will form a C^1 timelike vector field whose integral curves may be thought of as representing the flow lines of the matter. As the surface $s = s_0$ lies in the Cauchy development or horizon of $\widehat{H}$, all the flow lines that pass through it will intersect $\widehat{H}$. But then, as $\widehat{H}$ is homogeneous, all the flow lines through $\widehat{H}$ must pass through $s = s_0$. This is impossible as $\widehat{H}$ is three-dimensional and $s = s_0$ is two-dimensional. In fact, if all the flow lines passed through a 2-surface, the density would be infinite.

This theorem is useful as it shows that large scale effects like the rotation of the whole universe cannot prevent the occurrence of singularities in solutions which satisfy the Copernican principle. However, it is still unrealistic in that, by requiring exact spatial homogeneity, it makes no allowance for the effect of local irregularities. In fact Lifshitz and Khalatnikov [Lifshitz 1963] have claimed that, in general, singularities will not occur in solutions without exact symmetries. That is, the slightest perturbation of a solution with a singularity would prevent the occurrence of the singularity.

The first theorem to deal with solutions without exact symmetries was given in outline by Penrose [Penrose 1965b]. It was designed to prove the occurrence of a singularity in a collapsing star. As a star of mass greater than about twice that of the sun exhausts its nuclear fuel and cools, there is apparently no mechanism that can support it against its self-gravity, and so it will collapse. If the collapse was exactly spherical, the solution could be integrated explicitly and a singularity would always occur. However, it is not obvious that this would the case if there were irregularities or a small amount of angular momentum. Indeed, in Newtonian theory the smallest amount of angular momentum could prevent the occurrence of infinite density and cause the star to re-expand.

However, Penrose showed that the situation was very different in the general theory of relativity: once the star had passed within the Schwarzschild surface, i.e., the surface $r = 2m$, it could not come out again. In fact the Schwarzschild surface has been defined only for an exactly spherically symmetric solution, but the more general criterion used by Penrose is equivalent for such a solution and applicable also to solutions without exact symmetry. It is that there should exist a 'closed trapped surface' C. By this is meant a closed, i.e., compact and without boundary, spacelike 2-surface (normally $\mathbb{S}^2$) such that the two families of null geodesics orthogonal to C are converging at C, i.e., $\mathrm{Tr}(\underset{1}{X})$ and $\mathrm{Tr}(\underset{2}{X})$ negative, where $\underset{1}{X}$ and $\underset{2}{X}$ are the two null second fundamental forms of C. We may think of C as being in such a strong gravitational field that even the 'outgoing' light rays from it are dragged back and are in fact converging. One would expect that a sufficiently small perturbation from exact symmetry would not prevent the occurrence of a closed trapped surface. This has been verified by Doroshkevish et al. [Doroshkevish 1966]. Penrose's result is then:

Theorem 2. *M cannot be null geodesically complete if:*

1. *The weak energy assumption holds.*

2. *There is a non-compact Cauchy surface H.*
3. *There is a closed trapped surface C.*

By Lemma 20 of Section 5, M will be causally simple. This implies that the boundary of $\ll C)$, the chronological future of C, will be generated by null geodesic segments which have past end points on C. These segments will be orthogonal to C. Suppose that M were null geodesically complete. Then by Lemma 4 of Section 4, there would be a point conjugate to C along each future-directed null geodesic orthogonal to C, as $R_{ab}K^aK^b \geq 0$ and $\mathrm{Tr}\underset{1}{(X)} < 0$ and $\mathrm{Tr}\underset{2}{(X)} < 0$. By Lemma 16, points on a null geodesic beyond the point conjugate to C would lie in $\ll C)$. Thus each generating segment of $\ll \dot{C})$ would have a future end point at or before the point conjugate to C. Near C, we could assign, in a differentiable manner, an affine parameter on each null geodesic orthogonal to C. Consider the differentiable map

$$\beta = C \times [0, b] \times Q \longrightarrow M,$$

where Q is the discrete set $\{1, 2\}$, defined by taking each point of C an affine parameter distance $v \in [0, b]$ along the two families of future-directed null geodesics orthogonal to C. For some value of b, $\beta\big(C \times [0, b] \times Q\big)$ would contain $\ll \dot{C})$. Thus $\ll \dot{C})$ would be compact, being a closed subset of a compact set. We saw in Section 5.2 that M admitted a future-directed timelike C^3 vector field. Each integral curve of this field would intersect H once and once only, and $\ll \dot{C})$ at most once. They would define a map of $\ll \dot{C})$ into H. As $\ll \dot{C})$ was compact, its image would be also. But this is impossible as H is non-compact and $\ll \dot{C})$ is a three-dimensional manifold (without boundary). Thus M cannot be null geodesically complete.

If we interchange past and future in the definition of closed trapped surfaces, they occur also in Robertson-Walker solutions [Hawking 1965b]. One way of seeing this is to consider the past-directed null geodesics through the point $t = t_0$, $r = 0$. They have the equation

$$\mathrm{d}t = -S(t)\frac{\mathrm{d}r}{(1 - Kr^2)^{1/2}}.$$

For each constant value of t, the null geodesics will describe an S^2 whose area is $4\pi r^2 S^2(t)$. The expansion $\tilde{\theta}$ of the null geodesics will be $(\text{area})^{-1}\mathrm{d}(\text{area})/\mathrm{d}v$, where v is an affine parameter along the null geodesics. Thus,

$$\tilde{\theta} = \frac{2}{rS}\frac{\mathrm{d}}{\mathrm{d}t}(rS)\frac{\mathrm{d}t}{\mathrm{d}r},$$

where $\mathrm{d}t/\mathrm{d}r$ will be negative. By the Einstein equations,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}(rS) &= -\left(1 - Kr^2\right)^{1/2} + r\dot{S} \\ &= -\left(1 - Kr^2\right)^{1/2} + r\left(\frac{1}{3}\mu S^2 - K\right)^{1/2}.\end{aligned}$$

If $\mu > 0$ and $K = 0$ or -1, the second term will dominate for t less than some value $t_1 > 0$. Thus the past-directed null geodesics which were diverging initially will be made to start converging by the gravitational effect of the matter. For each value of t less than t_1, the surface $\mathbb{S}^2$ described by the null geodesics will be a closed trapped surface (the other family of null geodesics orthogonal to the surface will also be converging). If we take one of these surfaces at $t = t_2$, with $t_2 < t_1$, the null geodesics orthogonal to it will have a finite convergence. Thus there will also be a

closed trapped surface in any solution without exact symmetry which is sufficiently similar to a Robertson-Walker solution in the region bounded by the past null cone of the point $t = t_0$, $r = 0$, and by the 3-surface $t = t_2$. In other words, a sufficiently small perturbation of a Robertson-Walker solution could not prevent the occurrence of a closed trapped surface.

The time t_1 at which the past-directed null geodesics start converging will depend on which Robertson-Walker solution is being considered. For $K = 0$ and $p = 0$ (Einstein-de Sitter solution), $t_1 = 8t_0/27$. In fact, we think we can see objects such as quasars which are further back than this and there does not seem any sign of major inhomogeneity or anisotropy at this time (though admittedly the evidence is not strong enough to rule them out completely). Thus it might be reasonable to suppose that there would be a closed trapped surface in any solution which described the universe. If we also assumed that it had a non-compact Cauchy surface, we could conclude that it must have a singularity.

However, as a Cauchy surface is spacelike, it is difficult to see how one could tell by observation whether it was compact or not. Also one cannot be sure that the universe is really sufficiently similar to a Robertson-Walker that a closed trapped surface would exist. Thus it would be good to have a theorem which did not depend on the non-compactness of the Cauchy surface or on the solution being almost spherically symmetric. This will be given for an 'expanding' solution.

At the present time, the universe is expanding in the sense that, on average, the galaxies are moving apart. However, there are of course localised regions in which the matter is contracting. We therefore want a definition which says that the universe is expanding on average, but which does not exclude the existence of local contractions. The one we shall use is: there exists a Cauchy surface H such that $\mathrm{Tr}(\underline{X})$ has a positive lower bound on H, where $\underline{X}$ is the second fundamental tensor of H. In other words, the geodesics orthogonal to H are diverging at H. Of course, this would not be true for just any Cauchy surface, but if the solution was expanding in any sense, it would seem reasonable to suppose that we could deform a given Cauchy surface into one for which it was true.

Theorem 3. *M cannot be timelike geodesically complete if:*

1. *The energy assumption holds.*
2. *There is a Cauchy surface H.*
3. *$\mathrm{Tr}(\underline{X}) \geq b > 0$, where b is a constant.*

This follows directly from the lemmas established in Sections 4 and 5. By Lemma 26 of Section 5, there is a geodesic orthogonal to H of length $d(p, H)$ from any point $p \in (H \gg$ to H. By Lemma 12 of Section 4, there cannot be a point conjugate to H along this geodesic between p and H. Suppose that M were timelike geodesically complete. Then by conditions (1) and (3) and by Lemma 2 of Section 4, there would be a point conjugate to H on every past-directed geodesic orthogonal to H, within a distance $3/b$ from H. However, as we could continue any past-directed geodesic orthogonal to H to arbitrary length, there would be points p for which $d(p, H) > 3/b$. This shows that M cannot be timelike geodesically complete

Although it would seem reasonable to argue from the Copernican principle that, as the universe is expanding in our neighbourhood, it should be expanding everywhere, this is not an assumption that can be tested by observation. For all we can tell, there might be regions where the universe is contracting. One would therefore like to know whether one could avoid assuming condition (3) above. This is possible if the Cauchy surface is compact:

Theorem 4. *M cannot be timelike geodesically complete if:*

1. *The energy assumption holds.*

2. *There is a compact Cauchy surface H.*
3. *The strong energy assumption holds on H.*

Condition (3) might seem unrealistic, as it would rule out there being any empty regions on H. However, it is adopted simply for convenience. All that is really necessary is that every timelike geodesic should encounter some matter or some randomly orientated Weyl tensor field in the vicinity of H.

By condition (3), the energy-momentum tensor on H can be expressed as

$$T_{ab} = \mu V_a V_b + \sum_i p_i Y_{i\,a} Y_{i\,b},$$

where $\mu + p_i > 0$, $\mu + \sum_i p_i > 0$. As the Ricci tensor is C^1, the energy-momentum tensor will be also. Thus there will be a positive constant f such that, on H,

$$T_{ab} = f V_a V_b + S_{ab},$$

where the energy assumption holds for S_{ab}, i.e.,

$$S_{ab} W^a W^b \geq \frac{1}{2} W_a W^a S,$$

for any timelike vector W. For each $p \in H$, we can choose normal coordinates u^1, u^2, u^3, u^4 in a neighbourhood U of p such that $V = \partial/\partial u^4$ at p. Then we can find a positive constant b such that, for each p, the ball B of coordinate radius b about p is contained in U and such that in B the energy-momentum tensor can be expressed as

$$T_{ab} = \frac{1}{2} f \overline{V}_a \overline{V}_b + \overline{S}_{ab},$$

where $\overline{V}$ is a timelike unit vector field which coincides with V on H and where the energy assumption holds for $\overline{S}_{ab}$.

Suppose that M were timelike geodesically complete. Let $\gamma(s)$ be a timelike geodesic through $p \in H$ with $p = \gamma(0)$, and let E_1, E_2, E_3, E_4 be an orthonormal basis parallel transported along γ with $E_4 = (\partial/\partial s)_\gamma$. Let $Z_i = E_i \cos(\pi s/\ell)$, $i = 1, 2, 3$, where ℓ is a positive constant. Then the second variation of the length of γ between the points $q_1 = -\ell/2$ and $q_2 = \ell/2$, induced by the variation vector Z_i, is

$$L(Z_i, Z_i) = -\int_{-\ell/2}^{\ell/2} g\left(Z_i, \left\{\frac{\mathrm{D}^2 Z_i}{\partial s^2} + R(E_4, Z_i)E_4\right\}\right) \mathrm{d}s.$$

But

$$\sum_i g\big(E_i, R(E_4, E_i)E_4\big) = -R(E_4, E_4).$$

Therefore,

$$\sum_i L(Z_i, Z_i) = -\frac{3}{2\ell} + \int_{-\ell/2}^{\ell/2} R(E_4, E_4) \cos^2 \frac{\pi s}{\ell} \mathrm{d}s.$$

By the Einstein equations,

$$R(E_4, E_4) = T(E_4, E_4) - \frac{1}{2} T.$$

This will be non-negative, by the energy assumption. In $\gamma \cap B$,

$$R(E_4, E_4) \geq \frac{1}{4} f \big[g(E_4, V)\big]^2.$$

Thus if $\ell = \ell^* > 3b$,

$$\int_{-\ell/2}^{\ell/2} R(E_4, E_4) \cos^2 \frac{\pi s}{\ell} \mathrm{d}s \geq \frac{1}{16} f \int_{\gamma \cap B} \left[g\left((\partial/\partial s)_\gamma, V \right) \right]^2 \mathrm{d}s.$$

But

$$\int_{\gamma \cap B} \left[g\big((\partial/\partial s)_\gamma, V\big) \right]^2 \mathrm{d}s \geq \left[\int_{\gamma \cap B} g\big((\partial/\partial s)_\gamma, V\big) \mathrm{d}s \right]^2 \geq 2b^2.$$

Therefore, if ℓ^* was also greater than $24/fb$, the quantity $\sum_i L(Z_i, Z_i)$ would be positive and γ would not be the longest timelike curve from q_1 to q_2.

Let H_1, resp. H_2, be the set (not necessarily a spacelike surface) formed by moving each point of H a distance ℓ^* to the past, resp. future, along the geodesics orthogonal to H. Since H is compact, H_1 and H_2 would be compact. By the corollary to Lemma 26 of Section 5, there would be a timelike geodesic γ of length $d(H_1, H_2)$ from H_1 to H_2. Let p be the point where γ intersects H. By construction, $d(H_1, p)$ and $d(p, H_2)$ would be greater than or equal to ℓ^*. Thus γ could not be the longest timelike curve from H_1 to H_2. Therefore, M cannot be timelike geodesically complete.

So far we have been assuming that the solutions would have Cauchy surfaces. This again is not something that can be tested by observation. Thus maybe all that our theorems prove is that realistic solutions do not have Cauchy surfaces: as they evolve from a given surface, a Cauchy horizon always occurs and prevents the appearance of a singularity. However, the following theorem shows that, at least in some cases, this cannot happen.

Theorem 5. *M cannot be timelike geodesically complete if:*

1. *The energy assumption holds.*
2. *There is a compact slice (spacelike 3-surface without boundary) H.*
3. *$\mathrm{Tr}(\underline{X}) > 0$ (the geodesics orthogonal to H are diverging at H).*

These conditions are rather similar to those of Theorem 3, the difference being that H is not required to be a Cauchy surface but is required to be compact. This is necessary, as otherwise Minkowski space would satisfy the conditions.

By Lemma 21 of Section 5, M will have a covering manifold $\widehat{M}$ such that, in $\widehat{M}$, each connected component of the image of H will be diffeomorphic to H and will be a partial Cauchy surface in $\widehat{M}$. Suppose that $\widehat{M}$ was timelike geodesically complete and let $\widehat{H}$ be a connected component of the image of H. As $\mathrm{Tr}(\underline{X})$ has some positive lower bound b on $\widehat{H}$, there would be a point conjugate to $\widehat{H}$ on every past-directed geodesic orthogonal to $\widehat{H}$ within a distance $3/b$ from $\widehat{H}$. If $p \in J^-$ (the past Cauchy development of $\widehat{H}$), there would be a timelike geodesic orthogonal to $\widehat{H}$ of length $d(p, \widehat{H})$ from p to $\widehat{H}$. As this geodesic could not contain a point conjugate to $\widehat{H}$, the quantity $d(p, \widehat{H})$ could not be greater than $3/b$. This would imply that $\widehat{H}$ had a past Cauchy horizon L^-. Then for $q \in L^-$, $d(p, \widehat{H})$ would have an upper bound $3/b$, as otherwise there would be a timelike curve from L^- to $\widehat{H}$ of length greater than $3/b$, and this curve would contain points in J^- at distances greater than $3/b$ from $\widehat{H}$. As q is not in J^-, we cannot assume that $d(p, q)$ would be continuous. However, as $\widehat{H}$ is compact and $d(p, \widehat{H})$ finite, we can still use a method similar to that in Lemma 26 of Section 5 to show that there would be a geodesic orthogonal to $\widehat{H}$ of length $d(p, \widehat{H})$ from q to $\widehat{H}$.

Consider the function $d(p, q)$ for $p \in \widehat{H}$. There would be a point $r \in \widehat{H}$ such that any open neighbourhood of r in $\widehat{H}$ would contain a value of $d(q, p)$ arbitrarily close

to $d(q, \widehat{H})$. Let U be a normal coordinate neighbourhood of r and $B \subset U$ a ball of constant coordinate radius about r. Let y_n be a sequence of points in $\widehat{H}$ converging to r such that

$$d(q, y_n) > d(q, \widehat{H}) \left(1 - \frac{1}{n+2}\right),$$

and let λ_n be timelike curves of length greater than

$$d(q, \widehat{H}) \left(1 - \frac{1}{n+1}\right)$$

from q to y_n. Let ζ be a limit point of $\lambda_n \cap \dot{B}$. Then $\zeta \in \dot{B} \cap (r >$. Suppose that ζ did not lie on γ, the past-directed geodesic through r orthogonal to $\widehat{H}$. Then as $d(p, \widehat{H})$ is continuous for $p \in J^-$, there would be an open neighbourhood V of ζ, a $\delta > 0$, and an $N \in \mathbb{N}$ such that

$$d(p, \widehat{H}) > d(p, y_n) + \delta, \quad \text{for } n > N \text{ and } p \in V.$$

But there would be a curve λ_n with $n > N$ and $n > 1/\delta$ which intersected V. This could be deformed to give a timelike curve of length greater than $d(q, \widehat{H})$ from q to $\widehat{H}$. Thus ζ would like on γ. The relation

$$d(q, p) + d(p, r) = d(q, \widehat{H})$$

would hold for all points p on γ between r and ζ. But then it would hold for all points p on γ for which $d(p, r) \leq d(q, \widehat{H})$, for if X were the least bound of points for which it held, we could perform a construction similar to that in Lemma 26 of Section 5 and prove that it held for points beyond X. All the points for which $d(p, r) < d(q, \widehat{H})$ would lie in J^-. As d is continuous in J^- and as $\widehat{M}$ is supposed to be timelike geodesically complete, there would be points p on γ for which $d(p, r)$ was arbitrarily close to $d(q, \widehat{H})$. Thus, as d is lower semi-continuous, there would be a point p on γ for which $d(p, r) = d(q, \widehat{H})$. This point would either coincide with q or lie on the boundary of $< q)$ not at q. An application of a similar construction shows the latter to be impossible. Thus γ is a past-directed geodesic orthogonal to $\widehat{H}$ of length $d(q, \widehat{H})$ from q to $\widehat{H}$.

Let q^1 be another point on the same null geodesic generating segment λ of L^- with $q \nrightarrow q^1$. There would be a geodesic γ^1 of length $d(q^1, \widehat{H})$ from q^1 to $\widehat{H}$. This could be extended along λ to give a broken non-spacelike geodesic of length $d(q^1, \widehat{H})$ from q to $\widehat{H}$. One could find a variation of this that gave a timelike curve of length greater than $d(q^1, \widehat{H})$ from q to $\widehat{H}$. Thus $d(q, \widehat{H})$ would strictly decrease along each future-directed null geodesic generating segment of L^-.

Consider the differentiable map

$$\beta : \widehat{H} \times \left[0, \frac{3}{b}\right] \longrightarrow \widehat{M},$$

defined by taking each point of $\widehat{H}$ a distance $s \in [0, 3/b]$ along the past-directed geodesics orthogonal to $\widehat{H}$. The image

$$\beta \left(\widehat{H} \times \left[0, \frac{3}{b}\right]\right)$$

would be compact and would contain L^-. Thus, as L^- is closed, it would be compact. But then, as d is lower semi-continuous, $d(q, \widehat{H})$ ought to have a minimum on L^-. This would be impossible as $d(q, \widehat{H})$ would strictly decrease along each future-directed null geodesic generating segment of L^-, and such a segment could have no future end point. Therefore $\widehat{M}$, and hence M, cannot be timelike geodesically complete.

This theorem shows that neither the existence of a Cauchy surface nor the holding of any causality assumption are necessary to prove the occurrence of a singularity. However, it still involves a condition which cannot be tested observationally. It does not seem possible to do away with untestable conditions entirely, but in the following theorem, the only such condition is that the strong causality assumption should hold.

Theorem 6. *M cannot be timelike and null geodesically complete if:*

1. *The energy assumption holds.*
2. *The strong causality assumption holds.*
3. *There is a point p, a past-directed unit timelike vector $X \in T_p$ and positive constants b, c such that on each past-directed timelike geodesic $\gamma(s)$ through p, the expansion θ of these geodesics becomes less than $-cg\big(X, (\partial/\partial s)_\gamma\big)$ within a distance of $b\big[g\big(X, (\partial/\partial s)_\gamma\big)\big]^{-1}$ from p.*

Condition (3) implies that all the past-directed null geodesics through p will start converging again. This is a fairly severe requirement but one that could in principle be tested by observation. It would be satisfied in a Robertson-Walker solution. Thus it should also be satisfied by a solution which was sufficiently similar to a Robertson-Walker solution in the region bounded by the past null core of p and by a surface $t = t_2$.

Interchanging past and future, this theorem could also be used to prove the occurrence of a singularity in an approximately spherical collapsing star. This is important because Theorem 2 can be applied only when there is a non-compact Cauchy surface. However, the examples of the Reissner-Nordström and Kerr solutions show that the addition of a small electric charge or of a small amount of angular momentum can cause a Cauchy horizon to appear.

Suppose that M were timelike and null geodesically complete. Then conditions (1) and (3) would imply that there would be a point conjugate to p within a distance

$$\left(b + \frac{3}{c}\right)\left[g\big(X, (\partial/\partial s)_\gamma\big)\right]^{-1}$$

along each past-directed timelike geodesic $\gamma(s)$ through p. Take natural coordinates u^1, u^2, u^3, u^4 in T_p with X along the u^4 axis. Then within a coordinate distance of $2(b + 3/c)$ along each past-directed timelike ray in T_p from the origin, there would be a point whose image in M under the exponential map would be a point conjugate to p.

If D is a local causality neighbourhood of p, then $D \cap (p >$ will be in $C^-(p)$, the past compact region of p. By Lemma 26 of Section 5, there will be a timelike geodesic curve of length $d(p, q)$ from each point $q \in Q^-(p) \cap (p \gg$ to p. As this can contain no point conjugate to p, $Q^-(p) \cap (p \gg$ would be contained in $\exp(N)$, where N is the compact region in T_p consisting of all points on past-directed non-spacelike rays from the origin at coordinate distances not greater than $2(b + 3/c)$ from the origin. Since M was assumed timelike and null geodesically complete, $\exp(N)$ would be defined. Thus $\overline{Q^-(p) \cap (p \gg}$ would be compact. But $C^-(p)$ equals $Q^-(p) \cap (p >$, which is contained in $Q^-(p) \cap \overline{(p \gg}$. Therefore, as $Q^-(p)$ is open, $\overline{C^-}(p)$ and $\dot{C}^-(p)$ are contained in $\overline{Q^-(p) \cap (p \gg}$, and so are compact. We could therefore cover $\dot{C}^-(p)$

by a finite number of open balls B_i, for each of which $\overline{B_i}$ was contained in a local causality neighbourhood. As any past-directed timelike geodesic from p would have to leave the compact set $\overline{C^-}(p)$, we could find a point $q_1 \in (p \gg$ which lay in one of these balls B_1, but which was not in $C^-(p)$.

Every non-spacelike curve from q_1 to p would have to intersect $C^-(p)$. As $\dot{Q}^-(p)$ is a null horizon, no future-directed non-spacelike curve which intersects $\overline{Q^-}(p)$ can leave it again. Therefore, every non-spacelike curve from q_1 to p would enter $\overline{C^-}(p)$ and not leave it again. Suppose that $\langle\langle q_1, p\rangle\rangle \cap \dot{C}^-(p)$ was contained in B_1. Then $\overline{\langle\langle q_1, p\rangle\rangle}$ would be contained in $\overline{B_1} \cup \overline{C^-}(p)$ and so would be compact. This would be impossible, as it would imply that each point of the non-empty set $\langle\langle q_1, p\rangle\rangle$ would be in $C^-(p)$, which could not be the case as q_1 was not in $C^-(p)$. This shows that there would have to be a timelike curve λ_1 from q_1 to p which intersected $\dot{C}^-(p)$ in some other ball B_2. Then we could find a point $q_2 \in \lambda_1 \cap B_2$ which was not in $\overline{C^-}(p)$. Repeating the above, there would be an infinite sequence of points q_n, timelike curves λ_n from q_n to p, and balls $B_n \supset \lambda_{n-1} \cap \dot{C}^-(p)$. As a strong causality assumption holds, no curve λ_n could return to an earlier ball. Thus the B_n would be distinct. But this would be impossible, as there were only a finite number of balls B_i covering $\dot{C}^-(p)$. Thus M would not be timelike and null geodesically complete.

A further theorem on singularities has been given by Geroch [Geroch 1966]. As in Theorem 4, the aim is to dispense with the assumption that the normals to the Cauchy surfaces are diverging. It is replaced by the condition that there should be no particle or event horizons.

6.4 The description of singularities

The preceding theorems prove the occurrence of singularities in a large class of solutions, but give little information as to their nature. To investigate this in more detail, one would need to define what one meant by the size, shape, location, and so on, of a singularity. This would be fairly easy if the singular points were included in the spacetime manifold. However, as was said in Section 6.1, it would be impossible physically to determine the manifold structure at such points. In fact, there would probably be several manifold structures which agreed for the non-singular regions but which differed for the singular points. For example, the manifold at the $t = 0$ singularity in the Robertson-Walker solutions could be that described by the coordinates t, $r\cos\theta$, $r\sin\theta\cos\phi$, $r\sin\theta\sin\phi$, or that described by t, $Sr\cos\theta$, $Sr\sin\theta\cos\phi$, $Sr\sin\theta\sin\phi$. In the first case, the singularity would be a 3-surface, and in the second, a single point.

What is needed is a description based on measurements at non-singular points only. The one we shall give is probably not unique, but is based on the definition of singularities that we have adopted, namely, geodesic incompleteness.

The tangent bundle $T(M)$ may be thought of as consisting of the tangent spaces T_p for every $p \in M$. If u^1, u^2, u^3, u^4 are local coordinates in an open neighbourhood U of p, any vector $V \in T_p$ can be represented as $V^a \partial/\partial u^a$. Then

$$\{z^A\} = \{u^1, u^2, u^3, u^4, V^1, V^2, V^3, V^4\}$$

will be local coordinates in the open neighbourhood $\pi^{-1}(U)$, where $\pi : T(M) \to M$ is the natural projection which maps each point of T_p to p. The tangent space $T_q(T(M))$ to the tangent bundle at a point $q \in T(M)$ will have the coordinate basis $\{\partial/\partial z^A\}$. The subspace of $T_q(T(M))$ spanned by the vectors $\{\partial/\partial V^a\}$ is called the vertical subspace, denoted P_q, since it is tangent to the fibre $T_{\pi(q)}(M)$. It does not depend on

the choice of local coordinates in U. On the other hand, the subspace of $T_q(T(M))$ spanned by the vectors $\{\partial/\partial u^a\}$ does depend on the local coordinates. However, the subspace spanned by

$$\frac{\mathrm{D}}{\partial u^a} = \frac{\partial}{\partial u^a} - \Gamma^c_{ab} V^b \frac{\partial}{\partial V^c}$$

is coordinate independent. It is called the horizontal subspace, denoted by Q_q. Then $T_q = P_q + Q_q$.

If $\gamma(t)$ is a curve in M through $\pi(q)$, there will be a unique curve $\overline{\gamma}(t)$ called the lift of $\gamma(t)$ in $T(M)$ through q such that

$$\pi\big(\overline{\gamma}(t)\big) = \gamma(t),$$

and such that the tangent vector $(\partial/\partial t)_{\overline{\gamma}}$ is horizontal everywhere. Interpreted in M, $\overline{\gamma}(t) = \big\{u^a(t), V^a(t)\big\}$ represents both a curve $\gamma(t) = u^a(t)$ and a vector field $V(t) = V^a(t)\partial/\partial u^a$ along $\gamma(t)$. As

$$\left(\frac{\partial}{\partial t}\right)_{\overline{\gamma}} = \frac{\mathrm{d}u^a}{\mathrm{d}t}\frac{\partial}{\partial u^a} + \frac{\mathrm{d}V^b}{\mathrm{d}t}\frac{\partial}{\partial V^b}$$

is horizontal, we deduce that

$$\frac{\mathrm{d}V^b}{\mathrm{d}t} = -\Gamma^b_{ac} V^a \frac{\mathrm{d}u^c}{\mathrm{d}t}.$$

Thus $V(t)$ is parallel transported along $\gamma(t)$ in M. A point q in $T(M)$ with coordinates u^a, V^a may be thought of as a point in the tangent space $T_{\pi(q)}(M)$. Thus there is a natural correspondence which assigns to q the vector drawn from the origin in $T_{\pi(q)}$ whose components are V^a. This correspondence defines an intrinsic, global, vertical vector field $W = V^a\partial/\partial V^a$.

We can also define on $T(M)$ an intrinsic, global, horizontal vector field

$$X = V^a \frac{\mathrm{D}}{\mathrm{D}u^a}.$$

This may be interpreted as follows. Let $\gamma(t)$ be the unique geodesic in M through $\pi(q)$ such that

$$\left(\frac{\partial}{\partial t}\right)_{\gamma} = V^a \frac{\partial}{\partial u^a} \quad \text{at } \pi(q),$$

and let $\overline{\gamma}(t)$ be the lift of $\gamma(t)$ through q in $T(M)$. Then $(\partial/\partial t)_{\overline{\gamma}} = X$.

The exponential map may be thought of as a map $T(M) \to M$ which takes a point q in $T_{\pi(q)}(M)$ to the point in M a unit parameter distance along the geodesic $\gamma(t)$ through $\pi(q)$ for which

$$\left(\frac{\partial}{\partial t}\right)_{\gamma} = V^a \frac{\partial}{\partial u^a}.$$

Equivalently, one could go a unit parameter distance along $\overline{\gamma}(t)$ in $T(M)$ and then project down into M. A third way would be to proceed down the integral curves of the 'diagonal' vector field $X - W$ until one reached the cross-section of zero vectors which can be identified with M. These three equivalent procedures are illustrated in Figure 13 for the simple case of the tangent bundle to a one-dimensional manifold.

All points on a diagonal curve in $T(M)$ will have the same image in M under the exponential map. If M is four-dimensional, $T(M)$ will be eight-dimensional and there will be a three-dimensional family of diagonal curves in $T(M)$ corresponding to each

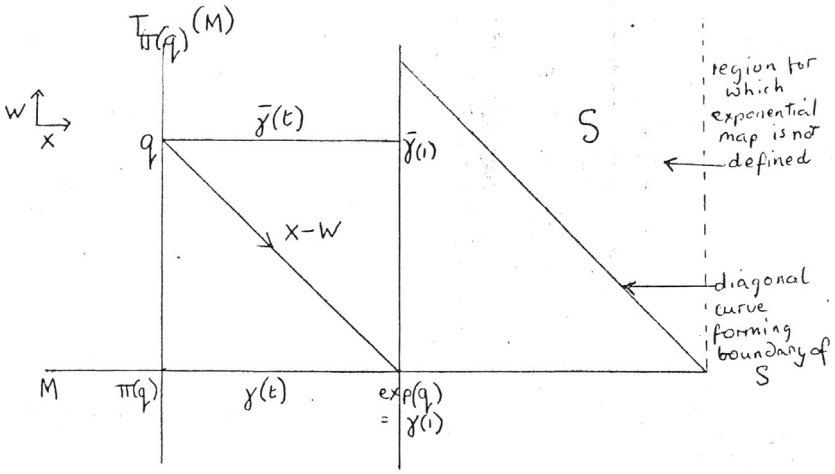

Fig. 13. The tangent bundle of a one-dimensional manifold M. The latter is identified with the cross-section of zero vectors in $T(M)$ (*horizontal line*). The *vertical line* through $\pi(q)$ represents $T_{\pi(q)}(M)$. Under the exponential map, the point q is taken down the diagonal integral curve of $X - W$ to where it intersects M. Also shown is the region S for which the exponential map is not defined.

point of M. If M is not geodesically complete, the exponential map will not be defined for some points of $T(M)$. The diagonal curves through these points will be defined, but they will not reach the cross-section of zero vectors in $T(M)$. We may think of the boundary of the region S of $T(M)$ for which the exponential map is not defined as representing the incompleteness and hence the singularities of M. This boundary will in general be seven-dimensional and will be generated by diagonal curves which just fail to reach the cross-section of zero vectors. In order to be able to talk about the dimension, etc., of the singularity, one would want to be able to say that some of these boundary diagonal curves correspond in some sense to the same point of some boundary of M. One possible way of establishing such a correspondence is given below.

The exponential map $T_p(M) \to M$ is of rank 4, except at points conjugate to p. However, such points are isolated [Milnor 1963, p. 98]. Thus the exponential map $T(M) \to M$ is open, i.e., it maps open sets into open sets. We shall say that diagonal curves λ_1 and λ_2 are in particular correspondence if $\exp(V_1)$ intersects $\exp(V_2)$ (where defined) for every pair of open neighbourhoods V_1 of λ_1 and V_2 of λ_2. Then we shall define a general correspondence class of the diagonal curves in some region as a set which cannot be divided into two disjoint subsets such that no curve of one subset is in particular correspondence with any curve of the other subset, and which is itself not a subset of any larger such set. Invoking the axiom of choice, each diagonal curve in a given region will belong to one and only one general correspondence class for that region. Let M_1 be the space consisting of the general correspondence classes of diagonal curves in $T(M) - S$. Each point of M_1 will represent all the diagonal curves in $T(M) - S$ whose image under the exponential map is some point in M. Thus there is a natural one-to-one correspondence between M_1 and M. This suggests that M_2, the space consisting of the general corresponence classes of diagonal curves in the boundary of S, could be thought of as representing the 'boundary' of M. We can define a topology on $M^+ = M_1 \cup M_2$ by requiring that a set of classes of curves which intersect an open set in $T(M)$, should be open in M^+. Restricted to M_1, this topology will agree with that of M under the natural correspondence. Restricted to

M_2, it will be Hausdorff. However, it would not be Hausdorff on M^+ if there were partially or totally imprisoned geodesics in M which were incomplete. This is because such a geodesic would seem to 'vanish' in a compact set and one could not add any boundary point for such a geodesic which would be Hausdorff with respect to this compact set.

M_1 can be given a natural manifold structure induced from M. If M^+ were Hausdorff, it might be possible to extend this to M_2 so that M^+ could be regarded as a paracompact manifold with boundary. This could be done for the Schwarzschild, Reissner-Nordström, and Robertson-Walker solutions, M_2 being three-dimensional in the first and third cases and one-dimensional in the second. However, more realistic solutions without exact symmetries might not have such well-behaved singularities as these. Nevertheless, the above procedure would enable one to describe the singularities and define a topological structure for them, even if not a manifold one.

6.5 Conclusion

The following seem to the author to be the important questions on singularities:

1. Not imposing any condition involving all points of spacetime other than that the energy assumption should hold, are there solutions which evolve from a non-singular state to an inevitable singularity and which are fully general in the sense that a sufficiently small perturbation of the initial state would not prevent the occurrence of the singularity?

This is answered affirmatively by Theorem 5, since a sufficiently small perturbation of the metric in the neighbourhood of a compact slice with diverging normals would leave the normals diverging.

2. Adopting the energy assumption, would there be a singularity in any solution which could represent a universe?

This question cannot be settled finally until we have a rather fuller knowledge of the nature of the universe. In fact, present observations do not completely rule out the possibility that the universe could be asymptotically flat. However, the accumulated weight of Theorems 1–5 would seem to indicate that a singularity would occur in any solution which was in accordance with some form of the Copernican principle. If one also adopted the strong causality assumption as being physically essential, a fairly small extension of the present observations would enable one to test whether the conditions of Theorem 6 would be satisfied in a solution representing the universe.

So far the constant λ in the Einstein equations has been assumed to be zero. One might hope that, if it had some suitable value, singularities would not occur. However, Theorem 2 does not depend on the value of λ, for what is required is that $R_{ab}K^aK^b \geq 0$ for any null vector K, and this is satisfied if $T_{ab}K^aK^b \geq 0$, no matter what λ is. In the other theorems, the energy assumption would have to be replaced by the condition that

$$T_{ab}W^aW^b \geq W_aW^a\left(\frac{1}{2}T + \lambda\right),$$

for any timelike vector W. In terms of the decomposition given before, this would hold if

$$\mu + p_i \geq 0, \qquad \mu + \sum_i p_i - 2\lambda \geq 0.$$

In a Robertson-Walker solution, the second expression is just $-3\ddot{S}/S$, where S is the scale factor of the universe. Observations of the present rate of change of the

expansion of the universe indicate that, at the present time, $3\ddot{S}/S$ is probably negative and, if positive, is certainly not more than fifty times the average density of observed matter [Sandage 1961]. Thus the energy condition in the theorems would have been satisfied if there had been a time in the past when the density was more than fifty times the present value. There seems to be quite a lot of observational evidence that this must have been the case. The conclusion therefore would seem to be that, even if λ were nonzero, it could not prevent the occurrence of a singularity in solutions representing a collapsing star, and it would probably not be large enough to cause solutions representing the universe to 'bounce' without a singularity.

3. The theorems given above are pieces of mathematics. They have physical significance only if the following are answered in the affirmative:
   - Does the general theory of relativity provide a correct description of all possible observations?
   - Does the energy assumption hold?

Taking the second question first, the conservation equations for a perfect fluid give

$$\dot{\mu} + (\mu + p)\theta = 0.$$

Suppose that the density μ was positive initially. Then, as the flow lines converged, μ could not become negative unless the pressure p became negative first. But if the pressure decreased as the density increased, there would be mechanical instability. For if a small region was compressed slightly, the pressure inside it would decrease and so it would be compressed further by the surrounding fluid. Of course, the situation is more complicated for an imperfect fluid with anisotropic pressures. Nevertheless, if energy conservation holds in any form, it is difficult to see how matter which satisfied the energy assumption at one time could fail to satisfy it at any other time.

There would still be the possibility that the energy assumption is not satisfied at any time. This would be the case in the C field theory of Hoyle. Hoyle and Narlikar [Hoyle 1964b] have shown that the C field could prevent the occurrence of singularities in certain circumstances. However, it is not clear to the author that the C field could prevent singularities in every situation, and the theory seems to suffer from the difficulties with negative energy mentioned in Section 6.2.

For the above reasons, the author feels that it is likely that the energy assumption holds. Returning to the first question, the general theory of relativity has so far been experimentally tested only in very weak fields. However, a good physical theory should not only correctly describe the currently experimental knowledge, but should also predict new results which can be tested by experiment. The further the predictions from the original experiments, the greater the credit to the theory if they are found to be correct. Thus observations of whether or not singularities actually occurred would provide a powerful test of the general theory of relativity in strong fields. Of course, general relativity is a classical theory. Thus one could not expect it to be correct if the curvature became so large that quantum effects had to be considered. Opinions seem to differ as to whether this would happen when the curvature was 10^{14} cm or when it was 10^{33} cm. In either case, it would be enormous compared to the value of about 10^{-13} cm at the earth's surface. For practical purposes, a region of such high curvature could be regarded as a singularity.

In the case of a spherically symmetric, uncharged, collapsing star, all the matter hits the singularity. However, the examples of the Reissner-Nordström and Kerr solutions show that the addition of a small amount of electric charge or angular momentum could completely alter the nature of the singularity, causing the matter to fall through a 'wormhole' and emerge into another universe. Similarly, it would seem reasonable to suppose that the singularity in the universe would not be the all-embracing kind in the Robertson-Walker solutions, but might consist of isolated singularities which

only a few worldlines hit. The author feels that the nature of the singularities to be expected in realistic solutions is the problem on which most future research in this field ought to be concentrated.

Appendix: The nature of a spatially homogeneous anisotropic solution near the singularity

By Theorem 1, a solution which is exactly spatially homogeneous and in which the strong energy assumption holds, will have a point of infinite density on each flow line of the matter. Although such singularities are probably not very realistic physically, it may nevertheless be of interest to study some of their properties. This could be done by obtaining exact solutions. However, although a number of particular spatially homogeneous solutions have been given [Heckmann 1962, Kantowski 1966], it has not been possible to obtain analytic expressions for the general class of such solutions. Instead, we shall give a number of properties which can be derived directly from the deviation equation derived in Section 4:

$$\frac{\mathrm{d}^2\underline{A}}{\mathrm{d}s^2} = -\underline{G} + \underline{F},$$

where $\underline{A}$ is the matrix representing the separation of neighbouring flow lines, $\underline{G}$ depends on the Riemann tensor, and $\underline{F}$ on the acceleration of the flow lines.

Suppose the matter was a perfect fluid with zero pressure. Then $\underline{F} = 0$ and we have the vorticity conservation law

$$\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A} = \text{const.}$$

If the vorticity $\underline{\omega}$ was nonzero, it would seem reasonable to suppose that the infinite density arose through the flow lines converging along the axis of vorticity. That is to say that, if the basis is chosen so that at $s = 0$,

$$\underline{\omega} = \begin{vmatrix} 0 & 0 & 0 \\ 0 & 0 & \omega \\ -\omega & 0 & 0 \end{vmatrix},$$

then as $s \to 0$, $\underline{A}$ would tend to

$$\begin{vmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix}.$$

The vorticity conservation law then shows that ω remains finite as $s \to 0$. We may write $\underline{A}$ as $\underline{Q} \cdot \underline{S}$. Let a, b, c be the eigenvalues of $\underline{S}$, where $a \to 0$, $b, c \to 1$ as $s \to 0$. Then,

$$\theta = \frac{1}{abc}\frac{\mathrm{d}}{\mathrm{d}s}(abc),$$

$$2\sigma^2 = \frac{1}{a^2}\left(\frac{\mathrm{d}a}{\mathrm{d}s}\right)^2 + \frac{1}{b^2}\left(\frac{\mathrm{d}b}{\mathrm{d}s}\right)^2 + \frac{1}{c^2}\left(\frac{\mathrm{d}c}{\mathrm{d}s}\right)^2 - \frac{1}{3}\theta^2.$$

By the Raychaudhuri and conservation equations,

$$\frac{\mathrm{d}\theta}{\mathrm{d}s} = -2\sigma^2 - \frac{1}{3}\theta^2 - \frac{1}{2}\mu + 2\omega^2, \qquad \mu = \frac{\mu_0}{abc}, \qquad \mu_0 = \text{const.}.$$

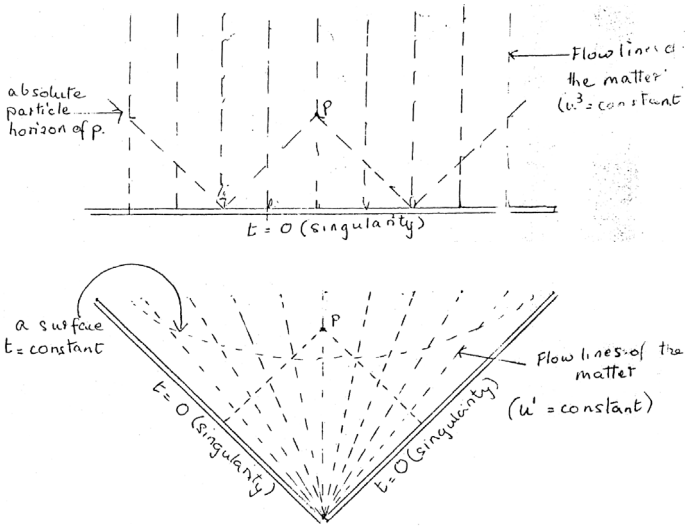

Fig. A.1. *Upper*: A section u^1, u^2 constant showing the particle horizon in the u^3 direction. *Lower*: A section u^2, u^3 constant showing the absence of a particle horizon in the u^1 direction

As $s \to 0$, the dominant term on the right of the Raychaudhuri equation will be

$$-\frac{1}{a^2}\left(\frac{\mathrm{d}a}{\mathrm{d}s}\right)^2.$$

Thus asymptotically, a is proportional to s and μ to s^{-1}. This should be contrasted with the behaviour of the Robertson-Walker solutions given in Section 3. In these, μ was asymptotically proportional to s^{-2}. This difference in time scale could have an important effect on processes in the early stages of the universe, such as the production of helium [Hawking 1966d].

The above calculation did not depend on spatial homogeneity. Thus it should hold whenever the flow lines of pressure-free matter converge in one direction. This case would seem to be more general than that in which they converged in two directions simultaneously.

In a spatially homogeneous solution, this difference in time scale can have another important effect. Consider, for example, the solution with metric

$$\mathrm{d}s^2 = \mathrm{d}t^2 - a^2(t)\left(\mathrm{d}u^1\right)^2 - b^2(t)\left(\mathrm{d}u^2\right)^2 - c^2(t)\left(\mathrm{d}u^3\right)^2,$$

where the curves u^1, u^2, u^3 constant are the flow lines of the pressure-free matter and $a \to t$, $b, c \to 1$ as $t \to 0$. A past-directed null geodesic from a point p could attain infinite values of the coordinate u^1, but only finite values of the coordinates u^2 and u^3. Thus an observer at p could see all the flow lines in the u^1 direction, but only a finite number in the u^2 and u^3 directions. In other words, p has a particle horizon in the u^2 and u^3 directions, but not in the u^1 direction. This is illustrated in Figure A.1, which shows Penrose diagrams of the two-dimensional sections u^1, u^2 constant and u^2, u^3 constant. This might be important for astrophysics as it is the existence of a particle horizon in every direction which severely limits the instabilities which can occur in a Robertson-Walker solution [Hawking 1966b].

To assume that the pressure is zero is hardly justified near the singularity. It is more likely that the matter will obey the relativistic equation of state $p = \mu/3$. Then

the vorticity conservation law becomes

$$\mu^{1/2}\underline{A}^{\mathrm{T}} \cdot \underline{\omega} \cdot \underline{A} = \text{const.}$$

If we assume as before that the flow lines converge along the axis of vorticity, we have the rather surprising result that $\omega \to 0$ as $\mu \to \infty$. If the solution is spatially homogeneous and the flow lines are orthogonal to the surfaces of homgeneity (this is possible only if the vorticity is zero), the acceleration of the flow lines will be zero. Then the asymptotic form of the eigenvalues of the matrix $\underline{S}$ will be the same as before and the density μ will be asymptotically proportional to $s^{-4/3}$. Thus in this case also there would be no particle horizon in the direction in which the flow lines are converging.

References

Bondi H. 1952. *Cosmology*. Cambridge University Press, Cambridge.

Bondi H. and T. Gold. 1948. The steady-state theory of the expanding universe. *Mon. Not. Roy. Ast. Soc.* **108**: 252–270.

Boyer R.H. and R.W. Lindquist. 1967. Maximal analytic extension of the Kerr solution. *J. Math. Phys.* **8**: 265–281.

Calabi E. and L. Markus. 1962. Relativistic space forms. *Ann. Math.* **75**: 63–76.

Carter B. 1966. The complete analytic extension of the Reissner-Nordström metric in the special case $e^2 = m^2$. *Phys. Lett.* **21**: 423–424.

Coxeter H.S.M. and G.J. Whitrow. 1950. World-structure and non-Euclidean honeycombs. *Proc. Roy. Soc. A* **201**: 417–437.

Doroshkevich A.G., Ya.B. Zeldovich, and I.D. Novikov. 1966. Gravitational collapse of non-symmetric and rotating masses. *J. Exp. Theor. Phys.* **22**: 122–130.

Ellis G.F.R. 2014. Stephen Hawking's 1966 Adams Prize Essay. *Eur. Phys. J. H*, DOI:10.1140/epjh/e2014-50014-x

Geroch R.P. 1966. Singularities in closed universes. *Phys. Rev. Lett.* **17**: 445–447.

Graves J.L. and D.R. Brill. 1960. Oscillatory character of the Reissner-Nordström metric for an ideal charged wormhole. *Phys. Rev.* **120**: 1507–1513.

Hawking S.W. 1965a. Properties of expanding universes. Ph.D. Thesis. Cambridge.

Hawking S.W. 1965b. Occurrence of singularities in open universes. *Phys. Rev. Lett.* **15**: 689–690.

Hawking S.W. and G.F.R. Ellis. 1965c. Singularities in homogeneous world models. *Phys. Rev. Lett.* **17**: 246–247.

Hawking S.W. 1966a. Singularities in the universe. *Phys. Rev. Lett.* **17**: 444–445.

Hawking S.W. 1966b. Perturbations of an expanding universe. *ApJ.* **145**: 544–554.

Hawking S.W. 1966c. The occurrence of singularities in cosmology. *Proc. Roy. Soc. Lond. A* **294**: 511–521.

Hawking S.W. and R.J. Tayler. 1966d. Helium production in anisotropic big bang universes. *Nature* **209**: 1278–1279.

Heckmann O. and E. Schücking. 1962. Relativistic cosmology. In: *Gravitation. An Introduction to Current Research*, ed. by L. Witten. Wiley, New York, pp. 438–469.

Hill E.L. 1955. Relativistic theory of discrete momentum space and discrete space-time. *Phys. Rev.* **100**: 1780–1783

Hocking J.G. and G.S. Young. 1961. *Topology*. Addison-Wesley Publishing Co. Inc., Reading, MA London. Reprinted, Dover Publications Inc., New York, 1988.

Hoyle F. 1948. A new model for the expanding universe. *Mon. Not. Roy. Ast. Soc.* **108**: 372–382.

Hoyle F. and J.V. Narlikar. 1964a. Time-symmetric electrodynamics and the arrow of time in cosmology. *Proc. Roy. Soc. A* **277**: 1–23

Hoyle, F. and J.V. Narlikar. 1964b. On the avoidance of singularities in C-field cosmology. *Proc. Roy. Soc. A* **278**: 465–478.

Kantowski R. 1966. Some relativistic cosmological models. Ph.D. Thesis. University of Texas.
Kobayashi S. and K. Nomizu. 1963. *Foundations of Differential Geometry.* Wiley Interscience, New York, Vol. I.
Kronheimer E.H. and R. Penrose. 1967. On the structure of causal spaces. *Proc. Camb. Phil. Soc.* **63**: 481–501.
Kruskal M.D. 1960. Maximal extension of Schwarzschild metric. *Phys. Rev.* **119**: 1743–1745.
Kundt W. 1963. Note on the completeness of spacetimes. *Z. f. Physik* **172**: 488–489.
Kundt W. and M. Trümper. 1962. Beiträge zur Theorie der Gravitations-Strahlungsfelder. *Akad. Wiss. Mainz* **12**: 967–1000.
Lichnerowicz A. 1955. *Global Theory of Connection and Holonomy Groups.* Noordhoff, Leyden.
Lifshitz E.M. and I.M. Khalatnikov. 1963. Investigations in relativistic cosmology. *Adv. Phys.* **12**: 185–249.
Markus L. 1955. Line element fields and Lorentz structures on differentiable manifolds. *Ann. Math.* **62**: 411–417.
Milnor J. 1963. *Morse Theory.* Ann. of Math. Studies no. 31, Princeton.
Misner C.W. 1965. Taub-NUT space as a counterexample to almost anything (preprint). In: *Relativity Theory and Astrophysics*, Vol. 1: Relativity and Cosmology. Lectures in Applied Mathematics, Vol. 8, edited by J. Ehlers. American Mathematical Society (Providence, Rhode Island, 1967), p. 160.
Penrose R. 1963. In: *Relativity, Groups, and Topology.* Gordon and Breach, New York.
Penrose R. 1965a. Zero rest-mass fields including gravitation: asymptotic behaviour. *Proc. Roy. Soc. A* **284**: 159–203.
Penrose R. 1965b. Gravitational collapse and space-time singularities. *Phys. Rev. Lett.* **14**: 57–59.
Rindler W. 1956. Visual horizons in world models. *Mon. Not. Roy. Ast. Soc.* **116**: 662–677.
Sandage A.R. 1961. The light travel time and the evolutionary correction to magnitudes of distant galaxies. *ApJ.* **134**: 916–926.
Trümper M. 1964. Contributions to actual problems in general relativity (preprint).
Walker A.G. 1944. Completely symmetric spaces. *J. Lond. Math. Soc.* **19**: 219–226.
Yano Y. and S. Bochner. 1953. *Curvature and Betti Numbers.* Ann. of Math. Studies no. 32, Princeton.
Zeeman E.C. 1964. Causality implies the Lorentz group. *J. Math. Phys.* **5**: 490–493.