Commun. math. Phys. 46, 135—152 (1976)

Communications in Mathematical Physics

An Application of Morse Theory to Space-Time Geometry

N. M. J. Woodhouse

Department of Mathematics, University of London, King's College, London WC2, England

Abstract. Milnor's treatment [6] of Morse's global theory of the calculus of variations for geodesics [7] is restated in the context of space-time geometry: it is seen as providing a link between the curvature and the causal structure of a stably causal globally hyperbolic Lorentzian manifold. An application is discussed.

Introduction

Morse's global theory of the calculus of variations is the basis of a number of theorems relating the curvature and topology of Riemannian manifolds [6, 7]. In this paper I shall describe a method whereby the theory can be restated in the context of space-time geometry and discuss its potential usefulness in dealing with global problems in general relativity.

The first three sections of the paper are an outline of the principal physical and mathematical ideas involved, leading up to a statement of the main theorem at the end of § 3: these sections can be regarded as an extended introduction (more detailed accounts of some of the material covered can be found in [1, 3, 4, 6, 9]). A large proportion of the argument consists of adapting standard elementary results from algebraic topology and Riemannian geometry. In order to keep the paper reasonably selfcontained, I have given outlines of the concepts involved and sketched proofs of the theorems before describing the necessary (but, for the most part, trivial) modifications. Only where the argument diverges radically from that used in Riemannian geometry have I gone into the full technical details.

The fourth section is a proof of the theorem.

Notation. Throughout, M denotes a smooth (C^∞) paracompact Hausdorff manifold of dimension greater than two in which is given a causal Lorentzian metric $\boldsymbol{g}$ with signature $(+, -, -\ldots)$. This means that $(M, \boldsymbol{g})$ is time oriented (the two halves of the light cone are labelled continuously throughout M as future and

past pointing) and contains no self intersecting causal paths (a causal path is a piecewise smooth map of a closed interval in $\mathbb{R}$ into M with future pointing timelike or null tangent vector). Thus M admits an antireflexive partial ordering $\ll$ (the *natural chronology*) defined by $p \ll q$ if there exists a timelike path from p to q and $p \neq q$, and a reflexive partial ordering $<$ (the *natural causality*) defined by $p < q$ if there exists a causal path from p to q or $p = q$. [In general, I shall distinguish between a path (which is a map from the real line into M) and a curve (which is the image of a path).]

The timelike and causal futures and pasts of an event $p \in M$ are denoted, respectively, $I^+(p)$, $J^+(p)$, $I^-(p)$ and $J^-(p)$ (for example, $I^-(p) = \{q \in M \mid q \ll p\}$).

A subset $S \subset M$ is said to be *açausal* if:

$$p, q \in S \Rightarrow p \not< q .$$

The *domain of dependence* of a closed acausal set S is defined by:

$$D(S) = \{q \in M \mid \text{every maximally extended timelike curve through } q \text{ intersects } S\} .$$

A closed acausal hypersurface without boundary is called a *partial Cauchy surface:* it is a Cauchy surface if $D(S) = M$.

In addition to being causal, in § 4 $(M, \boldsymbol{g})$ is required to be *globally hyperbolic* and *stably causal.* This means that each $\overline{I^+(p) \cap I^-(q)}$, $p, q \in M$, is compact and that M admits a second causal metric $\boldsymbol{h}$ such that every vector which is causal (that is, timelike or null) with respect to $\boldsymbol{g}$ is timelike with respect to $\boldsymbol{h}$: less formally, the light cones can be widened without destroying causality. (These are not the minimum axioms under which the theorem can be proved, but they make sense physically.) It follows that M is predictable (admits a Cauchy surface) and that every event lies in a *local causality neighbourhood*, that is a goedesically convex (normal) neighbourhood N with compact closure, satisfying:

If $q, r \in N$ then $q < r$ if, and only if, there exists a causal path from q to r in N.

More details of these concepts are given in [4, 9].

The metric connexion on M is denoted ∇ and the inner product in the tangent space at each event $\langle , \rangle$. The curvature tensor is defined by (X, Y and Z are vectors fields):

$$\nabla_X \nabla_Y Z - \nabla_Y \nabla_X Z - \nabla_{[X,Y]} Z = R(X, Y)Z .$$

A closed n-cell (that is a closed disc in $\mathbb{R}^n$) is denoted E^n and its boundary S^{n-1}. The boundary of a zero cell (a point) is empty.

The *index* of a symmetric bilinear form H on a vector space V is the maximum dimension of a subspace of V on which H is positive definite (this is the reverse of the normal definition: the reasons for this will emerge later). The *nullity* of H is the dimension of the subspace of V:

$$\{X \in V \mid H(X, Y) = 0 \forall Y \in V\} .$$

§ 1. Conjugate Points and Chronological Homotopy

The physical interpretation of the curvature tensor in space-time is based on the equation of geodesic deviation (the Jacobi equation) [11]. Briefly, the way this goes is this: a cloud of free falling test particles in space-time is represented by a timelike geodetic congruence. If T is the tangent vector field to this congruence and if X is some other vector field which is Lie propagated by the congruence, then (after a short calculation):

$$D^2X = R(T, X)T \quad \text{(or, in components: } D^2X^a = R^a_{bcd}T^bX^cT^d\text{)}, \tag{1.1}$$

where $D = \nabla_T$ is the covariant derivative in the direction of T. This is the Jacobi equation. An observer $\mathcal{O}$ moving along a geodesic γ_0 in the congruence can measure the position of a nearby particle $\mathcal{P}$ relative to a parallelly propagated orthonormal tetrad (T, Y_1, Y_2, Y_3) (Y_1, Y_2, and Y_3 can be thought of as non-rotating Cartesian axes in $\mathcal{O}$'s local rest frame). At each instant, $\mathcal{O}$ assigns a position vector $X = X^\alpha Y_\alpha$ ($\alpha = 1, 2, 3$) to $\mathcal{P}$. Provided that all the geodesics are parameterized by proper time, X will be Lie propagated by the congruence and [rewriting Eq. (1.1)] the frame components of the acceleration of $\mathcal{P}$ relative to $\mathcal{O}$ will be given by:

$$\frac{\partial^2 X^\alpha}{\partial t^2} = -\langle R(T, Y_\beta)T, Y_\alpha\rangle X^\beta, \tag{1.2}$$

where t is proper time. This equation gives direct physical meaning to the Riemann tensor components $\langle R(T, Y_\beta)T, Y_\alpha\rangle$.

Now consider the situation where all the particles in the cloud emerge in an explosion at an event $p \in M$. Initially, $\mathcal{O}$ will see $\mathcal{P}$ moving directly away from him. However, if the quantity on the right hand side of Eq. (1.2) is sufficiently negative, $\mathcal{P}$ will, after a time, start moving back towards $\mathcal{O}$ and, eventually, pass close by $\mathcal{O}$ again at some event[1] $q \in M$. In this case, q is said to be conjugate to p along γ_0; mathematically, q is characterized as a conjugate point to p by the existence of a nontrivial solution of Eq. (1.1) which vanishes at p and at q. Physically, conjugate points arise because of the focusing effect of the curvature which is a consequence of the attractive nature of gravity.

The theorem I shall prove below relates the number of conjugate points on the timelike geodesics from p to q to the structure of the space of timelike paths joining p and q, which is described within the framework of chronological homotopy theory [5, 13, 14]. This theory is purely global in the sense that, locally, all stably causal space-times have the same chronological homotopy type. The theorem, in conjunction with the equation of geodesic deviation, thus provides a direct link between the local properties of the curvature tensor and the global properties of the causal structure, that is between the small and large scale physical aspects of the gravitational field.

Before going into technical details of the theorem, I shall give a brief account of the main ideas of chronological homotopy theory and its potential usefulness in handling global problems.

---

[1] Only in the limiting case where $\mathcal{P}$ is infinitely close to $\mathcal{O}$ will the geodesics actually intersect again at q. The description of P's position by X is only accurate to first order in the X^α's.

Let $(M, \boldsymbol{g})$ be a causal space-time and let $p \ll q$ be two fixed events in M. By a chronological path from p to q is meant a continuous map $\alpha: [0, 1] \to M$ such that:

1) $\alpha(0) = p$, $\alpha(1) = q$.

2) If $s < t \in [0, 1]$ then $\alpha(s) \ll \alpha(t)$.

If α is piecewise smooth then its tangent vector must be future pointing and causal; if α is piecewise geodetic, then its tangent vector must actually be timelike. The space T_{pq} of all chronological paths from p to q has a natural topology (the *compact-open topology* [1]) generated by sets of the form:

$$\{\alpha \in T_{pq} | \alpha(K) \subset U\},$$

where $K \subset [0, 1]$ is compact and $U \subset M$ is open. For two paths to be close in the compact-open topology, not only must the corresponding curves in M be close, but also pairs of points with the same parameter values. However, the topology ignores smoothness: the tangent vectors (if they exist) need not be close in any sense. Two paths $\alpha, \beta \in T_{pq}$ are said to be *chronologically homotopic* if they lie in the same path connected component of T_{pq} (this means that α can be deformed into β through a sequence of chronological paths). Two space-times M and M' have the same *chronological homotopy type* if there exists a homeomorphism $\varphi: M \to M'$ such that T_{pq} and $T_{\varphi(p)\varphi(q)}$ have the same homotopy structure for each $p, q \in M$.

A consequence of the theorem is that, in general (in a globally hyperbolic, stably causal space-time), T_{pq} has a very simple structure: it has the homotopy type of a finite cell complex. Roughly speaking[2], this means that it can be deformed into a space K made up of a finite number of cells of dimension $0, 1, 2 \ldots k$ (that is, points, line segments, closed discs in $\mathbb{R}^2$, solid spheres in $\mathbb{R}^3$ etc.) glued together along their boundaries (to give a simple example, a sphere S^2 can be thought of as a cell complex made up of a point and a 2-cell). The number of cells of each dimension in K carries a great deal of information about the topology of T_{pq}. For instance the Euler characteristic of T_{pq} is given by:

$$\chi(T_{pq}) = \chi(K) = \sum_0^k (-1)^i \mu_i, \tag{1.3}$$

where μ_i is the number of cells of dimension i in K (for a more detailed explanation of this, see [1]). If p and q lie in a local causality neighbourhood, then K consists of a single 0-cell (a point), so that, locally, all stably causal space-times are of the same chronological homotopy type.

It must be emphasized that the structure of T_{pq} depends on the choice of p and q: chronological homotopy does not lead to any simple topological classification of the background manifold M (though, of course, the structure of T_{pq} is closely linked with the topology of M). However, the information carried by, for example, the two point integer valued function $\chi(p, q) = \chi(T_{pq})$ is physically relevant. For instance, consider the problem raised by Penrose's formulation of the cosmic censorship hypothesis [10] of defining black holes in closed universes. Loosely, this formulation is that no observer can ever see a singularity which was once in his past; thus no observer can see a singularity formed by collapse (either the local collapse of a star or the global collapse of the entire universe) until he actually runs into it. More formally, space-time must be globally hyperbolic (and

[2] Formal definitions are given in § 2.

hence predictable from a Cauchy surface). This contrasts with the conventional formulation (Hawking [3]) that no singularity formed by collapse can be seen by an observer at future null infinity ($\mathscr{I}^+$): such singularities are hidden inside black holes. In this context a black hole is defined to be a spatially connected region of space not in the past of $\mathscr{I}^+$ (more precisely, it is a connected component of $S \sim J^-(\mathscr{I}^+)$ where S is a partial Cauchy surface). This version states that space-time is (future) asymptotically predictable.

Penrose's formulation (which is stronger than Hawking's) has the advantage of being applicable to space-times (such as closed cosmological models) for which future null infinity is not defined. However, it leaves open the question of precisely what is meant by a black hole in such situations. One way of getting a handle on this problem is through chronological homotopy theory. Consider, first, a Schwarzschild black hole. Let p be a fixed event outside the horizon and consider what happens to T_{pq} as q moves into the future along some timelike path through p (also outside the horizon). At first T_{pq} will have a trivial structure: it will be equivalent to a single point. However, when q gets further into the future, there will exist timelike paths from p to q which "loop around the back" of the horizon. It is not hard to see that a hole appears in T_{pq}: its homotopy structure is that of a circle. As q gets still further into the future, the structure of T_{pq} gets more and more complicated. However, if the black hole is replaced by a star, the situation is qualitatively different: the structure of T_{pq} remains relatively simple for all points p and q.

Thus, the picture one would have of a closed universe is this: for points p and q near the initial singularity T_{pq} has a very simple structure: possible timelike paths have only one route from p to q or, possibly, they can wind round the back of the universe a few times. However if p and q are near the final singularity, which is made up of collapsed stars and the final collapsed state of the universe itself, T_{pq} has a vastly more complicated structure due to the presence of a large number of "black holes".

It is possible that a closer analysis of, for example, the way in which the two point function $\chi(p, q)$ behaves in exact black hole solutions will lead to a precise and workable definition of a black hole applicable in any situation: the theorem proved in this paper provides the technical machinery necessary for such an analysis.

In the next two sections, I shall review the elements of finite dimensional Morse theory and the calculus of variations for timelike geodesics: these form the basis of the theorem proved in § 4.

§ 2. Morse Theory: The Finite Dimensional Case

Definitions. Let A be a smooth paracompact Hausdorff manifold of dimension n and let $f: A \to \mathbb{R}$ be a smooth function on A. A *critical point* of f is a point $c \in A$ where the 1-form df vanishes; the value $f(c)$ of f at c is a *critical value* of f. At each critical point $c \in A$, f defines a symmetric bilinear form H_c (the *Hessian* of f) in the tangent space T_cA at c: if $X \in T_cA$, then:

$$H_c(X, X) = \left.\frac{\partial^2 f}{\partial x^2}\right|_{x=0}, \tag{2.1}$$

where the derivative is taken along any path $\xi: x \mapsto \xi(x)$ through $c = \xi(0)$ with tangent vector X at c. (That this definition is independent of the choice of ξ is most easily seen by rewriting Eq. (2.1) in the form:

$$H_c(X, X) = \frac{\partial^2 f}{\partial x^a \partial x^b} X^a X^b, \tag{2.2}$$

where $\{x^a\}$ are local coordinates at c.) The index μ_c and the nullity ν_c of f at c (or just of c if f is understood) are, respectively, the index and nullity of H_c. If $\mu_c = n$, then c is a local maximum of f. If $\nu_c \neq 0$, then c is a *degenerate* critical point: f is said to be *nondegenerate* if it has no degenerate critical points.

The idea is to relate the topology of A to the indices of the critical points of f. That such a relationship must exist is illustrated by the fact that though it is possible to find a smooth function on the sphere S^2 which has a maximum and a minimum, but no other critical points, it is not possible to do this on the two dimensional torus. Formally, the relationship is expressed in the theorem [6]:

2.1. Theorem. *Let $f: A \to \mathbb{R}$ be smooth and nondegenerate. If each set*

$$A_s = \{a \in A \mid f(a) \geqq s\}, \quad s \in \mathbb{R},$$

is compact then A is homotopically equivalent to a CW-complex K containing one cell of dimension μ for each critical point of index μ.

(Two topological spaces A_1 and A_2 have the same homotopy type (are *homotopically equivalent*, written $A_1 \sim A_2$) if there exist maps $\varphi_{12}: A_1 \to A_2$ and $\varphi_{21}: A_2 \to A_1$ such that $\varphi_{12} \circ \varphi_{21}$ and $\varphi_{21} \circ \varphi_{12}$ are homotopic with the identity maps on A_2 and A_1. In the application in § 4, the function considered has only a finite number of critical points. In this case K is a *finite cell complex*, that is it is the union of a finite number of closed sets C_i^μ (μ and i are integers) with the following properties: if $K^\mu = \bigcup_{\lambda \leqq \mu} C_i^\lambda$ and $B_i^\mu = K^{\mu-1} \cap C_i^\mu$, then:

K 1) $(C_i^\mu - B_i^\mu) \cap (C_j^\lambda - B_j^\lambda) = \emptyset$ unless $\mu = \lambda$ and $i = j$.

K 2) For each C_i^μ, there exists a map $\varphi_i^\mu: E^\mu \to K$ which takes $S^{\mu-1}$ onto B_i^μ and maps $E^\mu - S^{\mu-1}$ homeomorphically onto $C_i^\mu - B_i^\mu$.

For example, any compact triangulated manifold is a finite cell complex. More details are given in [1].)

In fact, in § 4, I shall use a slightly stronger version of the theorem. Before stating this, I shall outline a proof of Theorem 2.1 (a more detailed version is given by Milnor [6]).

The method is to investigate how the topology of A_s changes as s is decreased from the maximum value $s_{\max}$ of f to $-\infty$. To do this it is necessary to find a smooth vector field Y on A with the properties:

Y 1) $Y = 0$ only at the critical points of f.

Y 2) $Y(f) > 0$ except at the critical points of f.

Such a vector field always exists: for example, put $Y = \boldsymbol{g}^{-1}(df)$ where $\boldsymbol{g}$ is any Riemannian metric on A (this is not, in fact, how Y is constructed in § 4).

When $s = s_{\max}$, A_s is just a finite collection of points. Also if the interval $[s, t]$ contains no critical values of f, then $A_t \sim A_s$. The map $\varphi_{ts}: A_t \to A_s$ is just the natural inclusion map. To construct the map $\varphi_{st}: A_s \to A_t$, for each $a \in f^{-1}([s, t])$ let

$\alpha_a:[0, t-f(a)]\to A$ be an integral curve of Y *parameterized by* f with initial point a, that is:

1) $\alpha_a(0)=a$.
2) $f\circ\alpha_a(u)=u+f(a)$; $u\in[0, t-f(a)]$.
3) The tangent to α_a is parallel to Y.

Then $\varphi_{st}:A_s\to A_t$ is defined by:

$$\begin{aligned}\varphi_{st}(a)&=\alpha_a(t-f(a)) \qquad f(a)<t\\ &=a \qquad\qquad\quad f(a)\geqq t\,.\end{aligned} \tag{2.3}$$

This definition makes sense: because $Y\neq 0$ on the compact set $f^{-1}([s, t])$ and because $Y(f)>0$, an integral path of Y with initial point in $f^{-1}([s, t])$ must eventually reach $f^{-1}(t)$. Further φ_t is continuous and $\varphi_{st}\circ\varphi_{ts}$ is the identity on A_t. The homotopy $F:[0, 1]\times A_s\to A_t$ between $\varphi_{ts}\circ\varphi_{st}$ and the identity on A_s is defined by:

$$\begin{aligned}F(u, a)&=\alpha_a(u(t-f(a))) \qquad f(a)<t, \quad u\in[0, 1]\\ &=a \qquad\qquad\qquad\quad f(a)\geqq t, \quad u\in[0, 1]\,.\end{aligned} \tag{2.4}$$

Thus, when $s=s_{\max}$, A_s has the homotopy type of a finite collection of 0-cells (points) and, as s decreases, the homotopy type of A_s only changes when s passes through a critical value of f.

Suppose that s_c is a critical value of f and that $c_1, c_2\ldots c_m\in f^{-1}(s_c)$ are the corresponding critical points in A. Then, for small enough ε, $A_{s_c-\varepsilon}$ is equivalent to $A_{s_c+\varepsilon}$ with cells $E^{\mu_{c_1}}, E^{\mu_{c_2}}\ldots E^{\mu_{c_m}}$ attached. The proof of this is intuitively straightforward, though quite long when written out explicitly: one chooses coordinates $\{x_a\}$ in a neighbourhood N_i of each critical point c_i so that f has the local coordinate form:

$$f=s_c+x_1^2+x_2^2+\ldots+x_{\mu_1}^2-x_{\mu_{i+1}}^2-\ldots-x_n^2$$

(this is possible by a lemma of Morse [6]). Outside these neighbourhoods $A_{s_c-\varepsilon}$ is pushed along Y into $A_{s_c+\varepsilon}$ as before. A purely local argument is used to deal with what happens inside each N_i: for each i, the μ_{c_i}-cell:

$$\{(x_1, x_2\ldots x_{\mu_i}, 0, 0\ldots 0)\,|\,x_1^2+x_2^2+\ldots+x_{\mu_i}^2\leqq\varepsilon^2\}\,; \quad \mu_i=\mu_{c_i}$$

must be attached to $A_{s_c+\varepsilon}$ in N_i (this illustrated in Fig. 1 for $n=3$ and $\mu_{c_i}=2$).

Remark. Though it is clear from this outline that a space with the same homotopy type as A can be built up from a number of 0-cells by attaching a cell of the appropriate dimension for each critical point of f (such a space is called a spherical complex), it is not immediately clear that this space is a finite cell complex (that is that the cells are attached in such a way that the axioms K1 and K2 are satisfied): some sublety must be employed to prove this. However, for many applications [for example, the proof of Eq. (1.3)] this refinement is unnecessary.

Suppose now that $A\subset B$ is an open submanifold of B with compact closure and that, as before, f is a nondegenerate smooth function on A with a continuous extension to $\bar{A}$. Then, provided that there exists a smooth vector field Y on A

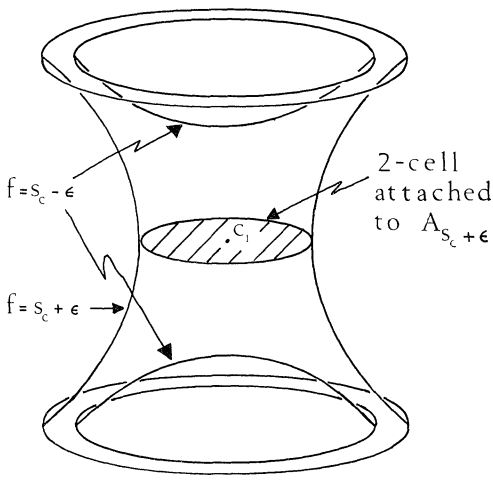

Fig. 1. The change in homotopy type at a critical point of index 2

which satisfies, in addition to Y1 and Y2:

Y3) If $\xi: x \mapsto \xi(x)$ is an integral path of Y with initial point $\xi(0) \in A$, then $\xi(x) \in A$ can be defined for all $x \geqq 0$, and $\xi(\overline{[0, \infty)}) \cap \dot{A} = \varphi$ (that is, ξ never reaches the boundary of A in B),

precisely the same proof can be used to show that the conclusion of the theorem still holds.

In its main application to Riemannian geometry, this theorem is used to calculate the structure of the space of paths joining two fixed points in a geodesically complete Riemannian manifold, and hence the homotopy invariants of the manifold. Naïvely, what is done is to treat the energy functional:

$$E = \int_0^1 \langle T, T \rangle dt$$

as a real function on this path space (all the paths are assumed to parameterized by the interval $[0, 1]$): its "critical points" are the geodesics. There are two ways of realizing this idea: the first, and older, approach is to approximate the path space by a finite dimensional manifold of the same homotopy type [6]. The second, and more sophisticated method, is to extend the finite dimensional theory outlined above to infinite dimensional Riemannian manifolds (that is, Hilbert manifolds) and then apply it directly to the full path space [8, 12]. It is possible that, with suitable modifications, this second approach could be made to work in Lorentzian manifolds: this has not yet been done. Here I shall show how the old method can be adapted to space-time geometry. What this approach lacks in mathematical sophistication, it makes up for with geometrical transparancy.

There are two points at which the Riemannian argument, as expounded by Milnor, breaks down when applied to metrics with Lorentzian signature (these same problems arise in a disguised form in trying to apply Palais and Smale's version of the theory). The first is that the energy functional is, locally, neither maximized nor minimized by affinely parameterized geodesics: for spacelike geodesics, this difficulty is intrinsic (a spacelike geodesic can be shortened by

deforming it in a timelike direction and lengthened by deforming it in a spacelike diection), but for timelike geodesics the problem can be trivially overcome by using a slightly different action functional. The second is in the construction of the approximation manifold. This is a more serious difficulty which arises from the noncompactness of the geodesic "spheres" even in geodesically complete space-times: two events can be far apart according to the manifold topology, but zero distance apart according to the metric. However, by concentrating on the timelike path space and by adopting a different (and, in fact, simpler) approximation procedure it is still possible to obtain physically interesting global information: the payoff is a means of relating local properties of the curvature to global properties of the causal structure. In this approach, the geodesic completeness condition is replaced by one of global hyperbolicity.

§ 3. The Calculus of Variations for Timelike Geodesics

Let $p \ll q$ be events in M and let $T^*_{pq} \subset T_{pq}$ denote the space of piecewise smooth timelike paths from p to q. An n-parameter variation of $\gamma_0 \in T^*_{pq}$ is a map $\gamma: U \times [0,1] \subset \mathbb{R}^n \times [0,1] \to M$ ($U \subset \mathbb{R}^n$ is an open neighbourhood of the origin) satisfying:

a) $\gamma(0,t) = \gamma_0(t)\ \forall t \in [0,1]$.

b) $\gamma_x : t \mapsto \gamma(\boldsymbol{x}, t)$ defines a path in T^*_{pq} for each $\boldsymbol{x} = (x^1, x^2, \ldots x^n) \in U$.

c) $\exists$ a partition $0 = t_0 < t_1 < \ldots < t_m = 1$ of $[0,1]$ such that each restriction $\gamma | U \times [t_i, t_{i+1}]$ is smooth.

According to emphasis, an n-parameter variation will either be written as a map $\gamma: U \times [0,1] \to M$ or as a collection of paths $\gamma_{\boldsymbol{x}}$.

Each n-parameter variation $\gamma_{\boldsymbol{x}}$ of a path $\gamma_0 \in T^*_{pq}$ generates a family $\{X_i\}$ of n continuous piecewise smooth vector fields on γ_0 which vanish at p and q. Explicitly, $X_i(t)$ is the tangent at $\gamma_0(t)$ to the path:

$$x^i \mapsto \gamma(0, \ldots 0, x^i, 0, \ldots 0, t)$$

for each $t \in [0,1]$.

The vector space of all continuous piecewise smooth vector fields on $\gamma_0 \in T^*_{pq}$ which vanish at the endpoints of γ_0 will be denoted by Γ; it is clear that any finite subset of Γ can be generated by a variation of γ_0. Naïvely, Γ can be thought of as the infinite dimensional tangent space to T^*_{pq} at γ_0 and, in fact, it is exactly this notion which is made precise in the modern treatment of Morse theory.

The action $j: T^*_{pq} \to \mathbb{R}$ is defined by:

$$j(\alpha) = \int_0^1 \langle T, T \rangle^{\frac{1}{4}} dt\,; \qquad \alpha \in T^*_{pq} \tag{3.1}$$

where T is the tangent vector to α. The exponent $\frac{1}{4}$ is chosen that the Schwarz inequality can be used directly to prove:

3.1. Lemma. *If $l(\alpha) = \int_0^1 \langle T, T \rangle^{\frac{1}{2}} dt$ is the proper length of the path $\alpha \in T^*_{pq}$ then $j(\alpha)^2 \leqq l(\alpha)$, with equality holding if and only if α is parameterized by a linear function of proper time.*

The point of this is that if p and q are sufficiently close (for instance if they lie in a local causality neighbourhood) then there is just one timelike geodesic curve from p to q; this geodesic maximizes proper time over all other timelike

curves from p to q [9]. It follows from the lemma that if the geodesic is parameterized by a linear function of proper time, then the resulting path maximizes j over all timelike *paths* in T^*_{pq}. In other words, locally j is maximized by affinely parameterized geodesics. The idea now is to investigate how this local behaviour is modified when p and q are not close together. The basic tools for this investigation are the first and second variation formulae (proofs are given in the appendix):

First Variation Formula. *If γ_x is a one parameter variation of $\gamma_0 \in T^*_{pq}$ generating the vector field $X \in \Gamma$ then:*

$$\frac{d}{\partial x}(j(\gamma_x))|_{x=0} = -\sum_i \tfrac{1}{2}\langle X, \triangle_i(\lambda^{-1}\cdot T)\rangle - \int_0^1 (\tfrac{1}{4}\cdot\lambda^{-1})\langle X, 3\perp DT - DT\rangle dt\,.$$

Second Variation Formula. *Let $\gamma_0 \in T^*_{pq}$ be an affinely parameterized geodesic. If $\gamma_{\boldsymbol{x}}$ is a two parameter variation of γ_0 generating the vector fields $X_1, X_2 \in \Gamma$ then:*

$$\frac{\partial^2}{\partial x^1 \partial x^2}(j(\gamma_{\boldsymbol{x}}))|_{\boldsymbol{x}=0} = \sum_i (\tfrac{1}{4}\cdot\lambda^{-1})\langle X_1 - 3\perp X_1, \triangle_i(DX_2)\rangle$$
$$- \int_0^1 (\tfrac{1}{4}\cdot\lambda^{-1})\langle X_1, 3\perp D^2X_2 - D^2X_2 - 2R(T, X_2)T\rangle dt\,.$$

Here T is the tangent to γ_0, $\lambda = \langle T, T\rangle^{\frac{1}{4}}$, $\perp$ is the projection orthogonal to T and $D = \nabla_T$. The points of $[0, 1]$ where $\gamma_{\boldsymbol{x}}$ fails to be smooth are labelled by i and $\triangle_i$ refers to the discontinuity at these points.

It follows from the first variation formula that the critical paths of j, that is the paths $\gamma_0 \in T^*_{pq}$ such that $\frac{d}{dx}(j(\gamma_x))|_{x=0} = 0$ for all one parameter variations of γ_0, are precisely the affinely parameterized geodesics in T^*_{pq} (this is because $3\perp DT - DT = 0$ if, and only if, $DT = 0$).

Just as in the finite dimensional case, j defines a symmetric bilinear form H (the *Hessian* of j) in the tangent space of its critical paths: if $\gamma_0 \in T^*_{pq}$ is an affinely parameterized geodesic and if $X_1, X_2 \in \Gamma$ then $H(X_1, X_2)$ is defined by choosing a two parameter variation $\gamma_{\boldsymbol{x}}$ of γ_0 which generates X_1 and X_2 and putting:

$$H(X_1, X_2) = \frac{\partial^2}{\partial x^1 \partial x^2}(j(\gamma_{\boldsymbol{x}}))|_{\boldsymbol{x}=0}\,. \tag{3.2}$$

The second variation formula implies that H is bilinear and independent of the choice of $\gamma_{\boldsymbol{x}}$; the symmetry follows directly from the definition.

A vector field $X_2 \in \Gamma$ in the null space of H at γ_0, is characterized by the condition:

$$H(X_1, X_2) = 0 \;\forall X_1 \in \Gamma\,. \tag{3.3}$$

This, in conjunction with the second variation formula, implies:

$$\text{a)}\;\; \triangle_i(DX_2) = 0 \;\forall i\,, \tag{3.4}$$

$$\text{b)}\;\; D^2X_2 = R(T, X_2)T\,. \tag{3.5}$$

That is, X_2 must be of class C^1 and it must satisfy Eq. (3.5) (the equation of geodesic deviation); in fact, a) and b) together imply that X_2 must be smooth. A vector field

satisfying Eq. (3.5) on a geodesic is called a *Jacobi field;* if n linearly independent Jacobi fields vanish at two distinct points p and q on a geodesic γ then p is said to be *conjugate to* q *along* γ with *multiplicity* n. What has been shown, therefore, is that the nullity of H for an affinely parameterized geodesic $\gamma \in T^*_{pq}$ is equal to the multiplicity of p as a conjugate point to q along γ. The crucial point is that the index of H can also be characterized in terms of the conjugate points of γ. This is the content of the index theorem:

Morse's Index Theorem. *Let $\gamma \in T^*_{pq}$ be an affinely parameterized geodesic. The index of H at γ is equal to the number of points $t \in [0, 1]$ such that $\gamma(t)$ is conjugate to q along γ, each such point being counted according to multiplicity.*

I will not give the proof here since it is rather long and it is identical to the proof of the corresponding theorem in Riemannian geometry, as given in a number of standard works on the calculus of variations (for example, Milnor [6]). Briefly, the idea is this: first a partition $0 = t_0 < t_1 < \ldots < t_n = 1$ of $[0, 1]$ is chosen so that each $\gamma(t_i, t_{i+1})$ is contained in a normal neighbourhood. Next, Γ is split into the direct sum of two subspaces Γ_1 and Γ_2 which are orthogonal with respect to H: Γ_1 consists of vector fields which vanish at $t_0, t_1, \ldots$ and t_n. Since j is maximized by affinely parameterized geodesics in a normal neighbourhood, H is negative semi-definite on Γ_1. The second subspace, Γ_2, is finite dimensional: it consists of broken Jacobi fields. A vector field $X \in \Gamma_2$ satisfies Eq. (3.5) in each interval (t_i, t_{i+1}) but its derivative DX need not be continuous at each t_i; such a vector field can be generated by a variation γ_x where each γ_x is a broken geodesic (that is it consists of a finite number of geodesic segments).

The index of H is equal to the index of its restriction $H|\Gamma_2$: this is computed by a slightly intricate argument in which p is allowed to move along γ. At each conjugate point to q the index of $H|\Gamma_2$ is shown to increase by the multiplicity of the conjugate point. □

The index of a geodesic $\gamma_0 : [0, 1] \to M$ is defined to be the index of the bilinear form H.

Enough technical machinery has now been assembled to state the main theorem (the proof is the content of the next section):

Theorem. *Let $p \ll q$ be two events in a stably causal globally hyperbolic space-time M. If p is not conjugate to q along any geodesic in T_{pq} then T_{pq} has the homotopy type of a CW-complex K containing one cell of dimension μ for each geodesic in T_{pq} of index μ.*

Remark. In general T_{pq} will contain only a finite number of geodesics and K will be a finite cell complex.

§ 4. The Proof of the Theorem

The proof proceeds in three stages. The first step is to construct an approximation manifold and to prove that it has the same homotopy type as T_{pq}. In the second stage, a certain vector field is defined on this manifold and in the third it is shown that this vector field has the properties necessary for the application of the finite dimensional theory outlined in § 2.

As before, $p \ll q$ are two fixed events in M. The open set $I^+(p) \cap I^-(q)$ is denoted I. Since M is globally hyperbolic, $\bar{I}$ is compact and equal to $J^+(p) \cap J^-(q)$.

By a theorem of Hawking's [2], M admits a continuous global time coordinate $\tau: M \to \mathbb{R}$ which is strictly increasing along every future directed causal path in M (the existence of such a time coordinate is an alternative characterization of stable causality). Each level surface of τ is an acausal C^0 hypersurface in M. Let $S = \tau^{-1}(s)$ be such a level surface. Since every point of S lies in a local causality neighbourhood, there exists a neighbourhood[3] N_S of S with the property:

> If $m \in S$ and $n \in N_S$ then $m \ll n$ ($m < n$) if, and only if, there exists exactly one timelike (causal) geodesic from m to n in N_S (and dually).

The collection of all neighbourhoods N_S, where S is a level surface of τ, forms an open cover of the compact set $\bar{I}$. Thus it is possible to find a finite set $\{S_0, S_1 \ldots S_{n+1}\}$ of level surfaces (labelled in order of increasing τ) such that:

a) $p \in S_0$, $q \in S_{n+1}$.
b) $\bar{I} \subset \bigcup_i N_{S_i}$.
c) $\bar{I} \cap S_i \subset N_{S_{i-1}} \cap N_{S_i}$; $i = 1, 2 \ldots n$.

[c) is a straight forward consequence of b).] Also, again since $\bar{I}$ is compact, there is no problem in assuming that each S_i is smooth near I.

Let W be the open subset of the product manifold $P = S_1 \times S_2 \ldots \times S_n$ consisting of n-tuplets $(m_1, m_2 \ldots m_n)$ which satisfy:

$$p \ll m_1 \ll m_2 \ll \ldots \ll m_n \ll q$$

and let V be the open subset of $\mathbb{R}^n$ of points $(t_1, t_2 \ldots t_n)$ satisfying:

$$0 < t_1 < t_2 < \ldots < t_n < 1$$

(for greater compactness in the following, I shall write $t_0 = 0$, $t_{n+1} = 1$, $m_0 = p$ and $m_{n+1} = q$). The approximation manifold A is defined to be the product $W \times V \subset P \times \mathbb{R}^n$.

4.1. Lemma. *A is homotopically equivalent to T_{pq}.*

Proof. There exist two maps $i: A \to T_{pq}$ and $r: T_{pq} \to A$. The first associates a broken geodesic $\gamma_a = i(a)$ with each point $a = (m_1, m_2 \ldots m_n, t_1, t_2 \ldots t_n) \in A$; γ_a is made up of the timelike geodesic segments joining m_0 to m_1, m_1 to m_2 and so on, the affine parameters on each segment being fixed by the condition:

$$\gamma_a(t_i) = m_i\,; \quad i = 0, 1 \ldots n+1\,. \tag{4.1}$$

The second map takes the path $\alpha \in T_{pq}$ to the point $r(\alpha) = a_\alpha = (m_1, m_2 \ldots m_n, t_1, t_2 \ldots t_n) \in A$ where m_i is the unique intersection point of α with S_i and $t_i = \alpha^{-1}(m_i)$.

Clearly $r \circ i: A \to A$ is the identity. To prove the lemma, it is sufficient to show that $i \circ r: T_{pq} \to T_{pq}$ is homotopic with the identity on T_{pq}. The required homotopy $F: [0, 1] \times T_{pq} \to T_{pq}$ is easily found: the image of $(u, \alpha) \in [0, 1] \times T_{pq}$ under F is the chronological path $F(u, \alpha): [0, 1] \to M: F(u, \alpha): t \mapsto F(u, \alpha)(t)$ which coincides with α for $0 \leqq t \leqq u$ and with a broken geodesic from $\alpha(u)$ to q for $u \leqq t \leqq 1$. To be explicit,

---

[3] For each $x \in S$, choose a local causality neighborhood $N_x \ni x$. Put $D_x = D(\bar{N}_x \cap S) \cap N_x$ and put $N_S = \bigcup_{x \in S} D_x$.

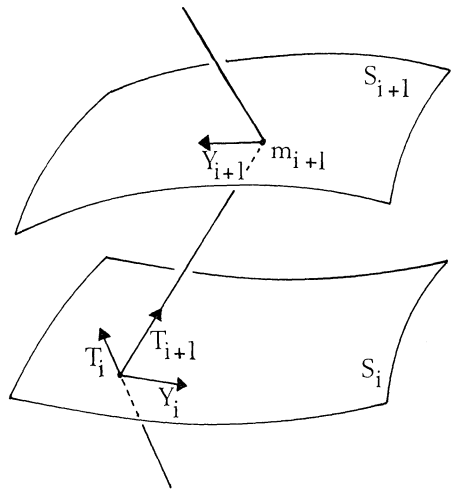

Fig. 2. The definition of Y_P

put $r(\alpha)=(m_1, m_2 \dots m_n, t_1, t_2 \dots t_n)$ and suppose that $u \in (t_{i-1}, t_i)$: the broken geodesic from $\alpha(u)$ to q is made up of the geodesic segment from $\alpha(u)$ to m_i, parameterized by the interval (u, t_i), together with the portion of $i \circ r(\alpha)$ between m_i and q. □

Let $a=(m_1 \dots m_n, t_1 \dots t_n) \in A$, let T_i be the tangent vector to γ_a on the geodesic segment between m_{i-1} and m_i and let $\lambda_i = \langle T_i, T_i \rangle^{\frac{1}{4}}$. The proper length l_i of the geodesic segment $m_{i-1} m_i$ is given by:

$$l_i = \int_{t_{i-1}}^{t_i} \langle T_i, T_i \rangle^{\frac{1}{2}} dt = \lambda_i^{\frac{2}{3}} \cdot (t_i - t_{i-1}) ; \quad i = 1, 2 \dots n+1 . \tag{4.2}$$

The quantities t_i, λ_i, and l_i and the action j can be regarded as real valued functions on A; they are all smooth and positive.

A tangent vector to A has the form $Y = Y_P + \boldsymbol{y}$ where $\boldsymbol{y} = (y_1, y_2 \dots y_n) \in \mathbb{R}^n$, $Y_P = (Y_1, Y_2 \dots Y_n)$ and, for each i, Y_i is tangent to the hypersurface S_i in M. Using this decomposition, one may define a vector field Y on A by specifying the values of Y_i and y_i at each $a=(m_1 \dots m_n, t_1 \dots t_n)$ as follows:

$$Y_i = \pi_i [(l_{i+1} l_i)^{\frac{3}{2}} \cdot (\lambda_{i+1}^{-1} \cdot T_{i+1} - \lambda_i^{-1} \cdot T_i)] , \tag{4.3}$$

$$y_i = (\lambda_i - \lambda_{i+1})(t_{i+1} - t)^{\frac{3}{2}} (t_i - t_{i-1})^{\frac{3}{2}} . \tag{4.4}$$

Here i runs over $1, 2 \dots n$. The right hand sides of these equations are evaluated on S_i and π_i denotes the orthogonal projection into S_i.

The aim is to compute the homotopy type of A by applying the finite dimensional theory to j and Y. The first step is to prove that j is monotonically increasing along the integral paths of Y. To do this, write $Y = Y_P + \boldsymbol{y}$ and compute $Y_P(j)$ and $\boldsymbol{y}(j)$ separately. If $\gamma_0 = i(a) \in T_{pq}$ is a broken geodesic and $\zeta : x \mapsto \zeta(x)$ is an integral path of Y_P through $\zeta(0) = a = (m_1 \dots m_n, t_1 \dots t_n)$ then $\gamma_x = i(\zeta(x))$ is a one parameter variation γ_0 which generates a broken Jacobi X on γ_0 satisfying:

$$X(t_i) = Y_i ; \quad i = 1, 2 \dots n . \tag{4.5}$$

It follows from the first variation formula that $Y_P(j) \geqq 0$.

An integral path of $\boldsymbol{y}$ through a defines a family of reparameterizations of $i(a)$. The function j can be written in the form:

$$j=\sum_{i=1}^{n+1}(l_i\cdot(t_i-t_{i-1}))^{\frac{1}{2}}. \tag{4.6}$$

Thus, since $\boldsymbol{y}(l_i)=0$:

$$\begin{aligned}\boldsymbol{y}(j)&=\sum_1^n \tfrac{1}{2}y_i\cdot(l_i^{\frac{1}{2}}\cdot(t_i-t_{i-1})^{-\frac{1}{2}}-l_{i+1}\cdot(t_{i+1}-t_i)^{-\frac{1}{2}})\\&=\sum_1^n \tfrac{1}{2}y_i\cdot(\lambda_i^{\frac{1}{2}}-\lambda_{i+1}^{\frac{1}{2}})\geqq 0.\end{aligned} \tag{4.7}$$

Next it must be shown that $Y(j)$ vanishes only at the critical points of j in A, that is at the points $a\in A$ for which $i(a)\in T_{pq}$ is geodetic. First note that if $i(a)$ is geodetic then $Y=0$ at a. Suppose that $Y(j)=0$ at a, that is $Y_p(j)=0=\boldsymbol{y}(j)$. By Eq. (4.7), $\lambda_i=\lambda_{i+1}$ and by the first variation formula, $\langle Y_i, Y_i\rangle=0$ (for each i). Thus $Y_i=0$ and $\lambda_i\cdot T_{i+1}-\lambda_{i+1}\cdot T_i$ is spacelike. But Y_i is a non-zero multiple of the projection of $\lambda_i\cdot T_{i+1}-\lambda_{i+1}\cdot T_i$ into the spacelike surface S_i. Hence, for each $i=1,2\ldots n$, $T_i=T_{i+1}$, that is $i(a)$ is a geodesic.

Now suppose that $a=(m_1\ldots m_n, t_1\ldots t_n)$ is a critical point of j, so that $i(a)\in T_{pq}$ is a geodesic. As in § 3, let Γ be the tangent to T_{pq}^* at $\gamma_0=i(a)$ and let H be the Hessian of j at γ_0. The following argument establishes that the index of j at a (as a function on A) is equal to the index of the bilinear form $H:\Gamma\times\Gamma\to\mathbb{R}$. Each one parameter variation γ_x of γ_0 generates a vector field $X\in\Gamma$ and defines a path ζ in A through a:

$$\zeta: x\mapsto r(\gamma_x).$$

Let $Z=(Z_1\ldots Z_n, z_1\ldots z_n)$ denote the tangent vector to this path at $a=\zeta(0)$. The one parameter family of broken geodesics $\tilde{\gamma}_x=i\circ r(\gamma_x)$ will not, in general, be a one parameter variation of γ_0 since the points of $[0,1]$ where $\tilde{\gamma}_x$ fails to be smooth are not fixed as x varies; it will however, generate a broken Jacobi field $\tilde{X}\in\tilde{\Gamma}$. Each restriction $\tilde{X}|_{(t_{i-1},t_i)}$ is smooth and:

$$\tilde{X}(t_i)=Z_i+z_i\cdot T;\qquad i=1,2\ldots n, \tag{4.8}$$

where T is the tangent to γ_0 in M. This equation identifies the tangent space T_aA at a with the subspace $\tilde{\Gamma}$ of Γ consisting of broken Jacobi fields which are smooth on each interval (t_{i-1}, t_i). If $X\in\tilde{\Gamma}$ then $X=\tilde{X}$ (though, in general, $\tilde{\gamma}_x\neq\gamma_x$) and:

$$H(X,X)=\left.\frac{\partial^2 j}{\partial x^2}\right|_{x=0}, \tag{4.9}$$

where the derivative is taken along ζ (see the remark following the proof of the second variation formula). Even when $X\neq\tilde{X}$, X can be written:

$$X=\tilde{X}+X' \tag{4.10}$$

where $X'(t_i)=0$, $i=1,2\ldots n$, and:

$$H(\tilde{X},X')=0 \tag{4.11}$$

(by the second variation formula). Thus (as in the proof of the index theorem [6]) the vector space Γ can decomposed into the direct sum of two H-orthogonal

subspaces:

$$\Gamma = \tilde{\Gamma} + \Gamma', \tag{4.12}$$

where the subspace Γ' consists of the vector fields in Γ which vanish at each t_i.

Now the section of γ_0 between m_{i-1} and m_i is contained in a local causality neighbourhood and so maximizes j over all other timelike paths form m_{i-1} to m_i [parameterized by the interval (t_i, t_{i-1})]. Thus γ_0 maximizes j over all variations which fix $p, m_1, m_2 \ldots m_n$ and q. Therefore H is negative semi-definite on Γ' and the index of H is the same as the index of its restriction to $\tilde{\Gamma}$; by Eq. (4.9), this is equal to the index at a of j (as a function on A). Further, if p is not conjugate to q along γ_0 then the nullity of j at a is zero.

The closure $\bar{A}$ of A in $P \times \mathbb{R}^n$ is compact, so that all that needs to be done to complete the proof of the theorem by applying the finite dimensional theory is to show that no integral path $\xi: x \mapsto A$ of Y reaches the boundary $\dot{A}$. More precisely, it must be shown that if $\xi: x \mapsto A$ is an integral path of Y with initial point $\xi(0) \in A$ then $\xi(x) \in A$ can be defined for all $x \in (0, \infty)$ and $\overline{\xi(0, \infty)} \cap A = \emptyset$. This involves examining the behaviour of Y on $\dot{A}$.

The functions l_i, t_i, and j have continuous extensions on $\bar{A}$ and, in fact $\dot{A}$ is characterized by the vanishing of either $t_i - t_{i-1}$ or of:

$$l_i^2 = \langle T_i, T_i \rangle (t_i - t_{i-1})^2 = (t_i - t_{i-1}) \cdot \int_{t_i}^{t_{i-1}} \langle T_i, T_i \rangle dt \tag{4.13}$$

for some $i = 1, 2 \ldots n+1$. The vector field Y is well defined and continuous on $\bar{A}$: this can be seen by rewriting Eqs. (4.3) and (4.4) in the forms:

$$Y_i = \pi_i(\varphi_i \cdot \tau_{i+1} \cdot T_{i+1} - \varphi_{i+1} \cdot \tau_i \cdot T_i), \tag{4.14}$$

$$y_i = \varphi_i \cdot \tau_{i+1} - \varphi_{i+1} \cdot \tau_i, \tag{4.15}$$

where:

$$\tau_i = (t_i - t_{i-1})^{\frac{3}{2}}, \tag{4.16}$$

$$\varphi_i = (l_i)^{\frac{3}{2}}, \tag{4.17}$$

and i runs over 1 to n. These equations are also formally valid for $i = 0$ and $i = n+1$ if:

$$\tau_0 = \varphi_0 = y_0 = Y_0 = \tau_{n+2} = \varphi_{n+2} = y_{n+1} = Y_{n+1} = 0.$$

From Eq. (4.15):

$$\begin{aligned} Y(t_i - t_{i-1}) &= \mathbf{y}(t_i - t_{i-1}) \\ &= \varphi_i(\tau_{i+1} + \tau_{i-1}) - \tau_i(\varphi_{i+1} + \varphi_{i-1}) \end{aligned} \tag{4.18}$$

while a short calculation along the lines of that used to prove the first variation formula yields:

$$\begin{aligned} Y((l_i)^2) &= Y_P((l_i)^2) \\ &= (t_i - t_{i-1}) \cdot (\langle T_i, Y_i \rangle_{t_i} - \langle T_i, Y_{i-1} \rangle_{t_{i-1}}) \\ &= (t_i - t_{i-1}) \cdot [\varphi_i(\tau_{i+1} \cdot \langle T_i, \pi_i T_{i+1} \rangle_{t_i} \\ &\quad + \tau_{i-1} \langle T_i, \pi_{i-1} T_{i-1} \rangle_{t_{i-1}}) - \tau_i(\varphi_{i+1} \langle T_i, \pi_i T_i \rangle_{t_i} \\ &\quad + \varphi_{i-1} \langle T_i, \pi_{i-1} T_i \rangle_{t_{i-1}})]. \end{aligned} \tag{4.19}$$

The suffices on the inner products indicate where they are to be evaluated. Both $\langle T_i, \pi_{i-1} T_i\rangle_{t_{i-1}}$ and $\langle T_i, \pi_i T_i\rangle_{t_i}$ are negative in A, so it follows from Eq. (4.19) that:

$$(l_i)^2 = 0 \Rightarrow Y((l_i)^2) \geqq 0\,; \qquad i = 1, 2 \ldots n+1$$

and from Eq. (4.18) that:

$$t_i - t_{i-1} = 0 \Rightarrow Y(t_i - t_{i-1}) \geqq 0\,; \qquad i = 1, 2 \ldots n+1\,.$$

Thus no integral path of Y can reach $\dot{A}$ at a point where $Y \neq 0$.

Unfortunately, there are points of $\dot{A}$ where $Y = 0$. The possibility of an integral path of Y reaching one of these points (after an infinite parameter distance) can be eliminated by the following rather messy argument: suppose that $\xi : x \mapsto A$ is an integral path of Y such that $\xi(x) \to a \in \dot{A}$ as $x \to \infty$ and that $Y = 0$ at a. Since j increases along ξ, $j(a) > 0$. Thus, for some value of i, $i = k+1$ say, $\varphi_{k+1}(a) \neq 0$ and $\tau_{k+1}(a) \neq 0$. Now $\boldsymbol{y} = 0$ at a, so it follows from Eq. (4.15) that either:

$$\varphi_i = \tau_i = 0 \tag{4.20}$$

or:

$$\varphi_i \neq 0 \quad \text{and} \quad \tau_i \neq 0 \tag{4.21}$$

for $i = k+2$ and $i = k$. Since Eq. (4.21) cannot hold for all values of i, it can be assumed, without loss of generality, that either $k \neq 0$ and $\varphi_k = \tau_k = 0$ or that $k \neq n$ and $\varphi_{k+2} = \tau_{k+2} = 0$. Suppose that the former statement is true and, for simplicity, suppose that $k \neq 1$ (the argument goes through in much the same way when $k = 1$).

As $x \to \infty$, $\tau_k(\xi(x)) \to 0$, so, for large values of x:

$$Y(t_k - t_{k-1}) < 0\,.$$

Substituting this into Eq. (4.18), one obtains that, for large values of x:

$$\begin{aligned} Y((l_k)^2) > \varphi_k(t_k - t_{k-1})[\mu(-\varphi_{k+1}\langle T_k, \pi_k T_k\rangle_{t_k} - \varphi_{k-1}\langle T_k, \pi_{k-1} T_k\rangle_{t_{k-1}}) \\ + \tau_{k+1}\langle T_k, \pi_k T_{k+1}\rangle_{t_k} + \tau_{k-1}\langle T_k, \pi_{k-1} T_{k-1}\rangle_{t_{k-1}}]\,, \end{aligned} \tag{4.22}$$

where:

$$\mu = (\tau_{k+1} + \tau_{k-1}) \cdot (\varphi_{k+1} + \varphi_{k-1})^{-1}\,.$$

As $\xi(x)$ approaches a, μ remains finite, $t_k - t_{k-1} \to 0$ and $\langle T_k, \pi_k T_k\rangle_{t_k}$ and $\langle T_k, \pi_{k-1} T_k\rangle_{t_{k-1}}$ become large and negative, behaving like $(t_k - t_{k-1})^{-2}$. The quantities:

$$(t_{k+1} - t_k)\langle T_k, \pi_k T_{k+1}\rangle_{t_{k+1}}$$

and:

$$(t_{k-1} - t_{k-2})\langle T_k, \pi_{k-1} T_{k-1}\rangle_{t_{k-1}}$$

also become infinite, but only as $(t_k - t_{k-1})^{-1}$. Thus, for large values of x, $Y((l_k)^2) > 0$, contradicting $(l_k(a))^2 = 0$.

To summarize, the smooth manifold A is homotopically equivalent to T_{pq}. The critical points of the function $j : A \to \mathbb{R}$ are in one to one correspondence with

the geodesics in T_{pq} and the index of j at each critical point is equal to the number of points conjugate to q on the corresponding geodesic in T_{pq}. If p is not conjugate to q along any geodesic in T_{pq} then the critical points of j are nondegenerate. The closure of A in $P \times \mathbb{R}^n$ is compact, j is strictly increasing along the integral paths of Y and no integral path of Y with initial point in A reaches the boundary of A. The theorem is now seen to be a corollary of the finite dimensional theory outlined in § 2.

Appendix: The Variation Formulae

Let $\gamma_0 \in T^*_{pq}$, let γ_x, $x \in (-\varepsilon, \varepsilon) \subset \mathbb{R}$, be a one parameter variation of γ_0 and let $0 = t_0 < t_1 < \ldots < t_m = 1$ be a partition of $[0, 1]$ such that $\gamma_x|_{(t_{i-1}, t_i)}$ is smooth for each $x \in (-\varepsilon, \varepsilon)$ and for each $i = 1, 2 \ldots m$. Let T be the tangent vector field to the γ_x's and let X be the vector field on γ_0 generated by the variation. Note that X can be extended off γ_0 as the tangent vector field to the family of paths:

$$x \mapsto \gamma_x(t); \quad t \in [0, 1].$$

The first variation formula is established by evaluating (at $x = 0$):

$$\frac{dj}{dx} = \int_0^1 \nabla_X(\langle T, T\rangle^{\frac{1}{4}}) dt \tag{A1}$$

(the integral is taken over γ_x), using the fact that $\mathcal{L}_X T = 0$, that is:

$$\nabla_X T = \nabla_T X. \tag{A2}$$

Now:

$$\begin{aligned} \nabla_X \langle T, T\rangle^{\frac{1}{4}} &= \tfrac{1}{2} \cdot \lambda^{-1} \cdot \langle T, \nabla_X T\rangle; \quad \lambda = \langle T, T\rangle^{\frac{3}{4}} \\ &= \tfrac{1}{2}[D(\lambda^{-1} \cdot \langle X, T\rangle) - \langle X, D(\lambda^{-1} \cdot T)\rangle] \end{aligned} \tag{A3}$$

and:

$$\begin{aligned} D(\lambda^{-1} \cdot T) &= \lambda^{-1}[DT - \tfrac{3}{2}\langle T, T\rangle^{-1} \cdot \langle T, DT\rangle \cdot T] \\ &= \tfrac{1}{2} \cdot \lambda^{-1}(3 \perp DT - DT), \end{aligned} \tag{A4}$$

where $\perp$ is the projection orthogonal to T. The proof is completed by substituting (A4) into (A3), integrating over each interval (t_{i-1}, t_i) and summing over i.

Now suppose that γ_0 is a geodesic. Let $\gamma_{\boldsymbol{x}}$ be a two parameter variation of γ_0, generating vector fields X_1 and X_2 on γ_0, and, as before, let $0 = t_0 < t_1 < \ldots < t_m = 1$ be a partition of $[0, 1]$ such that each $\gamma_{\boldsymbol{x}}|_{(t_{i-1}, t_i)}$ is smooth. Again, X_1 and X_2 can be extended off γ_0 so that:

$$\nabla_{X_1} T = \nabla_T X_1, \quad \nabla_{X_2} T = \nabla_T X_2, \quad \nabla_{X_1} X_2 = \nabla_{X_2} X_1. \tag{A5}$$

The second variation formula is established by calculating, at $\boldsymbol{x} = 0 = (0, 0)$:

$$\begin{aligned} \frac{\partial}{\partial x^2}\left(\frac{\partial j}{\partial x^1}\right) &= \frac{\partial}{\partial x^2}[-\tfrac{1}{2} \cdot \textstyle\sum_i \langle X_1, \triangle_i(\lambda^{-1} \cdot T)\rangle] - \int_0^1 \tfrac{1}{2}\langle X_1, \nabla_T(\lambda^{-1} \cdot T)\rangle dt \\ &= -\tfrac{1}{2}\textstyle\sum_i \langle X_1, \triangle_i(\nabla_{X_2}(\lambda^{-1} \cdot T)) \\ &\quad - \int_0^1 \langle X_1, \nabla_{X_2}(\nabla_T(\lambda^{-1} \cdot T))\rangle dt \end{aligned} \tag{A6}$$

since $\triangle_i(\lambda^{-1}\cdot T)=0$ and $\nabla_T(\lambda^{-1}\cdot T)=0$ on γ_0. Now:

$$\begin{aligned}\nabla_{X_2}(\lambda^{-1}\cdot T)&=\lambda^{-1}[DX_2-\tfrac{3}{2}\langle T,T\rangle^{-1}\cdot\langle T,DX_2\rangle\cdot T]\\&=\tfrac{1}{2}\cdot\lambda^{-1}[3\perp DX_2-DX_2]\end{aligned}\tag{A7}$$

and:

$$\begin{aligned}\nabla_{X_2}(\nabla_T(\lambda^{-1}\cdot T))&=\nabla_T(\nabla_{X_2}(\lambda^{-1}\cdot T))-R(T,X_2)\lambda^{-1}\cdot T\\&=\tfrac{1}{2}\cdot\lambda^{-1}[3\perp D^2X_2-D^2X_2-R(T,X_2)T]\,.\end{aligned}\tag{A8}$$

The proof is completed by substituting (A7) and (A8) into (A6).

Remark. In this second proof, it is not actually necessary that $\gamma_{\boldsymbol{x}}$ be a two parameter variation of γ_0: the proof works equally well for any two parameter family in T^*_{pq} containing γ_0 subject only to the conditions:

1) $\gamma:(\boldsymbol{x},t)\mapsto\gamma_{\boldsymbol{x}}(t)$ is smooth near $\gamma_0(t)$ for all but a finite number of values of t.
2) The vector fields X_1 and X_2 generated by $\gamma_{\boldsymbol{x}}$ on γ_0 are continuous.

Thus it is not essential that the points in [0, 1] where $\gamma_{\boldsymbol{x}}$ fails to be smooth are fixed as $\boldsymbol{x}$ varies.

Acknowledgements. I thank Stephen Hawking, Roger Penrose, and Felix Pirani for enlightening conversations.

This research was supported by the SRC.

An earlier version of this work appeared in the author's Ph.D. thesis [15].

References

1. Greenberg, M.J.: Lectures on algebraic topology. Menlo Park, California: Benjamin 1966
2. Hawking, S.W.: Proc. Roy. Soc. A **308**, 433 (1968)
3. Hawking, S.W.: The event horizon. In: DeWitt, C., DeWitt, B.S. (Eds.): Les astres occlus, les Houches 1972. New York-London-Paris: Gordon and Breach 1973
4. Hawking, S.W., Ellis, G.F.R.: The large scale structure of space-time. Cambridge: University Press 1973
5. Kronheimer, E.H.: G.R.G. **1**, 261 (1971)
6. Milnor, J.: Morse theory, annals of mathematics studies 51. Princeton: University Press 1963
7. Morse, M.: The calculus of variations in the large. New York: American Mathematical Society 1934
8. Palais, R.S.: Topology **2**, 299 (1963)
9. Penrose, R.: Techniques of differential topology in relativity. Philadelphia: SIAM 1974
10. Penrose, R.: Lectures at IAU symposia Nos. 63 and 64 (Warsaw, Cracow 1973) (to be published in the proceedings of the conference)
11. Pirani, F.A.E.: Introduction to gravitational radiation theory. In: Lectures on general relativity, Brandeis Summer Institute 1964. Englewood Cliffs, New Jersey: Prentice-Hall 1965
12. Smale, S.: Ann. Math. **80**, 382 (1964)
13. Smith, J.W.: Proc. Nat. Acad. Sci. **46**, 111 (1960)
14. Smith, J.W.: Amer. J. Math. **82**, 873 (1961)
15. Woodhouse, N.M.J.: Causal spaces and the structure of space-time. Thesis, London University 1973

Communicated by J. Ehlers

(Received August 5, 1974)