5 Exact Solutions

Any space-time metric can in a sense be regarded as satisfying Einstein's field equations

$$R_{ab} - \frac{1}{2} R g_{ab} + \Lambda g_{ab} = 8\pi T_{ab}, \qquad (5.1)$$

(where we use the units of chapter 3), because, having determined the left-hand side of (5.1) from the metric tensor of the space-time *(M,g),* one can *define* T_{ab} as the right-hand side of (5.1). The matter tensor so defined will in general have unreasonable physical properties; the solution will be reasonable only if the matter content is reasonable.

We shall mean by an ***exact solution*** of Einstein's equations, a spacetime *(M,g)* in which the field equations are satisfied with T_{ab} the energy-momentum tensor of some specified form of matter which obeys postulate *(a)* ('local causality') of chapter 3, and one of the energy conditions of §4.3. In particular, one may look for exact solutions for empty space *(Tab* = 0), for an electromagnetic ·field (T_{ab} has the form (3.7)) for a perfect fluid (T_{ab} has the form (3.8)), or for a space containing an electromagnetic field and a perfect fluid. Because of the complexity of the field equations, one cannot find exact solutions except in spaces of rather high symmetry. Exact solutions are also idealized in that any region of space-time is likely to contain many forms of matter, while one can obtain exact solutions only for rather simple matter content. Nevertheless, exact solutions give an idea of the qualitative features that can arise in General Relativity and so of possible properties of realistic solutions of the field equations. The examples we give will show many types of behaviour which will be of interest in later chapters. We shall discuss solutions with particular reference to their global properties. Many of these global properties have only recently been discovered, although the solutions have been known in a local form for some time.

In §5.1 and §5.2 we consider the simplest Lorentz metrics: those of constant curvature. The spatially isotropic and homogeneous cosmological models are described in §5.3, and their simplest anisotropic generalizations are discussed in §5.4. It is shown that all such simple models will have a singular origin provided that Λ does not take large positive values. The spherically symmetric metrics which describe

the field outside a massive charged or neutral body are examined in §5.5, and the axially symmetric metrics describing the field outside a special class of massive rotating bodies are described in §5.6. It is shown that some of the apparent singularities are simply due to a bad choice of coordinates. In §5.7 we describe the Godel universe and in § 5.8 the Taub-NUT solutions. These probably do not represent the actual universe but they are of interest because of their pathological global properties. Finally some other exact solutions of interest are mentioned in §5.9.

5.1 Minkowski space-time

Minkowski space-time (M,η) is the simplest empty space-time in General Relativity, and is in fact the space-time of Special Relativity. Mathematically, it is the manifold *R4* with a flat Lorentz metric η. In terms of the natural coordinates (x^1,x^2,x^3,x^4) on R^4, the metric η can be expressed in the form

$$ds^2 = -(dx^4)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)2 . \qquad (5.2)$$

If one uses spherical polar coordinates (t,r,θ,ϕ) where $x^4 = t$, $x^3 = r\cos\theta$, $x^2 = r\sin\theta\cos\phi$, $x^1 = r\sin\theta\sin\phi$, the metric takes the form

$$ds^2 = -dt^2 + dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \qquad (5.3)$$

This metric is apparently singular for $r = 0$ and sin $\theta = 0$; however this is because the coordinates used are not admissible coordinates at these points. To obtain regular coordinate neighbourhoods one has to restrict the coordinates, e.g. to the ranges $0 < r < \infty$, $0 < \theta < \pi$, $0 < \phi < 2\pi$. One needs two such coordinate neighbourhoods to cover the whole of Minkowski space.

An alternative coordinate system is given by choosing advanced and retarded null coordinates *v, W* defined by $v = t + r$, $W = t - r$ (=> $v \geq w$*)*. The metric becomes

$$ds^2 = -dvdw + \frac{1}{4}(v-w)^2(d\theta^2 + \sin^2\theta d\phi^2) \qquad (5.4)$$

where $-\infty < v < \infty$, $-\infty < w < \infty$. The absence in the metric of terms in dv^2, dw^2 corresponds to the fact that the surfaces $\{w = const\}$, $\{v = const\}$ are null (i.e. $w_{;a}w_{;b}g^{ab} = 0 = v_{;a}v_{;b}g^{ab}$)*);* see figure 12.

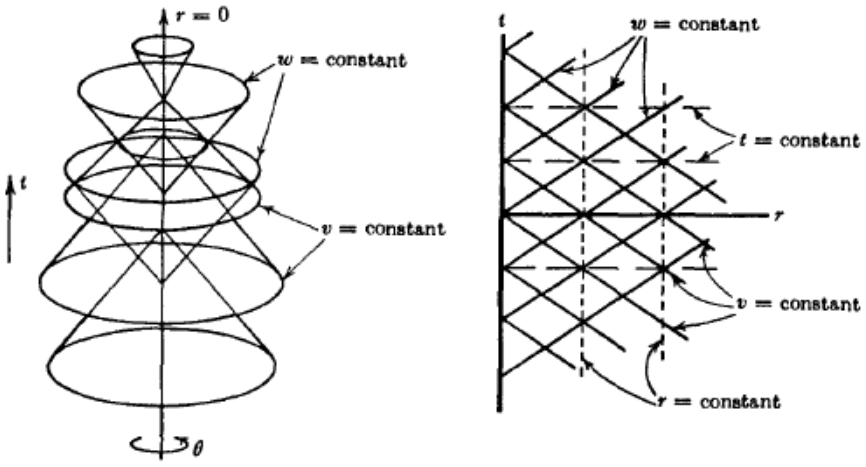


FIGURE 12. Minkowski space. The null coordinate $v(w)$ may be thought of as incoming (outgoing) spherical waves travelling at the speed of light; they are advanced (retarded) time coordinates. The intersection of a surface $\{v = \text{constant}\}$ with a surface $\{w = \text{constant}\}$ is a two-sphere.
(i) The v, w coordinate surfaces (one coordinate is suppressed).
(ii) The (t, r) plane; each point represents a two-sphere of radius r.

In a coordinate system in which the metric takes the form (5.2), the geodesics have the form $x^a(v) = b^a v + c^a$ where b^a and c^a are constants. Thus the exponential map $\exp_p : T_p \to M$ is given by

$$x^a(\exp_p \boldsymbol{X}) = X^a + x(p),$$

where X^a are the components of **X** with respect to the coordinate basis $\{\partial/\partial x^a\}$ of *Tp.* Since exp is one-one and onto, it is a diffeomorphism between *Tp* and M. Thus any two points of M can be joined by a unique geodesic curve. *As* exp is defined everywhere on *Tp* for all *p, (M, η)* is geodesically complete.

For a spacelike three-surface *S*, the future (past) Cauchy development *D+(S) (D-(S))* is defined as the set of all points qEM such that each past-directed (future-directed) inextendible non-spacelike curve through *q* intersects S, cf. §6.5. If $D^+(S) \cup D^-(S) = M$, i.e. if every inextendible non-spacelike curve in M intersects S, then S is said to be a Cauchy surface. In Minkowski space-time, the surfaces {x4 = constant} are a family of Cauchy surfaces which cover the whole *of M.* One can however find inextendible spacelike surfaces which are not Cauchy surfaces; for example the surfaces

$$S_\sigma : \{-(x^4)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 = \sigma = const.\},$$

where $u < 0$, x4 < 0, are spacelike surfaces which lie entirely inside the past null cone of the origin *0,* and so are not Cauchy surfaces (see figure 13). In fact the future Cauchy development of S_σ is the region bounded by S_σ and the past light cone of the origin. By lemma 4.5.2, the timelike geodesics through the origin *0* are orthogonal to the surfaces S_σ. If $r \in D^+(S_\sigma) \cup D^-(S_\sigma)$ then the timelike geodesic through *r* and *0* is the longest timelike curve between rand S_σ. If

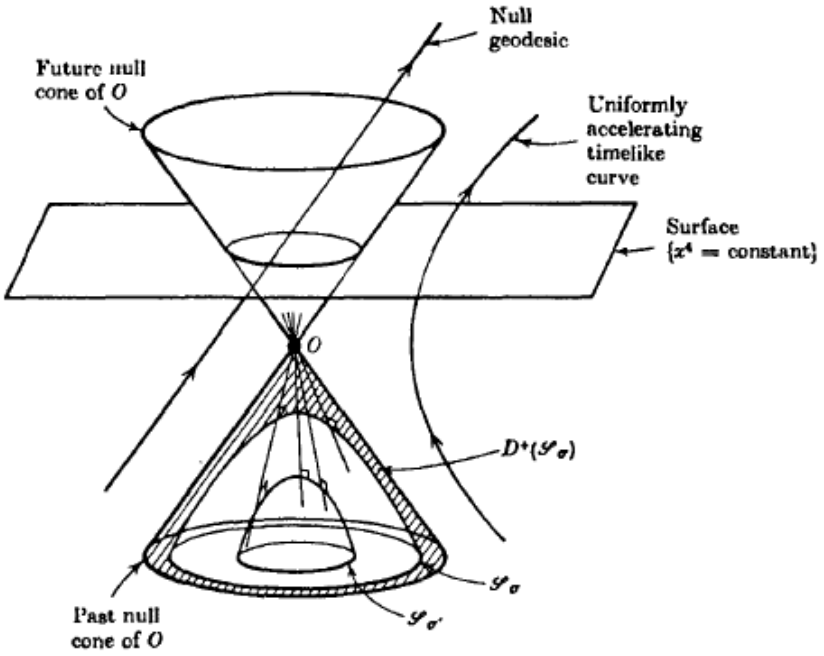


FIGURE 13. A Cauchy surface $\{x^4 = \text{constant}\}$ in Minkowski space–time, and spacelike surfaces $\mathcal{S}_\sigma, \mathcal{S}_{\sigma'}$ which are not Cauchy surfaces. The normal geodesics to the surfaces $\mathcal{S}_\sigma, \mathcal{S}_{\sigma'}$ all intersect at *O*.

however *r* does not lie in $D^+(S_\sigma) \cup D^-(S_\sigma)$ there is no longest timelike curve between *r* and S_σ: either *r* lies in the region $\sigma \geq 0$, in which case there is no timelike geodesic through *r* orthogonal to S_σ, or *r* lies in the region $\sigma < 0$, $x^4 \geq 0$, in which case there is a timelike geodesic through *r* orthogonal to S_σ but this geodesic is not the longest curve between *r* and S_σ as it contains a conjugate point to S_σ at *0* (cf. figure 13).

To study the structure of infinity in Minkowski space-time, we shall use the interesting representation of this space-time given by Penrose. From the null coordinates *v, W,* we define new null coordinates in which the infinities of *v, w* have been transformed to finite values; thus we define *p, q* by tanp = *v,* tanq = *w* where -1/2 π < p < 1/2 π ,

$-1/2\,\pi < q < 1/2\,\pi$ (and $p \geq q$). Then the metric of (M, η) takes the form

$$ds^2 = \sec^2 p \sec^2 q(-dpdq + \frac{1}{2}\sin^2(p-q)(d\theta^2 + \sin^2 d\phi^2)) .$$

The physical metric η is therefore conformal to the metric $\bar{g}$ given by

$$d\bar{s}^2 = -4dpdq + \sin^2(p-q)(d\theta^2 + \sin^2\theta d\phi^2) . \qquad (5.5)$$

This metric can be reduced to a more usual form by defining

$t' = p + q$, $r' = p - q$,

where

$$-\pi < t' + r' < \pi , \quad -\pi < t' - r' < \pi , \quad r' \geq 0 ; \qquad (5.6)$$

(5.5) is then

$$d\bar{s}^2 = -(dt')^2 + (dr')^2 + \sin^2 r'(d\theta^2 + \sin^2\theta d\phi^2) . \qquad (5.7)$$

Thus the whole of Minkowski space-time is given by the region (5.6) of the metric

$$ds^2 = \frac{1}{4}\sec^2(\frac{1}{2}(t'+r'))\sec^2(\frac{1}{2}(t'-r'))d\bar{s}^2$$

where $d\bar{s}^2$ is determined by (5.7); the coordinates *t, r* of (5.3) are related to *t′, r′* by

$$2t = \tan(\frac{1}{2}(t'+r')) + \tan(\frac{1}{2}(t'-r')) ,$$

$$2r = \tan(\frac{1}{2}(t'+r')) - \tan(\frac{1}{2}(t'-r')) .$$

Now the metric (5.7) is locally identical to that of the Einstein static universe (see §5.3), which is a completely homogeneous space-time. One can analytically extend (5.7) to the whole of the Einstein static universe, that is one can extend the coordinates to cover the manifold $R^1 \times S^3$ where - 00 < *t′* < 00 and *r′, θ , ¢* are regarded as coordinates on S^3 (with coordinate singularities at $r' = 0$, $r' = \pi$ and $\theta = 0, \theta = \pi$ similar to the coordinate singularities in (5.3); these singularities can be removed by transforming to other local coordinates in a neighbourhood of points where (5.7) is singular). On suppressing two dimensions, one can represent the Einstein static universe as the cylinder $x^2 + y^2 = 1$ imbedded in a three-dimensional Minkowski space with metric $ds^2 = -dt^2 + dx^2 + dy^2$ (the full Einstein static universe can be imbedded as the cylinder $x^2 + y^2 + z^2 + w^2 = 1$ in a five-dimensional Euclidean space with metric $ds^2 = -dt^2 + dx^2 + dy^2 + dz^2 + dw^2$, cf. Robertson (1933».

One therefore has the situation: the whole of Minkowski space-time is conformal to the region (5.6) of the Einstein static universe, that is, to the shaded area in figure 14. The boundary of this region may therefore be thought of as representing the conformal structure of infinity of Minkowski space-time. It consists of the null surfaces $p = \pi/2$ (labelled I^+) and $q = -\pi/2$ (labelled I^-) together with points $p = \pi/2$, $q = \pi/2$ (labelled i^+), $p = \pi/2$, $q = -\pi/2$ (labelled i^0) and $p = -\pi/2$, $q = -\pi/2$ (labelled i^-). Any future-directed timelike geodesic in

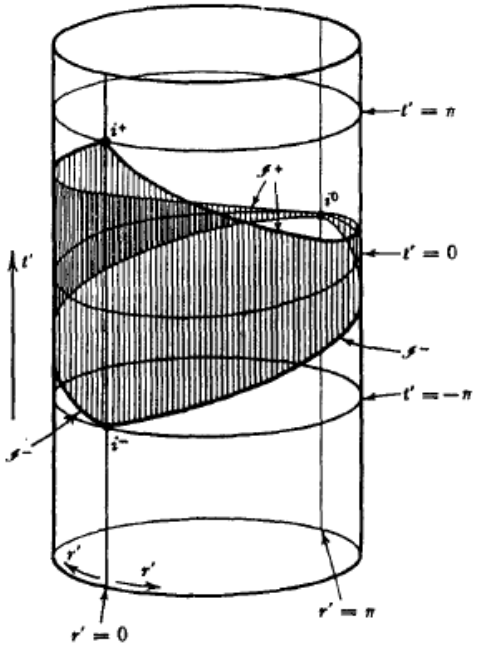


FIGURE 14. The Einstein static universe represented by an imbedded cylinder; the coordinates θ, ϕ have been suppressed. Each point represents one half of a two-sphere of area $4\pi \sin^2 r'$. The shaded region is conformal to the whole of Minkowski space-time; its boundary (part of the null cones of i^+, i^0 and i^-) may be regarded as the conformal infinity of Minkowski space-time.

Minkowski space approaches i^+ (i^-) for indefinitely large positive (negative) values of its affine parameter, so one can regard any timelike geodesic as originating at i^- and finishing at i^+ (cf. figure 15(i)). Similarly one can regard null geodesics as originating at I^- and ending at I^+, while spacelike geodesics both originate and end at i^0. Thus one may regard i^+ and i^- as representing future and past timelike infinity, I^+ and I^- as representing future and past null infinity, and i^0 as representing spacelike infinity. (However non-geodesic curves do not obey these rules; e.g. non-geodesic timelike curves may start on I^- and end on I^+.) Since any Cauchy surface intersects all timelike and null geodesics, it is clear that it will appear as a cross-section of the

space everywhere reaching the boundary at i^0.

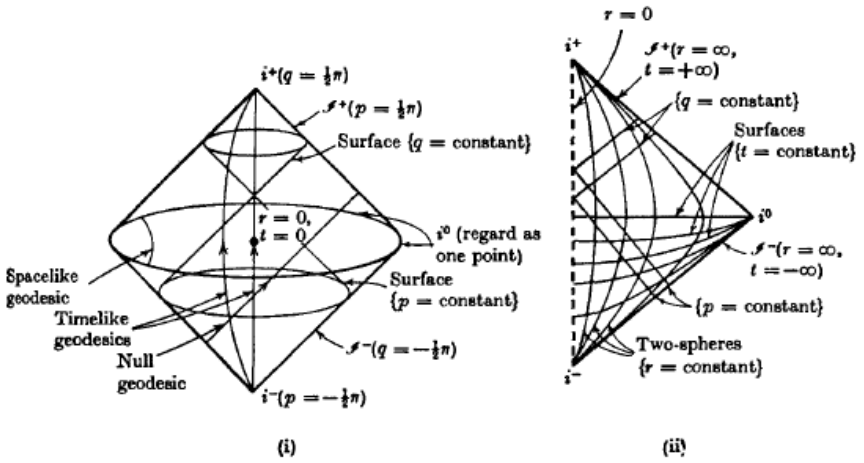


FIGURE 15

(i) The shaded region of figure 14, with only one coordinate suppressed, representing Minkowski space–time and its conformal infinity.

(ii) The Penrose diagram of Minkowski space–time; each point represents a two-sphere, except for i^+, i^0 and i^-, each of which is a single point, and points on the line $r = 0$ (where the polar coordinates are singular).

One can also represent the conformal structure of infinity by drawing a diagram of the *(t', r')* plane, see figure 15 (ii). As in figure 12 (ii), each point of this diagram represents a sphere S^2, and radial null geodesics are represented by straight lines at ±45. In fact, the structure of infinity in any spherically symmetric space-time can be represented by a diagram of this sort, which we shall call a *Penrose diagram.* On such diagrams we shall represent infinity by single lines, the origin of polar coordinates by dotted lines, and irremovable singularities of the metric by double lines.

The conformal structure of Minkowski space we have described is what one would regard as the 'normal' behaviour of a space-time at infinity; we shall encounter different types of behaviour in later sections.

Finally, we mention that one can obtain spaces locally identical to (M, η) but with different (large scale) topological properties by identifying points in *M* which are equivalent under a discrete isometry without a fixed point (e.g. identifying the point (x^1, x^2, x^3, x^4) with the point $(x^1, x^2, x^3, x^4 + c)$, where c is a constant, changes the topological structure from R^4 to $R^3 \times S^1$, and introduces closed timelike lines into the space-time). Clearly, (M, η) is the universal covering space for all such derived spaces, which have been studied in detail by Auslander and Markus (1958).

5.2 De Sitter and anti-de Sitter space-times

The space-time metrics of constant curvature are locally characterized by the condition $R_{abcd} = (1/12)R(g_{ac}g_{bd} - g_{ad}g_{bc})$. This equation is equivalent to $C_{abcd} = 0 = R_{ab} - (1/4)Rg_{ab}$; thus the Riemann tensor is determined by the Ricci scalar R alone. It follows at once from the contracted Bianchi identities that R is constant throughout space-time; in fact these space-times are homogeneous. The Einstein tensor is

$$R_{ab} - \frac{1}{2}Rg_{ab} = -\frac{1}{4}Rg_{ab}.$$

One can therefore regard these spaces as solutions of the field equations for an empty space with $\Lambda = (1/4)R$, or for a perfect fluid with a constant density $R/(32\pi)$ and a constant pressure $-R/(32\pi)$. However the latter choice does not seem reasonable, as in this case one cannot have both the density and the pressure positive; in addition, the equation of motion (3.10) is indeterminate for such a fluid.

The space of constant curvature with $R = 0$ is Minkowski spacetime. The space for $R > 0$ is ***de Sitter space-time***, which has the topology $R^1 \times S^3$ (see Schrodinger (1956) for an interesting account of this space). It is easiest visualized as the hyperboloid

$$-v^2 + w^2 + x^2 + y^2 + z^2 = \alpha^2$$

in flat five-dimensional space R^5 with metric

$$-dv^2 + dw^2 + dx^2 + dy^2 + dz^2 = ds^2$$

(see figure 16). One can introduce coordinates (t, χ, θ, ϕ) on the hyperboloid by the relations

$$\alpha \sinh(\alpha^{-1}t) = v,$$
$$\alpha \cosh(\alpha^{-1}t)\cos\chi = w,$$
$$\alpha \cosh(\alpha^{-1}t)\sin\chi\cos\theta = x,$$
$$\alpha \cosh(\alpha^{-1}t)\sin\chi\sin\theta\cos\phi = y,$$
$$\alpha \cosh(\alpha^{-1}t)\sin\chi\sin\theta\sin\phi = z.$$

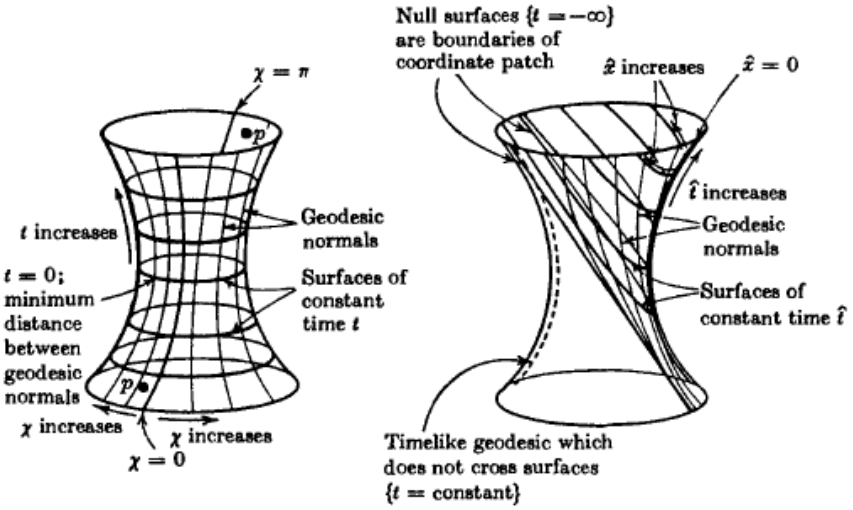


FIGURE 16. De Sitter space–time represented by a hyperboloid imbedded in a five-dimensional flat space (two dimensions are suppressed in the figure).
(i) Coordinates (t, χ, θ, ϕ) cover the whole hyperboloid; the sections {t = constant} are surfaces of curvature $k = +1$.
(ii) Coordinates $(\hat{t}, \hat{x}, \hat{y}, \hat{z})$ cover half the hyperboloid; the surfaces {$\hat{t}$ = constant} are flat three-spaces, their geodesic normals diverging from a point in the infinite past.

In these coordinates, the metric has the form

$$ds^2 = -dt^2 + \alpha^2 \cdot \cosh^2(\alpha^{-1}t) \cdot \left\{d\chi^2 + \sin^2\chi(d\theta^2 + \sin^2\theta d\phi^2)\right\}.$$

The singularities in the metric at $\chi = 0$, $\chi = \pi$ and at $\theta = 0$, $\theta = 0$, $\theta = \pi$, are simply those that occur with polar coordinates. Apart from these trivial singularities, the coordinates cover the whole space for $-\infty < t < \infty$, $0 \le \chi \le \pi$, $0 \le \theta \le \pi$, $0 \le \phi \le 2\pi$. The spatial sections of constant t are spheres S^3 of constant positive curvature and are Cauchy surfaces. Their geodesic normals are lines which contract monotonically to a minimum spatial separation and then re-expand to infinity (see figure 16 (i)).

One can also introduce coordinates

$$\hat{t} = \alpha \log\frac{w+v}{\alpha}, \quad \hat{x} = \frac{\alpha x}{w+v}, \quad \hat{y} = \frac{\alpha y}{w+v}, \quad \hat{z} = \frac{\alpha z}{w+v}$$

on the hyperboloid. In these coordinates, the metric takes the form

$$ds^2 = -d\hat{t}^2 + \exp(2\alpha^{-1}\hat{t})(d\hat{x}^2 + d\hat{y}^2 + d\hat{z}^2).$$

However these coordinates cover only half the hyperboloid as $\hat{t}$ is not defined for $w + v \le 0$ (see figure 16 (ii)).

The region of de Sitter space for which $v + w > 0$ forms the spacetime for the ***steady state*** model of the universe proposed by Bondi and

Gold (1948) and Hoyle (1948). In this model, the matter is supposed to move along the geodesic normals to the surfaces $\{\tilde{t} = const.\}$. As the matter moves further apart, it is assumed that more matter is continuously created to maintain the density at a constant value. Bondi and Gold did not seek to provide field equations for this model, but Pirani (1955), and Hoyle and Narlikar (1964) have pointed out that the metric can be considered as a solution of the Einstein equations (with $\Lambda = 0$) if in addition to the ordinary matter one introduces a scalar field of negative energy density. This 'C'-field would also be responsible for the continual creation of matter.

The steady state theory has the advantage of making simple and definite predictions. However from our point of view there are two unsatisfactory features. The first is the existence of negative energy, which was discussed in §4.3. The other is the fact that the space-time is extendible, being only half of de Sitter space. Despite these aesthetic objections, the real test of the steady state theory is whether its predictions agree with observations or not. At the moment it seems that they do not, though the observations are not yet quite conclusive.

de Sitter space is geodesica1ly complete; however, there are points in the space which cannot be joined to each other by any geodesic. This is in contrast to spaces with a positive definite metric, when geodesic completeness guarantees that any two points of a space can be joined by at least one geodesic. The half of de Sitter space which represents the steady state universe is not complete in the past (there are geodesics which are complete in the full space, and cross the boundary of the steady state region; they are therefore incomplete in that region).

To study infinity in de Sitter space-time, we define a time coordinate t' by

$$t' = 2\arctan(\exp(\alpha^{-1}t) - \pi/2, \qquad (5.8)$$

where

$$-\pi/2 < t' < \pi/2 .$$

Then

$$ds^2 = \alpha^2 \cosh^2(\alpha^{-1}t') \cdot d\bar{s}^2,$$

where $d\bar{s}^2$ is given by (5.7) on identifying $r' = \chi$. Thus the de Sitter space is conformal to that part of the Einstein static universe defined

by (5.8) (see figure 17 (i)). The Penrose diagram of de Sitter space is accordingly as in figure 17 (ii). One half of this figure gives the Penrose

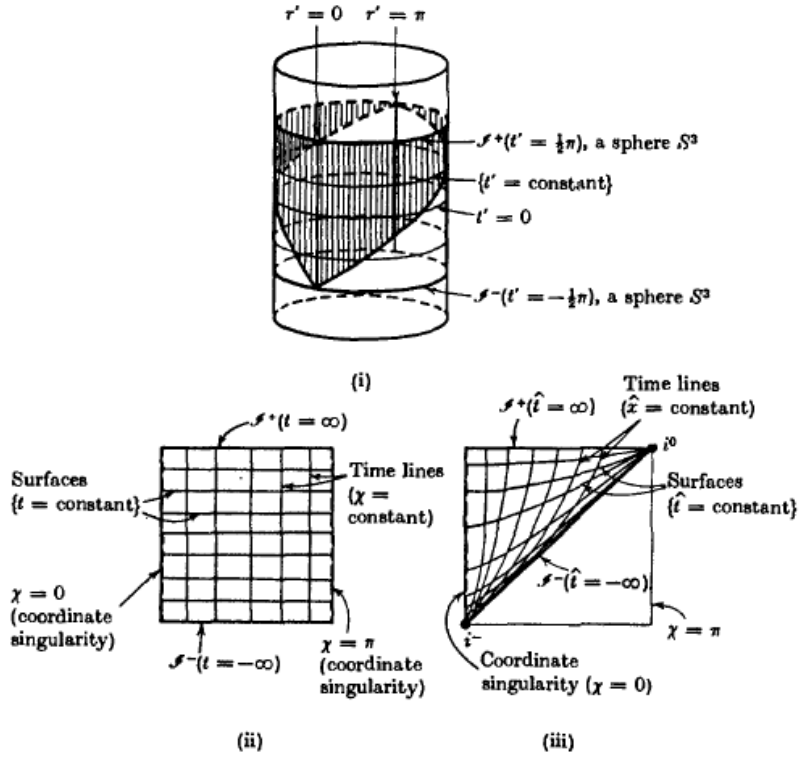


FIGURE 17

(i) De Sitter space–time is conformal to the region $-\frac{1}{2}\pi < t' < \frac{1}{2}\pi$ of the Einstein static universe. The steady state universe is conformal to the shaded region.

(ii) The Penrose diagram of de Sitter space–time.

(iii) The Penrose diagram of the steady state universe.

In (ii), (iii) each point represents a two-sphere of area $2\pi \sin^2\chi$; null lines are at 45°. $\chi = 0$ and $\chi = \pi$ are identified.

diagram of the half of de Sitter space-time which constitutes the steady state universe (figure 17(iii)).

One sees that de Sitter space has, in contrast to Minkowski space, a spacelike infinity for timelike and null lines, both in the future and the past. This difference corresponds to the existence in de Sitter space-time of both particle and event horizons for geodesic families of observers.

In de Sitter space, consider a family of particles whose histories are timelike geodesics; these must originate at the spacelike infinity I^{-1} and end at the spacelike infinity I^{+}. Let p be some event on the world-line of a particle *0* in this family, i.e. some time in its history (proper

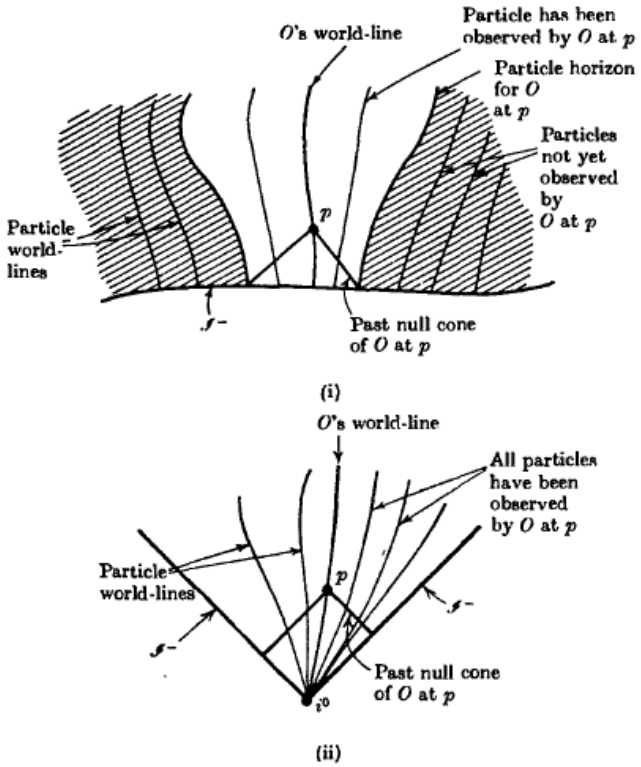


FIGURE 18

(i) The particle horizon defined by a congruence of geodesic curves when past null infinity $\mathscr{I}^-$ is spacelike.
(ii) Lack of such a horizon if $\mathscr{I}^-$ is null.

time measured along *O's* world-line). The past null cone of *p* is the set of events in space-time which can be observed by *O* at that time. The world-lines of some other particles may intersect this null cone; these particles are visible to *O*. However, there can exist particles whose world-lines do not intersect this null cone, and so are not yet visible to *O*. At a later time *O* can observe more particles, but there still exist . particles not visible to *O* at that time. We say that the division of particles into those seen by *O* at *p* and those not seen by *O* at *p,* is the ***particle horizon*** for the observer *0* at the event *p;* it represents the history of those particles lying at the limits of *O's* vision. Note that it is determined only when the world-lines of all the particles in the family are known. If some particle lies on the horizon, then the event *p* is the event at which the particle's creation light cone intersects *O's* world-line. In Minkowski space, on the other hand, all the other particles are visible at any event *p* on *O's* world-line if they move on timelike geodesics. As long as one considers only families of geodesic observers, one may think of the existence of the particle horizon as a

consequence of past null infinity being spacelike (see figure 18).

All events outside the past null cone of *p* are events which are not, and never have been, observable by *O* up to the time represented by the event *p*. There is a limit to *O's* world-line on I^+. In de Sitter space-time, the past null cone of this point (obtained by a limiting process in the actual space-time, or directly from the conformal space-time) is a boundary between events which will at some time be observable by *O* , and those that will never be observable by *O*. We call this surface the ***future event horizon*** of the world-line. It is the boundary of the past of the world-line. In Minkowski space-time, on the other hand, the limiting null cone of any geodesic observer includes the whole of space-time, so there are no events which a geodesic observer will never be able to see. However if an observer moves with uniform acceleration his world-line may have a future event horizon. One may think of the existence of a future event horizon for a geodesic observer as being a consequence of I^+ being spacelike (see figure 19).

Consider the event horizon for the observer *O* in de Sitter space-time and suppose that at some proper time (event *p)* on his world-line, his light cone intersects the world-line of the particle *Q*. Then *Q* is always visible to *O* at times after *p*. However there is on *Q's* world-line an event *r* which lies on *O's* future event horizon; *O* can never see later events on *Q's* world-line than *r*. Moreover an infinite proper time elapses on *O's* world-line from any given point till he observes *T,* but a finite proper time elapses along *Q's* world-line from any given event to *r,* which is a perfectly ordinary event on his world-line. Thus *O* sees a finite part of *Q's* history in an infinite time; expressed more physically, as *O* observes *Q* he sees a redshift which approaches infinity as *O* observes points on *Q's* world-line which approach *r*. Correspondingly, *Q* never sees beyond some point on *O's* world-line, and sees nearby points on *O's* world-line only with a very large redshift.

At any point on *O's* world-line, the future null cone is the boundary of the set of events in space-time which *O* can influence at and after that time. To obtain the maximal set of events in space-time that *O* could at any time influence, we take the future light cone of the limit point of *O's* world-line on past infinity I^-; that is, we take the boundary of the future of the world-line (which can be regarded as

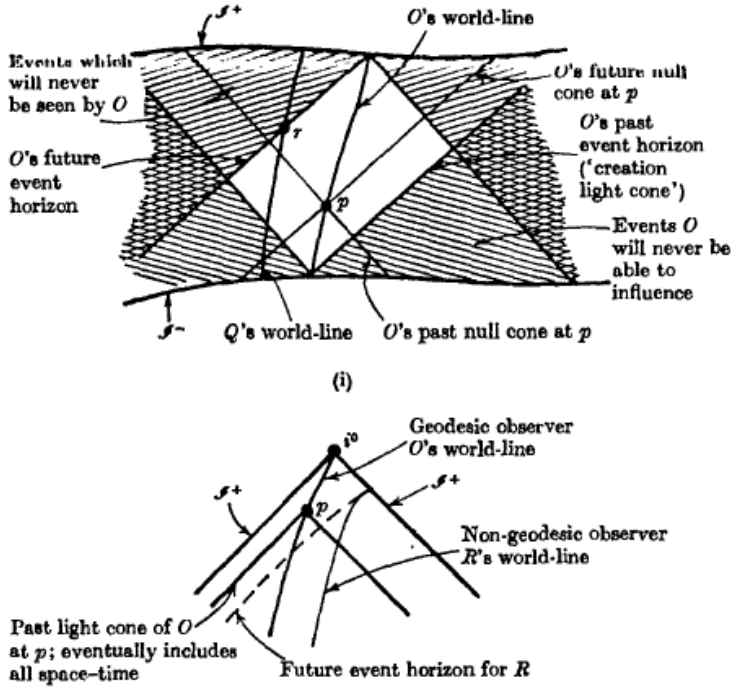


FIGURE 19

(i) The future event horizon for a particle O which exists when future infinity $\mathscr{I}^+$ is spacelike; also the past event horizon which exists when past infinity $\mathscr{I}^-$ is spacelike.

(ii) If future infinity consists of a null $\mathscr{I}^+$ and i^0, there is no future event horizon for a geodesic observer O. However an accelerating observer R may have a future event horizon.

O's creation light cone). This has a non-trivial existence for a geodesic observer only if the past infinity I^- is spacelike (and is in fact then *O's* past event horizon). It is clear from the above discussion that in the steady state universe, which has a null past infinity for timelike and null geodesics and a spacelike future infinity, any fundamental observer has a future event horizon but no past particle horizon.

One can obtain other spaces which are locally equivalent to the de Sitter space, by identifying points in de Sitter space. The simplest such identification is to identify antipodal points p, p' (see figure 16) on the hyperboloid. The resulting space is not time orientable; if time increases in the direction of the arrow at p, the antipodal identification implies it must increase in the direction of the arrow at p', but one cannot continuously extend this identification of future and past half null cones over the whole hyperboloid. Calabi and Markus (1962) have studied in detail the spaces resulting from such identifications; they show in particular that an arbitrary point in the resulting space can

be joined to any other point by a geodesic if and only if it is not time orientable.

The space of constant curvature with $R < 0$ is called ***anti-de Sitter space***. It has the topology $S^1 \times R^3$, and can be represented as the hyperboloid

$$-u^2 - v^2 + w^2 + x^2 + y^2 + z^2 = 1$$

in flat five-dimensional space R^5 with metric

$$ds^2 = -du^2 - dv^2 + dx^2 + dy^2 + dz^2 .$$

There are closed timelike lines in this space; however it is not simply connected, and if one unwraps the circle S^1 (to obtain its covering space R^1) one obtains the universal covering space of anti-de Sitter space which does not contain any closed timelike lines. This has the topology of R^4. We shall in future mean by 'anti-de Sitter space', this universal covering space.

It can be represented by the metric

$$ds^2 = -dt^2 + \cos^2 t\left\{d\chi^2 + \sinh^2 \chi(d\theta^2 + \sin^2 \theta d\phi^2)\right\}. \qquad (5.9)$$

This coordinate system covers only part of the space, and has apparent singularities at $t = \pm\pi/2$. The whole space can be covered by coordinates $\{t', r, \theta, \phi\}$ for which the metric has the static form

$$ds^2 = -\cosh^2 r dt'^2 + dr^2 + \sinh^2 r(d\theta^2 + \sin^2 \theta d\phi^2) .$$

In this form, the space is covered by the surfaces $\{t' = const.\}$ which have non-geodesic normals.

To study the structure at infinity, define the coordinate r' by

$$r' = 2\arctan(\exp r) - \pi/2 , \quad 0 \le r' < \pi/2 .$$

Then one finds $ds^2 = -\cosh^2 r d\bar{s}^2$, , where $d\bar{s}^2$ is given by (5.7); that is, the whole of anti-de Sitter space is conformal to the region $0 \le r' < \pi/2$ of the Einstein static cylinder. The Penrose diagram is shown in figure 20; null and spacelike infinity can be thought of as a timelike surfac.e in this case. This surface has the topology $R^1 \times S^2$,

One cannot find a conformal transformation which makes timelike infinity finite without pinching off the Einstein static universe to a point (if a conformal transformation makes the time coordinate finite it also scales the space sections by an infinite factor), so we represent timelike infinity by the disjoint points i^+, i^-.

The lines $\{\chi, \theta, \phi = const.\}$ are the geodesics orthogonal to the

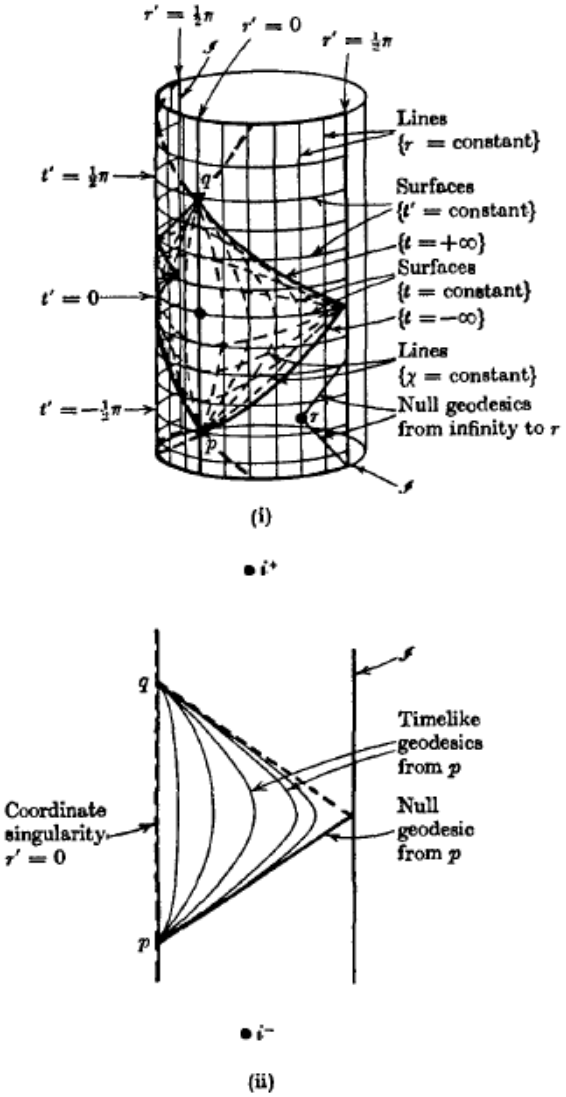


FIGURE 20

(i) Universal anti-de Sitter space is conformal to one half of the Einstein static universe. While coordinates (t', r, θ, ϕ) cover the whole space, coordinates (t, χ, θ, ϕ) cover only one diamond-shaped region as shown. The geodesics orthogonal to the surfaces $\{t = \text{constant}\}$ all converge at p and q, and then diverge out into similar diamond-shaped regions.

(ii) The Penrose diagram of universal anti-de Sitter space. Infinity consists of the timelike surface $\mathscr{I}$ and the disjoint points i^+, i^-. The projection of some timelike and null geodesics is shown.

surfaces $\{t = const.\}$; they all converge to points q (respectively, p) in the future (respectively, past) of the surface, and this convergence is the reason for the apparent (coordinate) singularities in the original metric form. The region covered by these coordinates is the region between the surface $t = 0$ and the null surfaces on which these normals become degenerate.

The space has two further interesting properties. First, as a consequence

of the timelike infinity, there exists no Cauchy surface whatever in the space. While one can find families of spacelike surfaces (such as the surfaces $\{t' = const.\}$ which cover the space completely, each surface being a complete cross-section of the spacetime, one can find null geodesics which never intersect any given surface in the family. Given initial data on any such surface, one cannot predict beyond the Cauchy development of the surface; thus from the surface $\{t = 0\}$, one can predict only in the region covered by the coordinates t, χ, θ, ϕ. Any attempt to predict beyond this region is prevented by fresh information coming in from the timelike infinity.

Secondly, corresponding to the fact that the geodesic normals from $t = 0$ all converge at p and q, all the past timelike geodesics from p expand out (normal to the surfaces $\{t = const.\}$) and reconverge at q. In fact, all the timelike geodesics from any point in this space (to either the past or future) reconverge to an image point, diverging again from this image point to refocus at a second image point, and so on. The future timelike geodesics from p therefore never reach I, in contrast to the future null geodesics which go to I from p and form the boundary of the future of p. This separation of timelike and null geodesics results in the existence of regions in the future of p (i.e. which can be reached from p by a future-directed timelike line) which cannot be reached from p by any geodesic. The set of points which can be reached by future-directed timelike lines from p is the set of points lying beyond the future null cone of p; the set of points which can be reached from p by future-directed timelike geodesics is the interior of the infinite chain of diamond-shaped regions similar to that covered by coordinates $\{t, \chi, \theta, \phi\}$. One notes that all points in the Cauchy development of the surface $t = 0$ can be reached from this surface by a unique geodesic normal to this surface, but that a general point outside this Cauchy development cannot be reached by any geodesic normal to the surface.

5.3 Robertson-Walker spaces

So far, we have not considered the relation of exact solutions to the physical universe. Following Einstein, we can ask: can one find space-times which are exact solutions for some suitable form of matter and

which give a good representation of the large scale properties of the observable universe? If so, we can claim to have a reasonable 'cosmological model' or model of the physical universe.

However we are not able to make cosmological models without some admixture of ideology. In the earliest cosmologies, man placed himself in a commanding position at the centre of the universe. Since the time of Copernicus we have been steadily demoted to a medium sized planet going round a medium sized star on the outer edge of a fairly average galaxy, which is itself simply one of a local group of galaxies. Indeed we are now so democratic that we would not claim that our position in space is specially distinguished in any way. We shall, following Bondi (1960), call this assumption the ***Copernican principle***.

A reasonable interpretation of this somewhat vague principle is to understand it as implying that, when viewed on a suitable scale, the universe is approximately spatially homogeneous.

By spatially homogeneous, we mean there is a group of isometries which acts freely on *M,* and whose surfaces of transitivity are spacelike three-surfaces; in other words, any point on one of these surfaces is equivalent to any other point on the same surface. Of course, the universe is not exactly spatially homogeneous; there are local irregularities, such as stars and galaxies. Nevertheless it might seem reasonable to suppose that the universe is spatially homogeneous on a large enough scale.

While one can build mathematical models fulfilling this requirement of homogeneity (see next section), it is difficult to test homogeneity directly by observation, as there is no simple way of measuring the separation between us and distant objects. This difficulty is eased by the fact that we can, in principle, fairly easily observe *isotropies* in extragalactic observations (i.e. we can see if these observations are the same in different directions, or not), and isotropies are closely connected with homogeneity. Those observational investigations of isotropy which have been carried out so far support the conclusion that the universe is approximately spherically symmetric about us.

In particular, it has been shown that extragalactic radio sources are distributed approximately isotropically, and that the recently observed microwave background radiation, where it has been examined,

is very highly isotropic (see chapter 10 for further discussion).

It is possible to write down and examine the metrics of all space-times which are spherically symmetric; particular examples are the Schwarzschild and Reissner-Nordstrom solutions (see §5.5); however these are asymptotically flat spaces. In general, there can exist at most two points in a spherically symmetric space from which the space looks spherically symmetric. While these may serve as models of space-time near a massive body, they can only be models of the universe consistent with the isotropy of our observations if we are located near a very special position. The exceptional cases are those in which the universe is isotropic about *every* point in space time; so we shall interpret the Copernican principle as stating that the universe is approximately spherically symmetric about every point (since it is approximately spherically symmetric about us).

As has been shown by Walker (1944), exact spherical symmetry about every point would imply that the universe is spatially homogeneous and admits a six-parameter group of isometries whose surfaces of transitivity are spacelike three-surfaces of constant curvature. Such a space is called a ***Robertson-Walker*** (or ***Friedmann)*** space (Minkowski space, de Sitter space and anti-de Sitter space are all special cases of the general Robertson-Walker spaces). Our conclusion, then, is that these spaces are a good approximation to the large scale geometry of space-time in the region that we can observe.

In the Robertson-Walker spaces, one can choose coordinates so that the metric has the form

$$ds^2 = -dt^2 + S^2(t)d\sigma^2,$$

where $d\sigma^2$ is the metric of a three-space of constant curvature and is independent of time. The geometry of these three-spaces is qualitatively different according to whether they are three-spaces of constant positive, negative or zero curvature; by rescaling the function *S,* one can normalize this curvature *K* to be +1 or -1 in the first two cases. Then the metric $d\sigma^2$ can be written

$$d\sigma^2 = d\chi^2 + f^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2),$$

where

$$f(\chi)=\begin{cases}\sin\chi\_if\_K=+1,\\ \chi\_if\_K=0,\\ \sinh\chi\_if\_K=-1.\end{cases}$$

The coordinate χ runs from 0 to ∞ if $K = 0$ or -1, but runs from 0 to 2π if $K = + 1$. When $K = 0$ or - 1, the three-spaces are diffeomorphic to R^3 and so are 'infinite', but when $K = + 1$ they are diffeomorphic to a three-sphere S^3 and so are compact ('closed' or 'finite'). One could identify suitable points in these three-spaces to obtain other global topologies; it is even possible to do this, in the case of negative or zero curvature, in such a way that the resulting three-space is compact (Löbell (1931)). However such a compact surface of constant negative curvature would have no continuous groups of isometries (Yano and Bochner (1953)) - although Killing vectors exist at each point, they would not determine any global Killing vector fields and the local groups of isometries they generate would not link up to form global groups. In the case of zero curvature, a compact space could only have a three-parameter group of isometries. In neither case would the resulting space-time be isotropic. We shall not make such identifications, as our original reason for considering these spaces was that they were isotropic (and so had a six-parameter group of isometries). In fact the only identifications which would not result in an anisotropic space would be to identify antipodal points on S^3 in the case of constant positive curvature.

The symmetry of the Robertson-Walker solutions requires that the energy-momentum tensor has the form of a perfect fluid whose density μ, and pressure p are functions of the time coordinate t only, and whose flow lines are the curves (χ,θ,ϕ) constant (so the coordinates are comoving coordinates). This fluid can be thought of as a smoothed out approximation to the matter in the universe; then the function $S(t)$ represents the separation of neighbouring flow lines, that is, of 'nearby' galaxies.

The equation of conservation of energy (3.9) in these spaces takes the form

$$\dot{\mu}=-3(\mu+p)S'/S\,. \qquad (5.10)$$

The Raychaudhuri equation (4.26) takes the form

$$4\pi(\mu+3p)-\Lambda=-3S''/S\,. \qquad (5.11)$$

The remaining field equation (which is essentially (2.35)) can be written

$$3S'^2 = 8\pi(\mu S^3)/S + \Lambda S^2 - 3K \,. \qquad (5.12)$$

Whenever $S' \neq 0$, (5.12) can in fact be derived, with an arbitrary value of the constant *K,* as a first integral of (5.10), (5.11); so the real effect of this field equation is to identify the integration constant as the curvature of the metric $d\sigma^2$ of the three-spaces $\{t = const.\}$.

It is reasonable to assume (cf. the energy conditions, §4.3) that μ is positive and *p* is non-negative. (In fact, present estimates are $10^{-29}\, gm \cdot cm^3 \geq \mu_0 \geq 10^{-31}\, gm \cdot cm^3$, $\mu_0 >> p_0 >> 0$. Then, if Λ is zero, (5.11) shows that *S* cannot be constant; in other words the field equations then imply the universe is either expanding or contracting. Observations of other galaxies show, as first found by Slipher and Hubble, that they are moving away from us, and so indicate that the matter in the universe is expanding at the present time. Current observations give the value of *S'*/*S* at the present time as

$$H \equiv (S'/S)\big|_0 \approx 10^{-10}\, year^{-1},$$

believed correct to within a factor 2. From this, (5.11) shows that if Λ is zero, *S* must have been zero a finite time *to* ago (that is, a time *to* measured along the world-line of our galaxy) where $t_0 < H^{-1} \approx 10^{10}\, years$.

From (5.10) it follows that the density decreases as the universe expands, and conversely that the density was higher in the past, increasing without bound as $S \to 0$. This is therefore not merely a coordinate singularity (as for example, in anti-de Sitter universe expressed in coordinates (5.9)); the fact that the density is infinite there shows that some scalar defined by the curvature tensor is also infinite. It is this that makes the singularity so much worse than in the corresponding Newtonian situation; in both cases the world-lines of all the particles intersect in a point and the density becomes infinite, but here space-time itself becomes singular at the point *S* = 0. We must therefore exclude this point from the space-time manifold, as no known physical laws could be valid there.

This singularity is the most striking feature of the Robertson-Walker solutions. It occurs in all models in which $\mu + 3p$ is positive and Λ is negative, zero, or with not too large a positive value. It would

imply that the universe (or at least that part of which we can have any physical knowledge) had a beginning a finite time ago. However this result has here been deduced from the assumptions of exact spatial homogeneity and spherical symmetry. While these may be reasonable approximations on a large enough scale at the present time, they certainly do not hold locally. One might think that, as one traced the evolution of the universe back in time, the local irregularities would grow and could prevent the occurrence of a singularity, causing the universe to 'bounce' instead. Whether this could happen, and whether physically realistic solutions with inhomogeneities would contain singularities, is a central question of cosmology and constitutes the principal problem dealt with in this book; it will turn out that there is good evidence to believe that the physical universe does in fact become singular in the past.

If some suitable relation between p and μ is specified, (5.10) can be integrated to give μ as a function of S. In fact the pressure is very small at the present epoch. If one takes it and Λ to be zero, one finds from (5.10)

$$\frac{4\pi}{3}\mu = \frac{M}{S^3},$$

where M is a constant, and (5.12) becomes

$$3S'^2 - 6M/S = -3K \equiv E/M \ . \qquad (5.13$$

The first equation expresses the conservation of mass when the pressure is zero, while the second (the ***Friedmann equation)*** is an energy conservation equation for a comoving volume of matter; the constant E represents the sum of the kinetic and potential energies. If E is negative (i.e., K is positive), S will increase to some maximum value and then decrease to zero; if E is positive or zero (i.e., K is negative or zero), S will increase indefinitely.

The explicit solutions of (5.13) have a simple form if given in terms of a rescaled time parameter $\tau(t)$, defined by

$$d\tau/dt = S^{-1}(t)\ ; \qquad (5.14)$$

they take the form

$$S = (E/3)(\cosh\tau - 1)\ , \quad t = (E/3)(\sinh\tau - \tau)\ , \quad \text{if} \quad K = -1\ ;$$

$$S = \tau^2\ , \quad t = \tau^3/3\ , \quad \text{if} \quad K = 0\ ;$$

$$S = (-E/3)(1 - \cos\tau)\ , \quad t = (-E/3)(\tau - \sin\tau)\ , \quad \text{if} \quad K = 1\ .$$

(The case $K = 0$ is the Einstein-de Sitter universe; clearly $S \propto t^{2/3}$.)

If p is non-zero but positive, the qualitative behaviour is the same. In particular if $p = (\gamma - 1)\mu$ where γ is a constant, $1 \le \gamma \le 2$, one finds $(4/3)\pi\mu = M / S^{3\gamma}$, and the solution of (5.12) near the singularity takes the form

$S \propto t^{(2/3)\gamma}$.

If Λ is negative, the solution expands from an initial singularity, reaches a maximum and then recollapses to a second singularity. If Λ is positive, then for K == 0 or -1 the solution expands forever and asymptotically approaches the steady state model. For K = + 1 there are several possibilities. If Λ is greater than some value Λ_{crit} ($\Lambda_{crit} = (-E/3M)^3 / (3M)^2$ if $p = 0$) the solution will start from an initial singularity and will expand forever asymptotically approaching the steady state model. If $\Lambda = \Lambda_{crit}$ there is a static solution, the ***Einstein static universe***. (The metric form (5.7) is that of the particular Einstein static solution for which $\mu + p = (4\pi)^{-1}$, $\Lambda = 1 + 8\pi p$.) There is also a solution which starts from an initial singularity and asymptotically approaches the Einstein universe, and one which starts from the Einstein universe in the infinite past and expands forever. If $\Lambda < \Lambda_{crit}$ there are two solutions - one expands from an initial singularity and then recollapses to a second singularity; the other contracts from an infinite radius in the infinite past, reaches a minimum radius, and then re-expands. This and the universe asymptotic to the static universe in the infinite past are the only solutions which could represent the observed universe and which do not have a singularity. In these models, S'' is always positive, and this seems to be in conflict with observations of redshifts of distant galaxies (Sandage (1961, 1968)). Also, the maximum density in these models would not have been very much larger than the present density. This would make it difficult to understand phenomena such as the microwave background radiation and the cosmic abundance of helium, which seem to point to a very hot dense phase in the history of the universe.

Just as in the previous cases we have studied, one can find conformal mappings of the Robertson-Walker spaces into the Einstein static space. We use the coordinate r defined by (5.14) as a time coordinate; then the metric takes the form

$$ds^2 = S^2(\tau)\{-d\tau^2 + d\chi^2 + f^2(\chi)(d\theta^2 + \sin^2\theta d\phi^2)\}. \qquad (5.15)$$

In the case $K = +1$, this is already conformal to the Einstein static space (put $\tau = t'$, $\chi = r'$ to agree with the notation of(5.7)). Thus these spaces are mapped into precisely that part of the Einstein static space determined by the values taken by τ. When $p = \Lambda = 0$, $p = \mathrm{A} = 0$, τ lies in the range $0 < \tau < \pi$, so the whole space is mapped into this region in the Einstein static universe while its boundary is mapped into the three-spheres $\tau = 0$, $\tau = \pi$. (If $p > 0$, it is mapped into a region for which τ takes values $0 < \tau < \pi$, for some number a.) In the case $K = 0$, the same coordinates represent the space as conformal to flat space (see (5.15)), so on using the conformal transformations of §5.1, one obtains these spaces mapped into some part of the diamond representing Minkowski space-time in the Einstein static universe (see figure 14); the actual region is again determined by the values taken by τ. When $\Lambda = 0$, $0 < \tau < \infty$, so this space (which is the Einstein-de Sitter space when $p = 0$) is conformal to the half $t' > 0$ of the diamond which represents Minkowski space-time. In the case $K = -1$, one obtains the metric conformal to part of the region of the Einstein static space for which $\pi/2 \geq t' + r' \geq -\pi/2$, $\pi/2 \geq t' - r' \geq -\pi/2$, on defining

$$t' = \arctan(\tanh\frac{1}{2}(\tau + \chi)) + \arctan(\tanh\frac{1}{2}(\tau - \chi)),$$

$$r' = \arctan(\tanh\frac{1}{2}(\tau + \chi)) - \arctan(\tanh\frac{1}{2}(\tau - \chi)).$$

The part of this diamond-shaped region covered depends on the range of τ; when $\Lambda = 0$, the space is mapped into the upper half.

One thus obtains these spaces and their boundaries conformal to some (generally finite) region of the Einstein static space, see figure 21 (i). However there is an important difference from the previous cases: part of the boundary is not 'infinity' in the sense it was previously, but represents the singularity when $S = 0$. (The conformal factor can be thought of as making infinity finite by giving an infinite compression, but making the singular point $S = 0$ finite by an infinite expansion.) In fact this makes little difference to the conformal diagrams; one can give the Penrose diagrams as before (see figures 21 (ii) and 21 (iii)) In each case when $p \geq 0$ the singularity at $t = 0$ is represented

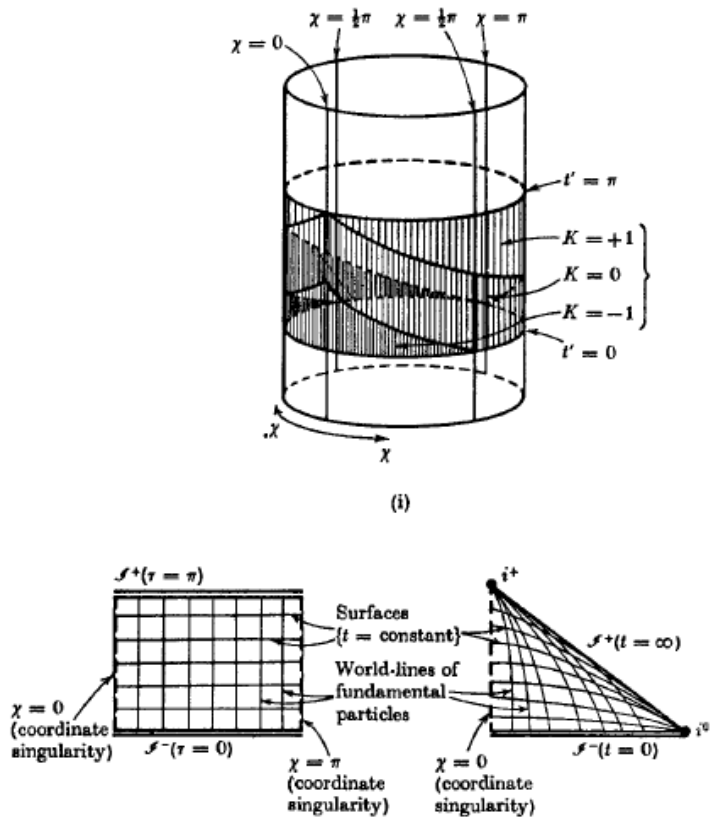


FIGURE 21

(i) The Robertson–Walker spaces ($p = \Lambda = 0$) are conformal to the regions of the Einstein static universe shown, in the three cases $K = +1, 0$ and -1.

(ii) Penrose diagram of a Robertson–Walker space with $K = +1$ and $p = \Lambda = 0$.

(iii) Penrose diagram of a Robertson–Walker space with $K = 0$ or -1 and $p = \Lambda = 0$.

by a spacelike surface; this corresponds to the existence of particle horizons (defined precisely as in §5.2) in these spaces. Also when $K = +1$ the future boundary is spacelike, implying the existence of event horizons for the fundamental observers; when $K = 0$ or - 1 and $\Lambda = 0$, future infinity is null and there are no future event horizons for the fundamental observers in these spaces.

At this stage, one should examine the following question: anti-de Sitter space could be expressed in the Robertson-Walker form (5.9) and then expressed conformally as part of the Einstein static universe. When one did so, one found that the Robertson-Walker coordinates covered only a small part of the full space-time. That is to say, the space-time described by the Robertson-Walker coordinates could be extended. One should therefore show that the Robertson-Walker universes in which there is matter are in fact inextendible. This

follows because one can show that if $\mu > 0$, $p \geq 0$ and $\boldsymbol{X}$ is any vector at any point *q,* the geodesic $\gamma(v)$ through $q = \gamma(0)$ in the direction of $\boldsymbol{X}$ is such that either

(i) $\gamma(v)$ can be extended to arbitrary positive values of *v,* or

(ii) there is some $v_0 > 0$ such that the scalar invariant

$$(R_{ij} - \frac{1}{2} R g_{ij})(R^{ij} - \frac{1}{2} R g^{ij}) = (\mu + \Lambda)^2 + 3(p - \Lambda)^2$$

is unbounded on $\gamma([0, v_0))$.

It is now clear that the surfaces $\{t = const\}$ are Cauchy surfaces in these spaces. Further one sees that the singularity is universal in the following sense: all timelike and null geodesics through any point in the space approach it for some finite value of their affine parameter.

5.4 Spatially homogeneous cosmological models

We have seen that there are singularities in any Robertson-Walker space-time in which $\mu > 0$, $p \geq 0$ and Λ is not too large. However one could not conclude from this that there would be singularities in more realistic world models which allow for the fact that the universe is *not* homogeneous and isotropic. In fact, one does not expect to find that the universe can be very accurately described by *any* attainable exact solution. However one can find exact solutions, less restricted than the Robertson-Walker solutions, which may be reasonable models of the universe, and see if singularities occur in them or not; the fact that singularities do occur in such models gives an indication that the existence of singularities may be a general property of *all* space-times which can be regarded as reasonable models of the universe.

A simple class of such solutions are those in which the requirement of isotropy is dropped but the requirement of *spatial homogeneity* (the strict Copernican principle) is retained (although the universe seems approximately isotropic at the present time, there might have been large anisotropies at an earlier epoch). Thus in these models one assumes there exists a group of isometries *Gr* whose orbits in some part of the model are spacelike hypersurfaces. (The orbit of a point *p* under the group *Gr* is the set of points into which *p* is moved by the action of all elements of the group.) These models may be constructed locally by

well-known methods; see Heckmann and Schücking (1962) for the case $r = 3$, and Kantowski and Sachs (1967) for the case $r = 4$ (if $r > 4$, the space-time is necessarily a Robertson-Walker space).

The simplest spatially homogeneous space-times are those in which the group of isometries is Abelian; the group is then of type I in the classification given by Bianchi (1918), so we call these *Bianchi 1* spaces. We discuss Bianchi I spaces in some detail, and then give a theorem showing singularities will occur in all non-empty spatially homogeneous models in which the timelike convergence condition (§ 4.3) is satisfied.

Suppose the spatially homogeneous space-time has an Abelian isometry group; for simplicity we assume $\Lambda = 0$ and that the matter content is a pressure-free perfect fluid ('dust'). Then there exist comoving coordinates *(t, x, y,* z) such that the metric takes the form

$$ds^2 = -dt^2 + X^2(t)dx^2 + Y^2(t)dy^2 + Z^2(t)dz^2 \qquad (5.16)$$

Defining the function *S(t)* by $S^3 = XYZ$, the conservation equations show that the density of matter is given by $(4/3)\pi\mu = M/S^3$, where *M* is a suitably chosen constant. The general solution of the field equations can be written

$$X = S(t^{2/3}/S)^{2\sin\alpha},$$
$$Y = S(t^{2/3}/S)^{2\sin(\alpha+2\pi/3)},$$
$$Z = S(t^{2/3}/S)^{2\sin(\alpha+4\pi/3)},$$

where S is given by

$$S^3 = \frac{9}{2}Mt(t+\Sigma);$$

Σ (> 0) is a constant determining the magnitude of the anisotropy (we exclude the isotropic case ($\Sigma = 0$), which is the Einstein-de Sitter universe (§ 5.3)), and α ($-\pi/6 < \alpha \le \pi/2$) is a constant determining the direction in which the most rapid expansion takes place. The average rate of expansion is given by

$$\frac{S'}{S} = \frac{2}{3t}\frac{t+\Sigma/2}{t+\Sigma};$$

the expansion in the x-direction is

$$\frac{X'}{X} = \frac{2}{3t}\frac{t+\Sigma(1+2\sin\alpha)/2}{t+\Sigma},$$

and the expansions *Y'/Y, Z'/Z* in the *y,* z directions are given by

similar expressions in which α is replaced by $\alpha + 2\pi/3$, $\alpha + 4\pi/3$, respectively.

The solution expands from a highly anisotropic singular state at $t = 0$, reaching a nearly isotropic phase for large t when it is nearly the same as the Einstein-de Sitter universe. The average length S increases monotonically as t increases, its initial high rate of change *(* $S \propto t^{1/3}$ for small *t)* decreasing steadily *(* $S \propto t^{2/3}$ for large *t)*. Thus the universe evolves more rapidly, at early times, than its isotropic equivalent.

Suppose one considers the time-reverse of the model, and follows this forward in time towards the singularity. The initially almost isotropic contraction will become very anisotropic at late times. For general values of α, i.e. $\alpha \neq \pi/2$, the term $1 + 2\sin(\alpha + 4\pi/3)$ will be negative. Thus the collapse in the z-direction would halt, and, for sufficiently early times, be replaced by an expansion, the rate of expansion becoming indefinitely large for early enough times. In the *x*- and y-directions, on the other hand, the collapse would continue monotonically towards the singularity. Thus if one considers the forward direction of time in the original model, one has a 'cigar' singularity: matter collapses in along the z-axis from infinity, halts, and then starts re-expanding, while in the *x*- and y-directions the matter expands monotonically at all times. If one could receive signals from early enough times in such a model, one would see a maximum redshift in the z-direction, at earlier times matter in this direction being observed with progressively smaller redshifts and then with indefinitely increasing *blue*-shifts.

The behaviour in the exceptional case $\alpha = \pi/2$ is rather different. In this case, the terms $1 + 2\sin(\alpha + 2\pi/3)$ and $1 + 2\sin(\alpha + 4\pi/3)$ both vanish. Thus the expansions in the axis directions are

$$\frac{X'}{X} = \frac{2}{3t}\frac{t + 3\Sigma/2}{t + \Sigma}, \qquad \frac{Y'}{Y} = \frac{Z'}{Z} = \frac{2}{3}\frac{1}{t + \Sigma}.$$

If one follows the time-reversed model, the rate of collapse in the *y*- and z-directions slows asymptotically down to zero, while the rate of collapse in the x-direction increases indefinitely. In the original model, one has a 'pancake' singularity; matter expands monotonically in all directions, starting from an indefinitely high expansion rate in the x-direction but from zero expansion rates in the *y*- and z-directions.

Indefinitely high redshifts would be seen in the x-direction, but there would be limiting redshifts in the *y*- and z-directions.

Further examination shows that in the general ('cigar') case, there is a particle horizon in every direction despite the anisotropic expansion. However in the exceptional ('pancake') case, no horizon occurs in the x-direction; in fact the particles that can be seen by an observer at the origin at time *to* are characterized by coordinate values *(x, y,* z) lying within the infinite cylinder

$x^2 + y^2 < \rho^2$

where

$$\rho = \frac{2}{3M}\left\{\left(\frac{9M}{2}(t_0 + \Sigma)\right)^{1/2} - \left(\frac{9M}{2}\Sigma\right)^{1/2}\right\}.$$

While we have here considered these models for vanishing pressure and Λ term only, properties of these spaces with more realistic matter contents can easily be obtained; for example if one has either a perfect fluid with $p = (\gamma - 1)\mu$, γ a constant ($1 < \gamma < 2$), or a mixture of a photon gas and matter with pressure $p \le \mu/3$, the behaviour near the singularity is the same as in the dust case.

An interesting consequence of the non-existence of a particle horizon in the x-direction in the exceptional ('pancake') case, is that one can extend the solution continuously across the singularity. We shall show this explicitly in the case of the dust solution.

The metric takes the form (5.16) where now

$$X(t) = t\left(\frac{9}{2}M(t+\Sigma)\right)^{-1/3}, \quad Y(t) = X(t) = \left(\frac{9}{2}M(t+\Sigma)\right)^{2/3}. \quad (5.17)$$

We now choose new coordinates τ, η which satisfy the equations

$$\tanh(2x/9M\Sigma) = \eta/\tau, \quad \exp\left(\frac{4}{9M}\int_0^t \frac{dt}{X(t)}\right) = \tau^2 - \eta^2,$$

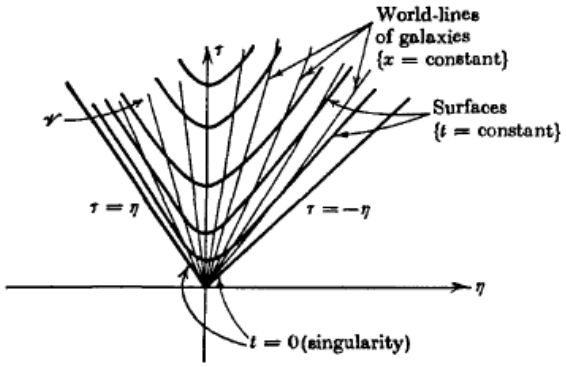
One then finds that the space with metric (5.16), (5.17) is given in the new coordinates by

$$ds^2 = A^2(t)(-d\tau^2 + d\eta^2) + B^2(t)(dy^2 + dz^2) \quad (5.18)$$

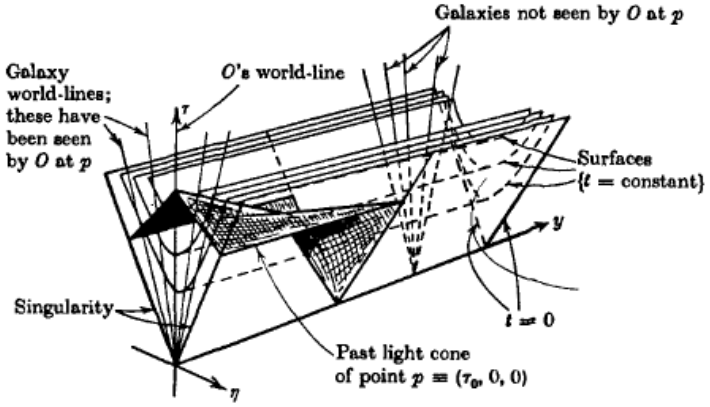
where

$$A(t) = \exp\left(-\frac{t+\Sigma}{\Sigma}\right)\left(\frac{9}{2}M(t+\Sigma)\right)^{-1/3},$$

$$B(t)=\left(\frac{9}{2}M(t+\Sigma)\right)^{2/3} \tag{5.19}$$



(i)



(ii)

FIGURE 22. Dust-filled Bianchi I space with a pancake singularity.
(i) The (τ,η) plane; null lines are at $\pm 45°$.
(ii) A half-section of the space in (τ, η, y) coordinates (the z-coordinate is suppressed), showing the past light cone of the point $p = (\tau_0, 0, 0)$. There is a particle horizon in the y-direction but not in the x- (i.e. η) direction.

the whole space (for t > 0) being mapped into the region *V* defined by $\tau > 0$, $\tau^2 - \eta^2 > 0$. The function $t(\tau,\eta)$ is now defined implicitly as the solution of the equation

$$\tau^2 - \eta^2 = \frac{9}{2}Mt^2 \exp\frac{2(t+\Sigma)}{\Sigma} \tag{5.20}$$

for which t>0. The (τ,η) plane is given in conformally flat coordinates. The region *V* in this plane, bounded by the surface $t = 0$, is shown in figure 22. In this diagram: the world-lines of the particles are straight lines diverging from the origin.

The functions *A(t), B(t)* are continuous as $t \to 0$ from above. One can therefore extend the solution continuously to the whole (τ,η) plane

by specifying that (5.19) holds everywhere, (5.20) holds inside V, and that

$$t(\tau,\eta) = 0$$

holds outside V. Then (5.18) is a C^0 metric which is a solution of the field equations equivalent to (5.16), (5.17) inside V, and is a flat space-time outside V. However the solution is not C^1 across the boundary of V, and in fact the density of matter becomes infinite on this boundary (as $S \to 0$ there). Since the first derivatives are not square integrable, the Einstein field equations cannot be interpreted on the boundary even in a distributional sense (see §8.4). While the extension onto the boundary is unique, it is in no way unique beyond the boundary. We have carried out the extension in the case of dust; a similar extension could be carried out if one had a mixture of matter and radiation.

Let us now return to considering general non-empty spatially homogeneous models. The existence of a singularity in these models will follow directly from Raychaudhuri's equation if the motion of the matter is geodesic and without rotation (as must be the case, for example, if the world-lines are orthogonal to the surfaces of homogeneity) and the timelike convergence condition is satisfied; however there exist such spaces in which the matter accelerates and rotates, and either of these factors could possibly prevent the existence of a singularity. The following result, which is an improved version of a theorem of Hawking and Ellis (1965), shows that in fact neither acceleration nor rotation can prevent the existence of singularities in these models.

Theorem

(M, g) cannot be timelike geodesically complete if:

(1) $R_{ab}K^aK^b > 0$ *for all timelike and null vectors K (this is true if the energy-momentum tensor is type I (§ 4.3) and* $\mu + p_i > 0$,

$\mu + \sum_i p_i - 4\pi\Lambda > 0$*);*

(2) there exist equations of motion for the matter fields such that the Cauchy problem has a unique solution (see chapter 7);

(3) the Cauchy data on some spacelike three-surface H is invariant

under a group of diffeomorphisms of H which is transitive on H.

Since the intrinsic geometry of H is invariant under a transitive group of diffeomorphisms, these are isometries and H is complete, i.e. cannot have any boundary. It can be shown (see §6.5) that if there is a non-spacelike curve which intersects H more than once, then there exists a covering manifold $\hat{M}$ of M in which each connected component of the image of H will not intersect any non-spacelike curve more than once. We shall assume that $\hat{M}$ is timelike geodesically complete, and show that this is inconsistent with conditions (1), (2) and (3).

Let $\hat{H}$ be a connected component of the image of H in $\hat{M}$. By (3), the Cauchy data on $\hat{H}$ is homogeneous. Therefore by condition (2), the Cauchy development of any region of $\hat{H}$ is isometric to the Cauchy development of any other similar region of $\hat{H}$. This implies that the surfaces $\{s =$constant$\}$ are homogeneous if they lie within the Cauchy development of $\hat{H}$, where s is the distance from $\hat{H}$ measured along the geodesic normals to $\hat{H}$. These surfaces must lie either entirely within or entirely outside the Cauchy development of $\hat{H}$, as otherwise there would be equivalent regions in $\hat{H}$ which had inequivalent Cauchy evolutions. The surfaces $\{s =$ constant$\}$ will lie in the Cauchy development of $\hat{H}$ as long as they remain spacelike, because the boundary of the Cauchy development of $\hat{H}$ (if it exists) must be null (§ 6.5).

The geodesics orthogonal to $\hat{H}$ will be orthogonal to the surfaces $\{s =$ constant$\}$, as a vector representing the separation of points equal distances along neighbouring geodesics will remain orthogonal to the geodesics if it is so initially. As in §4.1, one can represent the spatial separation of neighbouring geodesics orthogonal to $\hat{H}$ by a matrix A which is the unit matrix on $\hat{H}$. By homogeneity, it will be constant on the surfaces $\{s =$ constant$\}$ while these lie in the Cauchy development of $\hat{H}$. While A is non-degenerate, the map from $\hat{H}$ to a surface $\{s =$ constant$\}$ defined by the normal geodesics will be of rank three and so the surfaces will be spacelike three-surfaces contained within the Cauchy development of $\hat{H}$. The expansion
$\theta = (\det A)^{-1} d(\det A)/ds$
of these geodesics obeys Raychaudhuri's equation (4.26) with the

vorticity and acceleration zero. By condition (1), $R_{ab}V^aV^b$ is positive for all timelike vectors V^a. Thus θ will become infinite and A will be degenerate for some finite positive or negative value *s0* of *s*. The map from $\hat{H}$ to the surface *s = s0* can have at most rank two; there will therefore be at least one vector field Z on $\hat{H}$ such that AZ =0. The integral curves of this vector field are curves in $\hat{H}$ which are mapped by the geodesic normals to one point in the surface *s = s0*, Thus this surface will be at most two-dimensional. As the geodesics lie in the Cauchy development of $\hat{H}$ for $|s| < |s_0|$, the surface *s = s0* will lie in the Cauchy development or on the boundary of the Cauchy development of $\hat{H}$. By condition (1), the energy-momentum tensor has a unique timelike eigenvector at each point. These eigenvectors will form a C^1 timelike vector field whose integral curves may be thought· of as representing the flow lines of the matter. As the surface *s = s0* lies in the Cauchy development of $\hat{H}$ or on its boundary, all the flow lines that pass through it must intersect $\hat{H}$. But then as $\hat{H}$ is homogeneous, all the flow lines that pass through $\hat{H}$ must pass through *s = s0*, Thus the flow lines define a diffeomorphism between $\hat{H}$ and the surface *s = so*. This is impossible, as $\hat{H}$ is three-dimensional and *s = so* is two-dimensional.

In fact, if all the flow lines were to pass through a two-dimensional surface, one would expect the matter density to become infinite. We have now seen that a large scale rotation or acceleration cannot, by itself, prevent the occurrence of singularities in a universe model obeying the strict Copernican principle. In later theorems we shall see that irregularities are in general also unable to prevent the occurrence of singularities in world models.

5.5 The Schwarzschild and Reissner-Nordström solutions

While the spatially homogeneous solutions may be good models for the large scale distribution of matter in the universe, they are inadequate for describing, for example, the local geometry of space-time in the solar system. One can describe this geometry to a good approximation by the Schwarzschild solution, which represents the spherically symmetric empty space-time outside a spherically symmetric massive

body. In fact, all the experiments which have so far been carried out to test the difference between the General Theory of Relativity and Newtonian theory are based on predictions by this solution.

The metric can be given in the form

$$ds^2 = -\left(1-\frac{2m}{r}\right)dt^2 + \left(1-\frac{2m}{r}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2), \qquad (5.21)$$

where $r > 2m$. It can be seen that this space-time is static, i.e., $\partial/\partial t$ is a timelike Killing vector which is a gradient, and is spherically symmetric, i.e., is invariant under the group of isometries *S0(3)* operating on the spacelike two-spheres {*t, r* constant} (cf. appendix B). The coordinate *r* in this metric form is intrinsically defined by the requirement that $4\pi r^2$ is the area of these surfaces of transitivity. The solution is asymptotically flat as the metric has the form $g_{ab} = \eta_{ab} + O(1/r)$ for large *r*. Comparison with Newtonian theory (cf. §3.4) shows that *m* should be regarded as the gravitational mass, as measured from infinity, of the body producing the field. It should be emphasized that this solution is unique: if any solution of the vacuum field equations is spherically symmetric, it is locally isometric to the Schwarzschild solution (although it may of course look totally different if it is given in some other coordinate system; see appendix B and Bergmann, Cahen and Komar (1965)).

Normally one would regard the Schwarzschild metric for *r* greater than some value $ro > 2m$ as being the solution outside some spherical body, the metric inside the body $(r < ro)$ having a different form determined by the energy-momentum tensor of the matter in the body. However it is interesting to see what happens when the metric is regarded as an empty space solution for all values of *r*.

The metric is then singular when $r = 0$ and when $r = 2m$ (there are also the trivial singularities of polar coordinates when θ and $\theta = \pi$). One must therefore cut $r = 0$ and $r = 2m$ out of the manifold defined by the coordinates (t, r, θ, ϕ), since in §3.1 we took space-time to be represented by a manifold with a Lorentz metric. Cutting out the surface $r = 2m$ divides the manifold into two disconnected components for which $0 < r < 2m$ and $2m < r < \infty$. Since we took the space-time manifold to be connected, we must consider only one of these components and the obvious one to choose is the one for $r > 2m$, which

represents the external field. One must then ask whether this manifold M with the Schwarzschild metric g is extendible, i.e. whether there is a larger manifold M' into which M can be imbedded and a suitably differentiable Lorentz metric g' on M' which coincides with g on the image of M. The obvious place where M might be extended is where r tends to $2m$. A calculation shows that although the metric is singular at $r = 2m$ in the Schwarzschild coordinates (t, r, θ, ϕ), no scalar polynomials of the curvature tensor and the metric diverge as $r \to 2m$. This suggests that the singularity at $r = 2m$ is not a real physical singularity, but rather one which is a result of a bad choice of coordinates.
To confirm this, and to show that (M, g) can be extended, define

$$r^* \equiv \int \frac{dr}{1 - 2m/r} = r + 2m \log(r - 2m) .$$

Then

$$v \equiv t + r^*$$

is an advanced null coordinate, and

$$w \equiv t - r^*$$

is a retarded null coordinate. Using coordinates (v, r, θ, ϕ) the metric, takes the Eddington-Finkelstein form g' given by

$$ds^2 = -\left(1 - \frac{2m}{r}\right) dv^2 + 2dvdr + r^2 (d\theta^2 + \sin^2 \theta d\phi^2) . \qquad (5.22)$$

The manifold M is the region $2m < r < \infty$, but the metric (5.22) is non-singular and indeed analytic on the larger manifold M' for which $0 < r < \infty$. The region of (M', g') for which $0 < r < 2m$ is in fact isometric to the region of the Schwarzschild metric for which $o < r < 2m$. Thus by using different coordinates, i.e. by taking a different manifold, we have extended the Schwarzschild metric so that it is no longer singular at $r = 2m$. In the manifold M' the surface $r = 2m$ is a null surface, as can be seen from the Finkelstein diagram (figure 23). This is a section (θ, ϕ constant) of the space-time; each point represents a two-sphere of area $4\pi r^2$. Some null cones and radial null geodesics are indicated on this diagram. Surfaces $\{t = \text{constant}\}$ are indicated; one sees that t becomes infinite on the surface $r = 2m$.

This representation of the Schwarzschild solution has the odd feature that it is not time symmetric. One might expect this from the cross term (dvdr) in (5.22); it is qualitatively clear from the Finkelstein

diagram. The most obvious asymmetry is that the surface $r = 2m$ acts

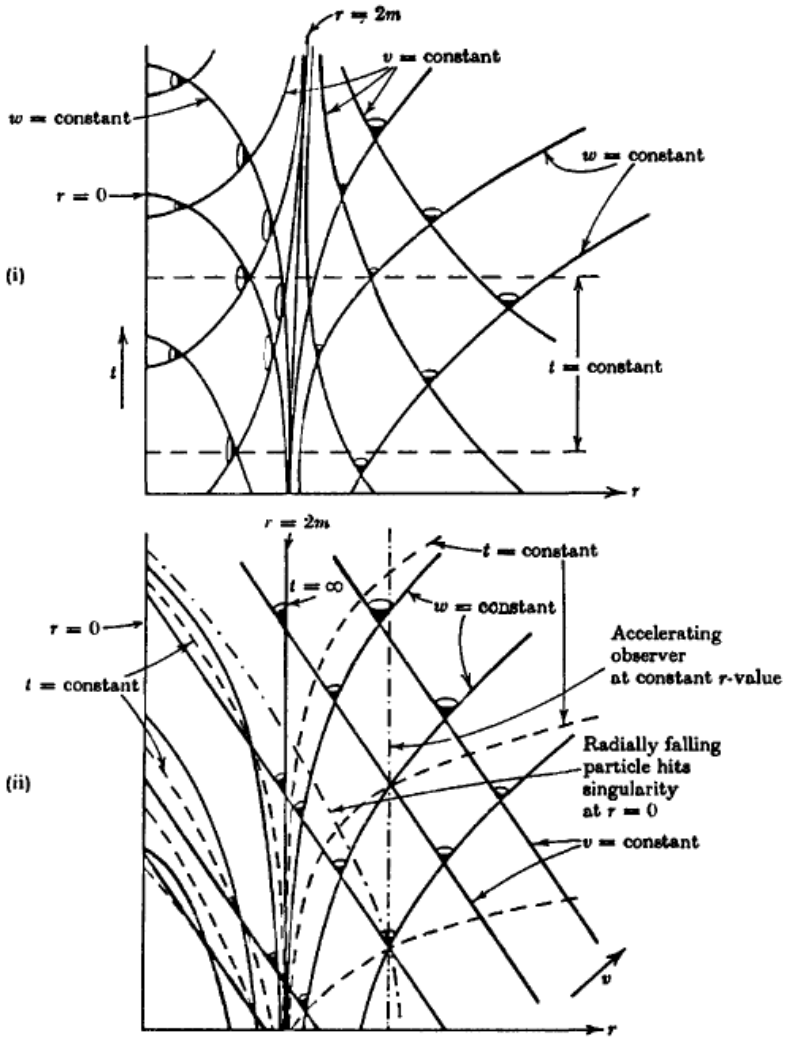


FIGURE 23. Section (θ, ϕ) constant of the Schwarzschild solution.
(i) Apparent singularity at $r = 2m$ when coordinates (t, r) are used.
(ii) Finkelstein diagram obtained by using coordinates (v, r) (lines at 45° are lines of constant v). Surface $r = 2m$ is a null surface on which $t = \infty$.

as a one-way membrane, letting future-directed timelike and null curves cross only from the outside (r> 2m) to the inside (r < *2m).* Any past-directed timelike or null curve in the outside region cannot cross into the inside region. No past-directed timelike or null curve within $r = 2m$ can approach $r = 0$. However any future-directed timelike or null curve which crosses the surface $r = 2m$ approaches $r = 0$ within a finite affine distance. As $r \to 0$, the scalar $R^{abcd}R_{abcd}$ diverges as m^2 / r^6 Therefore $r = 0$ is a real singularity; the pair (M', g') cannot be extended in a C^2 manner or in fact even in a C^0 manner across r =0.

If one uses the coordinate w instead of v, the metric takes the form

g'' given by

$$ds^2 = -\left(1-\frac{2m}{r}\right)dw^2 - 2dwdr + r^2(d\theta^2 + \sin^2\theta d\phi^2) .$$

This is analytic on the manifold M'' defined by the coordinates (w, r, θ, ϕ) for $0 < r < \infty$. Again the manifold M is the region $2m < r < \infty$ and the new region $0 < r < 2m$ is isometric to the region $o< r < 2m$ of the Schwarzschild metric, but the isometry reverses the direction of time. In the manifold M'', the surface $r = 2m$ is again a null surface which acts as a one-way membrane. However this time it acts in the other direction of time, letting only past-directed timelike or null curves cross from the outside $(r> 2m)$ to the inside $(r < 2m)$.

One can in fact make both extensions (M', g') and (M'', g'') simultaneously; that is to say, there is a still larger manifold M^* with metric g^* into which both (M', g') and (M'', g'') can be isometrically imbedded, so that they coincide on the region $r > 2m$ which is isometric to (M, g). A construction of this larger manifold has been given by Kruskal (1960). To obtain it, consider (M, g) in the coordinates (v, w, θ, ϕ); then the metric takes the form

$$ds^2 = -\left(1-\frac{2m}{r}\right)dvdw + r^2(d\theta^2 + \sin^2\theta d\phi^2) ,$$

where r is determined by

$$\frac{1}{2}(v - w) = r + 2m\log(r - 2m) .$$

This presents the two-space (θ, ϕ constant) in null conformally flat coordinates, as the space with metric $ds^2 = -dvdw$ is flat. The most general coordinate transformation which leaves this two-space expressed in such conformally flat double null coordinates is $v' = v'(v)$, $w' = w'(w)$ where v' and w' are arbitrary C^1 functions. The resulting metric is

$$ds^2 = -\left(1-\frac{2m}{r}\right)\frac{dv}{dv'}\frac{dw}{dw'}dv'dw' + r^2(d\theta^2 + \sin^2\theta d\phi^2) .$$

To reduce this to a form corresponding to that obtained earlier for Minkowski space-time, define

$x' = (v' - w')/2$, $\quad t' = (v' + w')/2$.

The metric takes the final form

$$ds^2 = F^2(t',x')(-dt'^2+dx'^2) + r^2(t',x')(d\theta^2 + \sin^2\theta d\phi^2)\,. \qquad (5.23)$$

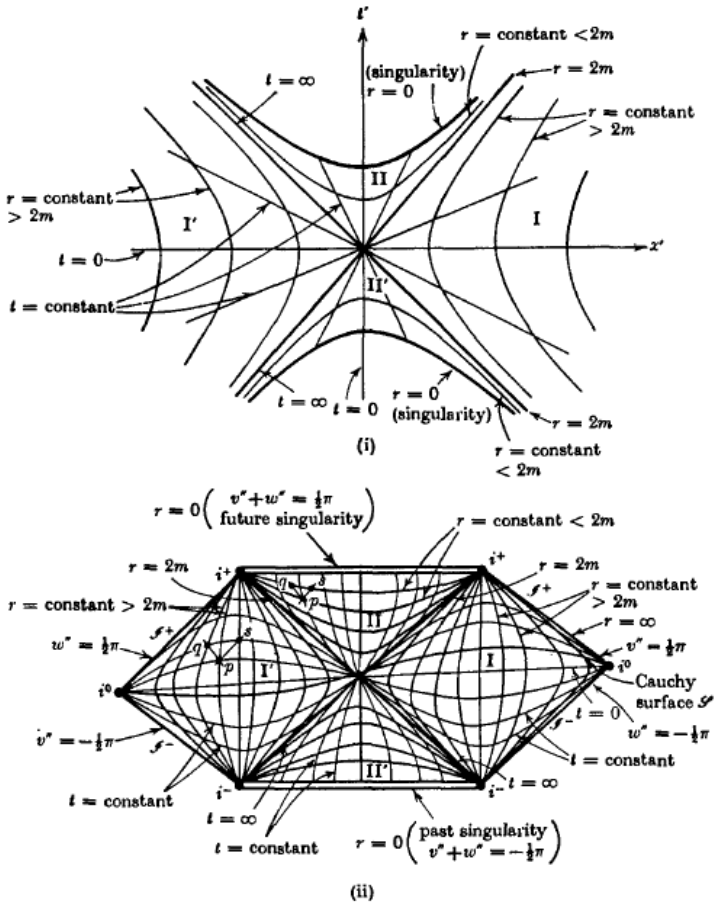


FIGURE 24. The maximal analytic Schwarzschild extension. The θ, ϕ coordinates are suppressed; null lines are at $\pm 45°$. Surfaces $\{r = \text{constant}\}$ are homogeneous.
(i) The Kruskal diagram, showing asymptotically flat regions I and I′ and regions II, II′ for which $r < 2m$.
(ii) Penrose diagram, showing conformal infinity as well as the two singularities.

The choice of the functions v', w' determines the precise form of the metric. Kruskal's choice was $v' = \exp(v/4m)$, $w' = -\exp(-w/4m)$.

Then r is determined implicitly by the equation

$$(t')^2 - (x')^2 = -(r-2m)\exp(r/2m) \qquad (5.24)$$

and F is given by

$$F^2 = \exp(-r/2m)\cdot 16m^2/r\,. \qquad (5.25)$$

On the manifold M^* defined by the coordinates (t', x', θ, ϕ) for $(t')^2 - (x')^2 < 2m$, the functions r and F (defined by (5.24), (5.25) are positive and analytic. Defining the metric g* by (5.23), the region I of (M^*, g^*) defined by $x' > |t'|$ is isometric to (M, g), the region of the Schwarzschild solution for which $r > 2m$. The region defined by $x' > -t'$ (regions I and II in figure 24) is isometric to the advanced

Finkelstein extension (M', g'). Similarly the region defined by $x' > t'$ (regions I and II' in figure 24) is isometric to the retarded Finkelstein extension (M'', g''). There is also a region I', defined by $x' < -|t'|$, which turns out to be again isometric with the exterior Schwarzschild solution (M, g). This can be regarded as another asymptotically flat universe on the other side of the Schwarzschild 'throat'. (Consider the section $t = 0$. The two-spheres $\{r = \text{constant}\}$ behave as in Euclidean space, for large r; however for small r, they have an area which decreases to the minimum value $16\pi m^2$ and then increases again, as the two spheres expand into the other asymptotically flat three-space.) The regions I' and II are isometric with the advanced Finkelstein extension of region I', and similarly I' and II' are isometric with the retarded Finkelstein extension of I', as can be seen from figure 24. There are no timelike or null curves which go from region I to region I'. All future-directed timelike or null curves which cross the part of the surface $r = 2m$ represented here by $t' = |x'|$ approach the singularity at $t' = (2m + (x')^2)^{1/2}$, where $r = 0$. Similarly past-directed timelike or null curves which cross $t' = -|x'|$ approach another singularity at $t' = -(2m + (x')^2)^{1/2}$, where again $r = 0$.

The Kruskal extension (M^*, g^*) is the unique analytic and locally inextendible extension of the Schwarzschild solution. One can construct the Penrose diagram of the Kruskal extension by defining new advanced and retarded null coordinates

$v'' = \arctan(v'(2m)^{-1/2})$, $\quad w'' = \arctan(w(2m)^{-1/2})$

for

$-\pi < v''+w'' < \pi$ and $-\pi/2 < v'' < \pi/2$, $\quad -\pi/2 < w'' < \pi/2$

(see figure 24 (ii)). This may be compared with the Penrose diagram for Minkowski space (figure 15 (ii)). One now has future, past and null infinities for each of the asymptotically flat regions I and I'. Unlike Minkowski space, the conformal metric is continuous but not differentiable at the points i^0.

If we consider the future light cone of any point outside $r = 2m$, the radial outwards geodesic reaches infinity but the inwards one reaches the future singularity; if the point lies inside $r = 2m$, both these geodesics hit the singularity, and the entire future of the point is ended by the singularity. Thus the singularity may be avoided by any

particle outside $r = 2m$ (so it is not 'universal' as it is in the Robertson-Walker spaces), but once a particle has fallen inside $r = 2m$ (in region II) it cannot evade the singularity. This fact will turn out to be closely related to the following property: each point inside region II represents a two-sphere that is a closed trapped surface. This means the following: consider any two-sphere p (represented by a point in figure 24) and two two-spheres q, s formed by photons emitted radially outwards, inwards at one instant from p. The area of q (which is given by $4\pi r^2$) will be greater than the area of p, but the area of s will be less than the area of p, if all three lie in a region $r > 2m$. However if they all lie in the region II where $r < 2m$, then the areas of *both* q and s will be less than the area of p (in the figure, r decreases as one moves from the bottom to the top of region II). In that case, we say that p is a closed trapped surface. Each point inside region II' represents a time-reversed closed trapped surface (the existence of trapped surfaces is a necessary consequence of the fact that the surfaces r = constant are spacelike), and correspondingly all particles in region II' must have come from the singularity in the past. We shall see in chapter 8 that the existence of the singularities is closely related to the existence of the closed trapped surfaces.

The Reissner-Nordström solution represents the space-time outside a spherically symmetric charged body carrying an electric charge (but with no spin or magnetic dipole, so this is not a good representation of the field outside an electron). The energy-momentum tensor is therefore that of the electromagnetic field in the space-time which results from the charge on the body. It is the unique spherically symmetric asymptotically flat solution of the Einstein-Maxwell equations and is locally rather similar to the Schwarzschild solution; there exist coordinates in which the metric has the form

$$ds^2 = -\left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)dt^2 + \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)^{-1} dr^2 + r^2(d\theta^2 + \sin^2\theta d\phi^2) \quad (5.26)$$

where m represents the gravitational mass and e the electric charge of the body. This asymptotically flat solution would normally be regarded as the solution outside the body only, the interior being filled in with some other suitable metric; but it is again interesting to

see what happens if we regard it as a solution for all *r*.

If $e^2 > m^2$ the metric is non-singular everywhere except for the irremovable singularity at *r* = 0; this may be thought of as the point charge which produces the field. If $e^2 \le m^2$, the metric also has singularities at r_+ and r_-, where $r_\pm = m \pm (m^2 - e^2)^{1/2}$; it is regular in the regions defined by $\infty > r > r_+$, $r_+ > r > r_-$ and $r_- > r > 0$ (if $e^2 = m^2$, only the first and third regions exist). As in the Schwarzschild case, these singularities may be removed by introducing suitable coordinates and extending the manifold to obtain a maximal analytic extension (Graves and Brill (1960), Carter (1966)). The major differences that arise are due to the existence of two zeros in the factor in front of dt^2, rather than one as in the Schwarzschild case. In particular this implies that the first and third regions are both static, whereas the second region (when it exists) is spatially homogeneous but is not static.

To obtain the maximally extended manifold, we proceed in steps analogous to those in the Schwarzschild case. Defining the coordinate r^* by

$$r^* = \int dr \Big/ \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right),$$

then for $r > r_+$,

$$r^* = r + \frac{r_+^2}{(r_+ - r_-)}\log(r - r_+) - \frac{r_-^2}{(r_+ - r_-)}\log(r - r_-) \qquad \text{if } e^2 < m^2,$$

$$r^* = r + m\log((r-m)^2) - \frac{2}{r-m} \qquad \text{if } e^2 = m^2,$$

$$r^* = r + m\log(r^2 - 2mr + e^2) + \frac{2}{e^2 - m^2}\arctan\left(\frac{r-m}{e^2 - m^2}\right) \qquad \text{if } e^2 > m^2.$$

Defining advanced and retarded coordinates *v, w* by

$$v = t + r^*, \quad w = t - r^*$$

the metric (5.26) takes the double null form

$$ds^2 = -\left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right)dvdw + r^2(d\theta^2 + \sin^2\theta d\phi^2) \qquad (5.27)$$

In the case $e^2 < m^2$, define new coordinates *v'', w''* by

$$v'' = \arctan\left(\exp\left(\frac{r_+ - r_-}{4r_+^2}v\right)\right), \quad w'' = \arctan\left(-\exp\left(\frac{-r_+ + r_-}{4r_+^2}w\right)\right).$$

Then the metric (5.27) takes the form

$$ds^2 = \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right) 64 \frac{r_+^4}{(r_+ - r_-)} \cos ec\, 2v'' \cos ec\, 2w'' dv'' dw'' + r^2 (d\theta^2 + \sin^2 \theta d\phi^2) \quad (5.28)$$

where r is defined implicitly by

$$\tan v'' \tan w'' = -\exp\left(\left(\frac{r_+ - r_-}{2r_+^2}\right) r\right) (r - r_+)^{1/2} (r - r_-)^{-\alpha/2}$$

and $\alpha = (r_+)^2 (r_-)^2$. The maximal extension is obtained by taking (5.28) as the metric g^*, and M^* as the maximal manifold on which this metric is C^2.

The Penrose diagram of the maximal extension is shown in figure 25. There are an infinite number of asymptotically flat regions, where $r > r_+$; these are denoted by I. These are connected by intermediate

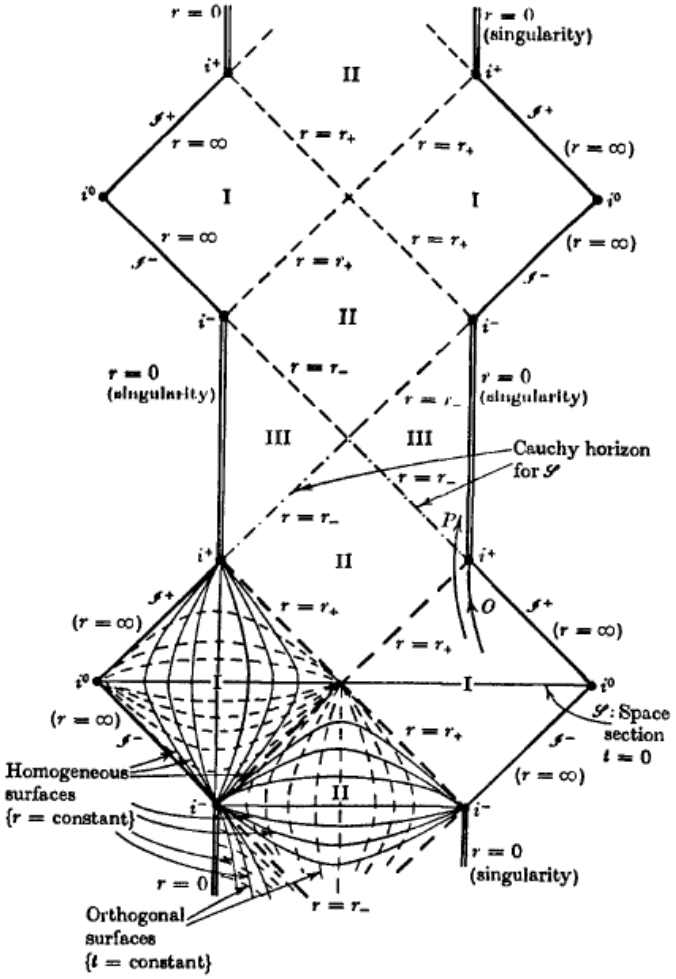


FIGURE 25. Penrose diagram for the maximally extended Reissner–Nordström solution ($e^2 < m^2$). An infinite chain of asymptotically flat regions I ($\infty > r > r_+$) are connected by regions II ($r_+ > r > r_-$) and III ($r_- > r > 0$); each region III is bounded by a timelike singularity at $r = 0$.

regions II and III where $r_+ > r > r_-$ and $r_- > r > 0$, respectively. There is still an irremovable singularity at $r = 0$ in each region III, but unlike in the Schwarzschild solution, it is timelike and so can be avoided by a future-directed timelike curve from a region I which crosses $r = r+$. Such a curve can pass through regions II, III and II and re-emerge into another asymptotically flat region I. This raises the intriguing possibility that one might be able to travel to other universes by passing through the 'wormholes' made by charges. Unfortunately it seems that one would not be able to get back again to our universe to report what one had seen on the other side.

The metric (5.28) is analytic everywhere except at $r = r_-$ where it is degenerate but one can define different coordinates v''' and w''' by

$$v''' = \arctan\left(\exp\left(\frac{r_+ - r_-}{2nr_-^2}v\right)\right), \quad w''' = \arctan\left(-\exp\left(\frac{-r_+ + r_-}{2nr_-^2}w\right)\right),$$

where n is an integer $\geq 2(r_+)^2(r_-)^2$. In these coordinates, the metric is analytic everywhere except at $r = r_+$, where it is degenerate. The coordinates v''' and w''' are analytic functions of v'' and w'' for $r \neq r_+$ or r_-. Thus the manifold M^* can be covered by an analytic atlas, consisting of local coordinate neighbourhoods defined by coordinates v'' and w'' for $r \neq r_-$ and by local coordinate neighbourhoods defined by v''' and w''' for $r \neq r_+$. The metric is analytic in this atlas.

The case $e^2 = m^2$ can be extended similarly; the case $e^2 > m^2$ is already inextendible in the original coordinates. The Penrose diagrams of these two cases are given in figure 26.

In all these cases, the singularity is timelike. This means that, unlike in the Schwarzschild solution, timelike and null curves can always avoid hitting the singularities. In fact the singularities appear to be repulsive: no timelike geodesic hits them, though non-geodesic timelike curves and radial null geodesics can. The spaces are thus timelike (though not null) geodesically complete. The timelike character of the singularity also means that there are no Cauchy surfaces in these spaces: given any spacelike surface, one can find timelike or null curves which run into the singularity and do not cross the surface. For example in the case *e2 < m2,* one can find a spacelike surface *S* which

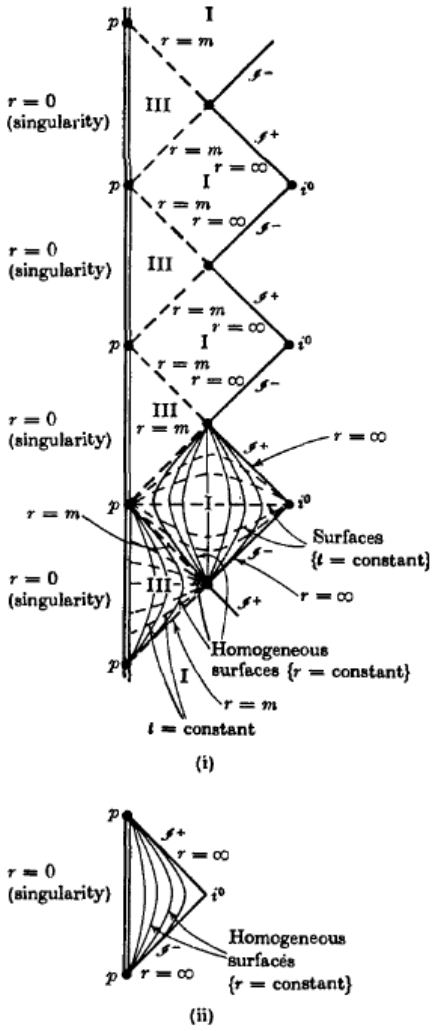


FIGURE 26. Penrose diagrams for the maximally extended Reissner–Nordström solutions:

(i) $e^2 = m^2$, (ii) $e^2 > m^2$.

In the first case there is an infinite chain or regions I ($\infty > r > m$) connected by regions III ($m > r > 0$). The points p are not part of the singularity at $r = 0$, but are really exceptional points at infinity.

crosses two asymptotically flat regions I (figure 25). This is a Cauchy surface for the two regions I and the two neighbouring regions II. However in the neighbouring regions III to the future there are past-directed inextendible timelike and null curves which approach the singularity and do not cross the surface $r = r_-$. This surface is therefore said to be the future Cauchy horizon for S. The continuation of the solution beyond $r = r_-$ is not determined by the Cauchy data on S. The continuation we have given is the only locally inextendible analytic one, but there will be other non-analytic C^∞ continuations which satisfy the Einstein-Maxwell equations.

A particle P crossing the surface $r = r+$ would appear to have

infinite redshift to an observer *O* whose world-line remains outside r = *r*+ and approaches the future infinity *i*+ (figure 25). In the region II between *r* = *r*+ and r = *r*_, the surfaces of constant *r* are spacelike and so each point of the figure represents a two-sphere which is a closed trapped surface. An observer *P* crossing the surface *r* = *r*_ would see the whole of the history of one of the asymptotically flat regions I in a finite time. Objects in this region would therefore appear to be infinitely blue-shifted as they approached *i*+. This suggests that the surface *r* = *r*_ would be unstable against small perturbations in the initial data on the spacelike surface *S*, and that such perturbations would in general lead to singularities on *r* = *r*_.

5.6 The Kerr solution

In general, astronomical bodies are rotating and so one would not expect the solution outside them to be exactly spherically symmetric. The Kerr solutions are the only known family of exact solutions which could represent the stationary axisymmetric asymptotically flat field outside a rotating massive object. They will be the exterior solutions only for massive rotating bodies with a particular combination of multipole moments; bodies with different combinations of moments will have other exterior solutions. The Kerr solutions do however appear to be the only possible exterior solutions for black holes (see §9.2 and §9.3).

The solutions can be given in Boyer and Lindquist coordinates (r,θ,ϕ,t) in which the metric takes the form

$$ds^2 = -\rho^2\left(\frac{dr^2}{\Delta}+d\theta^2\right)+(r^2+a^2)\sin^2\theta d\phi^2 - dt^2 + \frac{2mr}{\rho^2}(a\sin^2\theta d\phi - dt)^2 , \tag{5.29}$$

where

$$\rho^2(r,\theta)\equiv r^2+a^2\cos^2\theta \quad \text{and} \quad \Delta(r)\equiv r^2-2mr+a^2 .$$

m and *a* are constants, *m* representing the mass and *ma* the angular momentum as measured from infinity (Boyer and Price (1965»; when *a* = 0 the solution reduces to the Schwarzschild solution. This metric form is clearly invariant under simultaneous inversion of *t* and ϕ, i.e., under the transformation $t\to -t$, $\phi\to -\phi$, although it is not invariant under inversion of *t* alone (except when *a* = 0). This is what

one would expect, since time inversion of a rotating object produces an object rotating in the opposite direction.

When $a^2 > m^2$, $\Delta > 0$ and the above metric is singular only when *r=0.* The singularity at *r*=0 is not in fact a point but a ring, as can be seen by transforming to Kerr-Schild coordinates $(x, y, z, \bar{t})$, where

$$x + iy = (r + ia)\sin\theta \exp i\int (d\phi + a\Delta^{-1}dr),$$
$$z = r\cos\theta, \quad \bar{t} = \int (dt + (r^2 + a^2)\Delta^{-1}dr) - r.$$

In these coordinates, the metric takes the form

$$ds^2 = dx^2 + dy^2 + dz^2 + \frac{2mr^3}{r^4 + a^2z^2}\left(\frac{r(xdx + ydy) - a(xdy - ydx)}{r^2 + a^2} + \frac{zdz}{r} + d\bar{t}\right)^2, \tag{5.30}$$

where *r* is determined implicitly, up to a sign, in terms of *x, y, z.* by

$$r^4 - (x^2 + y^2 + z^2 - a^2)r^2 - a^2z^2 = 0.$$

For $r \neq 0$, the surfaces $\{r = \text{constant}\}$ are confocal ellipsoids in the (*x, y, z*) plane, which degenerate for *r* = 0 to the disc $z^2 + y^2 \leq a^2$, z=0. The ring $x^2 + y^2 = a^2$, z=0 which is the boundary of this disc, is a real curvature singularity as the scalar polynomial $R_{abcd}R^{abcd}$ diverges there. However no scalar polynomial diverges on the disc except at the boundary ring. The function *r* can in fact be analytically continued from positive to negative values through the interior of the disc $x^2 + y^2 < a^2$, z=0, to obtain a maximal analytic extension of the solution.

To do this, one attaches another plane defined by coordinates *(x',y',z')* where a point on the top side of the disc $x^2 + y^2 < a^2$, z=0 in the *(x, y, z.)* plane is identified with a point with the same *x* and *y* coordinates on the bottom side of the corresponding disc in the *(x',y',z')* plane. Similarly a point on the bottom side of the disc in the *(x, y, z.)* plane is identified with a point on the top side of the disc in the *(x',y',z')* plane (see figure 27). The metric (5.30) extends in the obvious way to this larger manifold. The metric on the *(x',y',z')* region is again of the form (5.29), but with negative rather than positive values of *r.* At large negative values of *r,* the space is again asymptotically flat but this time with negative mass. For small negative values of *r* near the ring singularity, the vector $\partial/\partial\phi$ is timelike, so the circles (*t*=constant, r=constant, θ=constant) are closed timelike curves. These closed timelike curves can be deformed to pass through any

point of the extended space (Carter (1968a)). This solution is geodesically incomplete at the ring singularity. However the only timelike and null geodesics which reach this singularity are those in the equatorial plane on the positive *r* side (Carter (1968a)).

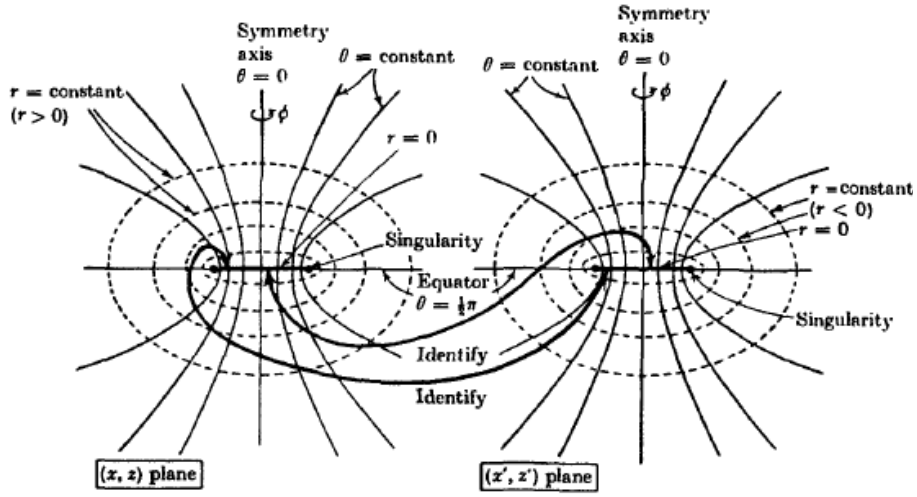


FIGURE 27. The maximal extension of the Kerr solution for $a^2 > m^2$ is obtained by identifying the top of the disc $x^2+y^2 < a^2, z = 0$ in the (x, y, z) plane with the bottom of the corresponding disc in the (x', y', z') plane, and vice versa. The figure shows the sections $y = 0, y' = 0$ of these planes. On circling twice round the singularity at $x^2+y^2 = a^2, z = 0$ one passes from the (x, y, z) plane to the (x', y', z') plane (where r is negative) and back to the (x, y, z) plane (where r is positive).

The extension in the case $a^2 < m^2$ is rather more complicated, because of the existence of the two values $r_+ = m + (m^2 - a^2)^{1/2}$ and $r_- = m - (m^2 - a^2)^{1/2}$ of r at which $\Delta(r)$ vanishes. These surfaces are similar to the surfaces $r = r_+$, $r = r_-$ in the Reissner-Nordström solution. To extend the. metric across these surfaces, one transforms to the Kerr coordinates (r, θ, ϕ_+, u_+), where

$du_+ = dt + (r^2 + a^2)\Delta^{-1}dr$, $d\phi_+ = d\phi + a\Delta^{-1}dr$.

The metric then takes the form

$$\begin{aligned} ds^2 &= \rho^2 d\theta^2 - 2a\sin^2\theta r d\phi + +2dr du_+ \\ &\quad + \rho^{-2}\left\{(r^2+a^2)^2 - \Delta a^2\sin^2\theta\right\}\sin^2\theta d\phi_+{}^2 \\ &\quad - 4a\rho^{-2}mr\sin^2\theta d\phi_+ du_+ - (1-2mr)du_+{}^2 \end{aligned} \tag{5.31}$$

on the manifold defined by these coordinates, and is analytic at $r = r_+$ and $r = r_-$. One again has a singularity at $r = 0$, which has the same ring form and geodesic structure as that described above. The metric can also be extended on the manifold defined by the coordinates (r, θ, ϕ_-, u_-) where

$$du_- = dt - (r^2 + a^2)\Delta^{-1}dr, \quad d\phi_- = d\phi - a\Delta^{-1}dr;$$

the metric again takes the form (5.31), with ϕ_+, u_+ replaced by $-\phi_-$, $-u_-$. The maximal analytic extension can be built up by a combination of these extensions, as in the Reissner-Nordström case (Boyer and Lindquist (1967), Carter (1968a)). The global structure is very similar to that of the Reissner-Nordström solution except that one can now continue through the ring to negative values of r. Figure 28 (i) shows the conformal structure of the solution along the symmetry axis. The regions I represent the asymptotically flat regions in which $r > r_+$. The regions II ($r_- < r < r_+$) contain closed trapped surfaces. The regions III ($-\infty < r < r_-$) contain the ring singularity; there are closed timelike curves through every point in a. region III, but no causality violation occurs in the other two regions.

In the case $a^2 = m^2$, r_+ and r_- coincide and there is no region II. The maximal extension is similar to that of the Reissner-Nordström solution when $a^2 = m^2$. The conformal structure along the symmetry axis in this case is shown in figure 28(ii).

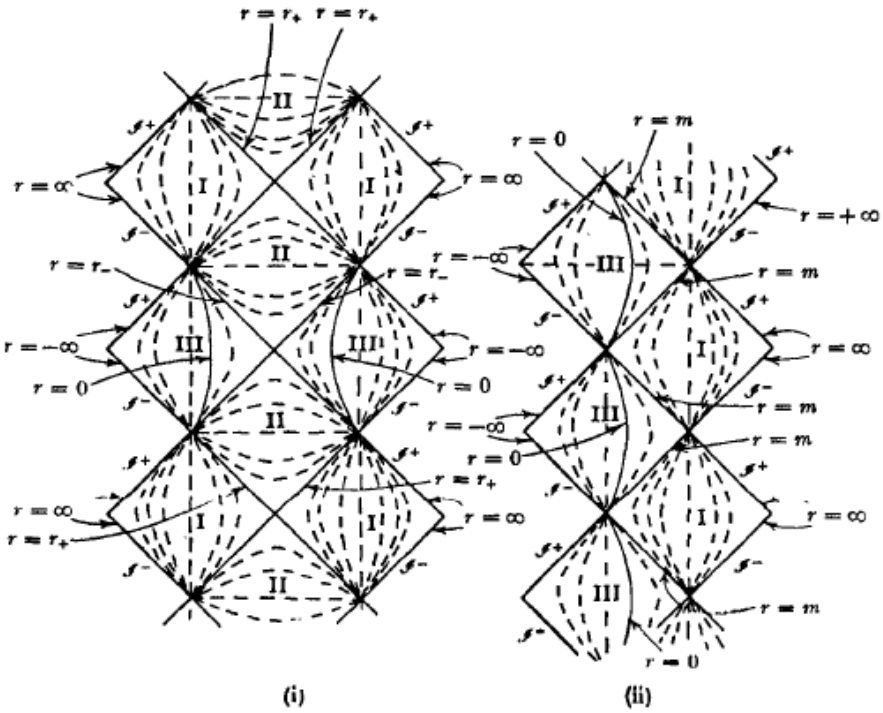


FIGURE 28. The conformal structure of the Kerr solutions along the axis of symmetry, (i) in the case $0 < a^2 < m^2$, (ii) in the case $a^2 = m^2$. The dotted lines are lines of constant r; the regions I, II and III in case (i) are divided by $r = r_+$ and $r = r_-$, and the regions I and III in case (ii) by $r = m$. In both cases, the structure of the space near the ring singularity is as in figure 27.

The Kerr solutions, being stationary and axisymmetric, have a two-parameter group of isometries. This group is necessarily Abelian

(Carter (1970)). There are thus two independent Killing vector fields which commute. There is a unique linear combination K^a of these Killing vector fields which is timelike at arbitrarily large positive and negative values of r. There is another unique linear combination $\tilde{K}^a$ of the Killing vector fields which is zero on the axis of symmetry. The orbits of the Killing vector K^a define the stationary frame, that is, an object moving along one of these orbits appears to be stationary with respect to infinity. The orbits of the Killing vector $\tilde{K}^a$ are closed curves, and correspond to the rotational symmetry of the solution.

In the Schwarzschild and Reissner-Nordström solutions, the Killing vector K^a which is timelike at large values of r is timelike everywhere in the region I, becoming null on the surfaces $r = 2m$ and $r = r+$, respectively. These surfaces are null. This means that a particle which crosses one of these surfaces in the future direction cannot return again to the same region. They are the boundary of the region of the solution from which particles can escape to the infinity . I^+ of a particular region I, and are called the *event horizons* of that I^+. (They are in fact the event horizon in the sense of§5.2 for an observer moving on any of the orbits of the Killing vector K^a in the region I.)

In the Kerr solution on the other hand, the Killing vector K^a is spacelike in a region outside $r = r+$, called the *ergosphere* (figure 29). The outer boundary of this region is the surface $r = m + (m^2 - a^2 \cos^2 \theta)^{1/2}$

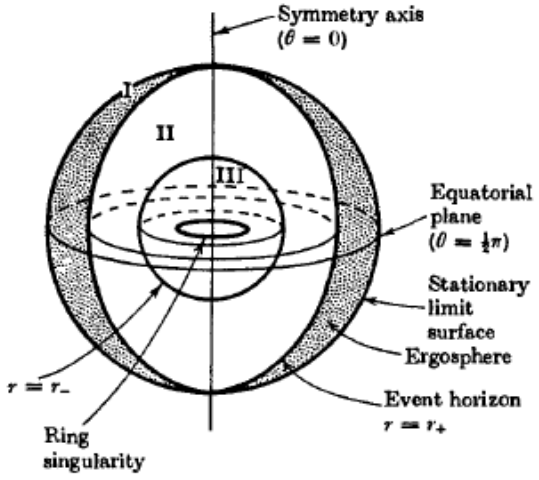


FIGURE 29. In the Kerr solution with $0 < a^2 < m^2$, the ergosphere lies between the stationary limit surface and the horizon at $r = r_+$. Particles can escape to infinity from region I (outside the event horizon $r = r_+$) but not from region II (between $r = r_+$ and $r = r_-$) and region III ($r < r_-$; this region contains the ring singularity).

on which K^a is null. This is called the *stationary limit surface* since it is the boundary of the region in which particles travelling on a timelike curve can travel on an orbit of the Killing vector K^a, and so remain at rest with respect to infinity. The stationary limit surface is a timelike surface except at the two points on the axis, where it is null (at these points it coincides with the surface $r = r+$). Where it is timelike it can be crossed by particles in either the ingoing or the outgoing direction. It is therefore not the event horizon for I^+. In fact the event horizon is the surface $r = r+ = m + (m^2 - a^2)^{1/2}$. Figure 30 shows why this is. It shows the equatorial plane $\theta = \pi / 2$; each point in this figure represents an orbit of the Killing vector K^a, i.e., it is stationary with respect to I^+. The small circles represent the position a short time later of flashes of light emitted from the points represented by the heavy black dots. Outside the stationary limit the Killing vector K^a is timelike and so lies within the light cone. This means that the point in figure 30 representing the orbit of emission lies within the wavefront of the light.

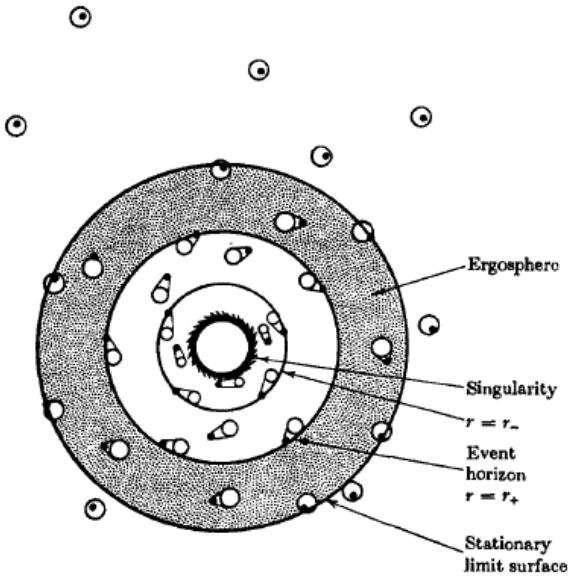


FIGURE 30. The equatorial plane of a Kerr solution with $m^2 > a^2$. The circles represent the position a short time later of flashes of light emitted by the points represented by heavy dots.

On the stationary limit surface, K^a is null and so the point representing the orbit of emission lies on the wavefront. However the wavefront lies partly within and partly outside the stationary limit surface; it is therefore possible for a particle travelling along a timelike curve to escape to infinity from this surface. In the ergosphere between the

stationary limit surface and $r = r+$, the Killing vector K^a is spacelike and so the point representing the orbit of emission lies outside the wavefront. In this region it is impossible for a particle moving on a timelike or null curve to travel along an orbit of the Killing vector and so to remain at rest with respect to infinity. However the positions of the wavefronts are such that the particles can still escape across the stationary limit surface and so out to infinity. On the surface $r = r+$, the Killing vector K^a is still spacelike. However the wavefront corresponding to a point on this surface lies entirely within the surface.
This means that a particle travelling on a timelike curve from a point on or inside the surface cannot get outside the surface and so cannot get out to infinity. The surface $r = r+$ is therefore the event horizon for I^+ and is a null surface.

Although the Killing vector K^a is spacelike in the ergosphere, the magnitude $K^a \tilde{K}^b K_{[a} \tilde{K}_{b]}$ of the *Killing bivector* $K_{[a} \tilde{K}_{b]}$ is negative everywhere outside $r=r+$, except on the axis $\tilde{K}^a = 0$ where it vanishes.
Therefore K^a and $\tilde{K}^a$ span a timelike two-surface and so at each point outside $r = r+$ off the axis there is a linear combination of K^a and $\tilde{K}^a$ which is timelike. In a sense, therefore, the solution in the ergosphere is locally stationary, although it is not stationary with respect to infinity. In fact there is no one linear combination of K^a and $\tilde{K}^a$ which is timelike everywhere outside $r = r+$. The magnitude of the Killing bivector vanishes on $r = r+$, and is positive just inside this surface.
On $r = r+$, both K^a and $\tilde{K}^a$ are spacelike but there is a linear combination which is null everywhere on $r = r+$ (Carter (1969)).

The behaviour of the ergosphere and the horizon we have discussed will play an important part in our discussion of black holes in § 9.2 and §9.3.

Just as the Reissner-Nordström solution can be thought of as a charged version of the Schwarzschild solution, so there is a family of charged Kerr solutions (Carter (1968 a)). Their global properties are very similar to those of the uncharged Kerr solutions.

5.7 Gödel's universe

In 1949, Kurt Gödel published a paper (Gödel (1949)) which provided a considerable stimulus to investigation of exact solutions more complex

than those examined so far. He gave an exact solution of Einstein's field equations in which the matter takes the form of a pressure-free perfect fluid ($T_{ab} = \rho u_a u_b$ where ρ is the matter density and u_a the normalized four-velocity vector). The manifold is R^4 and the metric can be given in the form

$$ds^2 = -dt^2 + dx^2 - \frac{1}{2}\exp(2\sqrt{2}\omega x)dy^2 + dz^2 - 2\exp(\sqrt{2}\omega x)dtdy ,$$

where $\omega > 0$ is a constant; the field equations are satisfied if $\boldsymbol{u} = \partial / \partial x^0$ (i.e., $u^a = \delta^a{}_0$) and

$$4\pi\rho = \omega^2 = -\Lambda .$$

The constant ω is in fact the magnitude of the vorticity of the flow vector u^a.

This space-time has a five-dimensional group of isometries which is transitive, i.e., it is a completely homogeneous space-time. (An action of a group is transitive on *M* if it can map any point of *M* into any other point of *M*.) The metric is the direct sum of the metric g_1 given by

$$ds_1{}^2 = -dt^2 + dx^2 - \frac{1}{2}\exp(2\sqrt{2}\omega x)dy^2 - 2\exp(\sqrt{2}\omega x)dtdy$$

on the manifold $M_1 = R^3$ defined by the coordinates (t, x, y), and the metric g_2 given by

$$ds_2{}^2 = dz^2$$

on the manifold $M_2 = R^1$ defined by the coordinate z. In order to describe the properties of the solution it is sufficient to consider only (M_1, g_1).

Defining new coordinates (t', r, ϕ) on M_1 by

$$\exp(\sqrt{2}\omega x) = \cosh 2r + \cos\phi \sinh 2r ,$$

$$\omega y \exp(\sqrt{2}\omega x) = \sin\phi \sinh 2r ,$$

$$\tan\frac{1}{2}(\phi + \omega t - \sqrt{2}t') = \exp(-2r)\tan\frac{\phi}{2} ,$$

the metric g_1 takes. the form

$$ds_1{}^2 = 2\omega^{-2}(-dt'^2 + dr^2 - (\sinh^4 r - \sinh^2 r)d\phi^2 + 2\sqrt{2}\sinh^2 r d\phi dt) ,$$

where $-\infty < t < \infty$, $0 \le r < \infty$, and $0 \le \phi \le 2\pi$, $\phi = 0$ being identified with $\phi = 2\pi$; the flow vector in these coordinates is $u = (\omega / \sqrt{2})\partial / \partial t'$. This form exhibits the rotational symmetry of the solution about the axis r = 0. By a different choice of coordinates the axis could be chosen

to lie on any flow line of the matter.

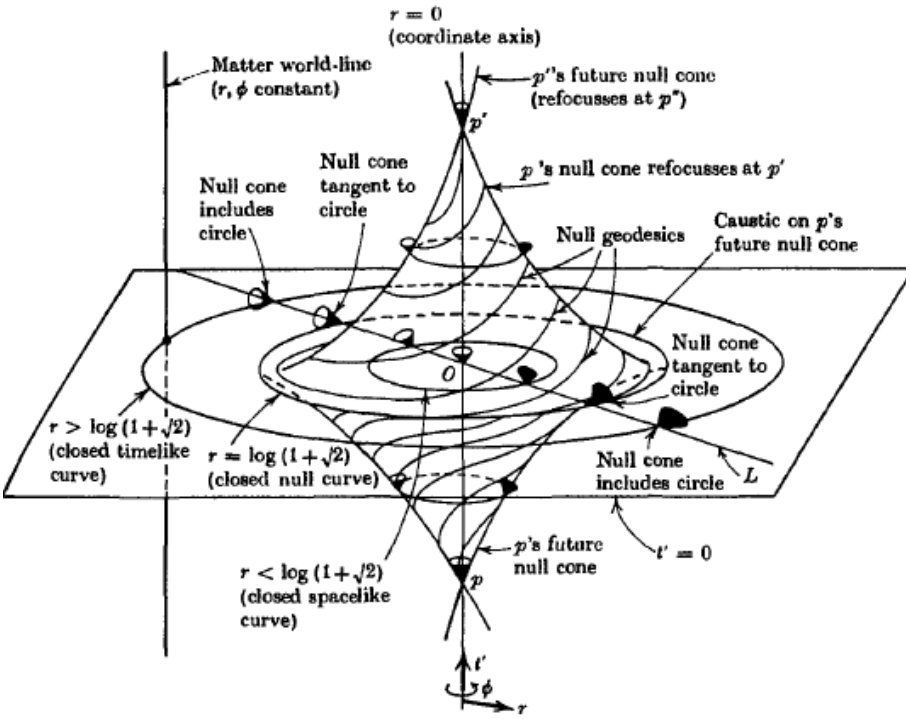


FIGURE 31. Gödel's universe with the irrelevant coordinate z suppressed. The space is rotationally symmetric about any point; the diagram represents correctly the rotational symmetry about the axis $r = 0$, and the time invariance. The light cone opens out and tips over as r increases (see line L) resulting in closed timelike curves. The diagram does not correctly represent the fact that *all* points are in fact equivalent.

The behaviour of (M_1, g_1) is illustrated in figure 31. The light cones on the axis $r = 0$ contain the direction $\partial / \partial t'$ (the vertical direction on the diagram) but not the horizontal directions $\partial / \partial r$ and $\partial / \partial \phi$. As one moves away from the axis, the light cones open out and tilt in the ϕ-direction so that at a radius $r = \log(1+\sqrt{2})$, $\partial / \partial \phi$ is a null vector and the circle of this radius about the origin is a closed null curve. At greater values of r, $\partial / \partial \phi$ is a timelike vector and circles of constant r, t' are closed timelike curves. As (M_1, g_1) has a four-dimensional group of isometries which is transitive, there are closed timelike curves through every point of (M_1, g_1), and hence through every point of the Gödel solution (M, g).

This suggests that the solution is not very physical. The existence of closed timelike curves in this solution implies that there are no imbedded three-dimensional surfaces without boundary in M (which are spacelike everywhere. For a closed timelike curve which crossed

such a surface would cross it an odd number of times. This would mean that the curve could not be continuously deformed to zero, since a continuous deformation can change the number of crossings only by an even number. This would contradict the fact that. M is simply connected, being homeomorphic to R^4. The existence of closed timelike lines also shows that there can be no cosmic time coordinate t in M which increases along every future-directed timelike or null curve.

The Gödel solution is geodesically complete. The behaviour of the geodesics can be described in terms of the decomposition into (M_1, g_1) and (M_2, g_2). Since the metric g_2 of M_2 is flat, the component of the geodesic tangent vector in M_2 is constant, i.e., the z-coordinate varies linearly with the affine parameter on the geodesic. It is sufficient therefore to describe the behaviour of geodesics in (M_1, g_1). The null geodesics from a point p on the axis of coordinates (figure 31) diverge from the axis initially, reach a caustic at $r = \log(1+\sqrt{2})$, and then reconverge to a point p' on the axis. The behaviour of timelike geodesics is similar: they reach some maximum value of r less than $\log(1+\sqrt{2})$ and then reconverge to p'. A point q at a radius r greater than $\log(1+\sqrt{2})$ can be joined to p by a timelike curve but not by a timelike or null geodesic.

Further details of Gödel's solution can be found in Gödel (1949), Kundt (1956).

5·8 Taub-NUT space

In 1951, Taub discovered a spatially homogeneous empty space solution of Einstein's equations with topology $R \times S^3$ and metric given by

$$ds^2 = -U^{-1}dt^2 + (2l)^2 U(d\psi + \cos\theta d\phi)^2 + (t^2 + l^2)(d\theta^2 + \sin^2 d\phi^2), \tag{5.32}$$

where

$$U(t) \equiv -1 + \frac{2(mt + l^2)}{t^2 + l^2},$$

m and l are positive constants.

Here θ, ϕ, ψ are Euler coordinates on S^3, so $0 \le \psi \le 4\pi$, $0 \le \theta \le \pi$, $0 \le \phi \le 2\pi$. This metric is singular at $t = t_{\pm} = m \pm (m^2 + l^2)^{1/2}$, where $U = 0$. It can in fact be extended across these surfaces to give a space found by Newman, Tamburino and Unti (1963), but before discussing

the extension we shall consider a simple two-dimensional example given by Misner (1967) which has many similar properties.

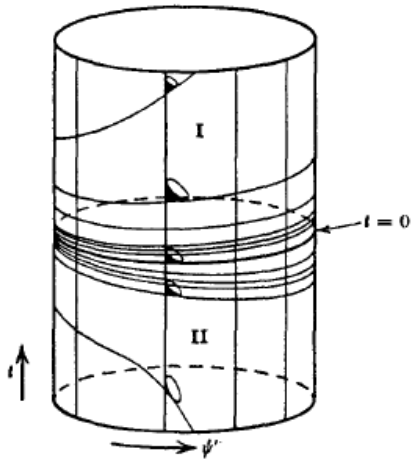
This space has the topology $S^1 \times R^1$ and the metric g given by

$ds^2 = -t^{-1}dt^2 + td\psi^2$

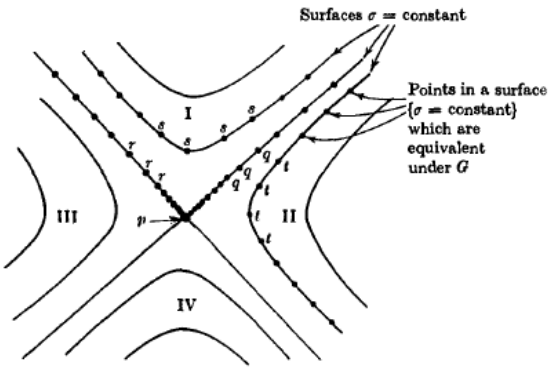
where $0 \le \psi \le 2\pi$. This metric is singular when $t = 0$. However if one takes the manifold M (defined by ψ and by $0 < t < \infty$, (M, g) can be extended by defining $\psi' = \psi - \log t$. The metric then takes the form g' given by

$ds^2 = +2d\psi' dt + t(d\psi')^2$.

This is analytic on the manifold M' with topology $S^1 \times R^1$ defined by ψ' and by $-\infty < t < \infty$. The region $t > 0$ of (M', g') is isometric with (M, g). The behaviour of (M', g') is shown in figure 32. There are closed timelike lines in the region $t < 0$, but there are none when



(i)



(ii)

FIGURE 32. Misner's two-dimensional example.
(i) Extension of region I across the boundary $t = 0$ into II. The vertical null geodesics are complete, but the twisted null geodesics are incomplete.
(ii) The universal covering space is two-dimensional Minkowski space. Under the discrete subgroup G of the Lorentz group, points s are equivalent; similarly points r, q and t are equivalent. (i) is obtained by identifying equivalent points in regions I and II.

$t > 0$. One family of null geodesics is represented by the vertical lines in figure 32; these cross the surface $t = 0$. The other family spiral round and round as they approach $t = 0$, but never actually cross this surface, and these geodesics have only finite affine length. Thus the extension (M', g') is not symmetric between the two families of null geodesics, although the original space (M, g) was. However one can define another extension (M', g') in which the behaviour of the two families of null geodesics is interchanged. To do so define ψ'' by $\psi'' = \psi + \log t$. The metric takes the form g'' given by

$$ds^2 = -2d\psi'' dt + t(d\psi'')^2 .$$

This is analytic on the manifold M'' with topology $S^1 \times R^1$ defined by ψ'' and $-\infty < t < \infty$. The region $t > 0$ of (M'', g'') is isometric with (M, g). In a sense, what we have done by defining ψ'' is to untwist the second family of null geodesics so that they become vertical lines, and can be continued beyond $t = 0$. However this twisting winds up the first family of null geodesics so that they spiral around and cannot be continued beyond $t = 0$. One has therefore two inequivalent locally inextendible analytic extensions of (M, g), both of which are geodesically incomplete. The relation between these two extensions can be seen clearly by going to the covering space of (M, g).

This is in fact the region I of two-dimensional Minkowski space $(\tilde{M}, \tilde{\eta})$ contained within the future null cone of a point p (figure 32(ii)) The isometries of $(\tilde{M}, \tilde{\eta})$ which leave p fixed form a one-dimensional group (the Lorentz group of $\tilde{\eta}$) whose orbits are the hyperbolae $\{\sigma = \text{constant}\}$ where $\sigma \equiv \tilde{t}^2 - \tilde{x}^2$ and $\tilde{t}, \tilde{x}$ are the usual Minkowski coordinates. The space (M, g) is the quotient of $(I, \tilde{\eta})$ by the discrete subgroup G of the Lorentz group consisting of A^n (n integer) where A maps $(\tilde{t}, \tilde{x})$ to

$$(\tilde{t}\cosh\pi + \tilde{x}\sinh\pi, \tilde{x}\cosh\pi + \tilde{t}\sinh\pi),$$

i.e., one identifies the points

$$(\tilde{t}\cosh n\pi + \tilde{x}\sinh n\pi, \tilde{x}\cosh n\pi + \tilde{t}\sinh n\pi)$$

for all integer values of n, and these correspond to the point

$$t = \frac{1}{4}(\tilde{t}^2 - \tilde{x}^2), \quad \psi = 2\,\text{arctanh}(\tilde{x}/\tilde{t}) \quad \text{in} \quad M .$$

The action of the isometry group G in the region I is properly discontinuous. The action of a group H on a manifold N is said to be

properly discontinuous if:

(1) each point $q \in N$ has a neighbourhood U such that $A(U) \cap U = \phi$ for each $A \in H$ which is not the identity element, and

(2) if $q, r \in N$ are such that there is no $A \in H$ with $Aq = r$, then there are neighbourhoods U and U' of q and r respectively such that there is no $B \in H$ with $B(U) \cap U' \neq \phi$.

Condition (1) implies that the quotient N/H is a manifold, and condition (2) implies that it is Hausdorff. Thus the quotient $(I, \tilde{\eta})/G$ is the Hausdorff space (M, g). The action of G is also properly discontinuous in the regions I + II $(\tilde{t} > -\tilde{x})$. Thus $(I + II, \tilde{\eta})/G$ is also a Hausdorff space; in fact it is (M', g'). Similarly $(I + III, \tilde{\eta})/G$ is the Hausdorff space (M'', g'') where I + III is the region $(\tilde{t} > \tilde{x})$. From this it can be seen how it is that one family of null geodesics can be completed in the extension (M', g') while the other family can be completed in the extension (M'', g''). This suggests that one might perform both extensions at the same time. However the action of the group on the region (I +II + III) (i.e., $\tilde{t} > -|\tilde{x}|$) satisfies condition (1) but condition (2) is not satisfied for points q on the boundary between I and II and points r on the boundary between I and III. Therefore the quotient $(I + II + III, \tilde{\eta})/G$ is not Hausdorff although it is still a manifold.

This kind of non-Hausdorff behaviour is different from that in the example given in §2.1. In that example, one could have continuous curves which bifurcate, one branch going into one region and another branch going into another region. Such behaviour of an observer's world-line would be very uncomfortable. However the manifold $(I + II + III)/G$ does not have any such bifurcating curves; curves in I can be extended into II or III but not into both simultaneously. Thus one might be prepared to relax the Hausdorff requirement on a space-time model to allow this sort of situation but not the sort in which one gets bifurcating curves. Further work on non-Hausdorff space-times can be found in the papers of Hajicek (1971).

Condition (1) is in fact satisfied by the action of G on. $\tilde{M} - \{p\}$. Thus the space $(\tilde{M} - \{p\}, \tilde{\eta})/G$ is in some sense the maximal non-Hausdorff extension of (M, g). However it is still not geodesically complete because there are geodesics which pass through the point p which has been left out. If p is included the action of the group does not satisfy

condition (1), and so the quotient $\tilde{M}/G$ is not even a non-Hausdorff manifold. However consider the bundle of linear frames $L(\tilde{M})$, i.e., the collection of all pairs $(X,Y), X,Y \in T_q$, of linearly independent vectors at all points $q \in \tilde{M}$. The action of an element *A* of the isometry group *G on* $\tilde{M}$ induces an action A_* on $L(\tilde{M})$ which takes the frame (X, Y) at *q* to the frame (A_*X, A_*Y) at *A(q).* This action satisfies condition (1) because even for $(X,Y) \in T_p$, $A_*X \neq X$ and $A_*Y \neq Y$ unless *A* = identity, and satisfies condition (2) even if X and Y lie on the null cone of *p.* Thus the quotient $L(\tilde{M})/G$ is a Hausdorff manifold. It is a fibre bundle over the non-Hausdorff non-manifold $\tilde{M}/G$. One could in a sense regard it as the bundle of linear frames for this space. The fact that the bundle of frames can be well behaved even though the space is not, suggests that it is useful to look at singularities by using the bundle of linear frames. A general procedure for doing this will be given in §8.3.

We shall now return to the four-dimensional Taub space (M,g) where M is $R^1 \times S^3$ and g is given by (5.32). As M is simply connected, one cannot take a covering space as we did in the two-dimensional example. However one can achieve a similar result by considering M as a fibre bundle over S^2 with fibre $R^1 \times S^1$; the bundle projection $\pi : M \to S^2$ is defined by $(t,\psi,\theta,\phi) \to (\theta,\phi)$. This is in fact the product with the t-axis of the Hopf fibering $S^3 \to S^2$ (Steenrod (1951)) which has fibre S^1. The space (M,g) admits a four-dimensional group of isometries whose surfaces of transitivity are the three-spheres $\{t = \text{constant}\}$. This group of isometries maps fibres of the bundle $\pi : M \to S^2$ into fibres, and so the pairs $(F,\tilde{g})$ are all isometric, where F is a fibre ($F \approx R^1 \times S^1$) and $\tilde{g}$ is the metric induced on the fibre by the four-dimensional metric g on *M*. The fibre F can be regarded as the (t,ψ) plane, and the metric $\tilde{g}$ on F is obtained from (5.32) by dropping the terms in $d\theta$ and $d\phi$; thus $\tilde{g}$ is given by

$$ds^2 = -U^{-1}dt^2 + 4l^2U(d\psi)^2. \qquad (5.33)$$

The tangent space T_q at the point $q \in M$ can be decomposed into a vertical subspace V_q which is tangent to the fibre and is spanned by the vectors $\partial/\partial t$ and $\partial/\partial\psi$, and a horizontal subspace H_q which is spanned by the vectors $\partial/\partial\theta$ and $\partial/\partial\phi - \cos\theta\partial/\partial\psi$. Any vector $X \in T_q$ can be split into a part X_V lying in V_q and a part X_H lying in H_q. The

metric g on T_q can then be expressed as

$$g(X,Y) = g_V(X_V, Y_V) + (t^2 + l^2) g_H(\pi_* X_H, \pi_* Y_H), \qquad (5.34)$$

where $g_V \equiv \tilde{g}$ and g_H is the standard metric on the two-sphere given by $ds^2 = d\theta^2 + \sin^2\theta d\phi^2$. Thus although the metric g is not the direct sum of g_V and $(t^2 + l^2) g_H$ (because $R^1 \times S^3$ is not the direct product of $R^1 \times S^1$ with S^2) it can nevertheless be regarded as such a sum locally.

The interesting part of the metric g is contained in g_V and we shall therefore consider analytic extensions of the pair (F, g_V). When combined with the metric g_H of the two-sphere as in (5.34), these give analytic extensions of (M, g).

The metric g_V, given by (5.33), has singularities at $t = t_\pm$ where $U = 0$. However if one takes the manifold F_0 defined by ψ and by $t_- < t < t_+$, (F_0, g_V) can be extended by defining

$$\psi' = \psi + \frac{1}{2l} \int \frac{dt}{U(t)}.$$

The metric then takes the form g_V' given by

$$ds^2 = 4l d\psi' (l U(t) d\psi' - dt).$$

This is analytic on the manifold F' with topology $S^1 \times R$ defined by ψ' and by $-\infty < t < \infty$. The region $t_- < t < t_+$ of $(F', g_{V'})$ is isometric with (F_0, g_V). There are no closed timelike curves in the region $t_- < t < t_+$ but there are for $t < t_-$ and for $t > t_+$. The behaviour is very much as for the space (M', g') we considered before, except that there are now two horizons (at $t = t_-$ and $t = t_+$) instead of the one horizon (at $t = 0$). One family of null geodesics crosses both horizons $t = t_-$ and $t = t_+$ but the other family spirals round near these surfaces and is incomplete.

As before, one can make another extension by defining the coordinate

$$\psi'' = \psi - \frac{1}{2l} \int \frac{dt}{U(t)}.$$

The metric then takes the form g_V'' given by

$$ds^2 = 4l d\psi'' (l U(t) d\psi' + dt)$$

which is analytic on the manifold F'' defined by ψ'' and by $-\infty < t < \infty$, and is again isometric to (F_0, g_V) on $t_- < t < t_+$.

Once again one can show the relation between the different extensions by going to the covering space. The covering space of F_0 is the

manifold $\widetilde{F}_0$ defined by the coordinates $-\infty < \psi < \infty$ and by $t_- < t < t_+$. On $\widetilde{F}_0$ the metric g_V can be written in the double null form

$$ds^2 = 4l^2 U(t) d\psi' d\psi'', \qquad (5.35)$$

where $-\infty < \psi' < \infty$, $-\infty < \psi'' < \infty$. One can extend this in a manner similar to that used in the Reissner-Nordström solution. Define new coordinates (u_+, v_+) and (u_-, v_-) on F_0 by

$u_\pm = \arctan(\exp(\psi'/\alpha_\pm))$, $v_\pm = \arctan(-\exp(-\psi''/\alpha_\pm))$,

where

$$\alpha_+ = \frac{t_+ - t_-}{4l(mt + l^2)} \quad \text{and} \quad \alpha_- = \frac{t_+ - t_-}{4nl(mt + l^2)};$$

n is some integer greater than $(mt_+ + l^2)/(mt_- + l^2)$. Then the metric $\widetilde{g}_V$ obtained by applying this transformation to (5.35) is analytic on the mnnifold $\widetilde{F}$ shown in figure 33, where the coordinates (u_+, v_+) are analytic coordinates except at $t = t_-$ where they are at least C^3, and the coordinates (u_-, v_-) are analytic coordinates except at $t = t_+$ where they are at least C^3. This is rather similar to the extension of the (*t*, *r*) plane of the Reissner-Nordström solution.

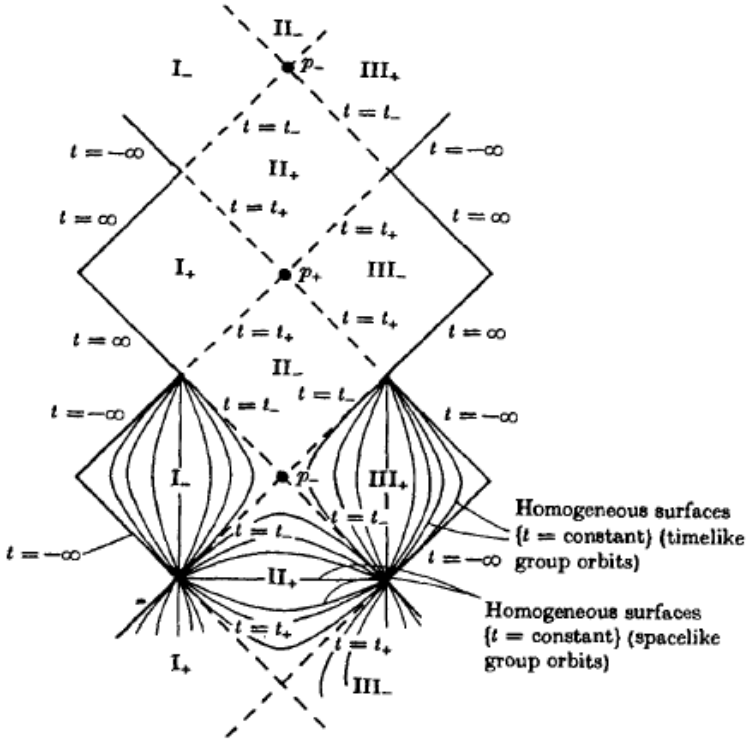


FIGURE 33. Penrose diagram of the maximally extended covering space of a two-dimensional section of Taub–NUT space, showing orbits of the isometry group. Taub–NUT space and its extensions are obtained from part of this space by identification of points under a discrete subgroup of the isometry group.

The space $(\widetilde{F},\widetilde{g}_V)$ has a one-dimensional group of isometries, the orbits of which are shown in figure 33. Near the points p_+, p_- the action of this group is similar to that of the Lorentz group in two-dimensional Minkowski space (figure 32 (ii)). Let *G* be the discrete subgroup of the isometry group generated by a non-trivial element *A* of the isometry group. The space (F_0, g_V) is the quotient of one of the regions $(II_+, \widetilde{g}_V)$ by *G*. The space (F', g_V') is the quotient $(I_- + II_+ + III_-, \widetilde{g}_V)/G$, and (F'', g_V'') is the quotient $(I_+ + II_+ + III_+, \widetilde{g}_V)/G$.

One would also obtain a Hausdorff manifold by taking the quotient of $I_+ + II_+ + I_-$: this corresponds to extending like (F', g_V') at the surface $t = t_+$ but extending like (F'', g_V'') at the surface $t = t_-$. By taking the quotient of the whole space $\widetilde{F}$ minus the points p_+ and p_- one obtains a non-Hausdorff manifold; and taking the quotient of $\widetilde{F}$ one obtains a non-Hausdorff non-manifold in a way analogous to that in the example above. As in that example, one can take the quotient of the bundle of linear frames over F and obtain a Hausdorff manifold.

By combining these extensions of the (t,ψ) plane with the coordinates (θ,ϕ) one can obtain corresponding extensions of the four-dimensional space (M, g). In particular, the two extensions (F', g_V') and (F'', g_V'') give rise to two different locally inextendible analytic extensions of (M, g), and both are geodesically incomplete.

Consider one of these extensions, say (M', g'). The three-spheres which are the surface of transitivity of the isometry group are spacelike surfaces in the region $t_- < t < t_+$ and are timelike for $t > t_+$ and $t < t_-$. The two surfaces of transitivity $t = t_-$ and $t = t_+$ are null surfaces and they form the Cauchy horizon of any spacelike surface contained in the region $t_- < t < t_+$, because there are timelike curves in the regions $t < t_-$ and $t > t_+$ which do not cross and $t = t_+$, respectively (for example, closed timelike curves exist in the regions $t < t_-$ and $t > t_+$). The region of space-time $t_- \le t \le t_+$ is compact yet there are timelike and null geodesics which remain within it and are incomplete. This kind of behaviour will be considered further in chapter 8.

Further details of Taub-NUT space may be found in Misner and Taub (1969), Misner (1963).

5.9 Further exact solutions

We have examined in this chapter a number of exact solutions and used them to give examples of the various global properties which we shall wish to discuss more generally later. Although a large number of exact solutions are known locally, relatively few have been examined globally. To complete this chapter, we shall mention briefly two other interesting families of exact solutions whose global properties are known.

The first of these are the *plane wave* solutions of the empty space field equations. These are homeomorphic to R^4, and global coordinates *(y, z, u, v),* which range from $-\infty$ to $+\infty$, can be chosen so that the metric takes the form

$$ds^2 = 2dudv + dy^2 + dz^2 + H(y,z,u)du^2,$$

where

$$H = (y^2 - z^2)f(u) - 2yzg(u);$$

$f(u)$ and $g(u)$ are arbitrary C^2 functions determining the amplitude and polarization of the wave. These spaces are invariant under a five-parameter group of isometries multiply transitive on the null surfaces $\{u = \text{constant}\}$; a special subclass, in which $f(u) = \cos 2u$, $g(u) = \sin 2u$, admit an extra Killing vector field, and are homogeneous space-times invariant under a six-parameter group of isometries. These spaces do not contain any closed timelike or null curves; however they admit no Cauchy surfaces (Penrose (1965a)). Local properties of these spaces have been studied in detail by Bondi, Pirani and Robinson (1959), and global properties by Penrose (1965a); Oszvath and Schücking (1962) have studied global properties of the higher symmetry space. The way in which two impulsive plane waves scatter each other and give rise to a singularity has been studied by Khan and Penrose (1971).

The other is the five-parameter family of exact solutions of the source-free Einstein-Maxwell equations found by Carter (1968b) (see also Demianski and Newman (1966)). These include the Schwarzschild, Reissner-Nordström, Kerr, charged Kerr, Taub-NUT, de Sitter and anti-de Sitter solutions as special cases. A description of some of their global properties is given in Carter (1967). Some cases closely related

to this family have been examined by Ehlers and Kundt (1962) and Kinnersley and Walker(1970).