On the Relationship between Spacetime Singularities, Holes, and Extensions

John Manchak
manchak@uw.edu

Abstract
Here, we clarify the relationship between three spacetime conditions of interest: geodesic completeness, hole-freeness, and inextendibility.

1 Introduction
In what follows, we consider three spacetime conditions of interest: geodesic completeness, hole-freeness, and inextendibility. How are these three conditions related? Here, we review what is known and contribute a few (minor) results of our own. We take no position as to which of these conditions are satisfied by any or all physically reasonable spacetimes. We seek only to shed light on the connections between them.

2 Background Structure
We begin with a few preliminaries concerning the relevant background formalism of general relativity. An \( n \)-dimensional, relativistic spacetime (for \( n \geq 2 \)) is a pair of mathematical objects \((M, g_{ab})\). \( M \) is a connected \( n \)-dimensional manifold (without boundary) that is smooth (infinitely differentiable). Here, \( g_{ab} \) is a smooth, non-degenerate, pseudo-Riemannian metric of Lorentz signature \((+, -, ..., -)\) defined on \( M \).

Note that \( M \) is assumed to be Hausdorff: for any distinct \( p, q \in M \), one can find disjoint open sets \( O_p \) and \( O_q \) containing \( p \) and \( q \) respectively. We say two spacetimes \((M, g_{ab})\) and \((M', g'_{ab})\) are isometric if there is a diffeomorphism \( \varphi : M \to M' \) such that \( \varphi_*(g_{ab}) = g'_{ab} \).

\( ^1 \)I am grateful to Erik Curiel, Bob Geroch, David Malament, and Jim Weatherall for helpful comments on previous drafts.

\( ^1 \)In the literature, geodesic completeness is usually taken to be violated by some physically reasonable spacetimes while hole-freeness and inextendibility are usually taken to be satisfied by all such spacetimes. See, for example, Clarke (1976, 1993) and Earman (1989, 1995).

\( ^2 \)The reader is encouraged to consult Hawking and Ellis (1973) and Wald (1984) for details. An outstanding (and less technical) survey of the global structure of spacetime is given by Geroch and Horowitz (1979).
For each point \( p \in M \), the metric assigns a cone structure to the tangent space \( M_p \). Any tangent vector \( \xi^a \) in \( M_p \) will be timelike if \( g_{ab}\xi^a\xi^b > 0 \), null if \( g_{ab}\xi^a\xi^b = 0 \), or spacelike if \( g_{ab}\xi^a\xi^b < 0 \). Null vectors create the “cone structure; timelike vectors are inside the cone while spacelike vectors are outside. A time orientable spacetime is one that has a continuous timelike vector field on \( M \). A time orientable spacetime allows one to distinguish between the future and past lobes of the light cone. In what follows, it is assumed that spacetimes are time orientable.

For some open (connected) interval \( I \subseteq \mathbb{R} \), a smooth curve \( \gamma : I \to M \) is timelike if the tangent vector \( \xi^a \) at each point in \( \gamma[I] \) is timelike. Similarly, a curve is null (respectively, spacelike) if its tangent vector at each point is null (respectively, spacelike). A curve is causal if its tangent vector at each point is either null or timelike. A causal curve is future-directed if its tangent vector at each point falls in or on the future lobe of the light cone.

We say a curve \( \gamma : I \to M \) is not maximal if there is another curve \( \gamma' : I' \to M \) such that \( I \) is a proper subset of \( I' \) and \( \gamma(s) = \gamma'(s) \) for all \( s \in I \). A curve \( \gamma : I \to M \) in a spacetime \((M, g_{ab})\) a geodesic if \( \xi^a \nabla_a \xi^b = 0 \) where \( \xi^a \) is the tangent vector and \( \nabla_a \) is the unique derivative operator compatible with \( g_{ab} \).

For any two points \( p, q \in M \), we write \( p << q \) if there exists a future-directed timelike curve from \( p \) to \( q \). We write \( p < q \) if there exists a future-directed causal curve from \( p \) to \( q \). These relations allow us to define the timelike and causal pasts and futures of a point \( p \): \( I^-(p) = \{ q : q << p \} \), \( I^+(p) = \{ q : p < q \} \), \( J^-(p) = \{ q : q < p \} \), and \( J^+(p) = \{ q : p < q \} \). Naturally, for any set \( S \subseteq M \), define \( J^+[S] \) to be the set \( \cup\{ J^+(x) : x \in S \} \) and so on. A set \( S \subseteq M \) is achronal if \( S \cap I^-[S] = \emptyset \). We say a spacetime \((M, g_{ab})\) is stably causal if there is a smooth function \( f : M \to \mathbb{R} \) such that, for any distinct points \( p, q, r \in M \), if \( p \in J^+(q) \), then \( f(p) > f(q) \).

A point \( p \in M \) is a future endpoint of a future-directed causal curve \( \gamma : I \to M \) if, for every neighborhood \( O \) of \( p \), there exists a point \( t_0 \in I \) such that \( \gamma(t) \in O \) for all \( t > t_0 \). A past endpoint is defined similarly. A causal curve is future inextendible (respectively, past inextendible) if it has no future (respectively, past) endpoint.

For any set \( S \subseteq M \), we define the past domain of dependence of \( S \), written \( D^-(S) \), to be the set of points \( p \in M \) such that every causal curve with past endpoint \( p \) and no future endpoint intersects \( S \). The future domain of dependence of \( S \), written \( D^+(S) \), is defined analogously. The entire domain of dependence of \( S \), written \( D(S) \), is just the set \( D^-(S) \cup D^+(S) \). The edge of an achronal set \( S \subseteq M \) is the collection of points \( p \in S \) such that every open neighborhood \( O \) of \( p \) contains a point \( q \in I^+(p) \), a point \( r \in I^-(p) \), and a timelike curve from \( r \) to \( q \) which does not intersect \( S \). A set \( S \subseteq M \) is a slice if it is closed, achronal, and without edge. A spacetime \((M, g_{ab})\) which contains a slice \( S \) such that \( D(S) = M \) is said to be globally hyperbolic.
3 Singularities and Holes

There are a number of ways to define a spacetime singularity but “none, with the exception of geodesic incompleteness, seems to have found any significant applications” (Geroch and Horowitz, 1979). Here, we take this position for granted and refer the reader to Curiel (1999) for thorough and convincing arguments in its favor. We now make the condition precise.

**Definition 3.1** A spacetime \((M, g_{ab})\) is geodesically complete (GC) if every maximal geodesic \(\gamma: I \rightarrow M\) is such that \(I = \mathbb{R}\). A spacetime is geodesically incomplete if it is not geodesically complete.

Physically, a timelike incomplete geodesic represents a freely falling observer who does not record all possible watch readings. If an incomplete geodesic is timelike or null, there is a useful distinction one can introduce. We say that a future-directed timelike or null geodesic \(\gamma: I \rightarrow M\) without future endpoint is future incomplete if there is an \(r \in \mathbb{R}\) such that \(s < r\) for all \(s \in I\). A past incomplete timelike or null geodesic is defined analogously.

The singularity theorems of Hawking and Penrose (1970) show that a large number of seemingly physically reasonable spacetimes fail to be geodesically complete. Thus, the condition is somewhat strong.

Of course, there are also a large number of geodesically incomplete spacetimes which seem to have “artificial” singularities. Indeed one can show that any spacetime with one point removed from the manifold is geodesically incomplete. In such a spacetime, a type of indeterminism is also present; the domain of dependence of some spacelike surface is not “as large as it could have been.” We turn our attention now to this type of indeterminism – due to spacetime “holes” – and its connections with geodesic incompleteness.

Initially, one defined (Geroch 1977) a spacetime \((M, g_{ab})\) to be hole-free if, for every spacelike surface \(S \subset M\) and every isometric embedding \(\varphi: D(S) \rightarrow M’\) into some other spacetime \((M’, g_{ab})\), we have \(\varphi(D(S)) = D(\varphi(S))\). The definition seemed to be satisfactory. Indeed, one can show that any spacetime with one point removed from the manifold is not hole-free. But surprisingly, it turns out the definition is too strong; Minkowski spacetime fails to be hole-free under this formulation (Krasnikov 2009). But one can make minor modifications to avoid this consequence (Manchak 2009).

Let \((K, g_{ab})\) be a globally hyperbolic spacetime. Let \(\varphi: K \rightarrow K’\) be an isometric embedding into a spacetime \((K’, g’_{ab})\). We say \((K’, g’_{ab})\) is an effective extension of \((K, g_{ab})\) if, for some Cauchy surface \(S\) in \((K, g_{ab})\), \(\varphi[K] \subseteq int(D(\varphi[S]))\) and \(\varphi[S]\) is achronal. Hole-freeness can then be defined as follows.

**Definition 3.3** A spacetime \((M, g_{ab})\) is hole-free (HF) if, for every set \(K \subseteq M\) such that \((K, g_{ab}|K)\) is a globally hyperbolic spacetime with Cauchy surface \(S\), if \((K’, g_{ab}|K’\)) is not an effective extension of \((K, g_{ab}|K)\) where \(K’ = int(D(S))\), then there is no effective extension of \((K, g_{ab}|K)\).
What is the relationship between hole-freeness and geodesic completeness? One can easily show that former does not imply the latter.

**Example 3.1** Let $(\mathbb{R}^2, \eta_{ab})$ be Minkowski spacetime and let $p$ be any point in $\mathbb{R}^2$ and let $q$ be any point in $I^-(p)$. Let $M$ be the manifold $I^-(p) \cap I^+(q)$. Clearly, the spacetime $(M, \eta_{ab}|_M)$ satisfies (HF) but not (GC).

One wonders if the implication relation holds in the other direction. And it has been conjectured by Geroch (private communication) that there even exists some intermediate completeness condition which is weaker than geodesic completeness but stronger than hole-freeness (see Figure 1).

\[(GC) \quad ? \quad (HF)\]

Figure 1: *Is there an intermediate condition which is implied by (GC) and implies (HF)?*

Why might such an intermediate condition be of interest? As noted above, geodesic completeness is a somewhat strong condition in the sense that not all seemingly physically reasonable spacetimes satisfy it. On the other hand, hole-freeness is a somewhat weak condition in the sense that some spacetimes (such as Example 3.1) which satisfy it can be constructed by removing points from otherwise geodesically complete spacetimes. An intermediate condition may be strong enough to rule out these seemingly artificial singularities but weak enough to allow the more physically reasonable, geodesically incomplete spacetimes guaranteed by the singularity theorems.

### 4 Singularities and Extensions

One way to rule out spacetimes which are constructed by removing points from the manifold is to require that spacetime be “as large as it could have been.” In other words, one can require that spacetime be inextendible. We have the following definition.

**Definition 4.1** A spacetime $(M, g_{ab})$ is *extendible* if there exists a spacetime $(M', g'_{ab})$ and an isometric embedding $\varphi : M \to M'$ such that $\varphi(M) \subset M'$. Here, the spacetime $(M', g'_{ab})$ is an *extension* of $(M, g_{ab})$. A spacetime is *inextendible* (I) if has no extension.

One can show that every extendible spacetime has a (not necessarily unique) inextendible extension. What is the relationship between geodesic completeness and inextendibility? One can show that the former implies the latter (Clarke 1993). And a simple example shows that the implication does not run in the other direction.
Example 4.1 Let \((\mathbb{R}^2, \eta_{ab})\) be Minkowski spacetime and let \(p\) be any point in \(\mathbb{R}^2\). Let \(M\) be the manifold \(\mathbb{R}^2 - \{p\}\) and let \((M', g_{ab})\) be the universal covering spacetime of \((M, \eta_{ab}|_M)\). Clearly, the spacetime \((M', g_{ab})\) satisfies (I) but not (GC).

The example above shows that inextendibility is a somewhat weak condition. Indeed, some inextendible spacetimes which are extraordinarily well-behaved (e.g. have flat metrics and manifolds diffeomorphic to \(\mathbb{R}^n\)) may nonetheless be geodesically incomplete. As before, one wonders if there an intermediate condition which is strong enough to rule out these seemingly artificial singularities but weak enough to allow the more physically reasonable, geodesically incomplete spacetimes guaranteed by the singularity theorems (see Figure 2).

\[
(GC) \implies ? \implies (I)
\]

Figure 2: Is there an intermediate condition which is implied by (GC) and implies (I)?

One such intermediate condition was thought to have been given by Hawking and Ellis (1973). A spacetime \((M, g_{ab})\) is said to be \textit{locally extendible} if there is an open set \(O \subset M\) with non-compact closure and an isometric embedding \(\varphi : O \to M'\) into some other spacetime \((M', g'_{ab})\) such that the closure of \(\varphi(O)\) is compact. A spacetime is \textit{locally inextendible} if it is not locally extendible. Clearly, local inextendibility implies inextendibility. And the problematic Example 4.1 given above is counted as locally extendible. But it turns out that the condition is not implied by geodesic completeness. Indeed, the condition is much too strong in the sense that Minkowski spacetime can be shown to be locally extendible (Beem 1980).

5 Holes and Extensions

Hole-freeness and inextendibility are independent conditions. Example 3.1 shows that hole-freeness does not imply inextendibility. And Example 4.1 shows that inextendibility does not imply hole-freeness. So, the conditions serve to rule two different types of seemingly artificial singularities. And therefore one routinely finds that \textit{both} hole-freeness \textit{and} inextendibility are assumed to be satisfied by all physically reasonable spacetimes. (See Clarke (1976, 1993) and Earman (1989, 1995) for examples.)

In the two previous sections we have wondered about the existence of two intermediate conditions: one between geodesic completeness and hole-freeness and another between geodesic completeness and inextendibility. Might there be a \textit{single} intermediate condition which is implied by geodesic completeness and implies both hole-freeness and inextendibility? (See Figure 3.) Such an intermediate condition may be strong enough to rule out, in one fell swoop,
both types of seemingly artificial singularities at issue (and possibly other types as well) but weak enough to allow the more physically reasonable, geodesically incomplete spacetimes guaranteed by the singularity theorems.

\[ \text{(HF)} \]
\[ \text{(GC)} \]  \[?\]  \[\text{(I)}\]

Figure 3: Is there an intermediate condition which is implied by (GC) and implies both (HF) and (I)?

6 An Intermediate Condition

Here, we show the existence of the intermediate condition mentioned in the previous section. We introduce the following.

**Definition 6.1** A spacetime \((M, g_{ab})\) is **effectively complete** (EC) if, for every future or past incomplete timelike geodesic \(\gamma : I \to M\), and every open set \(O\) containing \(\gamma\), there is no isometric embedding \(\varphi : O \to M'\) into some other spacetime \((M', g'_{ab})\) such that \(\varphi \circ \gamma\) has future and past endpoints.

The condition is a variation of one found in Clarke (1982) and Earman (1989). But these authors use the mathematically cumbersome and physically dubious concept of “b-incomplete” curves instead of incomplete timelike geodesics.\(^3\) The physical significance of effective completeness is the following: If a spacetime fails to be effectively complete, then there is a freely falling observer who never records some particular watch reading but who “could have” in the sense that nothing in her vicinity precludes it.

We note here that a variant of effective completeness can be formulated using arbitrary (instead of timelike) geodesics. The two conditions are not equivalent.\(^4\) But the stronger variant seems to be less significant physically and moreover is simply not needed to show that effective completeness implies hole-freeness and inextendibility (see below). And it follows immediately from our formulation.

\(^3\)See Schmidt (1971) and Ellis and Schmidt (1977) for details concerning b-incomplete curves. See Geroch, Can-bin, and Wald (1982) and Curiel (1999) for details concerning their physical significance – or lack thereof.

\(^4\)A counterexample can be constructed by considering Figure 8.3 in (Earman, 1989) and “turning it on its side” so that \(\gamma\) is spacelike and no timelike geodesic can reach the “apex” point.
that geodesic completeness implies effective completeness.

**Proposition 6.1** (GC) $\Rightarrow$ (EC).

A simple example shows that the two conditions are not equivalent.

**Example 6.1** Let $(\mathbb{R}^2, \eta_{ab})$ be Minkowski spacetime and let $p$ be any point in $\mathbb{R}^2$. Let $M$ be the manifold $\mathbb{R}^2 - \{p\}$. Let $\Omega : M \to \mathbb{R}$ be a smooth, strictly positive function which approaches zero as the missing point $p$ is approached. Let $g_{ab}$ be the conformally flat metric $\Omega^2 \eta_{ab}$. Clearly, the spacetime $(M, g_{ab})$ satisfies (EC) but not (GC).

Examples 3.1 and 4.1 above show that hole-freeness and inextendibility each do not imply effective completeness. (In fact, the various inequivalent but isometric spacetimes of Misner (1967) allow one to show that even the conjunction of hole-freeness and inextendibility does not imply effective completeness.)

It follows as a direct corollary to Proposition 1.3.1 in Clarke (1993) that a violation of inextendibility implies a violation of effective completeness. So, we have the following.

**Proposition 6.2** (EC) $\Rightarrow$ (I).

![Diagram](image-url)

**Figure 4:** The relationships between (EC), (GC), (HF), and (I).

Finally, we show here that effective completeness implies hole-freeness. (All of the implication relationships between the four conditions can be summarized in Figure 4.)

**Proposition 6.3** (EC) $\Rightarrow$ (HF).

**Proof.** Let $(M, g_{ab})$ be a spacetime which does not satisfy (HF). Then for some $K \subseteq M$ such that $(K, g_{ab}|_K)$ is a globally hyperbolic spacetime with Cauchy surface $S$, we know that (i) $\text{int}(D(S)) = K$ and (ii) there is a spacetime $(M', g_{ab}')$ and an isometric embedding $\varphi : K \to M'$ such that $\varphi[S]$ is achronal and $\varphi[K] \subseteq \text{int}(D(\varphi[S]))$. Without loss of generality, we may take $M' = \text{int}(D(\varphi[S]))$. So,
\((M', g_{ab})\) is globally hyperbolic with Cauchy surface \(S' = \varphi[S]\).

Let \(p'\) be a point in \(K'\) where \(K' = \varphi[K]\). Now, assume \(p' \in D^+(S')\). (A similar proof can be constructed if \(p' \in D^-(S')\).) Clearly, \(p' \in I^+[S']\). Let \(\gamma : I \to K'\) be a timelike geodesic with future endpoint \(p'\) and past endpoint in \(S'\). Either \(\varphi^{-1} \circ \gamma\) has a future endpoint or not.

First, assume \(\varphi^{-1} \circ \gamma\) does not have a future endpoint. It follows that \(\varphi^{-1} \circ \gamma\) is a timelike future incomplete geodesic. But by construction, \(\varphi \circ \varphi^{-1} \circ \gamma = \gamma\) has future and past endpoints. So in this case, \((M, g_{ab})\) does not satisfy (EC).

Second, assume that \(\varphi^{-1} \circ \gamma\) does have a future endpoint \(p\) in \(K\). Consider the open set \(U' = I^-(p') \cap I^+[S']\). Clearly, \(U' \subset K'\). Let \(U = \varphi^{-1}[U']\). Note that \(U \subset D^+(S)\). We also know that \(\overline{U'}\) is compact (Theorem 8.3.12, Wald, 1984). Now, either \(U\) is compact or not. Our next step is to show that the former case is impossible.

Assume that \(U\) is compact. Now, let \(\lambda : I \to M\) be any past inextendible causal curve with future endpoint \(p\). Either \(\lambda\) leaves the set \(U\) or not. Assume the latter case first. Since \(U \subset D^+(S)\), we can show that, for every \(q \in U\), either \(q \in \tilde{S}\) or every past inextendible timelike curve from \(q\) intersects \(S\) (see Proposition 8.3.2, Wald, 1984). But one can verify that, because \(\lambda\) never leaves \(U\), \(\lambda\) never intersects \(\tilde{S}\). So, for all \(s \in I\), there is a timelike curve \(\lambda_s : I' \to U\) with future endpoint \(\lambda(s)\) and past endpoint in \(S\). We know that, for each \(s \in I\), the timelike curve \(\varphi \circ \lambda_s\) has a future endpoint in \(\overline{U'}\) (Lemma 8.2.1, Wald, 1984). Let \(\lambda' : I \to M'\) be a smooth causal curve defined as follows: for each \(s \in I\), let \(\lambda'(s)\) be the future endpoint of \(\varphi \circ \lambda_s\). Since \(\lambda'\) is confined to \(\overline{U'}\), it has a past endpoint \(q' \in \overline{U'}\) (Lemma 8.2.1, Wald, 1984). Let \(\{s_i\}\) be a sequence of points in \(I\) such that the sequence \(\{\lambda'(s_i)\}\) has an accumulation point \(q'\). Now consider the sequence \(\{\lambda(s_i)\}\) in \(\overline{U}\). Since \(\overline{U}\) is compact by assumption, \(\{\lambda(s_i)\}\) has an accumulation point \(q \in \overline{U}\). But this implies that \(\lambda\) can be extended in the past: a contradiction.

Now, assume that \(\lambda\) leaves \(\overline{U}\) at point \(q \in \overline{U}\). For some \(I' \subset I\), let \(\overline{\lambda} : I' \to \overline{U}\) be the (unique) past-directed causal curve with future endpoint \(p\) and past endpoint \(q\) such that \(\lambda_{|I'} = \overline{\lambda}\). There are two subcases to consider: \(q\) is in \(\overline{S}\) or not. Assume the latter. Let \(\{q_i\}\) be a sequence in \(U\) which accumulates at \(q\). The compactness of \(\overline{U}\) ensures that \(\{\varphi(q_i)\}\) has an accumulation point \(q' \in \overline{U'}\). Clearly, \(q' \notin S'\). So, every past directed causal curve from \(q'\) must remain in \(\overline{U'}\) for some interval. But this implies that every past directed causal curve from \(q'\) must remain in \(\overline{U}\) for some interval: an impossibility since \(\lambda\) leaves \(\overline{U}\) at \(q\).

Now assume that \(q \in \overline{S}\). It is not hard to verify that \(q\) cannot be in \(\tilde{S}\). (If it were, one could find a sequence of points \(\{q_i\}\) in \(S \cap U\) which accumulate at \(q\). But the sequence \(\{\varphi(q_i)\}\) accumulates at a point \(q\) in \(S'\). Therefore, \(\varphi^{-1}(q') = q\) is in \(S\): a contradiction since \(S\) is open.) Thus, \(\lambda\) meets \(S\). And since \(\lambda\) was chosen arbitrarily, we have \(p \in D^+(S)\). Now, let \(\{p_i\}\) be a sequence of points in \(M - \overline{D(S)}\) with limit point \(p\). Let \(\{\lambda_i\}\) be a sequence of past inextendible causal curves with corresponding future endpoints \(\{p_i\}\) which also fail to meet \(S\). We know that there is a past inextendible causal curve through \(p\) which is a limit curve of the sequence (Lemma 8.1.6, Wald, 1984). Since \(p \in D^+(S)\), this
limit curve must intersect $S$. But $S$ is open and therefore some of the $\{\lambda_i\}$ must meet $S$ as well: a contradiction. So, $\overline{U}$ is not compact.

Finally, let $\{r_i\}$ be a sequence of points in $U$ without accumulation point in $\overline{U}$. Since $\overline{U}$ is compact, the sequence $\{\varphi(r_i)\}$ accumulates at some point $r' \in \overline{U}$. One can verify that $r$ must be in $I^+[S']$. Let $\zeta : I \rightarrow K'$ be a timelike geodesic with future endpoint $r'$ and past endpoint in $S'$. Clearly, $\varphi^{-1} \circ \zeta$ has no future endpoint. It follows that $\varphi^{-1} \circ \zeta$ is a timelike future incomplete geodesic. But by construction, $\varphi \circ \varphi^{-1} \circ \zeta = \zeta$ has future and past endpoints. So in this case as well, $(M, g_{ab})$ does not satisfy (SHF). □

7 Conclusion

One final note on how the causal structure of spacetime is connected with the preceding. Of course, under the assumption of any causal condition, the implication relations outlined in the previous section remain intact. And all the counterexamples given satisfy stable causality (and therefore any causal condition it implies). What about the stronger causal condition of global hyperbolicity?

One can easily find globally hyperbolic examples showing that effective completeness does not imply geodesic incompleteness, intextendibility does not imply effective completeness, hole-freeness does not imply effective completeness, and hole-freeness does not imply inextendibility. (All of the examples can be constructed using the manifold in Example 3.1 and adding various conformally flat metrics.)

\[
(GC) \implies (EC) \implies (I) \implies (HF)
\]

Figure 5: The relationships between $(GC)$, $(EC)$, $(I)$, and $(HF)$ under the assumption of global hyperbolicity.

But it turns out that under the assumption of global hyperbolicity, we find that inextendibility implies hole-freeness (Manchak 2009). It has been conjectured (Penrose 1979) that all physically reasonable spacetimes are globally hyperbolic. Thus, if the conjecture is true, we seem to have a useful hierarchy of conditions (see Figure 5).

References


