Commun. math. Phys. 49, 17—23 (1976)

Communications in
Mathematical
Physics

Space-Time Singularities

C. J. S. Clarke
Department of Mathematics, University of York, Heslington, York Y01 5DD, England

Abstract. A set of conditions for the reasonableness of space-time is proposed and investigated. Using these, together with strong causality and an assumption of genericness, it is shown that future timelike or null geodesically incomplete space-times contain either curvature or intermediate singularities, or primordial singularities.

1. What is a Reasonable Space-Time?

One would like to find acceptable physical grounds for excluding many of the "pathological" spacetimes that can be constructed as counter-examples to seemingly plausible conjectures. For instance, it might be thought that gravitational collapse would inevitably lead to a curvature or intermediate singularity [1]; it would, however, be mathematically possible for space-time simply to come to an end before any predicted singularity formed. To prevent this, I shall propose two physical conditions that space-time should satisfy. One (maximality) asserts that space-time does not arbitrarily stop; the other (hole-freeness) asserts that predictions, and perhaps retrodictions, made on the basis of formally adequate Cauchy data are not falsified by the spontaneous appearance of uncaused singularities.

A further condition, rather weaker than the Hausdorff conditions, requires that a non-quantum space-time (excluding the Wheeler-Everett picture) does not undergo arbitrary branching. This leads to the concept of a Hajicek space-time [2, 3].

In what follows "smooth" denotes some fixed sufficiently strong differentiability condition on the metric. "Singularity" is used in the sense of Schmidt [7].

Definition 1. *A Hajicek space-time (or simply: a space-time) is a pair* (M, g); *where* M *is a connected* C^∞ *4-manifold, not necessarily Hausdorff,* g *is a smooth pseudo-Riemannian metric on* M *of signature* $(- + + +)$, *and* M *has the Hajicek property: there exists no pair of curves* $c_i:(0, 1] \rightarrow M$ $(i=1, 2)$ *for which* $c_1|(0, g) = c_2|(0, g)$ *but* $c_1(g) \neq c_2(g)$ *for some* $g \in (0, 1]$.

Scholium. Such a pair $\{c_i\}$ constitute what Hajicek [2] calls "a bifurcate curve": that is, a curve which branches, not by splitting within an ordinary Hausdorff manifold [when $c_1(g)=c_2(g)$], but by participating in a *branching* of the whole space-time. If the c_i were past-directed timelike curves they would correspond to a pair of observes who persued a common path $c_i|(0,g)$ on a future segment of their world-lines, but who might totally disagree on what the universe had been doing when they compared notes about their past segments $c_i|[g,1]$. In a Hajicek space-time the universe is allowed to branch providing it does not thereby bifurcate any curves. As is well known (Lemma 1 and Theorem 2), this imposes a strong control on any branching.

Definition 2. *A space-time is maximal if it is not isometric to a proper subspace of any other space-time.*

Scholium. The class of maximal space-times excludes all those which are obtained by "cutting out" a closed set.

Definition 3. *A space-time is hole-free*[1] *if, for any spacelike submanifold S (without boundary), the domain of dependence*[2] *$D(S)$ has the property that there is no isometry $\phi:D(S)\to N$ into another space-time for which $D(\phi(S))\neq\phi(D(S))$.*

Scholium. This excludes examples such as the following. Let M be the universal covering space of Minkowski space with the 2-plane $\{t=0, x=0\}$ removed. This is maximal but not hole-free, since $D(\{t=-1\})$ (on any sheet of M) is "punctured" by the singularity at $t=x=0$ and its image under the natural map ϕ into Minkowski space is properly contained in $D(\phi(\{t=-1\}))$, which is the whole space. By using D, rather than D^+, the definition is made symmetric between retrodiction and prediction. This avoids the problem of having to determine what the appropriate "arrow of time" is either for M or for each S separately; but it has the possible drawback that examples such as the space-time in [6], where the singularity leaves no trace behind it, are not hole-free.

Theorem 1. *Any space-time has a maximal extension.*

This theorem is false for a non-Hausdorff space-time without the Hajicek condition, since there is then no limit to the extent to which additional branches can be grafted onto the space-time. We have, however the following:

Lemma 1. *A Hajicek space-time is second-countable.*

Proof of Lemma. As with the corresponding theorem for Hausdorff space-times, we can proceed via the bundle $L(M)$ of all frames on M (either pseudo-orthonormal or linear), showing first that $L(M)$ is Hausdorff (compare [3]).

1. There are no bifurcate curves in $L(M)$. For let $\{c_1, c_2\}$ ($c_i:(0,1]\to L(M)$) be a pair with $c_1|(0,g)=c_2|(0,g)$. Then $\pi\circ c_1|(0,g)=\pi\circ c_2|(0,g)$ [where $\pi:L(M)\to M$ is the canonical projection] and so, by the Hajicek property on M, $\pi c_1(g)=\pi c_2(g)=x$, say. Since both of $\pi\circ c_i$ $(i=1,2)$ are continuous, for any coordinate neighbourhood U of x there will be numbers h_1, h_2 with $\pi\circ c_i|(h_i,g]$ mapping into U. So $c_i|(h_i,g]$ maps into $\pi^{-1}U$, which is Hausdorff. Hence $c_1(g)=c_2(g)$.

1 I am indebted to J. Earman and N. Woodhouse for this definition (private communications)
2 The definition of $D(S)$ is as in [5], p. 201, except that I do not require S to be closed

2. $L(M)$ has a (positive definite) Riemannian metric $\tilde{g}$ [7]. Let $p, q \in L(M)$ and choose convex normal neighbourhoods P, Q of each with respect to $\tilde{g}$. For any choice of P, either there is a $\tilde{g}$-geodesic in Q ending at q which intersects P at points arbitrarily close to q, or else there is a least distance from q at which these geodesics intersect P and so, shrinking Q within this distance, p and q are Hausdorff-separated. So suppose the first possibility occurs. Take P, P' to be balls of radius ε, $\varepsilon/2$ respectively in some normal coordinate neighbourhood and let γ be a geodesic to q intersecting P' arbitrarily close to q. Consider a point r on γ, distant less than $\varepsilon/4$ from q along the geodesic, and lying in P'. Either $r=q$, or, since the intersection of the point set γ with P is open in γ, there is a positively-directed segment of γ from r lying in P. This must terminate in P'', the ball of radius $3\varepsilon/4$, since its length is less than $\varepsilon/4$; and, since curves – in particular, geodesics – cannot bifurcate, it must have q as its endpoint in $\overline{P''} \subset P$. Thus $q \in P$, for all ε. Hence $q=p$. So $L(M)$ is Hausdorff.

3. We can now implement a well-known proof ([7] p. 278) of second-countability for Hausdorff space-times. $L(M)$, as a Hausdorff connected Riemannian manifold, is second-countable ([4], p. 271) and has a countable dense set. This set projects to one in M whose second-countability then follows.

Proof of Theorem 1. We shall construct a maximal increasing chain of space-times whose "union" is to be the required maximal space. The construction fails in the general, non-Hajicek case because there are then "too many" space-times: I shall show that the class $\mathscr{H}$ of Hajicek space-times can be realized as a set, and is not only a class as in the general case. To represent $\mathscr{H}$ in concrete terms[3] so as to be able to apply set theory rigorously, note that any $M \in \mathscr{H}$ can, by Lemma 1, be specified by giving (i) a countable atlas $\{(U_i, \phi_i) | i=1,2,\ldots\}$ where, for simplicity, we may take the ϕ_i's to be *onto* $\mathbb{R}^4$; (ii) the transition functions $\psi_{ij}=\phi_i\phi_j^{-1}: \mathbb{R}^4 \to \mathbb{R}^4$; (iii) the metric coefficients $g^{(i)}_{\mu\nu}$ in each U_i. Then call $\mathscr{I}$ the *set* of all such specifications (ii) and (iii): that is, a member of $\mathscr{I}$ is a space-time which is concretely given as a countable collection of maps ψ_{ij} and coefficients $g^{(i)}_{\mu\nu}$ satisfying the usual metric conditions and transformation properties.

Since any $M \in \mathscr{H}$ is isometric to a concrete realisation in $\mathscr{I}$, it is now sufficient to prove maximality in $\mathscr{I}$. The problem is that the only natural inclusion of the elements of $\mathscr{I}$ as defined above depends on the numbering of the maps ψ_{ij}, and is not purely geometrical: We therefore must put in the inclusion maps. (Geroch [10] avoided this by taking the collection of *all* framed Hausdorff space-times, with geometrical inclusions. But this begs the question of whether or not this collection is a set or a proper class.)

We circumvent the difficulty by defining a *nest* to be a collection $\{M_\alpha, \chi_{\alpha\beta} | \alpha, \beta \in I; \alpha<\beta\}$ where I is a well-ordered index set, $M_\alpha \in \mathscr{I}$ and $\chi_{\alpha\beta}: M_\alpha \to M_\beta$ are isometries satisfying $\chi_{\beta\gamma}\chi_{\alpha\beta}=\chi_{\alpha\gamma}$ $(\alpha<\beta<\gamma)$. Nests on $\mathscr{I}$ are clearly partially ordered by

[3] The basic difficulty stems from the fact that a space-time is usually defined in terms of its internal properties and not in terms of a specific construction within set theory. Consequently the class of *all* space-times with a given property contains a huge number of isometric realisations that differ only in their incidental characteristics: An equivalence class of isometric space-times is then too big to be a set, and one cannot talk about "the set of equivalence classes". Either one postulates that there exists a *set* of spacetimes, within which one works (which begs the question); or, as here, one refers to some concrete construction in terms of classes of numerical functions which can be shown to be sets

inclusions and so we can apply the Kuratowski Lemma ([11], p. 33) to deduce the existence of a maximal nest containing any $M \in \mathscr{I}$.

Now a maximal nest $\{M_\alpha, \chi_{\alpha\beta}\}$ allows one to define the inductive limit M^* ([12], p. 255; nests must be ordered *inversely* by inclusion to apply this definition verbatim). The natural maps $M_\alpha \to M^*$ clearly define a unique space-time structure on M^*, and it is immediate that M^* is indeed a required maximal space-time.

It is false that any hole-free space-time has a maximal hole-free extension: there may be "latent holes" that are revealed by extending. For example, the metric

$$ds^2 = \Omega^2(-dt^2 + dx^2 + dy^2 + dz^2)$$

on the part of R^4 where $t < 2r$ $(r^2 = x^2 + y^2 + z^2)$ is not hole-free for

$$\Omega = \begin{cases} 1 & (t < r) \\ \sec \pi(t/r - 1)/2 & (r \leqq t < 2r) \end{cases}$$

because the singularity at the origin arises with no prior warning. However, if we take only the part of R^4 where, in addition to $t < 2r$, we have $1/2(\theta + \pi)^2 < r < 1/2\theta^2$, $x = \cos\theta$, $y = \sin\theta$ with $-\infty < \theta < \infty$, then the resulting space-time is hole-free and has no hole-free maximal extension. There would seem to be no reason why this space-time should not be modified to make it a solution of the vacuum Einstein equations, so that nothing would be gained by modifying the definition of "hole-free" to make the domain of dependence a solution to the corresponding Cauchy problem.

The power of the Hajicek condition is shown by the following:

Theorem 2. *A strongly causal space-time is Hausdorff.* This is a slight strengthening of the result of [2], and so we provide a new proof.

Proof. Suppose $p, q \in M$ are not Hausdorff separated, i.e. any pair of neighbourhoods of p, q intersect. As in the proof of Lemma 1, for any neighbourhood P of p, there is at least one geodesic γ to q which intersects P infinitely often, and which therefore has an accumulation point $p' \in \bar{P}$. If $\tilde{\gamma}$ is a horizontal lift of γ to the bundle $L(M)$ of pseudo-orthonormal frames, then, since this bundle is Hausdorff, $\tilde{\gamma}$ has no accumulation point in $\pi^{-1}(p')$: i.e. there is a sequence $\{x_i\}$ of points on $\tilde{\gamma}$ such that $\pi(x_i) \to p'$ but $\{x_i\}$ has no limit point in $\pi^{-1}(p')$.

We can now obtain a contradiction to strong causality by showing the existence of a timelike curve γ' with properties similar to γ; this γ' is chosen so as to stay "near" γ, both as seen from p' and as seen from q. The viewpoint of p' is investigated by examining the behaviour of the frame-curve $\tilde{\gamma}$ as it goes repeatedly past $\pi^{-1}p'$.

In a coordinate neighbourhood of p' define a local cross-section σ of $L(M)$, so that $x_i = l_i \sigma \pi x_i$ for a sequence of Lorentz transformations l_i. Write $l_i = r_i b_i r'_i$, where $r_i, r'_i \in SO(3)$, b_i is a boost along the x-axis with velocity v_i and, by choice of a subsequence of the x_i, $r_i \to r$, $r'_i \to r'$ and $v_i \to \infty$. Let $\zeta \in R^4$ be the null vector $(1, 1, 0, 0)$ for which $\|\zeta b_i\| \to \infty$.

Let X_i be the tangent vector to γ at πx_i and write $X_i = \xi^\mu e_{i\mu} = \xi^\mu (l_i \sigma \pi x_i)_\mu$, where $(e_{i0}, e_{i1}, e_{i2}, e_{i3}) = x_i$ and μ is a tetrad-component index. Since γ traverses any neighbourhood of p' infinitely often in finite proper time we must have $\|\xi l_i\| \to \infty$, i.e. $\|(\xi r_i) b_i\| \to \infty$.

Now, either (i) $(\xi r)_0 = (\xi r)_1 = 0$, or else (ii) the geodesics from πx_i with initial tangent vector $\pm(\xi r_i^{-1})^\mu e_{i\mu}$ (for an appropriate choice of sign) intersect the null cone through q in a sequence of points which tend to p'. In case (ii) we may, without loss of generality, assume that the "$-$" sign holds and that the geodesics intersect the past null cone. By construction the σ-components of their tangent vectors are bounded. Hence we can find a sequence of points on these geodesics which form a timelike chain tending to q and lying in a neighbourhood of p': joining these gives the required timelike curve. On the other hand, in case (i) this sequence of geodesics allows one to construct a rectifiable space-like curve, which can then be treated in the same way as γ: it will automatically yield case (ii), and a timelike curve to q is again obtained. □

2. The Existence of Curvature Singularities

In the preceding section the proofs assumed that g was at least C^{3-}, so that geodesics could be defined in $L(M)$ in the usual way. In fact this is unnecessary, since rectifiable curves could easily have been used instead of geodesics, and only some reasonably well-behaved measure of distance on such curves was needed. Indeed, the results still hold if the differentiability is lowered to the condition used in [1], where the metric is Lipshitz and the Riemann tensor locally bounded and locally integrable. We can restate the result obtained there in terms of maximality as follows.

Theorem 3. *In a globally hyperbolic space-time which is maximal (with the differentiability just stated) and nowhere D-specialised, every singularity that is accessible on a timelike or null curve is a curvature or intermediate singularity.*

Proof. This is simply the theorem of [1] with the inclusion of null curves – an addition that is desirable in view of the prediction of incomplete null curves in globally hyperbolic space-times by Hawking in Theorem 1 of [5], § 8.2.

Suppose, then, that $\kappa:[0, 1)\to M$ is a null curve leading to a singularity p, with horizontal lift $\tilde{\kappa}$ in $L(M)$. We may suppose κ to be a geodesic, since otherwise it is a straightforward manipulation to deform it to a timelike curve. Define a one-parameter family of geodesics by $\lambda(s, t) = \exp(\tilde{\kappa}(t)(-s, 0, 0, 0))$. Then, unless there is a curvature singularity, the curve $\lambda(a(1-t), t)$ is defined and timelike for small enough $a>0$ and t sufficiently close to 1, and leads to p. The argument is very similar to that employed in Lemma 3 of [1]: if $\lambda(1-t_1, t_1)$ were not defined, one could construct a set of causal curves between $\lambda(a, 0)$ and $\lambda(0, t')$ for $t' > t_1$, having non-compact closure and so violating global hyperbolicity. On the other hand, if $\lambda(a(1-t), t)$ failed to be timelike for t arbitrarily close to 1 then Proposition 1 of [1] could be used to construct a curve in the image of λ which led to p, but on which the components of the Riemann tensor became unbounded.

Having constructed a timelike curve, the result follows from [1]. □

If one has a situation of inhomogeneous gravitational collapse, where singularities may, in a sense, form earlier in some places than in others, then global hyperbolicity is very unlikely. Without this condition locally extensible (non-curvature) singularities may be present, as exemplified by the covering space of Minkowski space with a 2-plane removed: if the plane is space-like there is a "hole" (see the Scholium to Definition 3) while if it is timelike there is a primordial singularity. Theorem 4 below shows that these are the only possibilities.

Let M^* denote the set of all submanifolds of M of the form $I^-(\gamma)$ where γ is a timelike curve having a generalised affine parameter [5] that is bounded to the future. M^* is a subspace of the Geroch-Kronheimer-Penrose space $\hat{M}$ [8] and so inherits a natural causal structure with a past-relationship $J^-: A \in J^-(B) \Leftrightarrow A \subset B$. Write this as $A \leqq B$, and define $A < B \Leftrightarrow A \leqq B$ but $A \neq B$.

Note that any point q in M can be identified with the set $q_0 = I^-(q) \in M^*$; also any point p in the b-boundary $\dot{M}$ which is accessible along a future timelike curve γ can be mapped onto the point $p_0 = I^-(\gamma)$. Thus we have a map $x \to x_0$ from a subset of $\bar{M} = M \cup \dot{M}$ onto M^* which is injective on M, so that we can identify M with its image M_0 in M^*.

Definition 4. *An inextensible causal curve in M^* is a non-empty set $S \subset M^*$ such that*
(i) *for any $p, q \in S$ either $p = q$ or $p < q$ or $q < p$;*
(ii) *for any $p, q \in S$ with $p < q$ there is an $r \in S$ such that $p < r < q$;*
(iii) *S is maximal with respect to* (i) *and* (ii).

Lemma 2. *If M is a strongly causal space-time and S is a causal curve in M^*, then S with the order topology is homeomorphic to an interval of* $\mathbb{R}$.

Proof. For simplicity let us denote by S' the set S without its greatest and least members, if it has any. M has a countable dense set D; the subset $D' = \{x \in D | p \in S', x \in p\}$ is mapped into S' by setting $\phi(x) = \cup \{p \in S | x \notin p\} \subset M$. Clearly $T = \phi(D')$ is a countable dense subset of S', and hence ([9], p. 51) it is order-isomorphic to the rationals in (0, 1) by a map $\psi: T \to \mathbb{Q}$. It remains only to extend ψ to an order-isomorphism with (0, 1) by defining $\psi(x) = \sup \{\psi(t) | t \in T, t \leqq x\}$. Then ψ is certainly order-preserving and bijective; and it is surjective since for $r \in (0, 1)$ the set $\cup \{t \in T | \psi(t) \leqq r\}$ is easily seen to be an IP, and so it follows from (iii) that it is in S'. Finally, the greatest and least elements of S, if any, can be added, corresponding to 1 and 0, respectively. □

Definition 5. *A primordial singularity is a point $p \in \dot{M}$ such that*
(i) *p is the future endpoint of a timelike or null curve γ;*
(ii) *there is an inextensible causal curve S with $p_0 = I^-(\gamma) \in S$;*
(iii) $\{q \in S | q \leqq p_0\} \subset M^* \setminus M_0$.

For this definition to correspond to the intuitive picture M must be strongly causal.

Theorem 4. *If M is a strongly causal hole-free space-time that is nowhere D-specialised and p is a singularity in $\dot{M}$ accessible on a future-directed causal curve γ, then either p is a primordial singularity, or $\bar{M}$ contains a curvature or intermediate singularity.*

Proof. Suppose that $\bar{M}$ contains no curvature singularities. Let S_1 be a maximal chain in $M^* \setminus M$, simply ordered by $<$, containing $p_0 = I^-(\gamma)$. We show that S_1 can be extended to an inextensible causal curve.

1. Let q', p' be two points in S_1 with $q' < p'$, $p' = I^-(\gamma')$ where γ' is an inextensible future-incomplete curve. The sets $C_x = I^-(\{y | y \in I^-(x) \cap \gamma'\})$ for $x \in \gamma'$ form a nested sequence with q' properly contained in $\bigcup_{x \in \gamma'} C_x$. So for some x_0, q' is properly contained in C_{x_0}. Let γ_0 be the part of γ' to the future of x_0.

2. Suppose that for some $x \in \gamma_0$, $V = I^-(\gamma') \cap I^+(x)$ is globally hyperbolic. Then from the analysis of [1] we know that V is covered by the future timelike geodesics from x (provided that x is chosen near enough to p), and that V has an extension in some other space-time M' in which these geodesics continue without intersecting. Thus they define by their endpoints a natural map θ from $\bar{V}'$, the closure of V in M', onto $\bar{V}$, the closure of V in $\bar{M}$. Either (i) some of these geodesics have end-points in $\bar{M}$ on $\bar{V}$, or else (ii) by the argument of Lemma 5 of [1] θ is $1-1$ and onto and maps into M except for the point p. But this case (ii) implies that M is not hole-free, if we consider a partial cauchy surface which makes a compact intersection with $\bar{V}$.

3. Suppose, on the other hand, that V is not globally hyperbolic, for any x. Then, arbitrarily close to p', there will be pairs of points u, v with $u \in I^-(v) \cap \gamma'$; $v \in I^-(\gamma')$ such that the set $I^+(u) \cap I^-(v)$ is not compact. We can find a non-convergent sequence $\{x_i\}$ in this set and, if p is not a curvature singularity, Proposition 1 of [1] allows us to conclude that, for u near enough to p, there are geodesics joining u to x_i whose initial directions converge to an incomplete geodesic.

4. Thus by either 2 or 3 we find an incomplete geodesic in $I^+(x_0) \cap I^-(\gamma')$ which corresponds to some $r \in M^* \setminus M$ with $q' < r < p'$. Thus since S_1 is maximal either $r \in S_1$, or there is an $r' \in S_1$ such that $r \nless r'$ and $r' \nless r$. But then $q' < r' < p'$, so in any case there is a point between q' and p'. And, by the same argument, for any $p' \in S_1$ there is an $r' \in S_1$ with $r' < p'$.

5. Let S be a maximal extension of S_1 as a causal curve in M^*. Then S_1 is closed in S, since any $p \in S \setminus S_1$ is a *PIP* and so must have a neighbourhood of *PIP*'s [8]. Moreover by 4 above S_1 is order-dense and has no least member. Thus S_1 has the form $S_1 = \{t \in S | t \leqq u\}$ for some $u \in S$, and the result follows. □

Acknowledgements. In addition to the contributions already referred to, I have been helped by discussions with Dr. D. Robinson, Dr. B. Schmidt and Professor A. G. Walker, to whom I am most grateful.

References

1. Clarke, C. J. S.: Commun. math. Phys. **41**, 65—78 (1975)
2. Hajicek, P.: Commun. math. Phys. **21**, 75—84 (1971)
3. Hajicek, P.: J. math. Phys. **12**, 157—160 (1971)
4. Kobayaski, S., Nomizu, K.: Foundations of differential geometry, Vol. I. New York: Interscience 1969
5. Hawking, S. W., Ellis, G. F. R.: The large scale structure of space-time. Cambridge: University Press 1973
6. Steinmüller, B., King, A. R., Lasota, J. P.: Phys. Let. A **51**, 191 (1975)
7. Schmidt, B. G.: J. general relativity and gravitation **1**, 209—280 (1971)
8. Geroch, R. P., Kronheimer, E. H., Penrose, R.: Proc. Roy. Soc. Lond. A **327**, 545—567 (1972)
9. Hocking, J. G., Young, G. S.: Topology. Reading: Addison-Wesley 1961
10. Geroch, R. P.: "Singularities" in relativity, ed. Carmeli, M., Fickler, S. T., Witten, L.: New York, London: Plenum Press 1970
11. Kelley, J. L.: General topology. Princeton: Van Nostrand 1955
12. Kowalsky, H.-J.: Topologische Räume. Basel, Stuttgart: Birkhäuser 1961

Communicated by J. Ehlers

Received June 19, 1975; in revised form February 9, 1976