1
Introduction

A singularity, in the sense on which our later definitions will be based, is an exceptional or peculiar point in a space. For example, in global analysis a singularity in a smooth map from one manifold to another is (the image of) the places where the rank of the derivative is not a maximum – as contrasted with the "normal" situation of maximum rank. Or, to take a case closer to relativity, a singularity in a real-valued function which is everywhere else defined and continuous is an "exceptional" point at which the function cannot be given any value that makes it continuous throughout a neighbourhood of that point. As an example of this case, the electrostatic field of that hypothetical nineteenth century entity the "point charge"

$$\mathbf{E} = \text{const.} \times \mathbf{r}/|\mathbf{r}|^3$$

is singular, or "has a singularity", at $r = 0$.

In general relativity the term 'singularity' has undergone a succession of changes of meaning, which I shall sketch in historical sequence, introducing some of the basic definitions as we proceed.

1.1 The classical period

(I use the term 'classical' in its modern sense of 'before the author started doing research').

The first meaning follows the pattern of the case just described, of singularities in real-valued functions. If a metric is given in terms of components on a part of $\mathbb{R}^4$, then the singularities are the points of $\mathbb{R}^4$ at which one of the g_{ij} or g^{ij} cannot be continuously defined. One of the earliest known solutions to the vacuum Einstein equations contained singularities in this sense: the Schwarzschild metric in coordinates x, y, z, t on $\mathbb{R}^4$, having the

form (after rewriting so as to display these Cartesian coordinates)

$$-\left(1-\frac{2M}{r}\right)dt^2+dx^2+dy^2+dz^2+\frac{2M}{r(r-2M)}(xdx+ydy+zdz)^2 \tag{1}$$

(where $r=+\sqrt{x^2+y^2+z^2}$.)

This is singular on the 3-surface $r=2M$ and on the 2-surface $r=0$, in the sense that some components of g_{ij} cannot be defined there so as to give continuous functions.

Subsequently, singularities in this sense were found in the de Sitter metric

$$-dr^2-R^2\sin^2(r/R)\left[d\theta^2+\sin^2\theta d\phi^2\right]+\cos^2(r/R)\,c^2dt^2 \tag{2}$$

as it was given in 1917, and in the Friedmann metric (discovered in 1922, but given here in its modern form):

$$-dt^2+a(t)^2\left[dr^2+f(r)\left(d\theta^2+\sin^2\theta d\phi^2\right)\right] \tag{3}$$

where $f(r)=\sin^2 r$, r^2 or $\sinh^2 r$ and $a(t)\to 0$ as $t\to 0$. For both of these $\det(g_{ij})$ tends to zero on a 3-surface: for de Sitter, the surface $r=\pi R/2$, for Friedmann the surface $t=0$. So on these surfaces some component of g^{ij} cannot be defined, giving a singularity in the sense at present under discussion.

From the start, however, there was dissatisfaction felt with the this notion of singularity, because it clearly depended on a particular choice of coordinates. Consequently, the assertion that a metric was singular, in this sense, might not correspond to anything physically measurable in the spacetime represented by the metric in question.

This was stressed by Einstein (1918) in his discussion of the de Sitter metric, where he pointed out that two conditions had to be fulfilled for a singularity to be real. First, the singularity had to be accessible, in the sense that there was a timelike curve leading from a regular point to the singularity and having a finite proper-time. Secondly, it must not be possible to find a new coordinate system with respect to which the metric becomes regular at the singularity and capable of being continued past it. These two conditions will be expanded in the next two sections and will form the basis for our definition of a singular space-time. The first condition was actually ill-expressed by Einstein when he required merely that the singularity be reachable in a finite proper time. For the finite

time condition is actually no restriction at all: if it is possible to draw any timelike curve to the singularity, then, by wiggling the curve to make its speed close to the speed of light, it is possible to draw a curve of finite proper time. In the next section we shall see how to modify this condition so as to single out singularities at a "finite distance".

1.2 The idea of incompleteness

As a simple example consider the metric

$$-\left(1/t^2\right) dt^2 + dx^2 + dy^2 + dz^2 \tag{4}$$

which is singular (g_{ij} being undefined) on the plane $t = 0$ (in the $\mathbb{R}^4$ covered by the coordinates t, x, y, z). If an observer starting in the region $t > 0$ tries to reach the surface $t = 0$ by traveling, say, along the world-line $x = y = z = \text{const.}$ (which is clearly a geodesic), he will not reach $t = 0$ in any finite time – the surface is infinitely far into the future. Moreover, the fact that the singularity is not physically real can be seen by putting $t' = \log(-t)$ in $t < 0$ when the metric becomes

$$-dt'^2 + dx^2 + dy^2 + dz^2 \tag{5}$$

with $-\infty < t' < \infty$. In other words, the lower part of the space (and also the upper part) is just Minkowski space in disguise, and there is no singularity.

In his paper on the de Sitter metric just referred to, Einstein decided that the singularity was accessible (correctly), and that it was not possible to make the metric regular by a change of coordinates (incorrectly). He therefore deduced that there was a real singularity and, interestingly, he promptly rejected the solution as a consequence. The situation is the same with the singularity at $r = 2M$ in the Schwarzschild solution (1): it is accessible, but there is a change of coordinates for which it becomes regular.

In order to make Einstein's criterion for accessibility work, we can simply demand that there should be a timelike *geodesic* which reaches the singularity in a finite proper-time. Such a geodesic will have an endpoint on the singularity, in whatever coordinates are being used to describe the situation, but it will not have any endpoint in the regular part of the space-time. A space-time like this,

containing a timelike geodesic which (when maximally extended) has no endpoint in the regular space-time and which has finite proper length, is called *timelike geodesically incomplete*. Clearly this property of incompleteness, which now has no reference in it to a particular coordinate system, is independent of what coordinates we use to describe the space-time.

It would certainly be convenient to be able to use arbitrary curves to decide whether or not a singularity is accessible. Indeed, this seems to be physically reasonable, because if any fairly well-behaved observer (i.e. having bounded acceleration) can reach the singularity in a finite proper time, then the singularity should still count as physically accessible, even if no geodesic observer can reach it. We can capture this idea mathematically by using a different parameter on curves, in place of proper-time, so as to achieve a definition that includes the world lines of observers with bounded acceleration. This new parameter is called the generalised affine parameter.

1.2.1 Formalism

In the next sections we shall develop some of the mathematical machinery for dealing with these ideas. The notation will broadly follow Hawking and Ellis (1973). Briefly, the space-time manifold is denoted by M, its metric by g (regarded as a bilinear function from pairs of vectors at the same point to the reals). Boldface letters will be used for arrays of any sort. We suppose that all our curves are described by maps from an interval into the space-time that are differentiable almost everywhere and rectifiable. Then we make the following formulation

Definition. The generalised affine parameter length of a curve $\gamma : [0, a) \to M$ with respect to a frame

$$\mathbf{E} = (\underset{a}{E} : a = 0, \ldots, 3)$$

at $\gamma(0)$ is given by

$$\ell_{\mathbf{E}}(\gamma) = \int_0^a \left(\sum_{i=0}^{3} g\left(\dot{\gamma}, \underset{i}{E}(s)\right)^2 \right)^{1/2} ds \qquad (6)$$

where $\dot{\gamma}$ denotes the tangent vector $d\gamma/ds$ and $\mathbf{E}(s)$ is defined by parallel propagation along the curve, starting with an initial value $\mathbf{E}(0)$: that is, we impose

$$\begin{aligned} \nabla_{\dot{\gamma}} \mathbf{E}(s) &= 0 \\ \underset{a}{E}(0) &= \underset{a}{E} \end{aligned}$$

(We abbreviate this to g.a.p. length.)

Definition. A curve $\gamma : [0, a) \to M$ is *incomplete* if it has finite g.a.p. length with respect to some frame $\mathbf{E}$ at $\gamma(0)$. If $\ell_{\mathbf{E}}(\gamma) < \infty$, then if we take any other frame $\mathbf{E}'$ at $\gamma(0)$ we have that $\ell_{\mathbf{E}'}(\gamma) < \infty$. This is because the corresponding parallely propagated frames satisfy

$$\underset{i}{E'} = L_i^{\ j} \underset{j}{E}$$

for a constant Lorentz matrix L, and hence

$$\ell_{\mathbf{E}'} \leq \|\mathbf{L}\| \ell_{\mathbf{E}}$$

where $\|\mathbf{L}\|$ denotes the mapping norm:

$$\|\mathbf{L}\| = \sup \left(\sum_j \left(L_i^j X^i \right)^2 \right)^{1/2} \tag{7}$$

(with the supremum over all $\mathbf{X}$ with $|\mathbf{X}| = 1$ and $|\mathbf{X}|$ denotes the Euclidean norm of the components).

Definition. A curve $\gamma : [0, a) \to M$ is termed *inextendible* if there is no curve $\gamma' : [0, b) \to M$ with $b > a$ such that $\gamma'|[0, a) = \gamma$. This is equivalent to saying that there is no point p in M such that $\gamma(s) \to p$ as $s \to a$; i.e. that γ has no endpoint in M.

Definition. A space-time is *incomplete* if it contains an incomplete inextendible curve.

Discussion. We have now established definitions of geodesic incompleteness (which can be qualified by restricting to various sorts of geodesics) and incompleteness in the sense just defined. Clearly one can formulate many other possible definitions by restricting

the sort of curve used in the definition of incompleteness. For example, a space-time is called timelike incomplete if it contains an incomplete timelike inextendible curve. The definitions of geodesic incompleteness and incompleteness are concordant because, since the components $g(\dot{\gamma}, \mathbf{E})$ of the tangent vector to a geodesic are constant, the affine length is proportional to the generalised affine parameter length. So a geodesic is incomplete with respect to its affine parameter if, and only if, it is incomplete in the sense defined above.

The Friedmann "big bang" models (3) are geodesically incomplete (and hence incomplete) because the curve defined by

$$\gamma(s)^0 = a - s$$

$$\gamma(s)^i = \text{constant} \qquad (i = 1, 2, 3) \tag{8}$$

is a geodesic which is incomplete, having no endpoint in M as $s \to a$. Minkowski space is not incomplete (a result which is not trivial (Schmidt 1973)). The region $r > 2M$ in the Schwarzshild metric (1) is incomplete, while the region $r > 0$ in (1) is not a space-time, since g is not defined at $r = 2M$.

Finally, we note that many writers use "singular" as synonymous with "incomplete"; although as we have seen incompleteness is only one of the criteria which must be fulfilled for there to be a true singularity. Incompleteness corresponds to Einstein's accessibility criterion for a singularity. We must now consider the other requirement, needed to rule out an apparent singularity ("singularity" in the sense we were considering in 1.2) arising merely from a bad choice of coordinates.

1.3 Extendibility

In 1924 Eddington showed that there was an isometry between the space-time M defined by the regions $r > 2m$ in the Schwarzschild metric (1) and part of a larger space-time M'. Incomplete curves in M on which $r \to 2m$ were mapped by this isometry into curves that were extensible in M': the singularity at $r = 2m$ was no longer present. So if we identify the Schwarzschild space-time with the part of the Eddington space-time M' with which it is isometric, we see that it is not just incomplete in the formal sense defined above: it actually had a piece missing from it, a piece that is restored in

M'. The singularity at $r = 2m$ is thus a mathematical artifact, a consequence of the fact that the procedure used to solve the field equations had fortuitously produced only a part of the complete space.

We note that, despite this, there are still some authors who regard the Schwarzschild "singularity" at $r = 2m$ as genuine; but this is only justified if (as done by Rosen (1974)) one uses a non-standard physical theory in which there is some additional structure (such as a background metric) which itself becomes singular under the isometry of the metric into M', so that one structure, the metric or the background, is always singular at $r = 2m$.

The situation in Schwarzschild clearly contrasts with that of the Friedmann metrics (3). For these, on any of the incomplete curves (8) the Ricci scalar tends to infinity. For the smooth space-times that we are considering at the moment this is impossible on a curve which has an endpoint in the space-time, and so there can in this case be no isometric M' in which these curves have an endpoint. (Later we shall consider space-times in which the metric is not necessarily smooth, for which this does not hold.)

The singularity at $r = 2M$ in the Schwarzschild solution came to be called a "coordinate singularity", a term denoting any singularity in the sense of 1.2 which either did not give rise to incomplete curves, or which was such that incomplete curves tending to the singularity could be extended in some enlarged space-time. This larger space-time was constructed by applying a transformation to the coordinates specifying the original space-time, and extending the new coordinates (in modern terminology: by applying a diffeomorphism into a larger manifold). The Friedmann singularity, on the other hand, was termed a "physical" one, because a physically measurable quantity – the Ricci scalar – was unbounded on incomplete curves. On the whole I shall avoid the terms "coordinate" and "physical", since, while they convey important ideas, it is hard to give them precise definitions. Instead, I shall use the mathematical concept of extension to distinguish between the two types, the Schwarzschild space-time being extendible but the Friedmann one not so.

Definitions. An *extension* of a space-time (M, g) is an isometric embedding $\theta : M \to M'$, where (M', g') is a space-time and θ is

onto a proper subset of M'.

A space-time is termed *extendible* if it has an extension.

The relation between extendibility and incompleteness is then expressed by the following result, showing that extendible space-times are timelike incomplete.

Proposition 1.3.1

If M has an extension $\theta : M \to M'$ then there is an incomplete timelike geodesic γ in M such that $\theta \circ \gamma$ is extendible.

Proof

Let $x \in \partial\theta M \subset M'$ and let N be a convex normal neighbourhood of x in M'.

Case 1. Suppose there exists a point y in $\theta M \cap (I^+(x) \cup I^-(x)) \cap N$. Then let γ' be the closed geodesic segment in N with $\gamma' : [0,1] \to N$, $\gamma'(0) = y$, $\gamma'(1) = x$. Let I be a connected component of the set $\{s | \gamma'(s) \in \theta M\}$ and let $\gamma = \theta^{-1} \circ \gamma' | I$. Then I is relatively open, non-empty and connected in $[0,1]$ and $1 \in I$. Hence either $I = (a,b)$ or $I = [0,b)$, and so $\theta \circ \gamma$ can be extended to b, as required.

Case 2. Suppose $M \cap (I^+(x) \cup I^-(x)) \cap N = \emptyset$. Then we can choose $y \in M \cap N \setminus (I^+(x) \cup I^-(x))$ and $x' \in (I^+(y) \cup I^-(y)) \cap (I^+(x) \cup I^-(x)) \cap N$. Thus $x' \in M$. Let γ' join y to x', with $\gamma'(0) = y$ and $\gamma'(1) = x'$. Then define I as before and continue as in Case 1.

□

1.4 The maximality assumption

The forgoing result has shown that if M is extendible then there is some timelike curve (actually a geodesic) – i.e. a possible worldline of a particle – which could continue in some extension of M but which in M itself simply stops. This seems unreasonable: why should M be cut short in this way? It seems natural to demand that "if a space-time can continue then it will"; in other words to demand that any reasonable space-time should be inextendible. This is an assumption imposed upon space-time in addition to the field equations of Einstein.

It can easily be shown that any space-time can in fact be extended until no further extension is possible. At this point the space-time is called maximal, and so we are led to the idea that we need only consider maximal space-times. But this idea is not really as innocuous as it might seem, because of the problem that an extension of a space-time, when it exists, cannot usually be determined uniquely. In special cases there are unique extensions: an analytic space-time has (subject to some conditions) a unique maximal analytic extension; similarly a globally hyperbolic solution of the field equations (with a specified level of differentiability) is contained in a unique maximal solution. In both these cases a sort of "principle of sufficient reason" demands that the maximal solution be taken. But suppose one has a non-analytic space-time where Einstein's field equations fail to predict a unique extension (either because there is a Cauchy horizon or because there is some sort of failure of the differentiability needed for the existence of unique solutions). Or suppose a situation arises in which there is a set of incomplete curves, each one of which can be extended in some extension of the space-time, but where there is no extension in which they can all be extended. (There exist, admittedly artificial, examples of this (Misner, 1967).) In cases such as these the same principle of sufficient reason would not allow one extension to exist at the expense of another. Perhaps the space-time, like Buridan's ass between two bales of hay, unable to decide which way to go, brings the whole of history to a halt.

Any solution to these problems can only come through a greater understanding of the physics of situations which might give rise to non-uniqueness. In the absence of strong enough physical theories to enable us to decide, we can only note here that maximality, while a useful instrumental principle that seems very likely, is far from being absolutely certain. Nonetheless, we shall usually adopt the principle.

1.5 Singularities

We are now in a position to give a definition of a singular space-time, that incorporates the ideas we have just described. This will supersede the more primitive idea of a singularity with which we started in 1.2.

Definition. A space-time is *singular* if it contains an incomplete curve $\gamma : [0, a) \to M$ such that there is no extension $\theta : M \to M'$ for which $\theta \circ \gamma$ is extendible.

According to this definition, the region $r > 2m$ in the Schwarzschild solution (1) is not singular, merely incomplete. To say that a space-time is singular means that there is some positive obstacle that prevents an incomplete curve continuing: it is not just that the space-time is smaller than it might be. Note also that we have not yet defined "a singularity", only "singular". This will be rectified in the next chapter, when it will appear that any singular space-time contains a singularity, and so in a maximal space-time *all* incomplete inextendible curves end at a singularity. For the time being, I shall occasionally refer to a "singularity", when speaking loosely, in anticipation of its definition later.

1.5.1 Singularity theorems

It turns out that most physically reasonable known exact solutions, when maximally extended, are singular, in the sense of the definition just given. Of those mentioned so far, only the de Sitter metric (the maximal analytic extension of (2)) is not singular, but this is an exceptional case among isotropic cosmologies.

When it was realised that most cosmological solutions were singular, reactions varied. It would appear that at first Einstein and Hermann Weyl took the view that a singularity in the metric could be interpreted as the presence of a singular matter-source and should be rejected on the grounds that one should only be interested in regular matter-sources; though later they turned this argument on its head and regarded particles of matter as being non other than singularities. Others felt that, though singularities were inevitable in any description of cosmology and astrophysics by means of very symmetrical metrics, they were an artifact of the high symmetry – on the analogy with Newtonian gravitational theory where a singularity (in the sense of 1.1) is obtained when a cloud of particles collapses from an initially spherical state, but there is no singularity when a general initial state is assumed. There would be little point in devoting one's energy to the study of singularities if the only singular space-times were unrealistically symmetric.