PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The Occurrence of Singularities in Cosmology

S. W. Hawking

Proc. R. Soc. Lond. A 1966 **294**, 511-521
doi: 10.1098/rspa.1966.0221

The occurrence of singularities in cosmology

By S. W. Hawking

Gonville and Caius College, University of Cambridge

(*Communicated by H. Bondi, F.R.S.—Received* 15 *April* 1966)

It is shown that singularities of space-time are inevitable if the Einstein equations hold, if matter has normal properties and if the universe satisfies certain reasonable global conditions. The singularities would be in the past and would, in principle, be observable. Observation to determine whether such singularities actually occurred would provide a powerful test of the Einstein equations in strong fields. The singularity would not necessarily constitute a beginning of the universe.

1. Introduction

If it is assumed that there is no continual creation of matter, the observed recession of distant galaxies indicates that the density of the universe must have been higher in the past. There are two main possibilities: the universe might have contracted to some maximum density and then expanded again, or there might have been a time when the density was infinite. This would imply a singularity of space-time. Robertson (1933) showed that the universe must have had a singularity if it was exactly spatially homogeneous and isotropic, if the Einstein equations without cosmological constant held, and if matter had normal properties. However, it was suggested that maybe the singularity only occurred because of the high symmetry implied by homogeneity and isotropy and that a more realistic model with fewer symmetries would not have a singularity. Raychaudhuri (1955) and Komar (1956) showed that there must be a singularity in a model containing non-rotating dust and Shepley (1964) and Hawking & Ellis (1965) showed the same for a model that was spatially homogeneous but anisotropic. However, these models are not fully general, since in all of them there is some exact restriction or symmetry on the flow of matter. Lifshitz & Khalatnikov (1963) claim to have proved that there will not be a singularity in a fully general model that is, one that has the full number of arbitrary functions. However Penrose (1965) has shown that if the universe possesses a noncompact Cauchy surface a collapsing star must develop a singularity. The same method may be applied to prove the occurrence of a singularity in a universe model with a noncompact Cauchy surface which is approximately homogeneous and isotropic (Hawking 1965). However, the Cauchy surface of the universe might be compact (a 'closed' model) or it might even be that the universe possessed no global Cauchy surface. This paper will consider these possibilities and it will be shown that under certain reasonable conditions it can still be proved that singularities must occur.

2. Notation

Space-time is represented as a connected paracompact four-dimensional C^∞ manifold $\mathscr{M}$ with pseudo-Riemannian metric tensor g_{ij} of signature minus two. Covariant differentiation is indicated by a semi-colon, covariant differentiation along a timelike line by $\mathrm{D}/\mathrm{D}s$ and Lie differentiation by $\pounds$. Square brackets around indices indicate antisymmetrization. The conventions for the Riemann and Ricci tensors are

$$2V_{a;[bc]} = R^p{}_{abc}v_p,$$
$$R_{ab} = R^c{}_{acb}.$$

We assume the Einstein equations,

$$R_{ab} - \tfrac{1}{2}g_{ab}R = T_{ab},$$

where T_{ab} is the energy-momentum tensor of matter and units are such that the speed of light and the constant of gravitation equal one.

Space-time will be said to be singularity-free if the metric is a field of class C^2 and $\mathscr{M}$ is geodesically complete with respect to the metric.

3. Conjugate points

A timelike congruence will be said to be geodesic irrotational if its unit tangent vector v^a satisfies

$$v_{[a;b]} = 0.$$

The expansion of such a congruence is defined as $\theta = v^a{}_{;a}$. A point p will be said to be a singular point on a geodesic γ of a congruence if θ is infinite on γ at p. A point p will be said to be conjugate to a point q along a geodesic γ from q to p if p is a singular point on γ of the irrotational congruence of all timelike geodesics through q. A point p will be said to be conjugate to a spacelike hypersurface $\mathscr{H}$ if p is a singular point of the irrotational congruence of geodesics normal to $\mathscr{H}$.

A more precise description of conjugate points is as follows: let k^a be a vector connecting points corresponding distances along neighbouring geodesics in a congruence. Then k^a is dragged along by the congruence, that is,

$$\underset{v^a}{\pounds} k^a = 0. \tag{1}$$

Therefore
$$\mathrm{D}k^a/\mathrm{D}s = v_{a;b}k^b \tag{2}$$

and
$$\mathrm{D}^2k^a/\mathrm{D}s^2 = R_d{}^a{}_{bc}v^dv^ck^b. \tag{3}$$

Introducing an orthonormal tetrad $\underset{m}{e^a}(m = 1, 2, 3, 4)$ which is parallelly transported along the geodesic with $\underset{4}{e^a} = v^a$, equation (3) may be written

$$\frac{\mathrm{d}^2\underset{n}{k}}{\mathrm{d}s^2} = \underset{nr}{a}\,\overset{r}{k}, \tag{4}$$

where
$$\underset{n}{k} = \underset{n}{e^a}k_a,$$

$$\underset{nr}{a} = \underset{n}{e^b}\underset{r}{e^c}R_{abcd}v^av^d.$$

A solution of equation (4) will be called a Jacobi field. There are eight independent solutions. Since v^a and sv^a are solutions, the other six solutions may be taken orthogonal to v^a.

If q is a point on a geodesic γ with tangent vector v^a, the Jacobi fields along γ orthogonal to v^a, which vanish at q may be regarded as generating neighbouring geodesics in the irrotational congruence of all geodesics through q. Thus they obey

$$\mathrm{d}k_m/\mathrm{d}s = \underset{m:n}{v}\overset{n}{k}, \tag{5}$$

where $$\underset{m:n}{v} = e^a_m e^b_n v_{a;b} \quad (m,n = 1,2,3).$$

They may be written $$\underset{m}{k} = \underset{mn}{A}(s)\overset{n}{\frac{\mathrm{d}k}{\mathrm{d}s}}\Big|_q,$$

where $\underset{mn}{A}(s)$ is a 3×3 matrix which relates the fields at any point on γ to their derivatives at q.

For the rest of this paper we will omit tetrad indices from one to three and write product in matrix form, e.g.

$$\underset{mn}{A}\overset{n}{k} = \mathbf{A}.\mathbf{k}.$$

Note that because of the signature of the metric

$$\mathbf{A}.\mathbf{k} = -\sum_n \underset{mn}{A}\,\underset{n}{k}.$$

Then by (5) $$\mathrm{d}\mathbf{A}/\mathrm{d}s = \mathbf{V}.\mathbf{A},$$

where $$(\mathbf{V})_{mn} = \underset{m:n}{v}.$$

Thus $$\theta = \mathscr{T}r(\mathbf{V}) = \mathscr{T}r\left(\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}s}.\mathbf{A}^{-1}\right)$$

$$= [\det(\mathbf{A})]^{-1}\frac{\mathrm{d}}{\mathrm{d}s}[\det(\mathbf{A})].$$

But $$\mathrm{d}^2\mathbf{A}/\mathrm{d}s^2 = \mathbf{a}.\mathbf{A}.$$

Therefore $\mathrm{d}\mathbf{A}/\mathrm{d}s$ is finite. Hence θ is infinite where and only where $\det(\mathbf{A})$ is zero. But $\det(\mathbf{A})$ is zero if and only if there is a nonzero $\mathbf{h}$ such that

$$\mathbf{A}.\mathbf{h} = 0.$$

This gives an alternative definition of conjugate points. p is conjugate to q along γ if and only if there is a Jacobi field along γ, not identically zero which vanishes at p and q. This also shows that singular points of congruences are points where infinitesimally neighbouring geodesics meet.

4. The variation formulae

The length L of a timelike path γ from q to p is defined as

$$L = \int_q^p \mathrm{d}s.$$

We wish to investigate the behaviour of L under a variation of the path γ from q to p. A variation α of γ keeping end-points fixed is defined as a continuous map

$$\alpha: (-\epsilon, \epsilon) \times [0, L] \to \mathscr{M}$$

for some $\epsilon > 0$, such that

(1) $\alpha(0, s) = \gamma(s)$;

(2) there is a subdivision $0 = s_1 < s_2 \ldots < s_k = L$ of $[0, L]$ such that α is C^∞ on each strip $(-\epsilon, \epsilon) \times [s_i, s_{i+1}]$;

(3) the variation vector $w^a = \partial\alpha(u, s)/\partial u|_{u=0}$ is orthogonal to the tangent vector $v^a = \partial\alpha(0, s)/\partial s$;

(4) $\alpha(u, 0) = q, \quad \alpha(u, L) = p$.

The derivative of L with respect to the variation α is

$$\left.\frac{\partial L(\alpha(u))}{\partial u}\right|_{u=0} = -\Sigma w_a[v^a] - \int_0^L w_a \frac{\mathrm{D}v^a}{\mathrm{D}s}\,\mathrm{d}s,$$

where $[v^a]$ is discontinuity in v^a. Thus $\partial L/\partial u$ is zero under all variations if and only if $\mathrm{D}v^a/\mathrm{D}s = 0$ and there are no discontinuities in v^a. This expresses the well known fact that a geodesic (i.e. a path with $\mathrm{D}v^a/\mathrm{D}s = 0$) has stationary length. For a geodesic we may investigate the second derivative of L under a two parameter variation,

$$\alpha: U \times [0, L] \to \mathscr{M},$$

where U is a neighbourhood of $(0, 0)$ in E^2, with properties as before and

$$w_1^a = \left.\frac{\partial\alpha(u_1, 0, s)}{\partial u_1}\right|_{u_1=0},$$

$$w_2^a = \left.\frac{\partial\alpha(0, u_2, s)}{\partial u_2}\right|_{u_2=0}$$

are the two variation vectors.

Then the second derivative of L is

$$\left.\frac{\partial^2 L(\alpha(u_1, u_2))}{\partial u_1\,\partial u_2}\right|_{\substack{u_1=0\\u_2=0}} = -\int_0^L w_{2a}\left(\frac{\mathrm{D}^2 w_1^a}{\mathrm{D}s^2} + R^a{}_{bcd}\,v^b v^d w_1^c\right)\mathrm{d}s - \Sigma w_{2a}\left[\frac{\mathrm{D}w_1^a}{\mathrm{D}s}\right]$$

(for derivation see Milnor 1963, p. 75).

We may regard this second derivative of L as a bilinear functional $I(\mathbf{w}_1, \mathbf{w}_2)$ of the space of piecewise C^∞ vector fields along γ, orthogonal to v^a and vanishing at q and p.

Lemma 1. *If q has no conjugate points on γ in qp then $I(\mathbf{w}_1, \mathbf{w}_2)$ is negative definite, and therefore γ is maximal under variations α.*

Consider a set of three variation fields along γ orthogonal to v^a and vanishing at p and q. They may be written in the form

$$\mathbf{w}_i = \mathbf{C}(s).\mathbf{l}_i \quad (i = 1, 2, 3),$$

where $\mathbf{l}_i$ are a basis orthogonal to v^a at q. Let us take $\mathrm{d}\mathbf{w}_i/\mathrm{d}s$ to be linearly independent at p. The matrix $\mathbf{C}(s)$ vanishes at q and p. As there are no points conjugate to q in qp, the matrix $\mathbf{A}(s)$ of the previous section vanishes at q and is nonsingular elsewhere in qp. Thus $\mathbf{C}(s)$ may be written as

$$\mathbf{C}(s) = \mathbf{A}(s).\mathbf{B}(s),$$

where $\mathbf{B}(s)$ vanishes at p.
Then

$$\begin{aligned} I(\mathbf{w}_i, \mathbf{w}_j) &= \mathbf{l}_i^T.\left\{-\Sigma\mathbf{B}^T.\mathbf{A}^T.\left[\frac{\mathrm{d}}{\mathrm{d}s}(\mathbf{A}.\mathbf{B})\right] - \int_0^L \mathbf{B}^T.\mathbf{A}^T.\left(\frac{\mathrm{d}^2}{\mathrm{d}s^2}(\mathbf{A}.\mathbf{B}) - \mathbf{a}.\mathbf{A}.\mathbf{B}\right)\mathrm{d}s\right\}.\mathbf{l}_j \\ &= \mathbf{l}_i^T.\left\{-\Sigma\mathbf{B}^T.\mathbf{A}^T.\left[\frac{\mathrm{d}}{\mathrm{d}s}(\mathbf{A}.\mathbf{B})\right] - \int_0^L \mathbf{B}^T.\mathbf{A}^T.\left(\mathbf{A}.\frac{\mathrm{d}^2\mathbf{B}}{\mathrm{d}s^2} + 2\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}s}.\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}s}\right)\mathrm{d}s\right\}.\mathbf{l}_j \\ &= \mathbf{l}_i^T.\left\{\int_0^L \left(\frac{\mathrm{d}\mathbf{B}^T}{\mathrm{d}s}.\mathbf{A}^T.\mathbf{A}.\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}s} + \mathbf{B}^T.\left(\frac{\mathrm{d}\mathbf{A}^T}{\mathrm{d}s}.\mathbf{A} - \mathbf{A}^T.\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}s}\right).\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}s}\right)\mathrm{d}s\right\}.\mathbf{l}_j. \end{aligned}$$

But $\mathrm{d}\mathbf{A}/\mathrm{d}s = \mathbf{V}.\mathbf{A}$, where $\mathbf{V}$ is symmetric. Therefore,

$$I(\mathbf{w}_i, \mathbf{w}_j) = \mathbf{l}_i^T.\left\{\int_0^L \frac{\mathrm{d}\mathbf{B}^T}{\mathrm{d}s}.\mathbf{A}^T.\mathbf{A}.\frac{\mathrm{d}\mathbf{B}}{\mathrm{d}s}\,\mathrm{d}s\right\}.\mathbf{l}_j.$$

But $(\mathrm{d}\mathbf{B}^T/\mathrm{d}s)\,\mathbf{A}^T\mathbf{A}(\mathrm{d}\mathbf{B}/\mathrm{d}s)$ is negative semi-definite (because of the signature of the metric), and $\mathrm{d}\mathbf{B}/\mathrm{d}s$ is nonsingular at p. Therefore $I(\mathbf{w}_i, \mathbf{w}_j)$ is negative definite.

LEMMA 2. *If there is a point r on γ in qp, conjugate to q then γ is not maximal from p to q.*

Let $\mathbf{k}$ be a Jacobi field along γ vanishing at q and r. Let $\mathbf{h}$ be any field along γ vanishing at q and p and satisfying $\mathbf{h}^T.\mathrm{d}\mathbf{k}/\mathrm{d}s = 1$ at r. Extend $\mathbf{k}$ to p by putting it zero in rp. Let $\mathbf{w}$ be the field,

$$\mathbf{w} = c\mathbf{h} + c^{-1}\mathbf{k},$$

where c is a constant. Then

$$I(\mathbf{w}, \mathbf{w}) = 2 - c^2 b,$$

where b is a constant independent of c. Thus by taking c small enough $I(\mathbf{w}, \mathbf{w})$ can be made positive. Thus a geodesic γ from q to p is maximal if and only if q has no conjugate points in qp.

We may also consider variations of γ in which the end point q is not kept fixed but is allowed to vary over a surface $\mathscr{H}$ orthogonal to γ. Consider variations in L where L is the distance from $\mathscr{H}$ to p. Then

$$\begin{aligned} I(\mathbf{w}_1, \mathbf{w}_2) = &-\Sigma\mathbf{w}_1^T.\left[\frac{\mathrm{d}\mathbf{w}_2}{\mathrm{d}s}\right] - \int_0^L \mathbf{w}_1^T.\left(\frac{\mathrm{d}^2\mathbf{w}_2}{\mathrm{d}s^2} - \mathbf{a}.\mathbf{w}_2\right)\mathrm{d}s \\ &- \left(\mathbf{w}_2^T.\frac{\mathrm{d}\mathbf{w}_1}{\mathrm{d}s} - \mathbf{w}_2^T.\mathbf{U}.\mathbf{w}_1\right)\Big|_q, \end{aligned}$$

where $(\dot{\mathbf{U}})_{mn} = \underset{m:n}{u}$, and u^a is the unit normal to $\mathscr{H}$.

LEMMA 3. *A geodesic γ, orthogonal to $\mathscr{H}$, from q on $\mathscr{H}$ to p is maximal if and only if there is no point conjugate to $\mathscr{H}$ in qp.*

The proof is similar to that for lemmas 1 and 2.

5. THE ENERGY CONDITION

Equations of state are idealizations that are only approximately obeyed by the actual matter. Thus a singularity whose occurrence was dependent on the exact form of a particular equation of state would not be physically realistic. We shall therefore not assume anything about the detailed nature of the energy-momentum tensor T_{ab} but merely certain physically reasonable inequalities. The energy density in the frame of an observer with velocity u^a is $E = T_{ab}u^a u^b$. We shall assume that this is greater than or equal to $\frac{1}{2}T^a{}_a$ for any velocity, u^a. For a fluid with an isotropic pressure this assumption holds if the density is nonnegative and the pressure is greater than minus one-third the density. These are properties that any normal matter should have.

The conditions we have assumed imply $R_{ab}v^a v^b \geqslant 0$ for any timelike unit vector v^a. The significance of this is as follows: for any geodesic irrotational timelike congruence, the expansion θ obeys Raychaudhuri's equation,

$$\theta_{;b}v^b = -v_{a;b}v^{b;a} - R_{ab}v^a v^b.$$

Since $v_{[a;b]} = 0$, $v_{a;b}v^{b;a} \geqslant \frac{1}{3}\theta^2$. Therefore if θ is negative it will go to minus infinity in a finite distance. That is to say a congruence that is converging will have a singular point within a finite distance where neighbouring geodesics intersect. This result will be used in the next section.

6. SINGULARITIES

THEOREM 1. *Space-time cannot be singularity-free if,*

(1) *The energy-momentum tensor satisfies the condition of the previous section.*

(2) *The null-cones can be divided continuously into two classes 'past' and 'future'.* (*This means we can globally assign a time direction.*)

(3) *There is a Cauchy surface $\mathscr{H}$.* (*We define a Cauchy surface as a complete connected spacelike C^∞ surface which intersects every timelike and null line once and once only.*)

(4) *The expansion of v^a, the unit normal to $\mathscr{H}$, has a positive lower bound c on $\mathscr{H}$, i.e. $v^a{}_{;a} \geqslant c > 0$.*

Condition (2) is necessary if our ideas of time are valid. If condition (3) holds, condition (4) might be regarded as a precise statement of what we mean when we say the universe is expanding. Condition (3) will be further discussed in the next section.

To prove the theorem we will assume that space-time is singularity-free and show that this leads to an inconsistency if conditions (1) to (4) hold. The main lines of the proof will be given. It is necessary first to establish several lemmas.

Let $\mathscr{F}(\mathscr{S})$ and $\mathscr{P}(\mathscr{S})$ be the points that can be reached from a set $\mathscr{S}$ by respectively future and past directed timelike line. They have the property that if $p \in \mathscr{P}(\mathscr{S})$ then $\mathscr{P}(p) \subset \mathscr{P}(\mathscr{S})$. $\mathscr{P}(\mathscr{S})$ and $\mathscr{F}(\mathscr{S})$ are open sets.

PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

The boundary of the set $\mathscr{S}$ will be denoted by $\dot{\mathscr{S}} \equiv \overline{\mathscr{S}} \cap \overline{(\mathscr{M}-\mathscr{S})}$, where a bar indicates closure.

LEMMA 4. *If a nonempty set $\mathscr{N}$ is such that $p \in \mathscr{N}$ implies $\mathscr{P}(p) \subset \mathscr{N}$ and if $\mathscr{M}-\mathscr{N}$ is not empty then $\dot{\mathscr{N}}$ is locally homeomorphic to E^3 and no two points of $\dot{\mathscr{N}}$ have timelike separation.*

If $\mathscr{N}$ and $\mathscr{M}-\mathscr{N}$ are nonempty, then $\dot{\mathscr{N}}$ is non-empty. Let $q \in \dot{\mathscr{N}}$, then $\mathscr{P}(q) \subset \mathscr{N}$ and $\mathscr{F}(q) \subset \mathscr{M}-\mathscr{N}$. Thus there can be no point $r \in \dot{\mathscr{N}} \cap \mathscr{F}(q)$ since if $r \in \mathscr{F}(q)$ then there would be a neighbourhood $\mathscr{U}$ of r such that $\mathscr{U} \subset \mathscr{F}(q) \subset \mathscr{M}-\mathscr{N}$ which is impossible if $r \in \dot{\mathscr{N}}$. We may introduce a normal coordinate system x^a $(a = 1, 2, 3, 4)$ in a neighbourhood $\mathscr{U}$ of q with x^4 timelike, such that the lines $x^i =$ constant $(i = 1, 2, 3)$ intersect $\mathscr{P}(q)$ and $\mathscr{F}(q)$. Then each of these lines contains a point of $\dot{\mathscr{N}}$. Moreover, the x^4 coordinates of these points must be continuous since no two points of $\dot{\mathscr{N}}$ have a timelike separation. Thus $\dot{\mathscr{N}}$ is locally homeomorphic to E^3.

LEMMA 5. *If $\mathscr{N}$ satisfies the conditions of lemma 4 and if for a point $q \in \dot{\mathscr{N}}$ there is a normal coordinate neighbourhood $\mathscr{U}$ of q and a ball $\mathscr{B} \subset \mathscr{U}$ of constant coordinate radius about q which contains a sequence of points y_n converging to q such that there is a future directed null or timelike line λ_n from each point y_n which intersects $(\dot{\mathscr{B}} \cap \mathscr{N})$, then $\dot{\mathscr{N}}$ contains a future directed null geodesic segment from q.*

(The author is indebted to Dr R. Penrose for the following proof.)

Let z be a limit point of $\lambda_n \cap \overline{(\dot{\mathscr{B}} \cap \mathscr{N})}$. Select a subsequence y_n' such that $\lambda_n' \cap \overline{(\dot{\mathscr{B}} \cap \mathscr{N})}$ converges to z. Then if $\mathscr{V}$ is any neighbourhood of z, all y_n' for n large enough are contained in $\overline{\mathscr{P}}(\mathscr{V})$. Thus $q \in \overline{\mathscr{P}}(\mathscr{V})$. Therefore $\mathscr{V}$ intersects $\overline{\mathscr{F}}(q)$. This shows that $z \in \overline{\mathscr{F}}(q)$. However $z \in \overline{\mathscr{N}}$, and thus it is not contained in $\mathscr{F}(q)$. Therefore z lies on the future null cone from q and the null geodesic segment qz lies in $\dot{\mathscr{N}}$.

COROLLARY. If there is a past directed null geodesic segment from q in $\dot{\mathscr{N}}$ this must be a continuation of the future directed null geodesic segment from q since if it were in any other direction there would be points of $\dot{\mathscr{N}}$ with timelike separation. If there is more than one future-directed null geodesic segment from q in $\dot{\mathscr{N}}$ there can be no past directed null geodesic segment from q in $\dot{\mathscr{N}}$.

A boundary of a set satisfying lemmas 4 and 5 will be called a null horizon.

LEMMA 6. *For any p the set $\mathscr{J}(p) \equiv \mathscr{F}(p) \cap \mathscr{P}(\mathscr{H})$ has compact closure.*

Let $\mathscr{C}$ be the set of all points which have an open neighbourhood $\mathscr{W}$ such that $\overline{\mathscr{J}}(\mathscr{W})$ is compact or empty. Points near $\mathscr{H}$ will be in $\mathscr{C}$. Suppose $\mathscr{D} \equiv \mathscr{M}-\mathscr{C}$ is not empty. Then it satisfies the conditions of lemma 4. We shall see that the conditions of lemma 5 are satisfied at each point of $\dot{\mathscr{D}}$. Let $q \in \dot{\mathscr{D}}$, and let $\mathscr{B}$ be a ball of constant coordinate radius about q in a normal coordinate neighbourhood $\mathscr{U}$. Let y_n be a sequence of points in $\mathscr{P}(q)$ converging to q with $y_{n+1} \in \mathscr{F}(y_n)$. Let $\mathscr{V}_n$ be a set of open neighbourhoods in $\mathscr{P}(q)$ with $y_{n+1} \in \mathscr{V}_{n+1} \subset \mathscr{F}(y_n)$. Suppose that for some n, $(\overline{\mathscr{F}}(\mathscr{V}_n) \cap \dot{\mathscr{B}}) \subset \mathscr{C}$ then since $\overline{\mathscr{F}}(\mathscr{V}_n) \cap \dot{\mathscr{B}}$ is compact it can be covered by a finite number of open neighbourhoods $\mathscr{W}_i$ in $\mathscr{C}$ for which $\overline{\mathscr{J}}(\mathscr{W}_i)$ is compact. But

then $\overline{J}(\mathscr{V}_n)$ would be contained in $[\bigcup_i \overline{J}(\mathscr{W}_i)] \cup [\overline{F}(\mathscr{V}_n) \cap \overline{\mathscr{B}}]$ which is compact. This is impossible since it would imply $y_n \in \mathscr{C}$. Thus from each $\mathscr{V}_n$ and each y_{n-1} there must be a timelike or null line which intersects $\dot{\mathscr{B}} \cap \mathscr{D}$. This shows, by lemma 5, that $\dot{\mathscr{D}}$ is a null horizon, and that the null geodesic segments generating $\dot{\mathscr{D}}$ can have no future end-point. However, $\dot{\mathscr{D}}$ will be in $\mathscr{P}(\mathscr{H})$ and every future directed null geodesic in $\mathscr{P}(\mathscr{H})$ intersects $\mathscr{H}$ which is not in $\mathscr{D}$. This would imply that the generators of $\dot{\mathscr{D}}$ had future end-points which shows that the supposition that $\mathscr{D}$ was nonempty must be false. Thus $\overline{J}(p)$ is compact for all p.

For a point $p \in \mathscr{P}(\mathscr{S})$ we define $d(p, \mathscr{S})$ to be the least upper bound of the lengths of future directed timelike lines from p to $\mathscr{S}$. We define $d(p, \mathscr{S})$ to be zero if p is not in $\mathscr{P}(\mathscr{S})$. Near the Cauchy surface $\mathscr{H}$, there will be points in $\mathscr{P}(\mathscr{H})$ with $d(p, \mathscr{H})$ finite.

LEMMA 7. *If $p \in \mathscr{P}(\mathscr{H})$ and $d(p, \mathscr{H})$ is finite, then there is a future directed timelike geodesic from p to $\mathscr{H}$ of length $d(p, \mathscr{H})$.*

Let $\mathscr{G}(a)$ be the set of all points r such that $d(r, \mathscr{H}) = a$. Since $\overline{J}(p)$ is compact we can find a sufficiently small $a > 0$ such that every point $q \in (\mathscr{G}(a) \cap \overline{J}(p))$ is connected to $\mathscr{H}$ by a unique future directed timelike geodesic of length a. Then $\mathscr{G}(a) \cap \overline{J}(p)$ is compact. Let $q \in (\mathscr{G}(a) \cap \overline{J}(p))$ be such that $d(p, q)$ is maximized over $\mathscr{G}(a) \cap \overline{J}(p)$. Then $d(p, q) = d(p, \mathscr{H}) - a$. Let $\gamma(s)$ be the past directed timelike geodesic from $\mathscr{H}$ to q with $s = 0$ at $\mathscr{H}$ and $s = a$ at q. Let $\mathscr{U} \subset \mathscr{P}(\mathscr{H})$ be a normal coordinate neighbourhood in which $\gamma(s)$ lies on the x^4 axis and let $\mathscr{B} \subset \mathscr{U}$ be a ball of constant coordinate radius about q. Let $y \equiv \gamma(s) \cap \dot{\mathscr{B}}$ for $s < a$ and $y' \equiv \gamma(s) \cap \dot{\mathscr{B}}$ for $s > a$. Let $z \in \dot{\mathscr{B}} \cap \overline{\mathscr{P}}(q)$ be such that $d(p, z)$ is maximized. Then z must coincide with y' since if it did not $d(z, y)$ would be greater than $d(z, q) + d(q, y)$ and there would be a timelike line from $\mathscr{H}$ through y and z to p of length greater than $d(p, \mathscr{H})$. Thus the relation $d(p, \gamma(s)) = d(p, \mathscr{H}) - s$ holds along $\gamma(s)$ from $s = 0$ to z. But then it must hold for all values of s with $0 \leqslant s \leqslant d(p, \mathscr{H})$ because if s_0 were the least upper bound of the values of s for which it held we could apply the same construction at the point $\gamma(s_0)$ and prove it held for $s = s_0 + \delta$. Thus the point $\gamma(d(p, \mathscr{H}))$ either coincides with p or lies on $\dot{F}(p)$ not at p. An application of a similar construction shows the latter is impossible. Thus γ is a past directed timelike geodesic of length $d(p, \mathscr{H})$ from $\mathscr{H}$ to p. It will be orthogonal to $\mathscr{H}$.

LEMMA 8. $d(p, H)$ is *finite*.

Let $\mathscr{N}$ be the set of all points p for which $d(p, \mathscr{H})$ is infinite. Suppose $\mathscr{N}$ is nonempty. Then the conditions of lemmas 4 and 5 would be satisfied, and $\dot{\mathscr{N}}$ hence would be a null horizon. This is impossible since the generators of $\mathscr{N}$ would intersect $\mathscr{H}$. Hence $\mathscr{N}$ is empty.

We are now in a position to prove theorem 1. Conditions (1) and (4) imply that there is a point conjugate to $\mathscr{H}$ on every past directed geodesic normal to $\mathscr{H}$ within a distance $3/c$. But any point $p \in \mathscr{P}(\mathscr{H})$ can be joined to $\mathscr{H}$ by a future directed timelike geodesic orthogonal to $\mathscr{H}$ and of length $d(p, \mathscr{H})$. By lemma 3 this geodesic cannot contain a point conjugate to $\mathscr{H}$. But any past-directed timelike geodesic can be extended to a length greater than $3/c$. Thus there are points p for which $d(p, \mathscr{H}) > 3/c$.

PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

This is a contradiction which shows that the assumption that space-time is singularity free is incompatible with conditions (1) to (4).

This theorem depended on condition (4) which might not be observationally verifiable since if we were on the Cauchy surface ourselves we would not be able to investigate other parts of it to see whether the normals were indeed expanding. Thus a thoerem that depended on a condition that could be checked by observation might seem preferable.

THEOREM 2. *Space-time cannot be singularity-free if conditions* (1) *to* (3) *are satisfied, and*

(5) *There is a point q on $\mathscr{H}$ such that all past-directed timelike geodesics through q have a positive lower bound for convergence within an upper bounded distance from q. That is all past-directed geodesics through q start converging again in a finite distance.*

Condition (5) will be satisfied by a Robertson model. It will therefore be satisfied by a model that is sufficiently similar to a Robertson model at the present time. Since it only depends on the behaviour of geodesics inside the light cone of q, it could in principle be tested by observation.

Conditions (1) and (5) imply that all past-directed timelike geodesics through q have a point conjugate to q within an upper bound of distance from q. The proof is then similar to that for theorem 1.

Theorems 1 and 2 prove the occurrence of singularities in the past for expanding universe models. They could also be applied to contracting models to prove the occurrence of singularities in the future.

7. CAUCHY SURFACES

Theorems 1 and 2 depended on the existence of a Cauchy surface. Penrose has suggested (private communication) that the universe might not possess a global Cauchy surface. He instances the Kerr solution (cf. Carter 1966) as an example of a space-time where, although no spacelike surface intersects any timelike or null line more than once, it is impossible to find a spacelike surface that does intersect every line. Although it would mean modifying our ideas of determinism, we have no *a priori* reason for believing that the universe might not have similar properties. However, if our ideas of causality have any validity we ought to be able to find a complete connected spacelike surface $\mathscr{H}$ which although not necessarily intersecting every timelike and null line does not intersect any more than once. We shall call such a surface a partial Cauchy surface. $\mathscr{K}$, the set of points p such that every future directed timelike or null line from p intersects $\mathscr{H}$ will be called the past Cauchy development of $\mathscr{H}$. If $\mathscr{P}(\mathscr{H})-\mathscr{K}$ is not empty we can apply lemmas 4 and 5 to show that $\dot{\mathscr{K}}-\mathscr{H}$ is a null horizon which we shall call the past Cauchy horizon of $\mathscr{H}$. This does not imply any contradiction since the partial Cauchy surface, $\mathscr{H}$, does not intersect every null line.

THEOREM 3. *Space-time cannot be singularity free if conditions* (1) *and* (2) *are satisfied, and*

(6) *There exists a smooth compact partial Cauchy surface $\mathscr{H}$ whose unit normal v^a is expanding everywhere on $\mathscr{H}$, i.e. $v^a{}_{;a} > 0$ on $\mathscr{H}$.*

PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

(7) *There is an open neighbourhood $\mathscr{W}$ of each point $p \in \mathscr{P}(\mathscr{H})$ such that every future-directed null geodesic which intersects $\mathscr{W}$ leaves it at some point and never reenters it. (In effect this is a condition that there should not be closed null lines in $\mathscr{P}(\mathscr{H})$.)*

Note that it is necessary that $\mathscr{H}$ be compact since Minkowski space satisfies conditions (1), (2) and (7) and has a non-compact partial Cauchy surface whose normal is expanding.

LEMMA 9. *There exists a global nonvanishing past directed timelike C^2 vector field u^a on $\mathscr{M}$.* (The author is indebted to Dr R. Penrose and Professor C. W. Misner for this.)

By a standard theorem (cf. Bishop & Crittenden 1964, p. 126), $\mathscr{M}$ being paracompact may be endowed with a positive definite differentiable Riemannian metric m_{ab}. Consider the eigenvalue equation, $m_{ab}u^b = \lambda g_{ab}u^b$ at a point p. By transforming to coordinates in which m_{ab} is the unit matrix at p, it can be seen that the timelike eigenvector u^a is unique since it corresponds to a positive eigenvalue, whereas the other eigenvalues are negative. The eigenvector u^a may be normalized by $u^a u^b g_{ab} = 1$.

This vector field may be used to give a differomorphism β of $\mathscr{H} \times [0, t]$ into $\mathscr{M}$, for any $t > 0$,* by taking points of $\mathscr{H}$ a distance $s \in [0, t]$ along the past directed integral curves of u^a. If $d(r, \mathscr{H})$ had an upper bound t for $r \in \dot{\mathscr{K}} - \mathscr{H}$ then

$$\dot{\mathscr{K}} - \mathscr{H} \subset \beta(\mathscr{H} \times [0, t])$$

since every future directed timelike line through $\dot{\mathscr{K}} - \mathscr{H}$ intersects $\mathscr{H}$. Thus $\dot{\mathscr{K}} - \mathscr{H}$ would be compact and could be covered by a finite number of the neighbourhoods $\mathscr{W}_i$ of condition (7). A future directed null geodesic segment generating $\dot{\mathscr{K}} - \mathscr{H}$ must leave each $\mathscr{W}_i$ it enters and not reenter. This is impossible since it cannot have a future end-point. Hence $d(r, \mathscr{H})$ can have no upper bound for $r \in \dot{\mathscr{K}} - \mathscr{H}$.

By conditions (1) and (6) there is a point conjugate to $\mathscr{H}$ on each past directed geodesic normal to $\mathscr{H}$ within some finite upper bound b of distance from $\mathscr{H}$. Let $\mathscr{L}$ be the set of all points p for which $\mathrm{d}(p, \mathscr{H}) > b$. Then $\mathscr{L} \cap (\dot{\mathscr{K}} - \mathscr{H})$ is not empty. Since $\mathscr{L}$ is open, $\mathscr{L} \cap \mathscr{K}$ is not empty. If $r \in \mathscr{K}$ lemmas 6, 7 and 8 may be applied to show that $\mathrm{d}(r, \mathscr{H})$ is finite and that r may be joined to $\mathscr{H}$ by a future directed line of length $\mathrm{d}(r, \mathscr{H})$. This leads to the same contradiction as in theorem 1 since $p \in \mathscr{L} \cap \mathscr{K}$ would be joined to $\mathscr{H}$ by a line of length $\mathrm{d}(p, \mathscr{H}) > b$.

8. CONCLUSION

The preceding sections indicate that singularities are inevitable in certain models if condition (1) is satisfied and the Einstein equations hold. Unlike the singularities in collapsed stars which cannot be seen by an external observer, these singularities would in principle be observable. So far the Einstein equations have only been experimentally tested for very weak fields for which they give almost the same results as Newtonian theory. Thus observations to determine whether singularities actually occurred would provide a test of the equations for strong fields. Presumably

* *Note added in proof.* This is incorrect. A correct proof is possible using the differomorphism defined by the geodesics normal to $\mathscr{H}$.

it would be necessary to consider quantum effects in very strong fields. However, these would not become important until the radius of curvature became of the order of 10^{-14} cm which for practical purposes is pretty singular.

The view has been expressed that singularities are so objectionable that if the Einstein equations were to predict their occurrence, this would be a compulsive reason for modifying them. However, the real test of a physical theory is not whether its predicted results are aesthetically attractive but whether they agree with observation. So far there are no observations which would show that singularities do not occur.

Conditions (1) will be satisfied by normal matter. However, it will not be satisfied by the '*C*' field of Hoyle & Narlikar (1963). This is a field or negative energy density which can prevent the occurrence of singularities. However, there would seem to be a certain quantum-mechanical difficulty associated with the existence of fields of negative energy density. For there would not seem anything to prevent the creation, in a given volume of space-time, of an infinite number of particles of positive energy and a corresponding infinite number of quanta of the negative energy field. It might be possible to overcome this difficulty but so far no workable scheme has been proposed.

Theorems 1, 2 and 3 prove the existence of a singularity but give no information as to its nature. Thus it need not be like the all-embracing singularity in the Robertson models which every world-line hits: it might be an isolated singularity which only a few lines hit. Further research will be needed to determine the actual character of the singularity in each situation.

The author is grateful to Mr B. Carter, Dr G. R. F. Ellis, Dr R. Penrose and Dr D. W. Sciama for many useful discussions and helpful suggestions.

References

Bishop, R. L. & Crittenden, R. J. 1964 *The geometry of manifolds*. Academic Press.
Carter, B. 1966 *Phys. Rev.* **141**, 1242.
Hawking, S. W. 1965 *Phys. Rev. Lett.* **15**, 689.
Hawking, S. W. & Ellis, G. R. F. 1965 *Phys. Lett.* **17**, 246.
Hoyle, F. & Narlikar, J. V. 1963 *Proc. Roy. Soc.* A **278**, 465.
Komar, A. 1956 *Phys. Rev.* **104**, 544.
Lifshitz, E. M. & Khalatnikov, I. M. 1963 *Adv. Phys.* **12**, 183.
Milnor, J. 1963 *Morse theory*. Princeton University Press.
Penrose, R. 1965 *Phys. Rev. Lett.* **14**, 57.
Raychaudhuri, A. 1955 *Phys. Rev.* **98**, 1123.
Robertson, H. P. 1933 *Rev. Mod. Phys.* **5**, 62.
Shepley, L. C. 1964 *Proc. Nat. Acad. Sci. U.S.A.* **52**, 1403.

PROCEEDINGS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES