REMOVABLE SINGULARITIES AND THE CAUSALITY PRINCIPLE IN GENERAL RELATIVITY THEORY

É. B. GLINER

A. F. Ioffe Physico-technical Institute, USSR Academy of Sciences

Submitted January 3, 1967; resubmitted July 10, 1967

Zh. Eksp. Teor. Fiz. 54, 796–807 (March, 1968)

The conditions are established under which the space-time structure is compatible with the possibility of simultaneously satisfying the causality and equivalence principles ("causal completeness"). Metric forms with "removable" singularities analogous to the singularities on the gravitational radius in Schwarzschild's metric form are investigated from the point of view of satisfying these conditions. It turns out that in order to satisfy these conditions it is necessary to extend the Einstein-equation solutions represented by such forms, but the required extension is much narrower than that attained by Kruskal's method. The connection between both types of extensions is established.

1. In the general case, the Riemann space cannot be covered by a single system of coordinates, and can be represented only as a union of a finite or denumerable number of regions, each of which admits of such a covering. Therefore a solution obtained for Einstein's equations in a certain coordinate system does not of necessity define space time (S-T) as a whole. This raises two problems: a) establishment of a criterion of completeness of S-T on the basis of physical considerations, and b) extension of the solutions of Einstein's equations beyond the limits of the region of their initial definition, when this criterion is violated.

Usually[1-4] S-T is regarded as complete in the following sense: the geodesics can be continued to infinite values of the affine parameter or to a singular point - a point where the geometric invariants are singular. (We recall that the affine parameter for space-like geodesics has the meaning of length, and for time-like (TL) geodesics it has the meaning of the proper time of body moving on the geodesic, while for null geodesics it has no direct physical interpretation.) There are no known theoretical or observational justifications for the necessity of this type of completeness - affine completeness. The physical meaning of the requirement of affine completeness is not sufficiently distinct, owing to the fact that it implies equivalence of all types of geodesics, to the absence of a unified physical interpretation of the affine parameter for all the geodesics, and also to the uncertainty of the degree to which it guarantees the possibility of satisfying the causality principle.

In order for the criterion of completeness of S-T to have a clear-cut meaning, it must be defined in terms of concepts pertaining to local macroscopic processes ("observer"). We shall say that S-T is causally complete if its structure admits the possibility of satisfying the following physical principles.

I. ("Chronometric principle"). A complete experiment determining all the space-time relations between bodies includes only local measurements on moving bodies; in particular, the only direct measurement of intervals in S-T is measurement of the proper time of the bodies (see, e.g.,[5]).

II. ("Strong" equivalence principle). Everywhere in S-T where the geometric invariants are non-singular, a freely falling non-rotating body is acted upon, accurate to small deviations due to the finite dimensions of the body, by a system of physical laws independent of the curvature of the S-T along the world line of the body (see, e.g.,[6]). In particular, the body either has a history that extends in units of proper time into the past and the future without limit, or its world line reaches a singular point of S-T in a direction in which this does not take place.

III. (Causality principle). The irreversibility of the causal sequence of physical events is a general condition for physical realizability of bodies. In particular, in the S-T region encompassed by a macroscopic process (and therefore by a causal and irreversible process), the only physically realizable bodies (they can arise or can be produced from other bodies, they exist either "primordially" or are introduced from the outside) are those and only those on which the irreversible processes have the same direction as the aforementioned process.

In order to formulate a criterion that determines when the structure of the S-T agrees with principles I–III, it is convenient to regard the S-T as an open manifold with respect to time. This has physical justification in the fact that any concept of completeness of S-T must imply inevitably that the bodies cannot cross the boundary of the S-T. Therefore the points of the boundary of the S-T either cannot be reached by the matter, and it is then meaningless to regard them as an element of physical reality, or else are singular, and then there are no space-time relations in them in the usual sense. Thus, let us assume that the temporal boundary of the S-T, namely an aggregate of the initial and final points of TL geodesics, does not belong, by definition, to the S-T. Inasmuch as the geodesics are not continuable beyond the singular points of S-T, the latter belong to the boundary of S-T, so that the S-T does not contain its own singular points. By virtue of the latter circumstance, the S-T is locally Lorentzian at all its points, i.e., at each S-T point there exists a local light cone and in each direction inside the cone there passes strictly one TL geodesic that is continuable in both directions from the apex of the cone.

Under the indicated conditions, in order that the structure of S-T agree with I—III, it is necessary and sufficient to satisfy the following criterion of causal completeness.

A. At all points of the S-T there can be defined on the local light cones cavities of absolute future and absolute past and, by the same token, directions of TL geodesics from the past into the future, such that

a) the direction of the TL geodesic from the past into the future does not change over its entire extent;

b) each TL geodesic can be continued into the future (defined in accordance with condition a) either to the infinite values of the affine parameter, or to a singular point; if the TL geodesic is continued in the past only to a certain value s_0 of the affine parameter s, then no other TL or null geodesic reaching this point can be continued beyond the point where $s = s_0$ (this point is not necessarily singular), i.e., the point belongs to the boundary of S-T.

B. There exist TL geodesics that can be continued in the past either to infinite values of the affine parameter, or to a singular point of S-T.

In accordance with principle I, the properties of S-T are formulated here for TL geodesics. In view of the local Lorentzian behavior, the equivalence principle is valid for at least a sufficiently short interval of the proper time of the body. Then the conditions A, a) and A, b) are necessary and sufficient to permit satisfaction of the causality and equivalence principles in their complete formulation. All that needs explaining is the condition A, b). Let a TL geodesic γ be continuable only to a finite value s_0 of the affine parameter s, the point where $s = s_0$ not being singular. If it is located in the future with respect to other points on γ, then the principle of equivalence can be violated by producing (say, by collision between the bodies) a body having a geodesic γ as its world line. This geodesic would end at $s = s_0$, i.e., in contradiction to II, the body would cease to exist at a nonsingular S-T point. On the other hand, if the point where $s = s_0$ lies in the past, then, even when observing the body on γ, one cannot state that it was produced at $s = s_0$, since it can fall on γ as a result of a non-geodesic process (e.g., collision). Further, it is impossible to produce a body that would reach the point where $s = s_0$, by virtue of the causality principle: according to A, b), this body would have to be sent in the past. Finally, if the criterion B is not satisfied, then in S-T there are neither "sources" of matter - singular points of S-T lying in the past, nor geodesics on which the matter could exist "primordially." Therefore, without violating II, it is impossible to propose the existence of even trial bodies, thus making I—III physically meaningless.

Let us make a few remarks concerning TL geodesics that cannot be continued infinitely in the past or to a singular point. Their physical interpretation is possible only in connection with the causality principle. By the same token, supplementation of I—II by the causality principle III expands the class of Riemannian spaces that admit a physical interpretation, and even this in itself justifies the study of the consequences that follow from adding the causality principle to the general theory of relativity. Inasmuch as the initial points of these geodesics are not singular, matter can fall on them only as a result of nongeodesic processes. If there are no such processes, the geodesics are "empty." However, the general theory of relativity does not imply that all the TL geodesics are world lines of real bodies. Finally, in S-T with geodesics of this type, reversal of the direction of time would violate the equivalence principle. This imparts important irreversibility features to S-T. On the other hand, when there are no such geodesics, conditions A and B are not violated when the past is replaced by the future. In this case we shall say that S-T is causally reversible.

2. Let us apply the foregoing considerations to solutions of Einstein's equations, in connection with which the problem of the S-T has been discussed many times (e.g.,[1,3,4,9-13]). These are the well known vacuum solutions of Schwarzschild, de Sitter, Kottler, Reissner-Nordstrom, and also the recently published solutions of Newman, Tamburino, and Unti[7] and Kerr[8]. The S-T described by each of them is characterized by the existence of a family of geodesics with the following properties:

a) Consideration of the family is sufficient in the study of the completeness of S-T (see, e.g.,[3]).

b) Only two coordinates, which we shall denote by ξ and ζ, vary along the geodesics of the family; these coordinates can be chosen such that the metric form becomes (we omit terms that do not depend on $d\xi$ and $d\zeta$)

$$-ds^2 = A(\zeta)\,d\zeta^2 - B(\zeta)\,d\xi^2, \tag{1}$$

where the coefficients A and B have the following properties: 1) they depend only on ζ, 2) they reverse sign on the surface Γ defined by the equation $\zeta = 1$ and their order of magnitude at $\zeta \sim 1$ is

$$A(\zeta) \sim (\zeta - 1)^{-1}, \qquad B(\zeta) \sim \zeta - 1 \tag{2}$$

and 3) in the regions $\zeta_0 < \zeta < 1$ and $1 < \zeta < \zeta_k$ ($\zeta_0 < 1$ and $\zeta_k > 1$ are constants) they retain their sign and are monotonic differentiable functions of ζ. For example, the Schwarzschild metric form can be represented in the form

$$-ds^2 = r(r-1)^{-1}dr^2 - r^{-1}(r-1)\,dt^2, \quad r = \zeta,\ t = \xi.$$

We use below only the foregoing properties of the cited solutions, and the results pertain equally well to any solution of Einstein's equations that possesses such properties.

Integrating the equations of the geodesics for (1), we obtain $B\,d\xi = \nu^{-1}ds$, where ν is an arbitrary constant. Hence, by virtue of (1), we obtain along the geodesics

$$\xi = \text{const}, \quad \frac{d\zeta}{ds} = \pm\sqrt{-A} \text{ when } \nu = \infty, \tag{3}$$

$$\frac{d\zeta}{d\xi} = \varepsilon\,\text{sign}\,B\sqrt{\frac{B}{A}}\sqrt{1-\nu^2 B}, \quad \frac{d\zeta}{ds} = \varepsilon\frac{\sqrt{1-\nu^2 B}}{\nu\sqrt{AB}}, \quad \text{when } 0 < |\nu| < \infty, \tag{4}$$

where $|\epsilon| = 1$ and sign B is the sign of the coefficient B. The roots here and below are understood in the arithmetic sense. Integrating (4) under condition (2), we verify that, independently of the choice of the time direction, it is always possible to realize a body having the geodesic (4) as its world line and reaching the point where $\zeta = 1$ after a finite interval of proper time. But in this case $|\xi| = \infty$, i.e., the body is outside the S-T region covered by the system of coordinates ξ, and ζ, thus indicating the need for extending the solution (1).

In this typical example, the incompleteness of the solution is connected with the course of the geodesics near the surface on which the metric form (1) has singularities of the type (2). We shall consider extensions that eliminate the causal incompleteness or the irreversibility caused by precisely such singularities of the solutions of type (1). Many of the aforementioned solutions have singularities of type (2) on several non-intersecting surfaces. The extension procedure is analogous for each of them. It is therefore sufficient to consider the vicinity of only one of them.

Let us define our terminology and notation. The local map [14] of an open region of S-T is defined as a system of local coordinates covering this region. The latter is referred to as the region of the local map. The S-T as a whole is defined by a set of local maps, the regions of which in their aggregate cover all of S-T, and by metric forms specified on each of the local maps; these forms are solutions of the Einstein equations in the coordinates of the map. In a particular case, the set can consist of one map. Since the regions of the maps are open, it follows that, first, the S-T as a whole is open (in agreement with the statement made above) and, second, the S-T is connected only when the region of each map overlaps the region of at least one other map. With this, in order that each of the maps define in the overlap region the same topology, and in particular, in order that the convectiveness of the curves be conserved, the transformation from the coordinates of one map to the coordinates of another should be at least topological, i.e., mutually unique and mutually continuous. Third, infinite values of the map coordinates do not belong to the map. The direction of time in S-T should be specified in a consistent manner along each TL geodesic (the reference direction of the TL coordinate has no direct physical meaning). Implying time direction, we shall speak for brevity simply of the direction of TL geodesics.

The vicinity $\Omega(-\infty < \xi < \infty, \zeta_0 < \zeta < \zeta_k)$ of the surface Γ is assumed to be the region of the local map (ξ, ζ). Let R be the part of Ω where $1 < \zeta < \zeta_k$, and T the part of Ω where $\zeta_0 < \zeta < 1$ (this agrees with the notation in [12]). Then $\Omega = R + T + \Gamma$. In view of (2), $B > 0$ in R and $B < 0$ in T. We choose ζ_0 and ζ_k such that $B < 1$ in Ω.

The natural way of extending the solution (1) is to assume geodesics (4) with some fixed ν as the coordinate lines of a space-like coordinate, and the value of the parameter s along them as the TL coordinate. The region into which the geodesics go will then certainly belong, at least in part, to finite values of the new coordinates.

Let τ and ρ be the new coordinates; $d\tau = \alpha_1 d\xi + \alpha_2 d\zeta$, $d\rho = \alpha_3 d\xi + \alpha_4 d\zeta$, and the geodesics (4) at $\nu = 1$ are the coordinate lines ρ. By definition, $d\tau = ds$ and $d\rho = 0$ along the latter, thus imposing two limitations on the four quantities α_j. Setting $\alpha_1 = \alpha_3 = 1$ and finding with the aid of (4) that at $\epsilon = \pm 1$

$$\alpha_2 = \pm \operatorname{sign} B \sqrt{\frac{A}{B}} \sqrt{1-B}, \quad \alpha_4 = \pm \operatorname{sign} B \sqrt{\frac{A}{B}} \frac{1}{\sqrt{1-B}},$$

we represent the connection between τ, ρ and ξ, ζ in the region of overlap in integral form:

$$\tau = \xi \pm \int_{\zeta_0}^{\zeta} \operatorname{sign} B \sqrt{\frac{A}{B}} \sqrt{1-B}\, d\zeta, \quad \rho = \xi \pm \int_{\zeta_0}^{\zeta} \operatorname{sign} B \sqrt{\frac{A}{B}} \frac{d\zeta}{\sqrt{1-B}}. \quad (5)$$

Depending on the sign in front of the integrals, we get two systems of coordinates τ^+, ρ^+ (upper sign) and τ^-, ρ^-, analogous to the freely falling coordinate systems of Lemaitre for the Schwarzschild metric.

In $\Omega - \Gamma$, the correspondence between the coordinate systems (τ^+, ρ^+), (τ^-, ρ^-), and (ξ, ζ) is topological. In fact, a pair of finite values ξ, ζ of (5) in $\Omega - \Gamma$, where $\zeta \neq 1$, is in unique correspondence with a pair of finite $\tau^\pm$ and $\rho^\pm$. To the contrary, in view of (5) we have

$$\rho^\pm - \tau^\pm = \pm \int_{\zeta_0}^{\zeta} \frac{\sqrt{AB}}{\sqrt{1-B}} d\zeta. \quad (6)$$

The integral on the right side converges as a result of (2), and since the integrand does not reverse sign, the dependence of $\rho^\pm - \tau^\pm$ on ζ is monotonic. Therefore each pair of values of $\rho^\pm$ and $\tau^\pm$ from $\Omega - \Gamma$ corresponds strictly to one value $\zeta \neq 1$, and then any of the equations in (5) defines uniquely a finite ξ. Thus, the connection between (ξ, ζ) and any of the systems (τ^+, ρ^+) and (τ^-, ρ^-) in $\Omega - \Gamma$ is mutually unique, as is consequently also the connection between (τ^+, ρ^+) and (τ^-, ρ^-). The mutual continuity and moreover the differentiability of the transformations between the systems in question follow from the fact that all the partial derivatives of $\tau^\pm$ and $\rho^\pm$ with respect to ξ and ζ in $\Omega - \Gamma$ exist, and the Jacobians $J^\pm = \pm\sqrt{AB}/\sqrt{1-B}$ of the transformations (5) are of definite sign in Ω.

Since the integrals in the right side of (5) tend, by virtue of (2), to $-\infty$ as $\zeta \to 1$, the points on Γ belonging to the local maps $(\tau^\pm, \rho^\pm)$ correspond to $\xi = \pm\infty$. By the same token, each of the maps $(\tau^\pm, \rho^\pm)$ covers one of the "coordinate infinities" $\xi = \pm\infty$, $\zeta = 1$.

Substitution of (5) in (1) yields

$$-ds^2 = (1-B)(d\rho^\pm)^2 - (d\tau^\pm)^2. \quad (7)$$

Unlike (1), the forms (7) have no singularities on Γ. In this sense, the singularities of the form (1) on Γ are removable by the coordinate transformation (5). We note, incidentally, that the points on Γ where $|\xi| < \infty$, belonging to the map (ξ, ζ), do not belong to the maps $(\tau^\pm, \rho^\pm)$, inasmuch as $|\tau^\pm|$ and $|\rho^\pm|$ are infinite in them. Therefore the form (7) extends the metric given by (1) beyond the limits of the local map (ξ, ζ), but it does not remove essentially the singularities in the region of the map (ξ, ζ). In $\Omega - \Gamma$, the S-T metric defined by the form (7) coincides with the metric defined by (1), since in $\Omega - \Gamma$ the correspondence between the three coordinate systems under consideration is topological.

We shall show that the causal incompleteness of S-T, due to the course of the geodesics near Γ, can be eliminated if the solution (1) is extended by adding to the local map (ξ, ζ), with the metric form (1) specified on it, local maps $(\tau^{\pm}, \rho^{\pm})$ with metric forms (7) specified on them. To this end we trace the course of the TL geodesics in the S-T region covered by the indicated local maps (we retain the symbol Ω for this region), and specify their direction in suitable manner.

The geodesics (3) ξ = const near Γ belong completely to the map (ξ, ζ). They are time-like only in T, meaning that they are continuable as TL lines only up to Γ, but not into R. By virtue of (3) and (2) their length $|\int_1^{\infty}\sqrt{-A}\,ds|$ from the point with $\zeta = 1$ to any point with $\zeta < 1$ is finite. Therefore, only a time direction corresponding to decreasing ζ is compatible with causal completeness in T.

As already shown, the geodesics (4) do not reach Γ in the region of the map (ξ, ζ). Consequently, in the region of the map (ξ, ζ) the surface Γ consists of initial points of the TL geodesics emerging from it, and, by definition, does not belong to S-T. By the same taken, the local map (ξ, ζ) breaks up into two: with $\zeta > 1$ and with $\zeta < 1$. The physical meaning of this statement is that exchange of bodies or signals directly between the regions of S-T of the two maps is impossible.

Substituting (5) in (4) we obtain the equations of the geodesics (4) in terms of the coordinates $\tau^{\pm}$ and $\rho^{\pm}$:

$$\frac{d\rho^{\pm}}{d\tau^{\pm}} = \frac{1 \pm \varepsilon\sqrt{1-\nu^2 B}/\sqrt{1-B}}{1 \pm \varepsilon\sqrt{1-B}\sqrt{1-\nu^2 B}} \tag{8}$$

As above, the upper signs pertain to the coordinates τ^{+} and ρ^{+}, and the lower ones to τ^{-} and ρ^{-}. Near Γ, i.e., as $B \to 0$, we have

$$\frac{d\rho^{\pm}}{d\tau^{\pm}} \sim \begin{cases} 1 + B/2 & \text{if} \quad \pm\varepsilon = 1, \\ \dfrac{\nu^2-1}{\nu^2+1} + \dfrac{B}{2}\dfrac{3\nu^4 - 2\nu^2 - 1}{(\nu^2+1)^2} & \text{if} \quad \pm\varepsilon = -1. \end{cases} \tag{9}$$

It follows therefore that near Γ, the geodesics (4) break up on each of the maps $(\tau^{\pm}, \rho^{\pm})$ into two families: family γ_a of geodesics that approach Γ with a slope $\sim 1 + B/2$, and family γ_p of geodesics that approach Γ with a slope $\sim(\nu^2 - 1)/(\nu^2 + 1) \neq 1$. On the map (τ^{+}, ρ^{+}) (upper signs in (9)) the family γ_a consists of geodesics with $\epsilon = 1$, and γ_p of geodesics with $\epsilon = -1$, while on the map (τ^{-}, ρ^{-}) the relations are reversed: $\epsilon = 1$ for the family γ_p and $\epsilon = -1$ for the family γ_a. In other words, the geodesics with $\epsilon = 1$ ($\epsilon = -1$) on the map (τ^{+}, ρ^{+}) belong to the family $\gamma_a(\gamma_p)$, and on the map (τ^{-}, ρ^{-}) to the family γ_p (γ_a).

Since by virtue of (6) the lines ζ = const on the maps $(\tau^{\pm}, \rho^{\pm})$ are straight with unity slope, and since $B > 0$ in R and $B < 0$ in T, it follows from (9) that the geodesic families γ_a in R and T approach Γ asymptotically, i.e., they do not reach Γ within the limits of the corresponding map, going off to coordinate infinity whereas the geodesic families γ_p penetrate Γ, and consequently their sections close to Γ can be traced within the limits of the map. Inasmuch as the geodesic families γ_a on the other map belong to the family γ_p, the section of each of the TL geodesics (4) close to Γ can be traced on one of the maps: the geodesics (4) are continuable everywhere in Ω.

We now specified directions that are compatible with the criterion of causal completeness for the TL geodesics in Ω. We are essentially referring to compatibility of the directions of the time on the maps defining Ω, since the region of each of them, by virtue of the definition of the system of coordinates, is topologically isomorphic to part of a Euclidean plane, and the metric forms (1) and (7) specified on them do not have any singularities within the limits of the maps. By the same token, on each of the maps, accurate to curved geodesics, the picture of the geodesics is the same as in the open region of the Minkowski S-T, and consequently, the directions of the TL geodesics on each of the maps can be made compatible by specifying one of the two conceivable time directions.

Let us establish first several relations. By virtue of (7), the value of the local speed of light in coordinates τ^{+}, ρ^{+} and τ^{-}, ρ^{-} is the same and is equal to $(1 - B)^{-1/2}$. We denote by $\beta^{\pm}$ the ratio of the derivative $d\rho^{\pm}/d\tau^{\pm}$ along the world line of the body to the local speed of light. From (5) it follows that

$$\beta^{\pm} = \frac{1-B/2}{\sqrt{1-B}} - \frac{B/2}{\sqrt{1-B}}\frac{d\tau^{\mp}}{d\tau^{\pm}} = \sqrt{1-B} \pm \sqrt{AB}\frac{d\zeta}{d\tau^{\pm}} = \frac{1}{\sqrt{1-B}} - \frac{B}{\sqrt{1-B}}\frac{d\xi}{d\tau^{\pm}}. \tag{10}$$

Along the world line of a physically realizable body $\beta^{\pm} < 1$. Inasmuch as $(1 - B/2)/\sqrt{1 - B} > 1$ in R and in T, and $\sqrt{1 - B} > 1$ in T, we find from $|\beta^{\pm}| < 1$ that in R we have

$$\frac{d\tau^{+}}{d\tau^{-}} > 0, \quad \frac{d\xi}{d\tau^{\pm}} > 0, \tag{11a}$$

and in T

$$\frac{d\tau^{+}}{d\tau^{-}} < 0, \quad \frac{d\zeta}{d\tau^{+}} < 0, \quad \frac{d\zeta}{d\tau^{-}} > 0. \tag{11b}$$

As was already established, only one time direction in T is compatible with causal completeness, namely a direction corresponding to decreasing ζ. (We note that this result can be obtained also by considering the motion of bodies on the geodesics (4). When the time in T is in the direction of growing ζ, the causality principle is violated.) Therefore, by virtue of (11), the direction of the future in T corresponds to a growth of τ^{+} and a decrease of τ^{-}, and in R to either a growth or to a decrease of all the TL coordinates τ^{+}, τ^{-}, and ξ. There can be no additional limitations on the time direction in R in view of the invariance of the metric form (1), specified on the initial map (ξ, ζ), relative to the change of the sign of ξ.

Let us see how to reconcile, under the foregoing limitations, the time directions in R and T. It is obviously sufficient to regard each geodesic on that local map on which its section near Γ is represented completely. In other words, on each of the maps $(\tau^{\pm}, \rho^{\pm})$ it is necessary to consider the geodesic families γ_p $(\pm\epsilon = -1)$. By virtue of (10) and (8) we have

$$\sqrt{AB}\frac{d\zeta}{d\tau^{\pm}} = \frac{\varepsilon B\sqrt{1-\nu^2 B}}{1 \pm \varepsilon\sqrt{1-B}\sqrt{1-\nu^2 B}}, \tag{12}$$

whence we get $d\zeta/d\tau^{+} < 0$ and $d\zeta/d\tau^{-} > 0$ for the geodesic families γ_p.

Assume that in R the direction of time corresponds to the growth of τ^+, τ^-, and ξ. Then, by virtue of $d\zeta/d\tau^+ < 0$, on the local map (τ^+, ρ^+), the geodesic families γ_p cross Γ in the direction from R to T, i.e., the motion of physically realizable bodies through Γ is possible and proceeds from R into T. The situation is different in the region of the map (τ^-, ρ^-). By assumption, the time direction in R corresponds to growing τ^-, whereas in T, by virtue of $d\zeta/d\tau^- < 0$, it corresponds to decreasing τ^-. Therefore on the (τ^-, ρ^-) map the geodesic families γ_p are directed away from Γ both in R and in T. By the same token, the surface Γ on the map (τ^-, ρ^-), as on the map (ξ, ζ), consists of initial points of the TL geodesics outgoing from it, meaning that it does not belong to S-T. The map (τ^-, ρ^-) breaks up essentially into two maps, one of which pertains to R and the other to T.

When the time direction in R corresponds to decreasing τ^+, τ^-, and ξ, the result is similar, but the geodesics pass from R into T within the limits of the map (τ^-, ρ^-), and on the map (τ^+, ρ^+) the surface Γ does not belong to S-T.

Thus, there exist two combinations of time directions on the considered local maps, compatible with the causal completeness of S-T. This completes the proof of the statement that causal incompleteness of S-T is eliminated by the considered extension of the solution (1). The extended solution is defined by five local maps, of which only one, (τ^+, ρ^+) or (τ^-, ρ^-), connects the regions R and T. Both conceivable time directions are compatible with causal completeness in R, but only one in T. In this sense, R is causally reversible and T is causally irreversible. The latter circumstance leads to the physical nonrealizability in T of bodies that go over from T into R. The points of the surface Γ belong to S-T only in the region of the map connecting R with T. In the region of other maps it represents, as it were, a cut on the map (not in S-T!), ensuring causal completeness. (We note the analogy with the theory of complex variable, where a cut on the complex-variable plane ensures uniqueness of functions.)

It is interesting that the extension in question restores in a certain sense the affine completeness of S-T. Near Γ, namely from T into R, only the TL geodesics ξ = const are not continuable. However, if we disregard the difference in the physical interpretation of TL and space-like geodesics, then the geodesics ξ = const in R can be regarded as continuations of the TL geodesics ξ = const from T. Then all the geodesics are continuable in the union of the regions of all the maps (ξ, ζ), $(\tau^\pm, \rho^\pm)$.

3. The S-T of the extended solution (1) contains a causally irreversible part and in this sense is causally irreversible on the whole. Let us consider briefly the connection between the obtained extension and the extensions that lead to a causally reversible S-T (for example[3,10,13]).

We use a transformation analogous to that used in[15,3]. We introduce new variables χ and η with the aid of the relations

$$\operatorname{sh}\frac{1}{2}\left[\int_{\zeta_0}^{\zeta}\operatorname{sign}B\sqrt{\frac{A}{B}}\,d\zeta+\xi\right]=\varepsilon_1\operatorname{ctg}(\chi+\eta),$$
$$\operatorname{sh}\frac{1}{2}\left[\int_{\zeta_0}^{\zeta}\operatorname{sign}B\sqrt{\frac{A}{B}}\,d\zeta-\xi\right]=\varepsilon_2\operatorname{ctg}(\chi-\eta), \qquad (13)$$

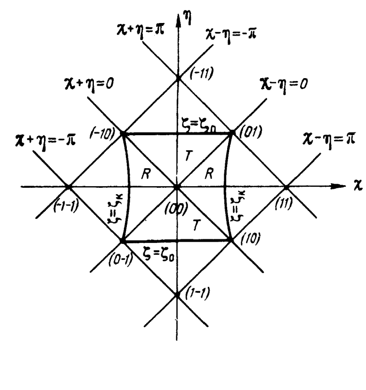
where $|\epsilon_1| = |\epsilon_2| = 1$. The form (1) is transformed in this case into

$$-ds^2=Q(d\chi^2-d\eta^2),\quad Q=\frac{4\varepsilon_1\varepsilon_2B}{|\sin(\chi+\eta)\sin(\chi-\eta)|}. \qquad (14)$$

The transformation (13) is mutually unique and mutually continuous in Ω in the intervals $\zeta_0 < \zeta < 1$ and $1 < \zeta < \zeta_k$. In view of the periodicity of the cotangent, it makes it possible to map topologically either R or T in the interior of any square separated on the (χ, η) plane by the lines $\chi - \eta = \nu_1\pi$, $\chi - \eta = (\nu_1 - 1)\pi$, $\chi + \eta = \nu_2\pi$, and $\chi + \eta = (\nu_2 + 1)\pi$, where ν_1 and ν_2 are integers. We denote by $K(\nu_1\nu_2)$ the interior of the square with given ν_1 and ν_2, by (kl) the vertex of the square lying on the intersection of the lines $\chi - \eta = k\pi$ and $\chi + \eta = l\pi$, and by (kl, mn) the side of the square joining the vertices (kl) and (mn) (without the vertices). If η is the ordinate axis, then $(\nu_1\nu_2)$ is the lower vertex of the square $K(\nu_1\nu_2)$.

We map T in K(00) at $\epsilon_1 = -1$ and $\epsilon_2 = 1$, and R in K(10) at $\epsilon_1 = \epsilon_2 = -1$ (see the figure). The region T is mapped on the part of K(00), which we shall call the map T, bounded by the diagonal $\zeta = \zeta_0$ (see the figure), while R is mapped on the part of K(10) - the map R - bounded by the lines $\zeta = \zeta_k$.

At the indicated values of ϵ_1 and ϵ_2 the maps T and R are "joined together" in the sense that in the open region V, consisting of the maps T and R and the boundary (00, 01) between them, the form (14) has no singularities, and in particular, on approaching the boundary (00, 01) from K(00) and from K(10), the coefficient Q tends to the same limit $(\cos\chi\cos\eta)^{-1}$, and the geodesics continue beyond (00, 01) without a change of slope. It is easy to see that the segment (00, 01) is the part of Γ with $\xi = \infty$, belonging to the map (τ^+, ρ^+). The vertex (00) is the part Γ with $|\xi| < \infty$ of the intial map (ξ, ζ). The transformation (14) compresses it in the point (00), but this leads to no physical consequences, since that part of Γ, as shown, does not belong to S-T (it is on the "cut"). The segments (00, −10) and (10, 00) are the part of Γ with $\xi = -\infty$, corresponding to the "cut" on the map (τ^-, ρ^-). The region V, regarded as a local map (η, χ) with a metric form (14) specified on it, encompasses in this manner the region of all maps (ξ, ζ), (τ^+, ρ^+) and (τ^-, p^-). It can be shown that, by considering the map (η, χ), we obtain also the physical consequences obtained above for the case when the "cut" is made on the map (τ^-, ρ^-) (a cut on the map (τ^+, ρ^+) corresponds to the mapping of R in K(0 − 1)). In particular, the region of the map (η, χ) can be part of a causally complete S-T.



The methods of constructing the maps $(\tau^{\pm}, \rho^{\pm})$ and the map (η, χ) are, however, greatly different. The former is natural in the sense that it is based on introducing a system of coordinates made up of bodies moving from the region of the initial local map. The resultant extension is the minimum required to be able to follow without limit the history of any body, and the local maps obtained are automatically overlapping. In the second method, different transformations (different ϵ_1 and ϵ_2) are used in R and in T, giving non-overlapping maps, although admitting of a "joining." It is not known whether both methods always lead to the same result. This has made it necessary to consider first the "natural" method, by obtaining the minimum extension. Henceforth, for the sake of brevity, we shall use "joining," corresponding to Kruskal's procedure.

In order to obtain a causally reversible extension, in addition to the mappings already performed, we map T in K(1 – 1) and R in K(0 – 1). These mappings can be "joined" with each other and with the previously obtained mappings by connecting the segments (10, 00), (0 – 1, 00), (–10, 00) and the points (00). The form (14) has no singularities in the obtained union W of the mappings, and W can be regarded as a local map of S-T. By virtue of the method used to construct W, the picture of the geodesics in W is symmetrical with respect to the plane $\chi = 0$. Consequently, all the relations on W do not change when the direction in which η is reckoned is changed, and by the same token the S-T region of the map W is causally reversible. Thus, causal reversibility is attained by symmetrical doubling of the causally irreversible map V.

We note that with the aid of joining it is possible to effect also more complicated constructions. Let us take, for example, two maps W superimposed one on the other, and instead of joining on each map along (–10, 00), we effect the joining crosswise: K(0 – 1) of the lower map with K(00) of the upper one, and vice-versa. We obtain a two-sheeted map (which can be mapped if desired on a plane), a region of which may belong to a causally reversible S-T. We see that formally nothing prevents us from obtaining multiply-sheeted maps with different joinings. Inasmuch as solutions of the type (1) with several surfaces similar to the surface Γ can be extended into solutions with an unlimited number of such surfaces (e.g.,[2,3]), it is clear that the joinings can lead to a very complicated structure.

In this connection it becomes doubtful whether the extension of the Schwarzschild S-T constructed by Kruskal[10] is maximal. Kruskal proves this stating that in the extension constructed by him there are no geodesics that can be continued beyond its limit. But it is also necessary to show that the permissible joinings do not lead to a larger extension with the same properties. (The problem of permissible joinings has so far not been considered at all in topologically complicated variants.) It is possible that a more correct view is that Kruskal's extension and an analogous extension for (1), represented by the map W, are the minimal causally complete and reversible extensions, whereas extension of solution (1), represented by the map V or the maps $(\tau^{\pm}, \rho^{\pm})$ and (ξ, ζ) is the minimal causally complete irreversible extension. It is possible that it is precisely the minimality which will make these extensions physically preferred.

4. An important conclusion that follows from the foregoing is that the physical requirement of simultaneous satisfaction of the principles of causality and equivalence does not lead to the requirement of geometrical completeness, but only to the requirement of causal completeness of S-T - geometric completeness in one of the two TL directions (into the "future"). In a causally complete S-T, a family of TL geodesics which are not continuable without limit in the past, is admissible. This was illustrated above by examples pertaining to the vacuum regions of S-T with spherical or cylindrical symmetry in connection with the fact that the number of exact solutions of Einstein's equations, for which extension methods have been developed, is small, and that almost all of them pertain to vacuum. The next example, however, shows that a geometrically incomplete S-T structure is physically conceivable also in the presence of large masses of matter.

Let us examine a world produced as a result of anticollapse (explosion) from a Schwarzschild singularity. Prior to the start of the anticollapse, this world could formally be described by the pure vacuum Schwarzschild solution continued without limit into the "past". However, such a "empty duration" lacking any physical events has no content. Essentially, a boundary of a physical world "in the past" is part of a light cone with a vertex at the point corresponding to the start of the anticollapse. The TL geodesics starting on this boundary are the ones forming a family of geodesics which cannot be continued into the "physical past" without limit. This is a physical fact, having no direct bearing on the formal geometry of the Riemannian manifold describing the S-T, because the geodesic families can be continued formally beyond the limit of the physical S-T either without limit or to the Schwarzschild singularity. A physical world represents only part of a geometrically complete manifold.

The same meaning is possessed also by the considered causally irreversible extensions. Formally, they are geometrically incomplete, admitting of further extension (symmetrization, "increase of multiplicity," "doubling," or Kruskal extension), which leads to a complete geometry. The physical meaning of the S-T boundary as a boundary separating the physical region from a region having no events (matter) is less lucid for vacuum solutions. The clarity is restored by assuming that only in the physical part of the S-T is there "trial" matter. The physical essence of such an assumption is ensured by condition B of Sec. 1.

The need for symmetrizing causally irreversible S-T—i.e., extending it to a complete geometry—to a causally reversible S-T does not follow from physical facts. To the contrary, physical observations serve as evidence in favor of the simultaneous satisfaction of the principles of causality and equivalence in macroscopic processes. Therefore the requirement that the structure of S-T admit of the possibility of time reversal (that it be formally geometrically complete), would be an independent postulate, introduced independently of experiment. In this sense, at the present stage of development of the theory and of observations, one must admit of a logical possibility of a physical world with incomplete geometry.

5. The considered exact solutions of Einstein's equations are vacuum solutions with spherical or cylindrical symmetry. Actually we considered only part of their region of definition: the vicinity Ω of the surface Γ. However, the method of extending the solutions is based essentially on the possibility of unlimited continuation of the surface Γ into the future and on relation (2), which are valid for the metric in vacuum. If these conditions are violated when joining the extended vacuum solutions to the non-vacuum ones, then the foregoing considerations regarding the direction of time in causally irreversible extensions are not applicable. That is to say, in anticollapse, the joining with the solution for matter cuts off part of the Schwarzschild "singular sphere" in the future. Therefore the direction of the TL geodesics from T into R, which can be realized in the vacuum region in the case of anticollapse, is compatible with the causality principle. This is due to the well known fact that a body from R cannot penetrate into T in the case of anticollapse - it will collide first with the front of the expanding matter, beyond which the T region no longer exists. To the contrary, the causality principle is not compatible with such a process as the discarding of a shell by a body situated "under the singular sphere," for in this case the structure of the "singular sphere" in the remote future remains the same.

The joining of causally irreversible extension with non-vacuum solutions outside the vicinity of Ω (e.g., a collapsing body in the metagalaxy) does not affect the considerations of compatibility with the causality principle of only time direction in T, corresponding to motion from R into T.

The author is sincerely grateful to L. É. Gurevich, A. Z. Dolginov, A. G. Doroshkevich, I. D. Novikov, V. A. Ruban, and A. D. Chernin for a discussion of different aspects of the work and for critical remarks.

[1] C. W. Misner, J. Math. Phys. 4, 924 (1963).
[2] A. D. Novikov, Astron. zh. 43, 911 (1966) [Soviet Atronomy AJ 10, 731 (1967)].
[3] B. Carter, Phys. Rev. 141, 1242 (1966).
[4] R. H. Boyer and R. W. Lindquist, J. Math. Phys. 8, 265 (1967).
[5] J. L. Synge, Relativity, The General Theory, Amsterdam, 1960.
[6] R. H. Dicke, Gravitation and Relativity, (Russ. Transl.), Mir, 1965.
[7] E. Newman, L. Tamburino, and T. Unti, J. Math. Phys. 3, 902 (1962).
[8] R. P. Kerr, Phys. Rev. Lett. 11, 238 (1963).
[9] R. Finkelstein, Phys. Rev. 110, 965 (1958).
[10] M. D. Kruskal, ibid. 119, 1743 (1960).
[11] V. A. Unt. Trudy, Inst. of Phys. and Astron. Estonian Acad. Sci. No. 16, 27 (1961).
[12] I. D. Novikov, Soobshch. GAISh (Commun. Main Astron. Inst.), Moscow State Univ. Press, 1964.
[13] J. C. Graves and D. R. Brill, Phys. Rev. 120, 1507 (1960).
[14] A. Lichnerowicz. Theories relativistes de la gravitation et de l'electromagnetisme, Paris, 1955.
[15] R. Penrose, in: Gravitatsiya i Topologiya (Gravitation and Topology), Mir, 1966.

Translated by J. G. Adashko
95