PHYSICAL REVIEW D VOLUME 17, NUMBER 10 15 MAY 1978

Energy conditions and spacetime singularities

Frank J. Tipler
Department of Mathematics, University of California at Berkeley, Berkeley, California 94720
(Received 7 June 1977)

In this paper, a number of theorems are proven which collectively show that singularities will occur in spacetime under weaker energy conditions than the strong energy condition. In particular, the Penrose theorem, which uses only the weak energy condition but which applies only to open universes, is extended to all closed universes which have a Cauchy surface whose universal covering manifold is not a three-sphere. Furthermore, it is shown that the strong energy condition in the Hawking-Penrose theorem can be replaced by the weak energy condition and the assumption that the strong energy condition holds only on the average. In addition, it is demonstrated that if the Universe is closed, then the existence of singularities follows from the averaged strong energy condition alone. It is argued that any globally hyperbolic spacetime which satisfies the weak energy condition and which contains a black hole must be null geodesically incomplete.

I. INTRODUCTION

The Hawking-Penrose singularity theorems collectively show that singularities must occur in spacetime provided three types of conditions are imposed:

(1) certain reasonable initial conditions, such as the existence of trapped surfaces or the existence of a compact spacelike hypersurface;

(2) various restrictions on the causal structure, such as the existence of a Cauchy surface or the absence of closed timelike curves;

(3) energy conditions.

Now at least one condition of the first type has been justified by astronomical observation,[1] and I have shown elsewhere[2,3] that, roughly speaking, conditions (1) and (3) imply that a violation of the causal structure conditions in unlikely to prevent the predicted singularities; indeed, such violation would tend to *make* singularities rather than remove them. In the present paper I shall address the problem of weakening the third type of conditions. I shall prove several singularity theorems which collectively show singularities will still occur even when the restrictions on the energy tensor are relaxed.

Recall that the singularity theorems use the energy conditions as sufficient conditions for the focusing of certain causal geodesic congruences whose presence in space-time is guaranteed by conditions of the first type. The existence of focal points on these congruences is then shown to be inconsistent with the causal structure restrictions. It is therefore concluded that although focusing occurs, it does not lead to focal points; i.e., some of the geodesics in the congruence must be incomplete. Since the energy conditions used in the standard singularity theorems are sufficient but not necessary conditions, we will still be able to assert that singularities must occur in spacetime if we can find weaker sufficient conditions for the occurrence of focal points on complete causal geodesics.

In Sec. II, we will discuss the possibility that the *strong energy condition*, which requires

$$(T_{ab} - \tfrac{1}{2}g_{ab}T)U^aU^b \geq 0 \tag{1}$$

for all causal vectors U^a, will be violated. This is the condition used in the best known singularity theorem, the Hawking-Penrose theorem (Ref. 1, p. 266), for the condition is a sufficient condition for the convergence of causal geodesic congruences, and it is various causal geodesic congruences whose continued focusing is required in the proof of this theorem. The strong energy condition is not a necessary condition for the continued focusing, however. In fact, it will be shown in Sec. II that if (1) holds only on the average along all causal geodesics, then generic closed universes must have an incomplete timelike or null geodesic. Thus more than a small, local violation of the strong energy condition will be required to avoid singularities in closed universes.

Many familiar matter fields, such as the massive scalar field, have a stress-energy tensor which can violate the strong energy condition but not the *weak energy condition*, which says that

$$T_{ab}U^aU^b \geq 0 \tag{2}$$

for all causal vectors U^a. Thus the prediction of singularities would be more believable if there were a singularity theorem which used only the weak energy condition. Penrose's theorem (Ref. 1, p. 262) uses only the weak energy condition, for only the convergence of *null* geodesic congruences is needed in this theorem, and the weak energy condition is sufficient to insure this. Unfortunately, this theorem is restricted to open universes. In Sec. III Penrose's theorem will be extended to all closed universes which possess a Cauchy sur-

face whose universal covering manifold is noncompact. It will be further shown that if black holes exist and the weak energy condition holds, then some null geodesics must be incomplete whatever the topology of the Cauchy surface. It will be demonstrated in this section that the strong energy condition in the Hawking-Penrose theorem can be replaced with the assumption that the weak energy condition holds everywhere and the strong energy condition holds on the average. The significance of these new singularity theorems will then be discussed. It will be pointed out that they imply that the "nonsingular" cosmology proposed by Murphy[4] is in fact singular, and that the nonsingular cosmology of Bekenstein[5] develops singularities when it is perturbed sufficiently to create black holes.

Section IV will conclude the paper with a discussion of the possibility that the weak energy condition is also violated at extremely high densities and pressures. It will be shown that there is no local energy condition which is weaker than the weak energy condition.

Conventions and notations of this paper are the same as those used in Hawking and Ellis[1] (HE) unless otherwise noted. The Einstein equations are $R_{ab} = 8\pi(T_{ab} - \frac{1}{2}g_{ab}T)$.

II. THE STRONG ENERGY CONDITION AND THE AVERAGED STRONG ENERGY CONDITION

The strong energy condition, when coupled to the space-time geometry via the Einstein equations, basically says that gravitation is always an attractive force. As emphasized by Hawking and Ellis[1] (p. 95), a violation of this condition has never been seen in the laboratory. In particular, the electromagnetic field and the massless scalar field cannot violate the strong energy condition under any circumstances. However, it is possible for the massive scalar field to violate the condition since for this field the stress-energy tensor gives

$$(T_{ab} - \tfrac{1}{2}g_{ab}T)U^aU^b = (\phi_{;a}U^a)^2 + \tfrac{1}{2}m^2\phi^2U_aU^a = (\phi_{;a}U^a)^2 + \tfrac{1}{2}\phi\phi_{;a}{}^{;a}U_bU^b, \qquad (3)$$

where U^a is any causal vector. Thus if the scalar field or its second derivative becomes sufficiently large, the absolute value of the second term in (3) can become larger than the first term, and this will cause a violation of the strong energy condition (since $U_aU^a \leq 0$). As emphasized by Bekenstein,[5] a violation of the strong energy condition by a massive scalar field may be physically significant because the strong interactions in nuclear matter can be regarded as mediated by a classical massive scalar field. Bekenstein has used the fact that (3) can become negative to construct a nonsingular Friedmann cosmology. His stress-energy tensor consists of incoherent radiation together with a classical conformal massless scalar field coupled to dust particles, and he finds that under these circumstances the S^3 topology Friedmann universe will "bounce" instead of terminating in a singularity. Interestingly, the weak energy condition is not violated, though the strong energy condition is.

Parker and Fulling[6] have obtained a bouncing S^3 Friedmann universe by using as their matter tensor the expectation value of the canonical stress-energy tensor arising from a quantized massive scalar field. A computer analysis of the case in which the cosmology is in a special quantum state showed that the universe could contract to a minimum size and then re-expand, but it is not known if this behavior would continue for an infinite number of cycles, or even if it would occur at all for a physically reasonable quantum state.

To my mind one of the strongest reasons for doubting the validity of the strong energy condition in the extremely high-density regime comes from a consideration of the spontaneously broken gauge symmetry theories. A typical action is

$$\int d^4x\,(-g)^{1/2}\{(1/16\pi)(R - F^{\alpha}{}_{ab}F_{\alpha}{}^{ab}) - (1/4\pi)[\tfrac{1}{2}D_a\phi^{\beta}D^a\phi_{\beta} + V(\phi)]\},$$

where

$$F^{\alpha}{}_{ab} = \partial_aA^{\alpha}{}_b - \partial_bA^{\alpha}{}_a + c^{\alpha}{}_{\beta\sigma}A^{\beta}{}_aA^{\sigma}{}_b,$$

$$D_a\phi^{\beta} = \partial_a\phi^{\beta} + A^{\alpha}{}_ac^{\beta}{}_{\alpha\delta}\phi^{\delta},$$

$$V(\phi) = \tfrac{1}{2}m^2\phi^{\alpha}\phi_{\alpha} + \tfrac{1}{4}\lambda(\phi^{\alpha}\phi_{\alpha})^2 + \tfrac{1}{4}m^4/\lambda \qquad (m^2 < 0,\ \lambda > 0)$$

and the c's are the structure constants of the gauge symmetry group. The canonical stress-energy tensor obtained from the above action is

$$T_{ab} = \tfrac{1}{4}\pi(F^{\alpha}{}_a{}^cF_{\alpha bc} - \tfrac{1}{4}g_{ab}F^{\alpha}{}_{de}F_{\alpha}{}^{de}) + \tfrac{1}{4}\pi[D_a\phi^{\alpha}D_b\phi_{\alpha} - \tfrac{1}{2}g_{ab}D_c\phi^{\alpha}D^c\phi_{\alpha} - g_{ab}V(\phi)]. \qquad (4)$$

Yasskin has pointed out[7] that, due to the presence of the $V(\phi)$ term, it is possible for (4) to violate the strong energy condition (but not the weak energy condition).

The important thing to notice in the above examples is the fact that the energy condition violation occurs only in a restricted set of circumstances; in the Parker-Fulling case, for example, the violation is *known* to occur only for a certain set of quantum states. It is quite possible that the violation would be quite local, with the strong energy condition holding in an average sense. The following theorem shows that if the strong energy condition holds on the average along all causal geo-

desics, then the occurrence of singularities is still inevitable in any closed universe on which the Einstein equations hold.

Theorem 1. Space-time (M,g) is not timelike and null geodesically complete if

(1) $\int_{-\infty}^{+\infty} R_{ab}U^aU^b dt \geq 0$ along every complete causal geodesic $\gamma(t)$, equality holding only if $R_{ab}U^aU^b \equiv 0$ over the entire history of $\gamma(t)$ [U^a is the tangent vector to $\gamma(t)$, and t is an affine parameter];

(2) every causal geodesic contains a point for which $U^cU^d U_{[a}R_{b]cd[e}U_{f]} \neq 0$ (i.e., the generic condition holds);

(3) the chronology condition holds;

(4) there exists a compact achronal set without edge.

Comment. Condition (2) is not needed if we assume that $\int R_{ab}U^aU^b dt$ is strictly positive along all causal geodesics. Note that $\int R_{ab}U^aU^b \geq 0$ and the Einstein equations imply $\int (T_{ab} - \frac{1}{2}g_{ab}T)U^aU^b \geq 0$. Thus condition (1) implies that the strong energy condition holds on the average, where the average is taken over the entire history of a causal geodesic.

Proof of Theorem 1. Hawking and Penrose have proven (Ref. 1, p. 267) that the following three conditions are inconsistent:

(a) every inextendible causal geodesic contains a pair of conjugate points;

(b) the chronology condition holds on (M,g);

(c) there exists an achronal set S such that $E^+(S)$ is compact.

We shall show that the assumption of causal geodesic completeness and assumptions (1)–(4) are inconsistent because together they imply conditions (a)–(c).

Assumption (3) is the same as condition (b) and the compact achronal set without edge of assumption (4) is the set S of condition (c), since for an achronal set without edge we have $E^+(S) = S$.

We will complete the proof by showing that (1) and (2) together with the assumption of causal geodesic completeness imply (a). Two points p and q on a causal geodesic $\gamma(t)$ are said to be *conjugate* along $\gamma(t)$ if the expansion θ of a geodesic congruence containing $\gamma(t)$ becomes infinite at p and q. The expansion θ satisfies

$$\frac{d\theta}{dt} = -R_{ab}U^aU^b - 2\sigma^2 - (1/n)\theta^2 , \tag{5}$$

where U^a is the tangent vector to the geodesic, t is an affine parameter along $\gamma(t)$, and $n = 3$ for timelike geodesics and $n = 2$ for null geodesics. The function σ^2 is positive definite. For null geodesics it satisfies

$$\frac{d\sigma_{mn}}{dt} = -R_{manb}U^aU^b - \theta\sigma_{mn} + \tfrac{1}{2}\delta_{mn}R_{ab}U^aU^b , \tag{6}$$

where $2\sigma^2 \equiv \sigma_{mn}\sigma^{mn}$ and $m, n = 1, 2$ label the two spacelike directions of a pseudo-orthonormal frame parallel propagated along $\gamma(t)$. For timelike geodesics σ^2 satisfies

$$\frac{d\sigma_{\alpha\beta}}{dt} = -R_{\alpha a\beta b}U^aU^b - \sigma_{\alpha\gamma}\sigma_{\gamma\beta} - \tfrac{2}{3}\sigma_{\alpha\beta}\theta + \tfrac{1}{3}\delta_{\alpha\beta}(R_{ab}U^aU^b + 2\sigma^2) , \tag{7}$$

where $2\sigma^2 \equiv \sigma_{\alpha\beta}\sigma_{\alpha\beta}$ and $\alpha, \beta = 1, 2, 3$ label the three spacelike directions of an orthonormal frame parallel propagated along $\gamma(t)$. [Equation (7) follows from equation (4.25) of HE,[1] and Eq. (6) follows from the equation for null geodesics analogous to (4.25).]

It can be shown (Ref. 1, pp. 97 and 100) that p and q are conjugate along $\gamma(t)$ if and only if a function y, defined by $\theta = (1/y)dy/dt$, satisfies $y = 0$ at q and p. If we define a new function x by the relation $x^n = y$, then $\theta = (n/x)dx/dt$, and (5) becomes

$$\frac{d^2x}{dt^2} + F(t)x = 0 , \tag{8}$$

where

$$F(t) = (1/n)(R_{ab}U^aU^b + 2\sigma^2) . \tag{9}$$

Since $x^n = y$, y will be zero at p and q if and only if $x = 0$ at p and q. Thus we have reduced the problem of finding conjugate points to the problem of discovering the location of zeros in solutions to (8): a complete causal geodesic $\gamma(t)$ will have a pair of conjugate points if and only if there is a solution to (8) which has at least two zeros on the interval $(-\infty, +\infty)$.

I have shown elsewhere[21] that (8) will have a solution with at least two zeros if $\int_{-\infty}^{+\infty} F(t)\,dt > 0$. For causal geodesics along which $R_{ab} \neq 0$, this integral is positive by assumption (1) and the fact that $\sigma^2 \geq 0$. If $\int_{-\infty}^{+\infty} R_{ab}U^aU^b = 0$ along a geodesic, then by assumption (1) $R_{ab}U^aU^b \equiv 0$ on the geodesic. By assumption (2), there exists a point p on every causal geodesic for which the first term of Eq. (6) or (7) is nonzero, depending on whether the geodesic is null or timelike respectively. This means that σ^2 is positive in some neighborhood of p. Thus

$$\int_{-\infty}^{+\infty} F(t)dt = \int_{-\infty}^{+\infty} (1/n)(R_{ab}U^aU^b + 2\sigma^2)dt = \int_{-\infty}^{+\infty} (2\sigma^2/n)dt > 0 ,$$

so (8) will have a solution with at least two zeros in $(-\infty, +\infty)$.

It follows that assumptions (1) and (2) together with causal geodesic completeness imply (a).

III. THE WEAK ENERGY CONDITION

The weak energy condition essentially says that the local energy density is positive definite; i.e., every local observer measures the mass density to be non-negative. As its name implies, the weak energy condition is a much weaker condition to impose on spacetime than the strong energy condition. In fact, except for the Parker-Fulling quantum field, all the matter fields mentioned in Sec. II as possible violators of the strong energy condition must obey the weak energy condition. Thus a singularity theorem which used only the weak energy condition would rest on a much firmer physical foundation than a singularity theorem which required use of the strong energy condition. Penrose's theorem (Ref. 1, p. 263) is a singularity theorem which uses only the weak energy condition. Unfortunately, it applies only to open universes. The following theorem extends Penrose's theorem to all closed universes whose Cauchy surfaces do not have a compact universal covering manifold.

Theorem 2. A spacetime (M,g) on which the Einstein equations hold cannot be null geodesically complete if

(1) the weak energy condition holds;

(2) (M,g) contains a spacelike Cauchy hypersurface S;

(3) the universal covering manifold to S is noncompact;

(4) there exists at least one of the following:

(i) a closed S^2 trapped surface,

(ii) a point p such that on every past (or every future) null geodesic from p the divergence θ of the null geodesics from p becomes negative (i.e., the null geodesics from p are focused by the matter or curvature and start to reconverge).

The proof of Theorem 2 will require the proof of two propositions. We first define the *universal covering spacetime* $(\tilde{M},\tilde{g})$ to a given spacetime (M,g) to be the universal covering manifold $(\tilde{M},f)$ to M and the unique metric $\tilde{g}$ induced on $\tilde{M}$ from g by the covering map f.

Proposition 1. Let (M,g) be a globally hyperbolic spacetime. Then the universal covering spacetime $(\tilde{M},\tilde{g})$ is also globally hyperbolic.

Proof: Since (M,g) is globally hyperbolic, $M=R^1\otimes S$, where S is a spacelike Cauchy hypersurface for (M,g), and for each $a\in R^1$, $\{a\}\otimes S$ is a spacelike Cauchy surface for (M,g) (Ref. 1, p. 212). Let $\tilde{S}$ be the universal covering manifold to S. Then $R^1\otimes\tilde{S}$ covers M, and further $R^1\otimes\tilde{S}$ is simply connected. Since $R^1\otimes\tilde{S}$ is simply connected, it is its own universal covering manifold. Thus $\tilde{M}=R^1\otimes\tilde{S}$ and the covering map f which maps $\tilde{M}$ onto M maps each $\tilde{S}$ onto a corresponding S. The metric $\tilde{g}$ induced on $\tilde{M}$ by f is unique, and together with $\tilde{M}$ it defines the universal covering spacetime $(\tilde{M},\tilde{g})$ to (M,g). Now each $\tilde{S}$ is spacelike is the metric $\tilde{g}$. For suppose some $\tilde{S}$ were not everywhere spacelike, say at the point $\tilde{x}_0$. Then since f is a diffeomorphism from some neighborhood U of $\tilde{x}_0$ onto $f(U)$ in M, S could not be spacelike at a point $x_0=f(\tilde{x}_0)$. But this contradicts the fact that S is a spacelike Cauchy hypersurface. Furthermore, all causal curves intersect a given $\tilde{S}$ at least once. For suppose there were a causal curve γ which did not intersect $\tilde{S}$. Then $f(\gamma)$ is a causal curve in M which does not intersect $f(\tilde{S})$. But all causal curves intersect $f(\tilde{S})$, since $f(\tilde{S})$ is a Cauchy surface for (M,g). Geroch has shown[8] that no timelike curve can intersect $f^{-1}(S)=\tilde{S}$ more than once, and if a null curve intersected $\tilde{S}$ more than once, there would be a timelike curve which intersected $\tilde{S}$ more than once. To see this, let $p<q$ be two points at which the null curve λ intersected $\tilde{S}$. If λ is not a null geodesic generator of $\dot{I}^+(p)$, then $p\ll q$. If λ is a null geodesic generator of $\dot{I}^+(p)$, then since $\tilde{S}$ is without boundary (because M and $\tilde{M}$ are without boundary) there is a point $w\in\lambda$ with $p<w<q$ such that $\tilde{S}\cap I^+(w)$ is nonempty. If $t\in\tilde{S}\cap I^+(w)$, then $p\ll t$. In summary, $\tilde{S}$ is a spacelike hypersurface in which every causal curve intersects exactly once. By definition (Ref. 1, pp. 201 and 205) $\tilde{S}$ is a Cauchy surface for $(\tilde{M},\tilde{g})$, and a spacetime which contains a Cauchy surface is globally hyperbolic (Ref. 1, p. 209).

Proposition 2. Let T be a spherical trapped surface in a spacetime (M,g). Then there exists a spherical trapped surface $\tilde{T}$ in the universal covering spacetime $(\tilde{M},\tilde{g})$ of (M,g) and $f(\tilde{T})=T$.

Proof. The spherical trapped surface T is an embedded two-dimensional submanifold of M; i.e., we can regard T as a map h: $S^2\to M$, where S^2 is a two-sphere (Ref. 1, pp. 101 and 262). Pick a point $\tilde{p}\in\tilde{M}$ such that $f(\tilde{p})=p\in T$. Let $x_0\in S^2$ be defined by $h(x_0)=p$. Then there exists a unique continuous map c: $S^2\to\tilde{M}$ such that $c(x_0)=\tilde{p}$ and $fc=h$. For by the homotopy lifting theorem[9] such a c will exist if and only if the image of h_*: $\pi_1(S^2,x_0)\to\pi_1(M,p)$ is contained in that of f_*: $\pi_1(\tilde{M},\tilde{p})\to\pi_1(M,p)$, where π_1 is the fundamental group. Now $\pi_1(S^2,x_0)=i$, the identity element, since S^2 is simply connected, and h_* maps $\pi_1(S^2,x_0)$ into the identity element of $\pi_1(M,p)$. Furthermore, the image of $f_*(\pi_1(\tilde{M},\tilde{p}))$ in $\pi_1(M,p)$ contains the identity element. Thus the image of h_* is contained in the image of f_*, so the map c exists. Since c is continuous, $c(S^2)\subset\tilde{M}$ is compact and without boundary. In addition, $c(S^2)$ is two-dimensional since f is locally a diffeomorphism. Defining $c(S^2)=\tilde{T}$, we have that $\tilde{T}$ is simply connected since if it were not, there would exist a loop $\tilde{n}$

$\in \tilde{T}$ which would not be homotopic to zero, but $f(\tilde{m}) = n \in T$. Because T is simply connected, n is homotopic to a loop $m \in T \cap U$ with n, m having a point in common, where U is a neighborhood with $f(\tilde{U})$ mapped diffeomorphically onto U. Let $\tilde{m} \in \tilde{T}$ be defined by $f(\tilde{m}) = m$. Then $\tilde{m}$ must be homotopic to zero since m is contained in U. By Corollary 2.4 of Ref. 10, $\tilde{m}$ is homotopic to $\tilde{n}$. Thus $\tilde{n}$ must be homotopic to zero, which is a contradiction. Thus $\tilde{T}$ is compact, two-dimensional. boundaryless, and simply connected. It is therefore a two-sphere.[11,12] Since f is locally a diffeomorphism, $\tilde{T}$ will be an imbedded spacelike submanifold of $\tilde{M}$, and if ${}_n\tilde{\chi}_{ab}$ is a null second fundamental form of $\tilde{T}$ at $\tilde{p} \in \tilde{T}$, then ${}_n\tilde{\chi}_{ab}|_{\tilde{p}} = {}_n\chi_{ab}|_p$, where $f(\tilde{p}) = p$. Thus ${}_1\tilde{\chi}_{ab}\tilde{g}^{ab}$ and ${}_2\tilde{\chi}_{ab}\tilde{g}^{ab}$ are both negative on $\tilde{T}$. It follows that $\tilde{T}$ is a spherical trapped surface.

Proof of Theorem 2. Suppose to the contrary that (M, g) *were* null geodesically complete. Let $(\tilde{M}, \tilde{g})$ be the universal covering space-time to (M, g). Then $(\tilde{M}, \tilde{g})$ would also be null geodesically complete (Ref. 8; Ref. 1, p. 181). Suppose condition (4ii) holds. Pick a point $\tilde{p} \in \tilde{M}$ such that $f(\tilde{p}) = p$. The set of directions at p and all null geodesics which move into the past from p is compact, and hence the set of directions at $\tilde{p}$ of all null geodesics which move into the past from $\tilde{p}$ will also be compact, since there is a neighborhood U of p which is mapped diffeomorphically onto a neighborhood $f^{-1}(U)$ of $\tilde{p}$ by f^{-1}. By conditions (1), (4ii) and proposition 4.4.4 of HE,[1] every past-directed null geodesic γ_p from p would have a point q conjugate to p within a finite affine parameter distance c. For each past-directed null geodesic γ_p, let γ_{pq} denote the null geodesic segment consisting of the portion of γ_p between q and p. By Corollary 2 of Ref. 13, there exists a unique path $\tilde{\gamma}_{pq}$ in $\tilde{M}$ which covers γ_{pq} and which has origin $\tilde{p}$. Since a neighborhood of each point of $\tilde{\gamma}_{pq}$ will be mapped diffeomorphically onto a neighborhood of each point of γ_{pq} by f, $\tilde{p}$ and $\tilde{q} = f^{-1}(q)$ will be conjugate points of $\tilde{\gamma}_p$, and the affine parameter length of $\tilde{\gamma}_{pq}$ must be less than or equal to c. By proposition 1, $(\tilde{M}, \tilde{g})$ is globally hyperbolic. Thus by proposition 6.6.1 of HE,[1] $\dot{J}^-(\tilde{p}) = E^-(\tilde{p})$, and hence $\dot{J}^-(\tilde{p})$ is generated by null geodesic segments with future endpoints at $\tilde{p}$. By proposition 4.5.12 of HE[1] and the achronality of $\dot{J}^-(\tilde{p})$, the set $\mathring{J}^-(\tilde{p})$ is compact, being the intersection of the closed set $\dot{J}^-(\tilde{p})$ with the compact set consisting of all the past-directed null geodesic segments from $\tilde{p}$ with length equal to c. By the proof of proposition 1, we can write $\tilde{M}$ as $R^1 \otimes \tilde{S}$, where $\tilde{S}$ is a Cauchy surface for $(\tilde{M}, \tilde{g})$ and $\tilde{S}$ is the universal covering manifold to S. By condition (3), $\tilde{S}$ is noncompact. Since $(\tilde{M}, \tilde{g})$ is a spacetime, it admits a future-directed C^1 timelike vector field which is everywhere nonvanishing. Each integral curve of this field will intersect $\tilde{S}$, and will intersect $\mathring{J}^-(\tilde{p})$ at most once. The set of integral curves thus defines a continuous injective map $\alpha: \mathring{J}^-(\tilde{p}) \to S$. If $\mathring{J}^-(\tilde{p})$ were indeed compact, its image $\alpha(\mathring{J}^-(\tilde{p}))$ would also be compact and would be homeomorphic or $\mathring{J}^-(\tilde{p})$. Since $\tilde{S}$ is noncompact, $\alpha(\mathring{J}^-(\tilde{p}))$ could not contain the whole of $\tilde{S}$ and would therefore have a boundary in $\tilde{S}$. This is impossible, since by proposition 6.3.1 of HE,[1] $\dot{J}^-(\tilde{p})$ and hence $\alpha(\mathring{J}^-(\tilde{p}))$ would be a three-dimensional manifold without boundary. We have a contradiction, and thus (M, g) cannot be null geodesically complete. If condition (4i) holds, then by this condition, conditions (2) and (3), and propositions 1 and 2, the universal covering spacetime $(\tilde{M}, \tilde{g})$ has a noncompact Cauchy surface and a spherical trapped surface. Thus the conditions imposed on the universal covering space-time are the same as those imposed on a spacetime in Penrose's theorem (Ref. 1, p. 263), and so $(\tilde{M}, \tilde{g})$ must be null geodesically incomplete. Thus (M, g) must be null geodesically incomplete.

Hawking and Ellis have shown (Ref. 1, Sec. 10.1) that condition (4ii) almost certainly holds in our universe. Therefore, if we can believe the weak energy condition, theorem 2 shows that singularities are inevitable in our Universe provided our Universe has a Cauchy surface whose universal covering manifold is noncompact. Now Poincaré's conjecture[14,15]—which almost all mathematicans believe but which is still unproven—states that a compact, boundaryless, simply connected three-manifold is a three-sphere. Hence, if we assume Poincaré's conjecture to be true, theorem 2 shows that our Universe must contain singularities unless our Universe contains a Cauchy surface whose universal covering manifold is topologically S^3. Unfortunately, it is quite possible that our Universe contains a Cauchy surface with topology S^3 (such a Cauchy surface will be its own universal covering manifold). Furthermore, condition (3) of theorem 2 cannot be replaced by a weaker condition. Consider, for example, the Robertson-Walker universes which satisfy Einstein's equations with a matter tensor that consists of a perfect fluid satisfying the weak energy condition plus a cosmological constant term with Λ positive; that is, the matter tensor is given by

$$T_{ab} = (\mu + p)V_a V_b + p g_{ab} - (\Lambda g_{ab})/8\pi .$$

Such a matter tensor will satisfy the weak energy condition. If $K = +1$ (S^3 topology) and $\Lambda = \Lambda_{crit} > 0$, then there is a static solution, the Einstein static universe. In this universe the null geodesics from *every* point first expand and then reconverge to a point, so condition (4ii) holds. Since each t = constant surface is a Cauchy surface, condition (2)

also holds. Thus every condition of theorem 2 holds except for condition (3), and the Einstein static universe is null geodesically complete. If $K=+1$ and $0<\Lambda<\Lambda_{crit}$, then there exists a solution which contracts from an infinite radius, reaches a minimum radius, and re-expands. It is easily verified[16] that this solution contains trapped surfaces in the contracting phase; for this solution all the conditions of theorem 2 hold except for condition 3, and this solution is also null geodesically complete (Ref. 1, p. 139).

These examples show that if we wish to extend theorem 2 to S^3 universes, we must use a stronger initial condition than condition (4). In particular, I will argue that if we replace condition (4) by the requirement that (M,g) contains a black hole, then we can remove condition (3) from theorem 2.

Black holes have been defined precisely only for asymptotically flat spacetimes: a black hole is a connected component of $S\cap\mathring{J}^-(\mathcal{I}^+)$, where S is a partial Cauchy surface from which the spacetime is strongly future asymptotically predictable (Ref. 1, p. 315). One must calculate the entire future history of S in order to determine the precise boundary of a black hole. In practice, no one bothers to do this. For astrophysical calculations relativists generally assume that the surface of a black hole is probably very near the outermost marginally trapped surface which is then said to be the surface of the black hole. This marginally trapped surface is defined in general by the sequence of trapped surfaces inside, so trapped surfaces really define black holes in practice. Thus two concepts, trapped surfaces and event horizons, are used to define black holes. The concepts are related in asymptotically flat predictable spacetimes since in these spaces trapped surfaces must lie inside event horizons. In closed universes, the relationship between these concepts is not so clear cut because there is no "natural observer" like $\mathcal{I}^+$ from which an absolute event horizon can be defined. However, it does seen intuitively clear that the definition of black holes in closed universes should involve trapped surfaces and event horizons, with the former inside the latter. I shall thus say that *a necessary condition for the existence of a black hole* is the existence of a trapped surface T and an inextendible timelike curve γ such that $T\cap J^-(\gamma)$ is empty. That is, there is at least *one* observer who can never see the trapped surface. We then have theorem 3.

Theorem 3. A spacetime (M,g) on which the Einstein equations hold cannot be null geodesically complete if

(1) the weak energy condition holds;

(2) (M,g) contains a spacelike Cauchy hypersurface;

(3) there exists a black hole in (M,g).

This theorem has essentially been proved in outline on page 265 of HE.[1] However, it was not, stated there very precisely.

This theorem indicates the the Bekenstein nonsingular Friedmann cosmology[5]—which has Cauchy surface topology S^3 and which obeys the weak (but not the strong) energy condition—would develop singularities if the inhomogeneities of a realistic version of this model caused the formation of black holes. Indeed, Bekenstein himself was well aware that inhomogeneities might possibly destroy the nonsingular nature of his model. In asserting that the homogeneity is responsible for the *absence* of singularities in the Bekenstein model, I am turning the pre-Penrose argument *against* singularities on its head.

It should be emphasized that the singularities predicted by theorems 2 and 3 are incomplete *null* geodesics; there need not be any incomplete *timelike* geodesics. There is a very simple criterion for the existence of incomplete null geodesics in open Robertson-Walker universes.

Theorem (Hawking[16]). Any expanding, open Robertson-Walker cosmology which satisfies the null convergence condition is null geodesically incomplete in the past direction. (The null convergence condition says that $R_{ab}K^aK^b\geq 0$ for all null vectors K^a.)

Murphy has constructed[4] a "nonsingular" $K=0$ Robertson-Walker universe which satisfies the weak energy condition. Since this model is expanding and satisfies the Einstein equations, it must, by Hawking's theorem, contain past incomplete null geodesics. The presence of such null geodesics is not surprising, since Murphy's universe asymptotically approaches the steady-state universe in the past direction, and the steady-state universe is known to be null geodesically incomplete in the past direction (Ref. 1, p. 126).

If we add the null convergence condition to theorem 1, we can extend theorem 1 to open universes.

Theorem 4 [Generalized Hawking-Penrose theorem (Ref. 1, p. 266)]. Spacetime (M,g) cannot be timelike and null geodesically complete if

(1) the null convergence condition holds;

(2) $\int_{-\infty}^{+\infty}R_{ab}U^aU^b dt\geq 0$ along every complete timelike geodesic $\gamma(t)$, equality holding only if $R_{ab}U^aU^b\equiv 0$ over the entire history of $\gamma(t)$:

(3) the generic condition holds;

(4) the chronology condition holds;

(5) there exists at least one of the following:

(i) a compact achronal set without edge,

(ii) a closed trapped surface,

(iii) a point p such that on every past (or every future) null geodesic from p the divergence θ of

the null geodesics from p becomes negative.

Proof. By (1) and (5), there exists an achronal set S such that $E^+(S)$ or $E^-(S)$ is compact [if (M,g) is null geodesically complete]. The argument is the same as the one on page 267 of HE.[1] By (1) and (3) every complete null geodesic will have a pair of conjugate points. By (2) and (3) and the proof of theorem 1, every complete timelike geodesic has a pair of conjugate points. We can now deduce the existence of an incomplete causal geodesic by repeating the argument on pages 267–269 of HE.[1]

IV. IS THE WEAK ENERGY CONDITION VIOLATED?

It is interesting that the weak energy condition is the weakest energy condition that can be defined locally; that is, it is the weakest energy condition which can be defined using the entire set of timelike vectors in T_p, where T_p is the set of all tangent vectors at a point p in M. To see this we prove proposition 3.

Proposition 3. If $T_{ab}U^aU^b$ at p is bounded below for all timelike vectors U^a in T_p, then the weak energy condition holds at p.

Proof. Suppose $T_{ab}U^aU^b = -c$, with $c>0$ and U^a timelike. Now $V^a = tU^a$, for all $t\in(0,+\infty)$, is also a timelike vector, Thus $T_{ab}=t^2T_{ab}U^aU^b=-t^2c$, which is not bounded below since t^2 can be made as large as we please. Thus there can be no such c; i.e., $T_{ab}U^aU^b\geq 0$ for all timelike U^a.

Therefore, if the weak energy condition can be weakened, it must be done by restricting in some way the causal vectors U^a in $T_{ab}U^aU^b$. The obvious restriction—requiring U^a to be a unit timelike vector—will not yield a weaker energy condition for the focusing of null geodesics provided the matter tensor is similar in form to the matter tensor of most known fields; that is, provided the matter tensor is type I, which means that at each point p there is an orthornomal frame $\vec{E}_1$, $\vec{E}_2$, $\vec{E}_3$, $\vec{E}_4$ for which the tensor takes the form (Ref. 1, p. 89)

$$T_{ab}=\begin{bmatrix} p_1 & & & \\ & p_2 & & \\ & & p_3 & \\ & & & \mu \end{bmatrix}, \tag{10}$$

where the p's denote the principal pressures and μ the energy density.

Proposition 4. If $T_{ab}U^aU^b$ is bounded below for all unit timelike vectors U^a in T_p, and if T_{ab} is type I, then $T_{ab}K^aK^b\geq 0$ at p for all null vectors K^a in T_p.

Proof. Since T_{ab} is type I, there is an orthornormal frame $\vec{E}_1$, $\vec{E}_2$, $\vec{E}_3$, $\vec{E}_4$ with T_{ab} represented as in (10). If we perform a Lorentz boost in the α direction ($\alpha=1,2,3$), we obtain a new unit timelike vector U^a which satisfies

$$T_{ab}U^aU^b=\gamma^2(\mu+\beta^2p_\alpha). \tag{11}$$

Since $\beta^2\leq 1$, but γ^2 can have any value greater than or equal to one, (11) will not be bounded below unless $p_\alpha\geq|\mu|$ (or unless $\mu\geq 0$, $\mu+p_\alpha\geq 0$, in which case the proof is in). Now any null vector at p can be written $K^a=K(a,b,c,1)$ in the $\vec{E}_a$ frame, where $a^2+b^2+c^2=1$. Thus

$$\begin{aligned} T_{ab}K^aK^b &= K^2(a^2p_1+b^2p_2+c^2p_3+\mu) \\ &\geq K^2[|\mu|(a^2+b^2+c^2)+\mu| \\ &= K^2(|\mu|+\mu)\geq 0. \end{aligned}$$

Although the weak energy condition is the weakest energy condition that can be defined locally, there is a very strong reason for wanting a singularity theorem that uses a weaker condition. In semiclassical gravitation theory, one assumes that the gravitational field is a classical object generated by quantized matter fields. These quantized matter fields couple to the gravitational field via the semiclassical Einstein equations:

$$G_{ab}=8\pi\left(\phi,\int f(x)T_{ab}(x)\,dx\,\phi\right)$$

where $(\phi,\int f(x)T_{ab}(x)dx\,\phi)$ is the expectation value for the state ϕ of the smeared stress-energy operator $T_{ab}(x)$. An important result of axiomatic quantum field theory is the following.

Theorem (Epstein, Glaser, and Jaffe[17]). Let f be a positive test function with compact support and let $T(x)$ be a local field[18] which satisfies $(\phi,\int T(x)f(x)\,dx\,\phi)\geq 0$ for all states ϕ, and $(\phi_0,\int T(x)f(x)\,dx\,\phi_0)=0$ for the vacuum state ϕ_0. Then $\int T(x)f(x)\,dx\equiv 0$.

In other words, there exists *some* state $\hat{\phi}$ for which the expectation value of the operator $\int T(x)f(x)dx$ is negative. Setting $T(x)=T_{ab}$ and $T(x)=T_{ab}-\frac{1}{2}g_{ab}T$, we find that this theorem implies that there must exist a state $\tilde{\phi}$ and a state $\hat{\phi}$ for which the weak and the strong energy conditions are violated respectively. Several authors[6,19,20] have suggested that this violation of the energy conditions might prevent the formation of singularities. However, the Epstein-Glaser-Jaffe theorem only guarantees the existence of *one* state violating an energy condition. There is no guarantee that many such states will exist, or that the Universe will ever be in such a state, or that the Universe will remain in such a state. Furthermore, the Universe may be in a state for which the strong energy condition but not the weak energy condition is violated. If this happens, then theorems 2 and 3 would im-

ply singularities. If the Universe remains in a strong-energy-condition-violating state only for a brief period, and if the amount of violation is not too great, then theorem 1 would imply singularities (provided the Universe is closed). Thus the existence of states for which the local energy conditions do not hold does not *ipso facto* invalidate the prediction of singularities. To show that in the actual Universe, singularities are prevented by the above-mentioned quantum-mechanical mechanism, it would be necessary to show that the Universe will be in such a negative-energy state sufficiently long to violate the averaged strong energy condition.

ACKNOWLEDGMENTS

I should like to thank M. H. Protter, R. K. Sachs, E. H. Wichmann, C. H. Woo, and P. B. Yasskin for helpful discussions. This work was supported by the National Science Foundation under Grant Number MCS-76-21525.

[1]S. W Hawking and G. F. R. Ellis, *The Large Scale Structure of Space-time* (Cambridge University Press, London, 1973), p. 354.
[2]F. J. Tipler, Phys. Rev. Lett. 37, 879 (1976).
[3]F. J. Tipler, Ann. Phys. (N.Y.) 108, 1 (1977). Ph. D. thesis (Department of Physics and Astronomy, University of Maryland, 1976) (unpublished).
[4]G. L. Murphy, Phys. Rev. D 8, 4231 (1973).
[5]J. D. Bekenstein, Phys. Rev. D 11, 2072 (1975).
[6]L. Parker and S. A. Fulling, Phys. Rev. D 7, 2357 (1973).
[7]P. Yasskin (private communication).
[8]R. P. Geroch, J. Math. Phys. 8, 782 (1967).
[9]T. S. Hu, *Homotopy Theory* (Academic, New York, 1959), p. 89.
[10]P. Hilton, *Algebraic Topology: An Introductory Course* (New York University Press, New York, 1969), p. 51.
[11]A. Wallace, *Differential Topology: First Steps* (Benjamin, New York, 1968), p. 107.
[12]M. Greenberg, *Lectures on Algebraic Topology* (Benjamin, New York, 1967), p. 116.
[13]I. M. Singer and J. A. Thorpe, *Lecture Notes on Elementary Topology and Geometry* (Scott, Foresman & Company, Glenview, 1967), p. 60.
[14]S. Smale, Bull. Amer. Math. Soc. 66, 373 (1960).
[15]C. D. Papakyriakopoulos, Bull. Amer. Math. Soc. 64, 317 (1958).
[16]S. W. Hawking, Phys. Rev. Lett. 15, 689 (1965).
[17]H. Epstein, V. Glaser, and A. Jaffe, Nuovo Cimento 36, 1016 (1965).
[18]I assume that the set of local fields satisfies the axioms of a local field theory. See R. F. Streater and A. S. Wightman, *PCT, Spin & Statistics, and All That* (Benjamin, New York, 1964), p. 101 for a detailed statement of these axioms.
[19]R. M. Wald, Commun. Math. Phys. 54, 1 (1977).
[20]L. Parker, in *Asymptotic Structure of Spacetime*, edited by F. P. Esposito and L. Witten (Plenum, New York, 1977), p. 182.
[21]F. J. Tipler, J. Diff. Eq. (to be published).