Global Structure of the Kerr Family of Gravitational Fields

BRANDON CARTER*
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, England
(Received 29 March 1968)

The Kerr family of solutions of the Einstein and Einstein-Maxwell equations is the most general class of solutions known at present which could represent the field of a rotating neutral or electrically charged body in asymptotically flat space. When the charge and specific angular momentum are small compared with the mass, the part of the manifold which is stationary in the strict sense is incomplete at a Killing horizon. Analytically extended manifolds are constructed in order to remove this incompleteness. Some general methods for the analysis of causal behavior are described and applied. It is shown that in all except the spherically symmetric cases there is nontrivial causality violation, i.e., there are closed timelike lines which are not removable by taking a covering space; moreover, when the charge or angular momentum is so large that there are no Killing horizons, this causality violation is of the most flagrant possible kind in that it is possible to connect any event to any other by a future-directed timelike line. Although the symmetries provide only three constants of the motion, a fourth one turns out to be obtainable from the unexpected separability of the Hamilton-Jacobi equation, with the result that the equations, not only of geodesics but also of charged-particle orbits, can be integrated completely in terms of explicit quadratures. This makes it possible to prove that in the extended manifolds all geodesics which do not reach the central ring singularities are complete, and also that those timelike or null geodesics which do reach the singularities are entirely confined to the equator, with the further restriction, in the charged case, that they be null with a certain uniquely determined direction. The physical significance of these results is briefly discussed.

INTRODUCTION

PROBABLY the most important problems in general relativity today are those concerning the singularities and other pathological features arising in gravitational collapse. Because of the scanty nature of the experimental evidence in its favor, the acceptability or the unacceptability of Einstein's theory must depend largely on whether its theoretical predictions seem reasonable or not.

A great deal is now known about the gravitational collapse to a curvature singularity of a spherically symmetric body. However, virtually nothing is known about collapse in more general circumstances, where angular momentum is present for example, except that by the results of Penrose[1] and Hawking[2–4] singular behavior of some sort must be expected to remain.

For this reason it is interesting to examine the properties of the Kerr family of gravitational fields from this point of view since these are the only solutions of Einstein's equations known at present which could represent the exterior field of a rotating body in asymptotically flat space.

It will be shown that the ringlike curvature singularities in the inner parts of the Kerr fields are comparatively innocuous (they are in fact invisible except in the equatorial direction) in contrast with the all-embracing curvature singularity in the Schwarzschild solution. On the other hand, there is a very complicated topological behavior and a complete and unavoidable breakdown of the causality principle.

The significance of these results depends on the as yet unanswered question whether exterior fields of the Kerr type could or would result as the final state in a dynamical treatment of the collapse of a rotating body. A hint that this question may have a positive answer comes from the recent demonstration by Israel[5] that the Schwarzschild solution is unique among asymptotically-flat static-vacuum solutions in being bounded by a simple nonsingular Killing horizon ("simple" meaning that the constant-time cross sections are topologically spherical), which suggests that the family of stationary axisymmetric asymptotically flat vacuum solutions with the same property may also be very restricted. It can be conjectured that the low-angular-momentum Kerr fields may be the only examples. If this is the case, or even if there are other examples provided that these also have pathological behavior similar to that of the Kerr fields, then grave doubt will have been cast on the validity of Einstein's theory in its present form.

1. PHYSICAL AND TOPOLOGICAL STRUCTURE

A. Metric Form

The original and, for many purposes, the most useful form of the Kerr family of solutions of the source-free Einstein-Maxwell equations is given in terms of coordinates u, r, θ, and φ which can be interpreted most simply and naturally on a manifold formed by taking the topological product of a 2-plane on which u and r are Cartesian coordinates running from $-\infty$ to $+\infty$ and a 2-sphere on which θ and φ are ordinary spherical coordinates (φ is periodic with period 2π, and θ runs from 0 to π). The covariant form of the metric tensor is

* Present address: Institute of Theoretical Astronomy, Cambridge, England.
[1] R. Penrose, Phys. Rev. Letters **14**, 57 (1965).
[2] S. W. Hawking, Proc. Roy. Soc. (London) **A294**, 511 (1966).
[3] S. W. Hawking, Proc. Roy. Soc. (London) **A295**, 490 (1966).
[4] S. W. Hawking, Proc. Roy. Soc. (London) **A300**, 187 (1967).
[5] W. Israel, Phys. Rev. **164**, 1776 (1967).

expressed in terms of three parameters, m, e, and a by

$$ds^2=\rho^2d\theta^2-2a\sin^2\theta dr d\varphi+2dr du$$
$$+\rho^{-2}[(r^2+a^2)^2-\Delta a^2\sin^2\theta]\sin^2\theta d\varphi^2$$
$$-2a\rho^{-2}(2mr-e^2)\sin^2\theta d\varphi du$$
$$-[1-\rho^{-2}(2mr-e^2)]du^2, \quad (1)$$

and the corresponding covariant form of the electromagnetic field tensor is

$$F=2e\rho^{-4}[(r^2-a^2\cos^2\theta)dr\wedge du-2a^2r\cos\theta\sin\theta d\theta\wedge du$$
$$-a\sin^2\theta(r^2-a^2\cos^2\theta)dr\wedge d\varphi$$
$$+2ar(r^2+a^2)\cos\theta\sin\theta d\theta\wedge d\varphi], \quad (2)$$

where the abbreviations

$$\rho^2=r^2+a^2\cos^2\theta, \quad (3)$$

$$\Delta=r^2-2mr+a^2+e^2, \quad (4)$$

have been used, and where the usual symbol, $\wedge$, has been used for the operation of taking the antisymmetrized tensor product. (When $e=0$ the electromagnetic field vanishes and the metric satisfies the vacuum Einstein equations.)

These solutions are clearly stationary and axisymmetric with Killing vectors $\partial/\partial u$ and $\partial/\partial\varphi$, and it is also apparent that they are invariant under the discrete transformation of inversion about the equatorial hyperplane $\theta=\frac{1}{2}\pi$. Both the metric and the electromagnetic field forms are analytic except on the stationary ring $r=0$, $\theta=\frac{1}{2}\pi$, where ρ^2 vanishes. In fact, the curvature itself becomes singular as $\rho^2\to 0$ except in the special case where e and m both vanish. In this special case there must still be a singularity of the geometry at $\rho^2=0$, although the metric is then flat everywhere else. In all cases the metric and the electromagnetic field are well behaved throughout the rest of the manifold, except for the usual trivial degeneracy of spherical coordinates at $\theta=0$, $\theta=\pi$.

In all these spaces the Weyl tensor is of type D in the Petrov-Pirani classification, the two double principal null vectors being given by

$$-\partial/\partial r,$$
$$(r^2+a^2)\partial/\partial u+a\partial/\partial\varphi+\Delta\partial/\partial r. \quad (5)$$

By the Kundt and Trumper generalization[6] of the Goldberg-Sachs theorem[7] these are integrable to give two shear-free null geodesic congruences. The first of these (which is ingoing in the sense that r decreases in the time direction determined by increasing u) consists simply of the curves on which u, θ, and φ, are all constant, while the outgoing congruence is less simple in these coordinates. The principal null congruences have nonzero rotation (except when a vanishes, in which case the solutions are spherically symmetric) and therefore are not hypersurface-orthogonal. It was by making use of these structural properties of the Weyl tensor, and specifically looking for non-hypersurface-orthogonal solutions, that the empty space metrics of the family were derived by Kerr.[8] Subsequently these metrics were derived by Kerr and Schild,[9] from a systematic study of empty solutions whose metric tensor is (locally) the sum of a flat-space metric tensor and the tensor product of a null vector with itself. The charged solutions are also of this form, as can be seen by making the coordinate transformation

$$x+iy=(r+ia)e^{i\varphi}\sin\theta, \quad z=r\cos\theta, \quad t=u-r, \quad (6)$$

which gives the metric tensor as

$$ds^2=dx^2+dy^2+dz^2-dt^2+\frac{2mr-e^2}{r^4+a^2z^2}r^2$$
$$\times\left(\frac{r(xdx+ydy)-a(xdy-ydx)}{r^2+a^2}+\frac{zdz}{r}+dt\right)^2, \quad (7)$$

where r is determined implicitly in terms of x, y, z, by

$$r^4-(x^2+y^2+z^2-a^2)r^2-a^2z^2=0. \quad (8)$$

However, this Kerr-Schild form of coordinate system is rather awkward for studying global structures, because (as the price of imposing a flat-space background metric on a manifold with the topology described above) each set of values of the x, y, z, t coordinates corresponds to two different points, distinguished by the two different real values of r determined by (7). These coordinates have the further disadvantage that the axis symmetry is no longer manifest, but there is the compensating advantage that the degeneracy on the axis itself is removed.

The generalization of the solutions to include an electromagnetic field was originally achieved not by a systematic logical method but by an algebraic trick discovered by Newman and Janis[10] who succeeded in obtaining the empty-space Kerr solutions by transformation from the Schwarzschild solution (to which they reduce in the case when a vanishes). The charged generalization of the empty-space Kerr metrics was obtained by Newman, Couch, Chinnapared, Exton, Prakash, and Torrence[11] who applied an analogous transformation to the charged spherical solution of Reissner and Nordström (which is likewise the limiting case to which the charged solutions reduce when a vanishes).

Alternative systematic derivations of these solutions from different points of view, and with more explicit

[6] W. Kundt and M. Trumper, Akad. Wiss. Lit. Mainz **12** (1962).
[7] J. N. Goldberg and R. K. Sachs, Acta Phys. Polon. **22**, Suppl. 13 (1962).
[8] R. P. Kerr, Phys. Rev. Letters **11**, 237 (1963).
[9] R. P. Kerr and A. Schild, Am. Math. Soc. Symposium, New York, 1964.
[10] E. T. Newman and A. I. Janis, J. Math. Phys. **6**, 915 (1965).
[11] E. T. Newman, E. Couch, R. Chinnapared, A. Exton, A. Prakash, and R. Torrence, J. Math. Phys. **6**, 918 (1965).

information about the curvature, have been given by Carter[12] and Ernst.[13]

B. Rotating Body Interpretation

Despite its many advantages the coordinate system (1) (which will subsequently be referred to as the Kerr-Newman form) has the drawback that it does not display the full symmetry of the space.

Papapetrou[14] has shown that any connected stationary axisymmetric solution of Einstein's empty-space equations must have an additional discrete symmetry under simultaneous inversion of the axial and stationary Killing vectors, while Boyer and Lindquist[15] have independently discovered a specific transformation which casts the empty-space Kerr metrics into a form which is manifestly invariant under such an inversion. Carter[16] has shown that Papapetrou's result can be generalized to include cases where the space is nonempty, provided that the matter tensor is itself invariant under simultaneous inversion of the time and axial angle and that this situation holds automatically if the only contribution to the matter tensor comes from a source-free electromagnetic field. (It would not necessarily hold in the presence of a perfect fluid.) Thus Papapetrou's result generalizes directly to the solutions of the source-free Maxwell-Einstein equations, and hence applies to the charged Kerr solutions.

The specific transformation needed to obtain a manifestly invertible form is an immediate generalization of the one given by Boyer and Lindquist; thus introducing new time and angle coordinates $\hat{t}$ and $\hat{\varphi}$ defined by

$$\begin{aligned} d\hat{t} &= du-(r^2+a^2)\Delta^{-1}dr\,, \\ d\hat{\varphi} &= d\varphi-a\Delta^{-1}dr\,, \end{aligned} \tag{9}$$

we obtain the metric tensor form as

$$\begin{aligned} ds^2=\rho^2\Delta^{-1}dr^2+\rho^2d\theta^2+\rho^{-2}\sin^2\theta[ad\hat{t}-(r^2+a^2)d\hat{\varphi}]^2 \\ -\rho^{-2}\Delta[d\hat{t}-a\sin^2\theta d\hat{\varphi}]^2\,, \end{aligned} \tag{10}$$

where the cross terms between the ignorable coordinates and the others have been eliminated. The electromagnetic field tensor now takes the form

$$\begin{aligned} F=2e\rho^{-4}(r^2-a^2\cos^2\theta)dr\wedge[d\hat{t}-a\sin^2\theta\, d\hat{\varphi}] \\ -4e\rho^{-4}ar\cos\theta\sin\theta d\theta\wedge[ad\hat{t}-(r^2+a^2)d\hat{\varphi}]. \end{aligned} \tag{11}$$

In this system (which will be subsequently referred to as the Boyer-Lindquist form) it is immediately clear how the metrics reduce to the familiar forms of the Schwarzschild and Reissner-Nordström solutions when a vanishes. (That the metrics are flat when both e and m vanish can be seen more easily from the Kerr-Schild form unless one is familiar with spheroidal coordinates.) It is also clear that the spaces are asymptotically flat, in both the local and the global sense, in the limits of large positive or negative values of r. The Boyer-Lindquist form is ideal for the examination of the asymptotic behavior of the fields, on which the physical interpretation of the parameters is based.

It can be easily seen from (10) and (11) by analogy with the Schwarzschild and Reissner-Nordström solutions that (in unrationalized units with Newton's constant G and the speed of light c both set equal to unity) m represents the mass and e the charge in the limit of large positive r, and that the mass and charge are, respectively, $-m$ and $-e$ in the limit of large negative r. There is no loss of generality in assuming, as we shall do from now on, that m is positive; this is simply equivalent to choosing which of the two asymptotically flat regions we shall label with positive values of r.

The interpretation of the parameter a requires more care, since its effects are of asymptotically higher order. In confirmation of a remark originally made by Kerr,[8] Boyer and Price[17] have shown, by a careful examination of the geodesics in the equatroial plane in the uncharged case, that it gives rise to Coriolis-type forces which are asymptotically identical to those which one would expect from a rotating body with angular momentum ma in the weak-field limit (cf. also Cohen[18]). As the effects of the charge on the metric are of asymptotically higher order than those of the mass, it can be seen that this conclusion still stands in the charged case. Thus a is what may be called the specific angular momentum. The metric form used here has been adjusted so that a positive value of a corresponds to a positive sense of rotation (it turned out to be the other way round in the form used by Boyer and Price).

It is the presence of rotational effects which gives the Kerr solutions their importance. This family includes all solutions yet known which could represent the exterior fields of rotating charged or uncharged bodies, other asymptotically flat solutions such as those of Weyl[19] or Papapetrou[20] being either static or massless. Whether physically natural interior material solutions exist (e.g., a simply rotating perfect-fluid body) of which these are the exterior fields is not yet known. Boyer[21] has given conditions which the surface of a perfect-fluid interior would have to satisfy. Zel'dovich and Novikov[22] attempted to argue from the apparent absence of simultaneous inversion symmetry (at a time when the Boyer-Lindquist transformation had not yet been published) that such a body must contain meridional circulation. Now that the inversion symmetry is

[12] B. Carter, J. Math. Phys. (to be published).
[13] F. J. Ernst, Phys. Rev. **167**, 1175 (1968); **168**, 1415 (1968).
[14] A. Papapetrou, Ann. Inst. H. Poincaré 4, 83 (1966).
[15] R. H. Boyer and R. W. Lindquist, J. Math. Phys. **8**, 265 (1967).
[16] B. Carter, Comm. Math. Phys. (to be published).
[17] R. H. Boyer and T. G. Price, Proc. Camb. Phil. Soc. **61**, 531 (1965).
[18] J. M. Cohen, J. Math. Phys. **8**, 1477 (1967).
[19] W. Weyl, Ann. Physik **54**, 117 (1917).
[20] A. Papapetrou, Ann. Physik **12**, 309 (1953).
[21] R. H. Boyer, Proc. Camb. Phil. Soc. **61**, 527 (1965).
[22] Ya. B. Zel'dovich and I. D. Novikov, Zh. Eksperim i Teor. **49**, 170 (1965) [English transl.: Soviet Phys.—JETP **22**, 122 (1966)].

known, one might be tempted to argue the other way round. However, the results of Papapetrou and Carter which have just been mentioned show that no such deductions can be made at all since the invertibility of the exterior is inevitable in any case.

Just as the parameter a couples with the mass to give the angular momentum, also [as can be seen from the form (11) of the field] it couples with the charge to give an asymptotic magnetic dipole moment ea. There is no freedom of variation of the gyromagnetic ratio which is simply e/m. It is noteworthy that this is exactly the same as the gyromagnetic ratio predicted for a spinning particle by the simple Dirac equation, which is obeyed to quite a high accuracy by the electron. Therefore, despite the fact that the parameters of the solutions contain only two adjustable ratios, it is possible to choose them in such a way that they agree with the corresponding parameters for an electron, for which, in units with $\hbar=1$, the mass, angular momentum, and squared charge are given by $m\approx 10^{-22}$, $ma=\frac{1}{2}$, $e^2\approx 1/137$ from which we obtain $a\approx\frac{1}{2}m^{-1}\approx 10^{22}$, $e\approx\frac{1}{12}$. The value of the length scale determined by a is therefore quite large, in fact, about the same as the Compton radius. On the other hand, the value of m is so small that the field differs very little from the limiting case $m=0$, with e and a as the only parameters.

Despite its great elegance the Boyer-Lindquist form unfortunately fails where Δ vanishes. This will occur whenever m is greater than the critical value

$$m^2=a^2+e^2, \tag{12}$$

in which case Δ has a zero at each of the two values of r (both positive) defined by

$$r_\pm=m\pm(m^2-a^2-e^2)^{1/2} \tag{13}$$

and is negative in between them. In the intermediate region the solution changes character. It can be seen clearly from the form (10) that this region cannot be regarded as stationary in the strict sense since there are no longer any timelike vectors in the planes ($r=$ const, $\theta=$ const) of the Killing vectors, but instead r has taken over the role of a timelike variable. As a consequence of the simultaneous inversion symmetry, the hypersurfaces bounding this region must be null, by a theorem given by Carter[16] and must, moreover, satisfy the strict definition of a Killing horizon given in that reference. In the limiting case when a and e both vanish, the inner horizon collapses onto the central singularity and the outer horizon becomes the well-known Schwarzschild horizon at $r=2m$. When the equality (12) is satisfied the two horizons coalesce at $r=m$. When $a^2+e^2>m^2$ there are no Killing horizons and, as will subsequently be proved, the manifold is geodesically complete except for those geodesics which reach the central singularity at $\rho^2=0$. However, when $a^2+e^2\leqslant m^2$, although the local failure of the metric can be cured by reverting to the Kerr-Newman coordinate system, the manifold defined above remains incomplete as r tends to $r_\pm$ since there are geodesics for which the coordinate u becomes unbounded within a finite affine distance. In Sec. C, the analytic extensions required to remedy this defect will be discussed and, subsequently, when the geodesic equations have been integrated, it will be shown that the extended manifolds so obtained are in fact geodesically complete, again with the exception of the geodesics which reach the singularities at $\rho^2=0$.

C. Maximal Analytic Extension

The most suitable basic unit for building up the extended manifold is the original Kerr-Newman coordinate patch (1) which is connected to the invertible Boyer-Lindquist form (10) by the transformation (9). The starting point for the extension is the remark that the invertible form can be extended in a symmetric manner in an inverted direction in terms of new time and angle coordinates w and $\tilde{\varphi}$ by the transformation

$$\begin{aligned} d\hat{t}&=-dw+(r^2+a^2)\Delta^{-1}dr\,,\\ d\hat{\varphi}&=-d\tilde{\varphi}+a\Delta^{-1}dr\,. \end{aligned} \tag{14}$$

The resulting form for the metric is

$$\begin{aligned} ds^2=\rho^2d\theta^2&-2a\sin^2\theta dr\,d\tilde{\varphi}+2dr\,dw\\ &+\rho^{-2}[(r^2+a^2)^2-\Delta a^2\sin^2\theta]\sin^2\theta d\theta\\ &-2a\rho^{-2}(2mr-e^2)\sin^2\theta d\tilde{\varphi}\,dw\\ &-[1-\rho^{-2}(2mr-e^2)]dw^2\,, \end{aligned} \tag{15}$$

which is formally identical to the original Kerr-Newman form. The transformation between the two Kerr-Newman forms can be given directly as

$$du+dw=2(r^2+a^2)\Delta^{-1}dr\,, \tag{16}$$

$$d\varphi+d\tilde{\varphi}=2a\Delta^{-1}dr\,. \tag{17}$$

In the case when the zeros of Δ coincide, i.e., when $a^2+e^2=m^2$ so that $r_+=r_-=m$, we can proceed directly to the extended manifold. The transformations (16) and (17) give rise to complete ranges of the new coordinates in each of the regions $r>m$ and $r<m$, and they therefore describe two distinct extensions applying to each of these regions separately. By performing these two types of extension alternately, one can build up an extended manifold consisting of an infinite sequence of (u,r,θ,φ) patches, labelled $\cdots$, $(n,-)$, $(n+1,-)$, $\cdots$ linked transversely by a symmetrically arranged sequence of $(w,r,\theta,\tilde{\varphi})$ patches, labelled $\cdots$, $(-,n)$, $(-,n+1)$, $\cdots$ with overlaps alternately in the regions $r<m$ and $r>m$. The overlap region between $(n,-)$ and $(-,l)$ will be denoted (n,l). By adjusting the relative values of n and l one can arrange that the nonempty overlaps (n,l) are such that $n=l$ (for a region $r<m$) or $n=l+1$ (for a region $r>m$). It will be shown in a subsequent section that the manifold so obtained is in fact geodesically complete, except for those geodesics which

reach the singularities at $\rho^2=1$, i.e., it is a maximal extension.

In the general case when the zeros of Δ are distinct, i.e., when $a^2+e^2<m^2$, the extension is more difficult. One can start in the same way as in the previous case, except that (16) and (17) now give rise to three distinct transformations instead of two, corresponding to the regions I; $r>r_+$; II: $r_+>r>r_-$; and III: $r_-<r$; it is therefore natural to build up an extended manifold again consisting of an infinite sequence of (u,r,θ,φ) patches labelled $\cdots$, $(n,-)$, $(n+1,-)$, $\cdots$ overlapping a symmetric sequence of $(w,r,\theta,\tilde{\varphi})$ patches labelled $\cdots$, $(-,n)$, $(-,n+1)$, $\cdots$ is such a way that the overlap region (n,l) is nonempty only if $n=l$ (in which case it is of the type II) or if $n=l\pm1$ (in which case it is of the type I if n is odd and l even, and of type III if it is the other way around) (cf. the illustrations given by Carter[23] in a discussion of the restriction of this manifold to the axis of symmetry).

However, the manifold just described still has an incompleteness associated with the Killing horizons; there are 2-surfaces missing where u and w both tend to infinity together. The crux of the extension program is the construction of new coordinate patches to include these missing 2-surfaces. In preparation for this construction we introduce the u and w coordinates simultaneously, and drop r as a coordinate, instead treating it as a function of u and w given implicitly (once the region I, II, or III has been specified) by

$$F(r)=u+w, \tag{18}$$

where by (16) we have

$$F(r)=2r+\kappa_+^{-1}\ln|r-r_+|+\kappa_-^{-1}\ln|r-r_-| \tag{19}$$

with the constants $\kappa_\pm$ defined by

$$\kappa_\pm=\tfrac{1}{2}(r_\pm^2+a^2)^{-1}(r_\pm-r_\mp). \tag{20}$$

We also use the device, introduced by Boyer and Lindquist[15] for dealing with the uncharged case, of defining a new angle variable, constant along the trajectories of that particular Killing vector field which coincides with the null generators on the Killing horizon. From (5) we see that this Killing vector field is $(r_\pm^2+a^2)\partial/\partial u+a\partial/\partial\varphi$ in terms of the original Kerr-Newman coordinates (1), which unfortunately depends on which of the Killing horizons $r=r_\pm$ is under consideration. Thus we shall need two alternative new angle coordinates which we shall denote by $\varphi^\pm$ and which can be defined by

$$2d\varphi^\pm=d\varphi-d\tilde{\varphi}-a(r_\pm^2+a^2)^{-1}(du-dw). \tag{21}$$

Thus we obtain the symmetric double quasi-null-metric form:

$$ds^2=\rho^{-2}\Delta\left(\frac{\rho^2}{r^2+a^2}+\frac{\rho_\pm^2}{r_\pm^2+a^2}\right)\frac{(r^2-r_\pm^2)a^2\sin^2\theta}{(r^2+a^2)(r_\pm^2+a^2)}\tfrac{1}{4}(du^2+dw^2)+\rho^{-2}\Delta\left[\frac{\rho^4}{(r^2+a^2)^2}+\frac{\rho_\pm^4}{(r_\pm^2+a^2)^2}\right]\tfrac{1}{2}(du\,dw)+\rho^2d\theta^2$$
$$-\rho^{-2}\Delta a\sin^2\theta\left[a\sin^2\theta d\varphi^\pm-\frac{\rho_\pm^2}{r_\pm^2+a^2}(du-dw)\right]d\varphi^\pm+\rho^{-2}\sin^2\theta\left[a\frac{r_\pm^2-r^2}{r_\pm^2+a^2}\tfrac{1}{2}(du-dw)-(r^2+a^2)d\varphi^\pm\right]^2, \tag{22}$$

where the obvious abbreviation $\rho_\pm^2=r_\pm^2+a^2\cos^2\theta$ has been introduced.

This form is in itself even more limited in range than the Kerr-Newman form from which we started; in fact, it covers the same patches as the Boyer-Lindquist form, depending on which of the regions I, II, and III the solution of (18) is specified to lie in. However, we are now in a position to give a direct generalization of the method used by Carter[23] for the symmetry axis. When this method was originally devised a rather complicated transformation was used whose purpose was not only to cover the missing regions separately, but also to cover the whole manifold by a single coordinate patch. This was worthwhile when the symmetry axis alone was under consideration; it could also be done here (except that there would remain the trivial degeneracy at $\theta=0$, $\theta=\pi$ and the curvature singularities at $\rho^2=0$) but it would be a messy process because the angular coordinates $\varphi^\pm$ required at the two horizons are different, so that the angular coordinate would have to be gradually changed in between, which would destroy the manifest axisymmetry of the manifold. Instead of doing this we shall be content with covering the missing pieces one at a time. With this more limited objective it is possible to choose a coordinate transformation which is very simple indeed.

We introduce new coordinates x, y and construct a patch with coordinates $(x,y,\theta,\varphi^\pm)$ to cover the four $(u,w,\theta,\varphi^\pm)$ patches adjacent to a missing region at $r=r_\pm$. [Two of these patches will have the form $(n,n+1)$, $(n+1,n)$ and will cover regions of type II, and the other two will have the form (n,n), $(n+1,n+1)$ and will cover regions of type I or type III according to whether r_+ or r_- is under consideration.] The new coordinates are defined by the simple transformation

$$x=(\pm)e^{\kappa_\pm u},\quad y=(\pm)e^{\kappa_\pm w}, \tag{23}$$

where the sign $(\pm)$ in the definition of x changes between the two (u,r,θ,φ) patches involved, and the sign $(\pm)$ in the definition of y changes between the two $(w,r,\theta,\tilde{\varphi})$ patches involved. We shall choose the signs so that the product xy is positive in the two regions where

[23] B. Carter, Phys. Rev. **141**, 1242 (1966).

$r-r_\pm$ is positive, and negative in the other two. (This still leaves an arbitrary choice of sign in the definition of x and y but because of the inversion symmetry it will not be necessary to make it explicitly.)

By (18) and (19), r will now be determined in terms of x and y by

$$xy=(r-r_\pm)G_\pm^{-1}(r)\,, \tag{24}$$

where $G_\pm(r)$ is defined by

$$G_\pm(r)=e^{-2\kappa_\pm r}|r-r_\pm|^{|\kappa_\pm/\kappa|_\mp}. \tag{25}$$

Thus r is an analytic function of x and y since in the whole (x,y) plane r lies between $r_\pm$ and $\pm\infty$. Thus we obtain the new metric form

$$ds^2=\rho^{-2}\left(\frac{\rho^2}{r^2+a^2}+\frac{\rho_\pm^{\,2}}{r_\pm^{\,2}+a^2}\right)\frac{(r-r_\mp)(r+r_\pm)a\sin^2\theta}{(r^2+a^2)(r_\pm^{\,2}+a^2)}\kappa_\pm^{\,2}G_\pm^{\,2}(r)\tfrac{1}{4}(y^2dx^2+x^2dy^2)+\rho^{-2}\left(\frac{\rho^4}{(r^2+a^2)^2}+\frac{\rho_\pm^{\,4}}{(r_\pm^{\,2}+a^2)^2}\right)$$
$$\times(r-r_\mp)\kappa_\pm^{\,2}G_\pm(r)\tfrac{1}{2}(dx\,dy)+\rho^2d\theta^2-\rho^{-2}a\sin^2\theta\left(\Delta a\sin^2\theta d\varphi^\pm-\frac{\rho_\pm^{\,2}}{r_\pm^{\,2}+a^2}(r-r_\mp)G_\pm(r)\kappa_\pm(ydx-xdy)\right)d\varphi^\pm$$
$$+\rho^{-2}\sin^2\theta\left((r^2+a^2)d\varphi^\pm+a\frac{r+r_\pm}{r_\pm^{\,2}+a^2}\kappa_\pm G_\pm(r)\tfrac{1}{2}(ydx-xdy)\right)^2. \tag{26}$$

This metric is clearly analytic everywhere on the $(x,y,\theta,\varphi^\pm)$ patch except at the curvature singularities $\rho^2=0$. It must also be checked that it is nondegenerate on the Killing horizons at $x=0$ and $y=0$ since the transformation we have used is singular there. This is also immediately verifiable. It follows that it is nondegenerate everywhere except at the curvature singularities and (trivially) at $\theta=0$, $\theta=\pi$.

With the additional points on the 2-surfaces $x=0$, $y=0$ in the new coordinate patches, the extension that we have obtained is maximal, since it is now geodesically complete except for geodesics which reach the ring singularity. We shall be able to prove this in Sec. 3B after the integrals of the geodesic equations have been obtained.

Although the fact that it exists is of importance, the form (26) is too complicated to have much practical use. However, it does clearly show the existence of spacelike hypersurfaces $x=K^2y$, where K is any real nonvanishing constant, which extend right across the manifold, one such hypersurface passing through each point of the regions $r>r_+$. Analogous hypersurfaces in the regions $r<r_-$ do not exist, because of the curvature singularity at $\rho^2=0$.

2. CAUSALITY

A. Causally Well-Behaved Parts of the Kerr Solutions

In this discussion the causality principle means the condition that there exist no closed causal (i.e., timelike or null) curves in the space under consideration. This condition can be violated in two ways: We shall refer to the violation as trivial if none of the closed causal curves are homotopic to zero, since in this case we may construct a covering space in which the causality principle is satisfied, and we may use the covering space for purposes of physical interpretation; we shall refer to causality violation where there exist closed timelike lines homotopic to zero as unavoidable since in this case it could only be removed by altering the local structure of the space, not merely its global connectivity properties. Some of the possibilities of trivial causality violation in members of this family have been discussed previously. Fuller and Wheeler[24] considered the possibility of causality violation resulting from multiconnectedness introduced when the two asymptotically flat backgrounds in the analytically extended Schwarzschild space are identified so that the Kruskal throat becomes a wormhole; they showed that, in fact, causality violation cannot arise in this way. On the other hand, the author[23,25] pointed out that identifications of this kind could lead to causality violations in the Reissner-Nordström and Kerr solutions (the argument in the latter case depending purely on the properties of the symmetry axis). However, in all these cases the causality violation being contemplated results from unnecessary identifications which produce multiconnectedness. In this section, we shall be considering causality violation of the unavoidable kind first studied in Godel's universe.

We have seen in Sec. 1 that, by the mode of its construction, the extended spaces consist of a combination of patches of type I ($r>r_+$) or III ($r<r_-$) in which the surfaces of transitivity are everywhere timelike and of type II ($r_-<r<r_+$) in which the surfaces of transitivity are everywhere spacelike, and that these patches are separated by null hypersurfaces—the Killing horizons. Provided we do not unnecessarily identify some of these patches, but piece them together exactly in the manner described in Sec. 1 and illustrated in the diagrams of Ref. 23, a causal curve which leaves one of these patches can never re-enter. It follows that insofar as we are considering only nontrivial causality

[24] R. W. Fuller and J. A. Wheeler, Phys. Rev. **128**, 919 (1962).
[25] B. Carter, Phys. Letters **21**, 423 (1966).

violation we can consider each of these patches separately.

A very useful criterion for the nonexistence of closed causal curves in a time-oriented space is the presence of a spacelike hypersurface which is a properly immersed submanifold in the sense of Sternberg[26] since it is impossible for a closed causal curve to intersect such a hypersurface if the curve is homotopic to zero and therefore impossible altogether in a simply connected space. (This follows from the fact that since the immersion is proper, the number of times the curve crosses the hypersurface can change only by two at a time under the homotopy, so that if the homotopy starts from zero there will at all stages be a one-one correspondence between the crossings in the forward and backward time directions, whereas a causal curve can cross only in the forward time direction.) Hawking[4] and Geroch[27] have given (different) constructions by which a covering space can be constructed which preserves the topology of such a spacelike hypersurface but at the same time removes all closed causal curves through it by unwinding them. This shows explicitly that a space with a properly immersed spacelike hypersurface through each point cannot have nontrivial causality violation.

We can apply this criterion to the Kerr solutions. Thus we saw in Sec. 1 that when $a^2+e^2<m^2$ such a spacelike hypersurface exists through any point in region I ($r>r_+$) and it is clear from the Boyer-Lindquist form (10) that the hypersurfaces $r=$const through any point of the regions II ($r_-<r<r_+$) also satisfy the required conditions. Therefore each connected region $r>r_-$ in the manifold we have constructed can have no nontrivial causality violation; furthermore, since each such region is simply connected, the possibility of trivial causality violation does not arise and hence the whole of each region $r>r_-$ is causally well behaved. If $a^2+e^2=m^2$ the surfaces $\hat{t}=$const in the Boyer-Lindquist form also satisfy the required condition, so that we can conclude in this case also that each region $r>r_-$ ($=m$) is causally well behaved.

On the other hand, even when $a^2+e^2\leqslant m^2$ we cannot draw such conclusions for the regions $r<r_-$, and we cannot apply this criterion anywhere when $a^2+e^2>m^2$.

B. Causality Violation in the Kerr Solutions

In a separate paper[28] the author has derived a criterion for causal bad behavior which is applicable to any space with an Abelian isometry group which is everywhere transitive over timelike surfaces. This criterion states that if there does not exist a Lie algebra covector (i.e., a linear map of the Lie algebra onto the real numbers) such that the corresponding differential form in each surface of transitivity is everywhere spacelike or null, then the whole space is a single nontrivially vicious set. In a terminology which generalizes the concept of a closed timelike curve (considered as a vicious cycle), a vicious set is defined as a set in which any point can be connected to any other point by both a future and a past directed timelike curve, i.e., it is one in which the causality principle is violated in the most flagrant conceivable manner; by nontrivially vicious it is meant that the same property holds in any covering space so that the implied causality violation is nontrivial in the sense used in the previous section.

Now the group generated by $\partial/\partial\hat{t}$, $\partial/\partial\hat{\varphi}$ in the Boyer-Lindquist form satisfies the required conditions for applying this criterion in the regions $r<r_-$ when $a^2+e^2\leqslant m^2$ and in the whole manifold when $a^2+e^2>m^2$, i.e., in the regions where the previous criterion failed. Choosing a Lie algebra covector means in effect choosing a differential form $\omega=Kd\hat{t}+Ld\hat{\varphi}$ on the surfaces of transitivity where K and L are arbitrary constants. The criterion will be satisfied if it is not possible to choose K, L so that ω is everywhere spacelike with respect to the induced metric in the surfaces of transitivity. It is easy to see that the most obvious choice, $\hat{t}$ itself, is spacelike everywhere except in the subregion where

$$r^2+a^2+\rho^{-2}(2mr-e^2)a^2\sin^2\theta<0. \tag{27}$$

However, it can easily be checked that no choice of ω satisfies the required conditions over the whole region except in the spherically symmetric cases. Therefore in the case when $a^2+e^2\leqslant m^2$ ($a\neq 0$) each region $r<r_-$ is a vicious set and the boundaries $r=r_-$ are causality horizons; in the case when $a^2+e^2>m^2$ ($a\neq 0$) the whole space is a single vicious set. The essential details of this causality violation may be understood as follows.

In the uncharged case, condition (27) is satisfied in a small region of negative r in the immediate neighborhood of the singularity $\rho^2=0$, and in the charged case it is satisfied in a larger region including positive values of r, although never extending beyond a point where r^2 is equal to e^2 on the positive r side, or where r^2 is equal to the greatest of a^2, e^2, or $4m^2$ on the negative r side. In this region the vector $\partial/\partial\hat{\varphi}$ is timelike, so the circles $\hat{t}=$const, $r=$const, $\theta=$const, are themselves closed timelike lines. However, although it is necessary that any closed timelike lines should enter the region defined by (27), our application of the criterion of Ref. 28 shows that they are by no means restricted to it but can extend to any part of one of the regions $r<r_-$ or over the whole space when $a^2+e^2>m^2$. This criterion also implies that they cannot be removed by taking a covering space.

Actually, since the manifold as a whole, as it has been described so far, is simply connected, there is no proper covering space, but because the geometry is singular at $\rho^2=0$, one might just as well consider the manifold from which these ring singularities have been excluded,

[26] S. Sternberg, *Lectures on Differential Geometry* (Prentice-Hall, Inc., Englewood Cliffs, N. J., 1964).
[27] R. Geroch, J. Math. Phys. 8, 782 (1967).
[28] B. Carter, Ph.D. thesis, University of Cambridge, England, 1967 (unpublished).

as far as the physics is concerned. Since there would then be curves not homotopic to zero, looping around the ring singularities, it would be possible to construct numerous covering spaces by partially or totally unwinding them. When $a^2+e^2>m^2$ the universal covering space will consist of a simple infinite linear sequence, but when $a^2+e^2\leqslant m^2$ it will have an extremely complicated unendingly branching topology. However, since the closed timelike lines do not need to loop round the ring, their existence is not affected by this process.

A more drastic way of obtaining a covering space would be to cut out the symmetry axis $\theta=0$, $\theta=\pi$ and unwind the remaining space by treating $\hat{\varphi}$ as a nonperiodic coordinate. This would not be a physically reasonable process because it would create an artificial singularity in the limit $\theta\rightarrow 0$ or $\theta\rightarrow\pi$. However, even this would not be sufficient to remove the closed timelike lines, because the impossibility of finding a suitable combination ω does not depend on the presence of the axis. (The simple circles described above would of course cease to exist.)

It is fairly easy to understand the general nature of the more complicated closed timelike lines. Outside the region (27) the coordinate $\hat{t}$ must increase continually along a timelike line, although r and θ may be varied in any direction at will in the region under consideration ($r<r_-$ or the whole space for $a^2+e^2>m^2$) as may $\hat{\varphi}$ also except in a limited region satisfying the condition

$$\rho^2+e^2-2mr<0\,, \tag{28}$$

where $\partial/\partial\hat{t}$ becomes spacelike. In order to make up literally for lost time the path must enter the region (27). Here time can be gained, but only at the expense of clocking up a large change (negative for $a>0$) in the angle $\hat{\varphi}$. It can be seen that in all cases the least upper bound to the time that can be saved per unit change in angle is $|a|$. However, this does not prevent the line from being closed even when the symmetry axis is removed and the coordinate $\hat{\varphi}$ made nonperiodic, because by letting the line proceed to within a sufficiently small but finite distance from the symmetry axis, all the lost angle can be made up at a very small cost in time.

To sum up, in the case when $a^2+e^2>m^2$, the central region has the properties of a time machine. It is possible, starting from any point in the outer regions of the space, to travel into the interior, move backwards in time ($\hat{t}$) as far as desired, at a rate up to $2\pi|a|$ per revolution about the axis, and then return to the original position. (By keeping the motion at all stages sufficiently close to the light cone, the proper time involved in the process could be kept below any given nonzero limit, although this would not be possible if some sort of bound were to be placed on the allowed acceleration.)

In the case when $a^2+e^2\leqslant m^2$ the outer parts of the space on the positive r side are causally well behaved, and there is even a partial Cauchy surface. However, the null hypersurface $r=r_-$ is a causality horizon, for by (irreversibly) crossing it a timelike path can enter a region where causality is violated just as in the previous case.

3. GEODESICS AND ORBITS

A. Integration of Geodesic and Orbit Equations

The equations of motion of a test particle of mass μ and charge ϵ are given by

$$D^2x^i/D\tau^2=(\varepsilon/\mu)F_k{}^i(Dx^k/D\tau)\,, \tag{29}$$

where $D/D\tau$ denotes covariant differentiation with respect to the proper time τ, and F is the electromagnetic field tensor. These equations may be derived from the Lagrangian

$$L=\tfrac{1}{2}g_{ij}\dot{x}^i\dot{x}^j+\varepsilon A_i\dot{x}^i\,, \tag{30}$$

where the covariant vector potential A has been introduced, satisfying

$$F=2dA\,, \tag{31}$$

and where a dot over a symbol denotes ordinary differentation with respect to an affine parameter λ. In order to obtain (29), λ must be related to the proper time by

$$\tau=\mu\lambda \tag{32}$$

which is equivalent to imposing the normalizing condition

$$g_{ij}\dot{x}^i\dot{x}^j=-\mu^2. \tag{33}$$

By taking zero and negative values of μ^2 in (33) and setting $\epsilon=0$, the same Lagrangian (30) can be used to give null and spacelike geodesics. [When there is no charge, the actual value of the mass has no significance, and so we may obtain timelike and spacelike geodesics with λ as a metric parameter by setting $\mu^2=\pm 1$ in (33).]

In order to transform to a Hamiltonian formulation, we introduce the momenta obtained from (30) as

$$p_i=g_{ij}\dot{x}^j+\varepsilon A_i \tag{34}$$

and thus obtain the Hamiltonian

$$H=\tfrac{1}{2}g^{ij}(p_i-\epsilon A_i)(p_j-\epsilon A_j)\,. \tag{35}$$

Since it does not depend explicitly on λ, the Hamiltonian is automatically a constant of the motion, and it is apparent that it is this constant which is determined by the normalizing condition (33). Thus we have

$$H=-\tfrac{1}{2}\mu^2. \tag{36}$$

We shall work with the Kerr-Newman form of the metric since it is simple and since the corresponding coordinate patches cover the whole manifold except for the 2-surfaces ($x=0$, $y=0$) in the form (26).

The simplest vector potential giving rise to the field (2) by (31) is

$$A=e\rho^{-2}r(du-a\sin^2\theta d\varphi)\,. \tag{37}$$

Thus from (1) we obtain the momenta

$$p_u=-[1-\rho^{-2}(2mr-e^2)]\dot{u}-a\rho^{-2}(2mr-e^2)\times\sin^2\theta\,\dot{\varphi}+\dot{r}+e\rho^{-2}r, \quad (38)$$

$$p_\varphi=-a\rho^{-2}(2mr-e^2)\sin^2\theta\,\dot{u}+\rho^{-2}[(r^2+a^2)^2-\Delta a\sin^2\theta]\sin^2\theta\,\dot{\varphi}-a\sin^2\theta\,\dot{r}-e\rho^{-2}ar\sin^2\theta, \quad (39)$$

$$p_r=\dot{u}-a\sin^2\theta\,\dot{\varphi}, \quad (40)$$

$$p_\theta=\rho^2\dot{\theta}. \quad (41)$$

The inverse of the metric (1) is

$$(\partial/\partial s)^2=\rho^{-2}(\partial/\partial\theta)^2+2\rho^{-2}(r^2+a^2)(\partial/\partial r)(\partial/\partial u)+2\rho^{-2}a(\partial/\partial r)(\partial/\partial\varphi)+2\rho^{-2}a(\partial/\partial u)(\partial/\partial\varphi)+\rho^{-2}a\sin^2\theta(\partial/\partial u)^2+\rho^{-2}\sin^2\theta(\partial/\partial\varphi)^2+\rho^{-2}\Delta(\partial/\partial r)^2 \quad (42)$$

from which we obtain the Hamiltonian

$$H=\tfrac{1}{2}\rho^{-2}\{\Delta p_r^2+2[(r^2+a^2)p_u+ap_\varphi-\varepsilon er]p_r+p_\theta^2+[a\sin\theta p_u+\sin^{-1}\theta p_\varphi]^2\}. \quad (43)$$

From the symmetries we immediately obtain two constants of the motion corresponding to conservation of energy, E, and angular momentum about the symmetry axis, Φ; thus we have

$$p_u=-E, \quad (44)$$

$$p_\varphi=\Phi. \quad (45)$$

In addition, we automatically have the constant of the motion given by (36), corresponding to conservation of rest mass.

These three first integrals are sufficient to determine the motion only when some restriction is imposed which reduces the problem effectively to three or fewer dimensions. This situation holds for the spherical cases, for which a thorough analysis in the uncharged case has been carried out by Darwin,[29,30] and for which a discussion of the charged cases has been given by Graves and Brill.[31] It also applies to suitable subspaces in the fully general cases, namely, the symmetry axis, which has been analyzed by Carter,[23] and the equatorial symmetry plane, which has been analyzed by Boyer and Price[17] in the asymptotically flat limit, and by Boyer and Lindquist[15] in the inner regions (all these applying to the empty-space cases only).

In order to tackle the general case, a fourth first integral of the motion is needed which cannot come from the obvious symmetries of the metric. However, it turns out that it is possible to obtain such an integral by taking advantage of the unexpected fact that the Hamilton-Jacobi equation can be solved by separation of variables in the coordinate system (1), and with the choice of gauge (37). (The method would also work in in the Boyer-Lindquist coordinates with the analogous choice of vector potential, but a transformation involving the nonignorable coordinates r, θ would destroy the separability.)

By (35) the general form of the Hamilton-Jacobi equation is

$$\partial S/\partial\lambda=\tfrac{1}{2}g^{ij}[(\partial S/\partial x^i)-\varepsilon A_i][(\partial S/\partial x^j)-\varepsilon A_j], \quad (46)$$

where S is the Jacobi action.

If there is a separable solution, then in terms of the already known constants of the motion it must take the form

$$S=-\tfrac{1}{2}\mu^2\lambda-Eu+\Phi\varphi+S_\theta+S_r, \quad (47)$$

where S_θ and S_r are, respectively, functions of θ and r only. Inserting this in (43), we see that the equation can in fact be separated in the form

$$(dS_\theta/d\theta)^2+a^2\mu^2\cos^2\theta+(aE\sin\theta-\Phi\sin^{-1}\theta)^2=-\Delta(dS_r/dr)^2+2[(r^2+a^2)E-a\Phi+\varepsilon er]dS_r/dr-\mu^2r^2. \quad (48)$$

Thus both sides must be equal to a new constant of the motion, which we shall denote by $\mathcal{K}$. It can be seen from the form of the right-hand side that $\mathcal{K}$ must be positive whenever μ is real, i.e., for all particle orbits and timelike or null geodesics. Using the relations $p_\theta=\partial S/\partial\theta$ and $p_r=\partial S/\partial r$, it may be related directly to the momenta in the form

$$p_\theta^2+(aE\sin\theta-\Phi\sin^{-1}\theta)^2+a^2\mu^2\cos^2\theta=\mathcal{K}, \quad (49)$$

$$\Delta p_r^2-2[(r^2+a^2)E-a\Phi+\varepsilon er]p_r+\mu^2r^2=-\mathcal{K}. \quad (50)$$

These together with (44) and (45) provide a complete set of first integrals of the motion. [It is easy to verify directly, without considering the action, that expressions (49) and (50) are indeed constant, since it is almost immediately apparent that their Poisson brackets with the Hamiltonian (43) vanish.]

Equation (48) can be solved completely by quadratures. It splits up to give two ordinary differential equations:

$$dS_\theta/d\theta=\sqrt{\Theta}, \quad (51)$$

$$dS_r/dr=\Delta^{-1}(P+\sqrt{R}), \quad (52)$$

where the functions $\Theta(\theta)$, $P(r)$, $R(r)$ are defined by

$$\Theta=Q-\cos^2\theta[a^2(\mu^2-E^2)+\Phi^2\sin^{-2}\theta], \quad (53)$$

$$P=E(r^2+a^2)-\Phi a+\varepsilon er, \quad (54)$$

$$R=P^2-\Delta(\mu^2r^2+\mathcal{K}), \quad (55)$$

and where it has been convenient to define a new constant Q related to the others by

$$Q=\mathcal{K}-(\Phi-aE)^2. \quad (56)$$

[29] S. C. Darwin, Proc. Roy. Soc. (London) **A249**, 180 (1958).
[30] S. C. Darwin, Proc. Roy. Soc. (London) **A263**, 39 (1961).
[31] J. C. Graves and D. R. Brill, Phys. Rev. **120**, 1507 (1960).

Thus the final solution for the Jacobi action is

$$S=-\tfrac{1}{2}\mu^2\lambda-Eu+\Phi\varphi + \int^{\theta}(\sqrt{\Theta})d\theta+\int^{r}\Delta^{-1}Pdr+\int^{r}\Delta^{-1}(\sqrt{R})dr, \quad (57)$$

where the signs of the two square roots are independent of each other, and where the lower limits of integration need not be specified, since only changes of the action are important.

The integrated forms of the geodesic and orbit equations can now be obtained automatically by using the fact that the partial derivatives of the Jacobi action with respect to the constants of the motion are themselves constant.

Thus by differentiating with respect to $\mathcal{K}$, μ, E, Φ, we obtain, respectively,

$$\int^{\theta}\frac{d\theta}{\sqrt{\Theta}}=\int^{r}\frac{dr}{\sqrt{R}}, \quad (58)$$

$$\lambda=\int^{\theta}\frac{a^2\cos^2\theta d\theta}{\sqrt{\Theta}}+\int^{r}\frac{r^2dr}{\sqrt{R}}, \quad (59)$$

$$u=\int^{\theta}\frac{-a(aE\sin^2\theta-\Phi)d\theta}{\sqrt{\Theta}} +\int^{r}\frac{r^2+\alpha^2}{\Delta}\left(1-\frac{P}{\sqrt{R}}\right)dr, \quad (60)$$

$$\varphi=\int^{\theta}\frac{-(aE-\Phi\sin^{-2}\theta)d\theta}{\sqrt{\Theta}} +\int^{r}\frac{a}{\Delta}\left(1-\frac{P}{\sqrt{R}}\right)dr, \quad (61)$$

where $\sqrt{\Theta}$ and $\sqrt{R}$ may take either sign independently, but where, once a choice has been made, it must be used consistently in all four equations, and where the lower limits of integration may be chosen quite independently in each term.

For many purposes this information is more conveniently expressed in terms of the first-order differential system:

$$\rho^2\dot{\theta}=\sqrt{\Theta}, \quad (62)$$

$$\rho^2\dot{r}=\sqrt{R}, \quad (63)$$

$$\rho^2\dot{u}=-a(aE\sin^2\theta-\Phi)+(r^2+a^2)\Delta^{-1}[(\sqrt{R})-P], \quad (64)$$

$$\rho^2\dot{\varphi}=-(aE-\Phi\sin^{-2}\theta)+a\Delta^{-1}[(\sqrt{R})-P], \quad (65)$$

which may be obtained either from the explicitly integrated form [(58) to (61)] or else directly from (44), (45), (48), and (50), and where again the signs of $\sqrt{\Theta}$ and $\sqrt{R}$ may be chosen independently, but once chosen must be used consistently.

B. Geodesic Completeness

We are now in a position to demonstrate that the analytic extensions obtained in Sec. 1C are indeed maximal, in the sense that the only geodesics which are incomplete are those which reach the ring singularity, so that they cannot possibly be imbedded as subspaces of any larger manifold.

A geodesic is complete if it can be extended to unbounded values of the affine parameter λ. It is apparent that any geodesic can be extended indefinitely unless it reaches the singularity or unless one of the integrals in the Eqs. (58), (60), or (61) diverges. The latter can occur only where Δ has a zero, or where Θ or R has a double zero.

If Θ or R has a double zero, then the integrals for λ will diverge, and λ itself will be unbounded except in the cases of geodesics which reach the singularity, for which the divergent integrals for λ may be able to cancel each other out. This can be seen more easily from the form

$$d\lambda=\rho^2(d\theta/\sqrt{\Theta}), \quad (66)$$

$$d\lambda=\rho^2(dr/\sqrt{R}), \quad (67)$$

of the Eqs. (62) and (63), than from (58) and (59) directly. Thus although this coupled form is not suitable for explicit evaluation, it shows clearly that no question of incompleteness can arise where Θ or R has a double zero except for geodesics reaching the singularity $\rho^2=0$.

Therefore in considering incompleteness away from singularity we need only consider the cases where Δ has a zero, which can occur only for strictly positive values of r. Possible divergences occur only in the equations for u and φ, which may be written in differential form as

$$du=\frac{-a(aE\sin^2\theta-\Phi)}{\sqrt{\Theta}}d\theta+\frac{r^2+a^2}{\Delta}\left(1-\frac{P}{\sqrt{R}}\right)dr, \quad (68)$$

$$d\varphi=\frac{-(aE-\Phi\sin^{-2}\theta)}{\sqrt{\Theta}}d\theta+\frac{a}{\Delta}\left(1-\frac{P}{\sqrt{R}}\right)dr. \quad (69)$$

These equations can be reexpressed in terms of the $(w,r,\theta,\tilde{\varphi})$ coordinates given by (16) and (17) as

$$dw=\frac{+a(aE\sin^2\theta-\Phi)}{\sqrt{\Theta}}d\theta+\frac{r^2+a^2}{\Delta}\left(1+\frac{P}{\sqrt{R}}\right)dr, \quad (70)$$

$$d\tilde{\varphi}=\frac{+(aE-\Phi\sin^{-2}\theta)}{\sqrt{\Theta}}d\theta+\frac{a}{\Delta}\left(1+\frac{P}{\sqrt{R}}\right)dr. \quad (71)$$

Now provided that P is nonzero where $\Delta=0$, we obtain from (54) and (55) the expansion

$$\frac{P}{\sqrt{R}}=\pm\left[1\pm\frac{\mu^2r^2+\Lambda^2}{2P}\left(\frac{\Delta}{P}\right)+O\left(\frac{\Delta}{P}\right)^2\right], \quad (72)$$

where the sign depends on which choice of $\sqrt{R}$ is under consideration. It can be seen from this that only one pair of the expressions (68) and (69), or (70) and (71), contains a genuine divergence at $\Delta=0$; in the other pair the divergent terms cancel out. Thus, although the geodesic leaves one of the Kerr-Newman coordinate patches, it can be continued on an overlapping patch.

The case in which P vanishes where $\Delta=0$ remains to be considered. In this case it follows from (55) that R must have at least a single zero there. If R has a double zero, there is no problem, because as we have seen the integral for λ will then diverge anyway; this must necessarily be the case if Δ has a double zero, which shows why an analytic extension consisting only of Kerr-Newman patches is sufficient in this case.

Thus we can restrict our attention to the case where P vanishes where $\Delta=0$ and where Δ and R have only a single root there. In this case we have

$$P/\sqrt{R}=O(\Delta^{1/2}) \tag{73}$$

in the limit as the zero of Δ is approached, so the coordinates u and w diverge to $+\infty$ or $-\infty$ together. This simply means that the geodesic reaches one of the points $x=0$, $y=0$ in one of the $(x, y, \theta, \varphi^{\pm})$ patches (26). Since the immediate neighborhoods of these points are well behaved, such a geodesic can straightforwardly be continued on the other side.

By working with the geodesic equations in the Kerr-Newman coordinate system, we have left out of account the possibility that there may be goedesics confined entirely to the 2-surfaces $x=0$, $y=0$. In fact, it is obvious from the symmetry of the form (26) that such geodesics do exist. Nevertheless no question of incompleteness arises because the surfaces $x=0$, $y=0$ are topologically 2-spheres, and, as can be seen at once from (26), they are spacelike. It is well known that a compact spacelike manifold cannot possibly be incomplete.

This completes our demonstration that the analytic extensions of Sec. 1C. are maximal, and that only geodesics which strike the singularity are incomplete. As a by-product we have shown that the charged-particle orbits have the same property. It is widely conjectured that this ought to follow automatically from the completeness of the geodesics, but no rigorous theorem about this question is known to the author, so it is perhaps worth mentioning that this is not a counterexample.

C. Some Qualitative Properties of Geodesics and Orbits

It is possible to see quite easily how the θ coordinate varies during the geodesic and orbital motions, due to the remarkable simplicity of the function Θ by which the variation of θ is governed. It can be seen from (53) that not only is the form of Θ quite independent of the presence of electric charge either on the test particle or in the field, but it is even independent of the mass parameter of the field. In other words, Θ is identical with the function obtained in the limit of field-free flat space.

Useful information about the orbits, and restrictions on the values which can be taken by the constants of the motion, may be obtained by examining the extent of the allowed regions where Θ is non-negative. The results may be summarized as follows:

Case (1), $Q>0$

In this case there are always real solutions in which θ ranges over a region straddling the equator, $\cos\theta=0$. This region extends to the axis of symmetry $\sin\theta=0$, if and only if $\Phi=0$ and $Q+a^2(E^2-\mu^2)\geqslant 0$.

In addition there is a solution in which θ is constant at the axis value, $\sin\theta=0$, when $\Phi=0$ and

$$Q+a^2(E^2-\mu^2)=0.$$

Case (2), $Q=0$

In this case there are always real solutions in which θ is constant at the equatorial value, $\cos\theta=0$.

There are real solutions in which θ varies if and only if the energy is sufficiently high, i.e.,

$$a^2(E^2-\mu^2)>\Phi^2. \tag{74}$$

If this is satisfied, θ varies over a range touching the equator on one side or other. The range extends to the axis of symmetry if and only if $\Phi=0$.

The only other case where there are real solutions is that in which $\Phi=0$ and $a^2(E^2-\mu^2)=0$, when θ may take any constant value whatsoever.

Case (3) $Q<0$

In this case there are no real solutions at all unless (74) is satisfied, and, in addition,

$$Q\geqslant -\{[a^2(E^2-\mu^2)]^{1/2}-|\Phi|\}^2. \tag{75}$$

If (75) is satisfied as a strict inequality, θ varies over a range which does not touch the equatorial plane and which extends to the symmetry axis if, and only if, $\Phi=0$. If equality holds in (75), then θ takes a fixed value which lies strictly between the equatorial plane and the symmetry axis, except when $\Phi=0$ in which case it lies on the symmetry axis.

It is not easy to give such a complete description of the motion of the r coordinate, because the corresponding governing function $R(r)$ is a quartic in the full sense: The odd-power terms do not drop out as they do for $\Theta(\theta)$. Nevertheless, it is possible without much trouble to reach some interesting conclusions.

The function $R(r)$ may be expanded in the form

$$\begin{aligned} R=&(E^2-\mu^2)r^4+2(\mu^2m+\varepsilon eE)r^3\\ &+[a^2E^2-\Phi^2+e^2(\varepsilon^2-\mu^2)-a^2\mu^2-Q]r^2\\ &+2[m(aE-\Phi)^2+\varepsilon ea(aE-\Phi)+mQ]r\\ &-e^2(aE-\Phi)^2-(a^2+e^2)Q. \end{aligned} \tag{76}$$

From the form of the quartic term it can be seen that no orbit or geodesic can escape to the asymptotically flat regions of large positive or negative r if it has less than the escape energy, i.e., if $E^2<\mu^2$, as one would expect.

From the form of the constant term in (76) it can be seen that no geodesic or orbit can possibly cross the hypersurface $r=0$ which extends across the mouth of the singular ring if Q is positive. Nor can it do so if it is confined to the equator since the way would be blocked by the ring singularity itself. Therefore the hypersurface $r=0$ cannot be crossed unless the inequality (74) is satisfied.

Thus we reach the conclusion that no orbit or geodesic can pass through the ring between regions of positive and negative r unless its energy is greater than some minimum which is certainly not less than the escape energy.

This repulsive property of the gravitational field across the mouth of the ring has already been noted by Carter[23] insofar as it applies to the symmetry axis, and the exact height of the energy barrier on the symmetry axis is calculated in this reference. In general, the minimum energy for passing through the ring will depend on the angular momentum, etc., possibly in a complicated way. We shall not investigate the matter further here, but only remark that geodesics and particles with sufficiently high energy can clearly pass through without difficulty.

D. Geodesic Structure of Ring Singularity

The results of Sec. 3C can be used to give very stong restrictions on the geodesics and orbits which may reach the singularity $\rho^2=0$. Thus we have seen that r cannot reach the value zero if Q is positive and that $\cos\theta$ cannot reach the value zero if Q is negative, and hence a necessary condition for an orbit or geodesic to reach the singularity is

$$Q=0. \tag{77}$$

Moreover, from the form of the constant term in (76) it is apparent that when Q is zero it will still be impossible for r to reach the value zero unless either the equality

$$\Phi=aE \tag{78}$$

is satisfied, or alternatively the charge e of the solution vanishes. In the timelike and null cases, i.e., when $\mu^2\geqslant 0$, (78) is incompatible with (74), and therefore if the motion is not to be confined to the equator (78) may not hold, but instead the charge e must vanish. Now under these circumstances the remaining coefficients in (76) are all strictly positive, and therefore the parameter r can only reach zero by approaching from and returning to asymptotically large values on the positive side, and the integral on the right-hand side of (58) remains finite during this process; on the other hand, the integral on the left-hand side of (58) diverges as $\cos\theta$ approaches zero, with the implication that the geodesic or orbit only approaches the equator asymptotically as r tends to infinity.

Thus we reach the conclusion that a timelike or null geodesic or orbit cannot reach the singularity under any circumstances except in the case where it is confined to the equator, $\cos\theta=0$.

The restriction can be carried even further than this. An examination of the equatorial geodesics in the case where the solution is uncharged ($e=0$) has already been made by Boyer and Lindquist,[15] who have shown that there is in general a finite range of angular momentum within which a geodesic of a given sufficiently high energy from a general point on the positive-r side of the equator can reach the singularity. However, when the solution is charged (or if one is considering approach from the negative-r side) the restriction is considerably more severe because (78) must be satisfied. In other words, a geodesic or charged particle orbit with a given energy can only reach the singularity if it has a uniquely determined angular momentum. Even this is not quite sufficient as can be seen from the form to which (76) reduces when (77) and (78) are satisfied, which is

$$R=(E^2-\mu^2)r^4+2(\mu^2m+\varepsilon eE)r^3 +[e^2(\varepsilon^2-\mu^2)-a^2\mu^2]r^2. \tag{79}$$

It is clear that an additional necessary condition for the singularity to be attainable is that the coefficient of the quadratic term be non-negative; in other words, the charge on the test particle must be large enough to satisfy

$$\varepsilon^2\geqslant(1+a^2/e^2)\mu^2. \tag{80}$$

If this holds as a strict inequality, it is a sufficient condition for the singularity to be attainable from sufficiently close points on the equator on either side, whereas if it holds as an equality the singularity will in general be attainable from one side only (although there will be exceptional cases when the energy is such that either the quartic or the cubic term vanishes). In the particular case of timelike geodesics ($\varepsilon=0$, $\mu^2>0$), the inequality (80) cannot be satisfied at all except in the Schwarzschild limit when e and a both vanish. In the case of null geodesics ($\varepsilon=0$, $\mu=0$), strict equality holds in (80) but the singularity can be reached from either side because the cubic term in (79) vanishes.

Thus we conclude that when the solution is charged, no timelike geodesics can reach the singularity, while null geodesics reach the singularity if, and only if, they lie in the equator and have a uniquely determined angular momentum given by (78). Even when the solution is uncharged, the only timelike or null geodesics which can reach the singularity are those confined to the equator, but as Boyer and Lindquist have shown, in this case both null and timelike ones reach the singularity and their angular momentum may lie in a finite range.

The significance of this for an observer studying the singularity visually, i.e., by receiving photons which have come out from the singularity along null geodesics, is as follows: If he observes from a point on the equator then when the field is uncharged the singularity is visible as a finite one-dimensional line, as one would expect for a ring seen edgewise on (except that by the results of Boyer and Lindquist[15] the line may sometimes consist of two disconnected parts); however, if the field is charged then the singularity is visible from the equator only as a point, and in either case if the observer moves off the equator the singularity will become totally invisible to him.

4. IMPLICATIONS

The fact that there are closed timelike lines looping through the interior does not affect the reasonableness of interpreting the Kerr solutions as the exterior fields of rotating bodies, since a source body might be expected to block off these regions in any case. However, it does hint (although it certainly does not prove) that causality breakdown may be expected to result from the collapse of a rotating body. The theorems of Penrose[1] and Hawking[2–4] indicate that something pathological must be expected to occur in a situation of gravitational collapse of a rotating body, but different opinions may be held about the nature of the breakdown. From a physical point of view the least serious kind of breakdown would be the local development of density or curvature singularities, and this is also the kind of breakdown which has been considered most widely in the past. The reason why this would not be very serious is that one would expect in any case that general relativity would need to be modified in conditions of extreme curvature, in order to accommodate quantum theory, and one might be sanguine enough to hope that the necessary modifications would cure the trouble. However, it is also conceivable, as has been suggested by Lifshitz and Khalatnikov,[32] that the curvature singularities which are familiar in highly symmetric solutions do not exist in more general cases.

The Kerr solutions have a lower symmetry group than any other solutions in which (as far as the author knows) an analytic study of a curvature singularity has been made (although the separability which has made this possible is itself a symmetry of a kind), and the results of the previous section seem to lend a certain amount of support to this last idea. Thus as the symmetry is progressively reduced, starting from the Schwarzschild solution, the extent of the class of geodesics reaching the singularity is steadily reduced likewise, until in the case with both charge and rotation there are almost none at all, which suggests that after further reduction of the symmetry, incomplete geodesics might cease to exist altogether. Even if a few incomplete geodesics remain in the fully general case, their importance is overshadowed by the causal pathology, which seems to increase as the symmetry is reduced. [However, in spite of this apparent effect the existence or lack of symmetry may not be as important as it appears, for the Taub-N.U.T. space which has been discussed by Misner[33] is much more highly symmetric than the Kerr solution, and yet its global behavior is in some ways worse: It has geodesics which are incomplete in a region which is locally nonsingular, and it also has closed timelike lines confined within any neighborhood, no matter how thin, of the causality horizon of the well-behaved part, whereas in the Kerr solutions, in the cases when there are causality horizons (at $r=r_-$), any closed timelike line must at some stage penetrate deeply into the bad part beyond these horizons.]

All these things suggest that the breakdown in general relativity may be of a global rather than (or as well as) of a local nature, in which case it is very serious indeed. If this turns out to be the case then one will not be able to expect to cure the trouble by minor modifications significant only in regions of high curvature, so that the whole theory might have to be abandoned, or at least drastically reformulated.

ACKNOWLEDGMENTS

The author would like to thank Dr. S. W. Hawking, Dr. R. W. Lindquist, Professor C. W. Misner, Dr. R. Penrose, and Dr. D. W. Sciama for many useful discussions and ideas. He is also grateful for the encouragement and suggestions of the late Dr. R. H. Boyer.

[32] E. M. Lifshitz and I. M. Khalatnikov, Zh. Eksperim i Teor. Fiz. **39**, 149 (1960) [English transl.: Soviet Phys.—JETP **12**, 108 (1961)].

[33] C. W. Misner, in *Relativity Theory and Astrophysics: Relativity and Cosmology*, edited by J. Ehlers (American Mathematical Society, Providence, 1967), Vol. 8, p. 160.