Annals of Mathematics Studies

Number 51

MORSE THEORY

BY

J. Milnor

Based on lecture notes by

M. SPIVAK and R. WELLS

PRINCETON, NEW JERSEY

PRINCETON UNIVERSITY PRESS

L.C. Card 63-13729
ISBN 0-691-08008-9

Third Printing, with corrections
and a new Preface, 1969
Fourth Printing, 1970
Fifth Printing, 1973

Printed in the United States of America

19 18 17 16 15 14

PREFACE

This book gives a present-day account of Marston Morse's theory of the calculus of variations in the large. However, there have been important developments during the past few years which are not mentioned. Let me describe three of these

R. Palais and S. Smale have studied Morse theory for a real-valued function on an infinite dimensional manifold and have given direct proofs of the main theorems, without making any use of finite dimensional approximations. The manifolds in question must be locally diffeomorphic to Hilbert space, and the function must satisfy a weak compactness condition. As an example, to study paths on a finite dimensional manifold M one considers the Hilbert manifold consisting of all absolutely continuous paths $\omega\colon [0,1] \to M$ with square integrable first derivative. Accounts of this work are contained in R. Palais, Morse Theory on Hilbert Manifolds, Topology, Vol. 2 (1963), pp. 299-340; and in S. Smale, Morse Theory and a Non-linear Generalization of the Dirichlet Problem, Annals of Mathematics, Vol. 80 (1964), pp. 382-396.

The Bott periodicity theorems were originally inspired by Morse theory (see part IV). However, more elementary proofs, which do not involve Morse theory at all, have recently been given. See M. Atiyah and R. Bott, On the Periodicity Theorem for Complex Vector Bundles, Acta Mathematica, Vol. 112 (1964), pp. 229-247, as well as R. Wood, Banach Algebras and Bott Periodicity, Topology, 4 (1965-66), pp. 371-389.

Morse theory has provided the inspiration for exciting developments in differential topology by S. Smale, A. Wallace, and others, including a proof of the generalized Poincaré hypothesis in high dimensions. I have tried to describe some of this work in Lectures on the h-cobordism theorem, notes by L. Siebenmann and J. Sondow, Princeton University Press, 1965.

Let me take this opportunity to clarify one term which may cause confusion. In §12 I use the word "energy" for the integral

$$E = \int_0^1 \left\| \frac{d\omega}{dt} \right\|^2 dt$$

along a path $\omega(t)$. V. Arnol'd points out to me that mathematicians for the past 200 years have called E the "action" integral. This discrepancy in terminology is caused by the fact that the integral can be interpreted, in terms of a physical model, in more than one way.

Think of a particle P which moves along a surface M during the time interval $0 \leq t \leq 1$. The action of the particle during this time interval is defined to be a certain constant times the integral E. If no forces act on P (except for the constraining forces which hold it within M), then the "principle of least action" asserts that E will be minimized within the class of all paths joining $\omega(0)$ to $\omega(1)$, or at least that the first variation of E will be zero. Hence P must traverse a geodesic.

But a quite different physical model is possible. Think of a rubber band which is stretched between two points of a slippery curved surface. If the band is described parametrically by the equation $x = \omega(t)$, $0 \leq t \leq 1$, then the potential energy arising from tension will be proportional to our integral E (at least to a first order of approximation). For an equilibrium position this energy must be minimized, and hence the rubber band will describe a geodesic.

The text which follows is identical with that of the first printing except for a few corrections. I am grateful to V. Arnol'd, D. Epstein and W. B. Houston, Jr. for pointing out corrections.

J.W.M.

Los Angeles, June 1968.

CONTENTS

PREFACE v

PART I. NON-DEGENERATE SMOOTH FUNCTIONS ON A MANIFOLD

§1. Introduction. 1

§2. Definitions and Lemmas. 4

§3. Homotopy Type in Terms of Critical Values 12

§4. Examples. 25

§5. The Morse Inequalities. 28

§6. Manifolds in Euclidean Space: The Existence of Non-degenerate Functions. 32

§7. The Lefschetz Theorem on Hyperplane Sections. 39

PART II. A RAPID COURSE IN RIEMANNIAN GEOMETRY

§8. Covariant Differentiation 43

§9. The Curvature Tensor. 51

§10. Geodesics and Completeness. 55

PART III. THE CALCULUS OF VARIATIONS APPLIED TO GEODESICS

§11. The Path Space of a Smooth Manifold 67

§12. The Energy of a Path. 70

§13. The Hessian of the Energy Function at a Critical Path . . 74

§14. Jacobi Fields: The Null-space of E_{**}. 77

§15. The Index Theorem 83

§16. A Finite Dimensional Approximation to Ω^c 88

§17. The Topology of the Full Path Space 93

§18. Existence of Non-conjugate Points 98

§19. Some Relations Between Topology and Curvature 100

CONTENTS

PART IV. APPLICATIONS TO LIE GROUPS AND SYMMETRIC SPACES

§20. Symmetric Spaces 109

§21. Lie Groups as Symmetric Spaces 112

§22. Whole Manifolds of Minimal Geodesics 118

§23. The Bott Periodicity Theorem for the Unitary Group . . . 124

§24. The Periodicity Theorem for the Orthogonal Group 133

APPENDIX. THE HOMOTOPY TYPE OF A MONOTONE UNION 149

PART I

NON-DEGENERATE SMOOTH FUNCTIONS ON A MANIFOLD.

§1. Introduction.

In this section we will illustrate by a specific example the situation that we will investigate later for arbitrary manifolds. Let us consider a torus M, tangent to the plane V, as indicated in Diagram 1.

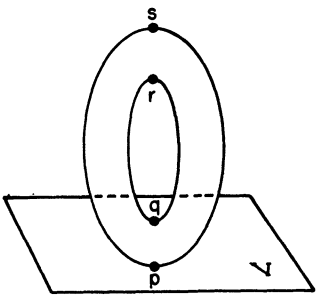


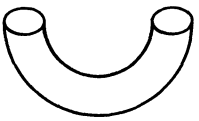
Diagram 1.

Let $f: M \to \mathbf{R}$ ($\mathbf{R}$ always denotes the real numbers) be the height above the V plane, and let M^a be the set of all points $x \in M$ such that $f(x) \leq a$. Then the following things are true:

(1) If $a < 0 = f(p)$, then M^a is vacuous.

(2) If $f(p) < a < f(q)$, then M^a is homeomorphic to a 2-cell.

(3) If $f(q) < a < f(r)$, then M^a is homeomorphic to a cylinder:



(4) If $f(r) < a < f(s)$, then M^a is homeomorphic to a compact manifold of genus one having a circle as boundary:

(5) If $f(s) < a$, then M^a is the full torus.

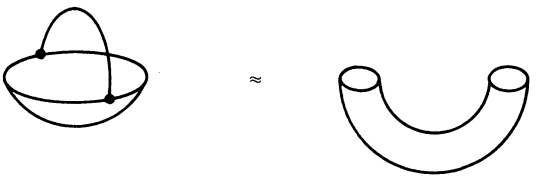
In order to describe the change in M^a as a passes through one of the points $f(p), f(q), f(r), f(s)$ it is convenient to consider homotopy type rather than homeomorphism type. In terms of homotopy types:

(1) → (2) is the operation of attaching a 0-cell. For as far as homotopy type is concerned, the space M^a, $f(p) < a < f(q)$, cannot be distinguished from a 0-cell:

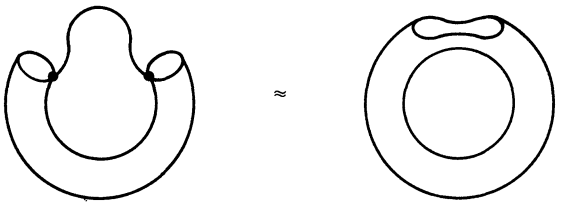
$\approx$ 

Here "$\approx$" means "is of the same homotopy type as."

(2) → (3) is the operation of attaching a 1-cell:



(3) → (4) is again the operation of attaching a 1-cell:



(4) → (5) is the operation of attaching a 2-cell.

The precise definition of "attaching a k-cell" can be given as follows. Let Y be any topological space, and let

$$e^k = \{x \in \mathbf{R}^k : \|x\| \leq 1\}$$

be the k-cell consisting of all vectors in Euclidean k-space with length ≤ 1.

The boundary

$$\dot{e}^k = \{x \in \mathbf{R}^k : \|x\| = 1\}$$

will be denoted by S^{k-1}. If $g: S^{k-1} \to Y$ is a continuous map then

$$Y \cup_g e^k$$

(Y with a k-cell attached by g) is obtained by first taking the topological sum (= disjoint union) of Y and e^k, and then identifying each $x \in S^{k-1}$ with $g(x) \in Y$. To take care of the case $k = 0$ let e^0 be a point and let $\dot{e}^0 = S^{-1}$ be vacuous, so that Y with a 0-cell attached is just the union of Y and a disjoint point.

As one might expect, the points p,q,r and s at which the homotopy type of M^a changes, have a simple characterization in terms of f. They are the critical points of the function. If we choose any coordinate system (x,y) near these points, then the derivatives $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are both zero. At p we can choose (x,y) so that $f = x^2 + y^2$, at s so that $f = \text{constant} - x^2 - y^2$, and at q and r so that $f = \text{constant} + x^2 - y^2$. Note that the number of minus signs in the expression for f at each point is the dimension of the cell we must attach to go from M^a to M^b, where $a < f(\text{point}) < b$. Our first theorems will generalize these facts for any differentiable function on a manifold.

REFERENCES

For further information on Morse Theory, the following sources are extremely useful.

M. Morse, "The calculus of variations in the large," American Mathematical Society, New York, 1934.

H. Seifert and W. Threlfall, "Variationsrechnung im Grossen," published in the United States by Chelsea, New York, 1951.

R. Bott, The stable homotopy of the classical groups, Annals of Mathematics, Vol. 70 (1959), pp. 313-337.

R. Bott, Morse Theory and its application to homotopy theory, Lecture notes by A. van de Ven (mimeographed), University of Bonn, 1960.

§2. Definitions and Lemmas.

The words "smooth" and "differentiable" will be used interchangeably to mean differentiable of class C^∞. The tangent space of a smooth manifold M at a point p will be denoted by TM_p. If $g: M \to N$ is a smooth map with $g(p) = q$, then the induced linear map of tangent spaces will be denoted by $g_*: TM_p \to TN_q$.

Now let f be a smooth real valued function on a manifold M. A point $p \in M$ is called a critical point of f if the induced map $f_*: TM_p \to T\mathbf{R}_{f(p)}$ is zero. If we choose a local coordinate system $(x^1,\dots,x^n)$ in a neighborhood U of p this means that

$$\frac{\partial f}{\partial x^1}(p) = \dots = \frac{\partial f}{\partial x^n}(p) = 0 .$$

The real number $f(p)$ is called a critical value of f.

We denote by M^a the set of all points $x \in M$ such that $f(x) \leq a$. If a is not a critical value of f then it follows from the implicit function theorem that M^a is a smooth manifold-with-boundary. The boundary $f^{-1}(a)$ is a smooth submanifold of M.

A critical point p is called non-degenerate if and only if the matrix

$$\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(p)\right)$$

is non-singular. It can be checked directly that non-degeneracy does not depend on the coordinate system. This will follow also from the following intrinsic definition.

If p is a critical point of f we define a symmetric bilinear functional f_{**} on TM_p, called the Hessian of f at p. If $v,w \in TM_p$ then v and w have extensions $\tilde{v}$ and $\tilde{w}$ to vector fields. We let * $f_{**}(v,w) = \tilde{v}_p(\tilde{w}(f))$, where $\tilde{v}_p$ is, of course, just v. We must show that this is symmetric and well-defined. It is symmetric because

$$\tilde{v}_p(\tilde{w}(f)) - \tilde{w}_p(\tilde{v}(f)) = [\tilde{v},\tilde{w}]_p(f) = 0$$

where $[\tilde{v},\tilde{w}]$ is the Poisson bracket of $\tilde{v}$ and $\tilde{w}$, and where $[\tilde{v},\tilde{w}]_p(f) = 0$

---

* Here $\tilde{w}(f)$ denotes the directional derivative of f in the direction $\tilde{w}$.

since f has p as a critical point.

Therefore f_{**} is symmetric. It is now clearly well-defined since $\tilde{v}_p(\tilde{w}(f)) = v(\tilde{w}(f))$ is independent of the extension $\tilde{v}$ of v, while $\tilde{w}_p(\tilde{v}(f))$ is independent of $\tilde{w}$.

If $(x^1,\dots,x^n)$ is a local coordinate system and $v = \Sigma\, a_i \left.\frac{\partial}{\partial x^i}\right|_p$, $w = \Sigma\, b_j \left.\frac{\partial}{\partial x^j}\right|_p$ we can take $\tilde{w} = \Sigma\, b_j \frac{\partial}{\partial x^j}$ where b_j now denotes a constant function. Then

$$f_{**}(v,w) = v(\tilde{w}(f))(p) = v(\Sigma\, b_j \frac{\partial f}{\partial x^j}) = \sum_{i\,j} a_i\, b_j \frac{\partial^2 f}{\partial x^i \partial x^j}(p)\ ;$$

so the matrix $\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(p)\right)$ represents the bilinear function f_{**} with respect to the basis $\left.\frac{\partial}{\partial x^1}\right|_p, \dots, \left.\frac{\partial}{\partial x^n}\right|_p$.

We can now talk about the index and the nullity of the bilinear functional f_{**} on TM_p. The index of a bilinear functional H, on a vector space V, is defined to be the maximal dimension of a subspace of V on which H is negative definite; the nullity is the dimension of the null-space, i.e., the subspace consisting of all $v \in V$ such that $H(v,w) = 0$ for every $w \in V$. The point p is obviously a non-degenerate critical point of f if and only if f_{**} on TM_p has nullity equal to 0. The index of f_{**} on TM_p will be referred to simply as the index of f at p. The Lemma of Morse shows that the behaviour of f at p can be completely described by this index. Before stating this lemma we first prove the following:

LEMMA 2.1. Let f be a C^∞ function in a convex neighborhood V of 0 in $\mathbf{R}^n$, with $f(0) = 0$. Then

$$f(x_1,\dots,x_n) = \sum_{i=1}^{n} x_i g_i(x_1,\dots,x_n)$$

for some suitable C^∞ functions g_i defined in V, with $g_i(0) = \frac{\partial f}{\partial x_i}(0)$.

PROOF:

$$f(x_1,\dots,x_n) = \int_0^1 \frac{df(tx_1,\dots,tx_n)}{dt}\,dt = \int_0^1 \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(tx_1,\dots,tx_n)\cdot x_i\,dt\ .$$

Therefore we can let $g_i(x_1,\dots,x_n) = \int_0^1 \frac{\partial f}{\partial x_i}(tx_1,\dots,tx_n)\,dt$.

LEMMA 2.2 (Lemma of Morse). Let p be a non-degenerate critical point for f. Then there is a local coordinate system $(y^1,\dots,y^n)$ in a neighborhood U of p with $y^i(p) = 0$ for all i and such that the identity

$$f = f(p) - (y^1)^2 - \dots - (y^\lambda)^2 + (y^{\lambda+1})^2 + \dots + (y^n)^2$$

holds throughout U, where λ is the index of f at p.

PROOF: We first show that if there is any such expression for f, then λ must be the index of f at p. For any coordinate system $(z^1,\dots,z^n)$, if

$$f(q) = f(p) - (z^1(q))^2 - \dots - (z^\lambda(q))^2 + (z^{\lambda+1}(q))^2 + \dots + (z^n(q))^2$$

then we have

$$\frac{\partial^2 f}{\partial z^i\,\partial z^j}(p) = \begin{cases} -2 & \text{if } i = j \le \lambda, \\ 2 & \text{if } i = j > \lambda, \\ 0 & \text{otherwise,} \end{cases}$$

which shows that the matrix representing f_{**} with respect to the basis $\frac{\partial}{\partial z^1}\Big|_p,\dots,\frac{\partial}{\partial z^n}\Big|_p$ is

$$\begin{pmatrix} -2 & & & & \\ & \ddots & & & \\ & & -2 & & \\ & & & 2 & \\ & & & & \ddots \\ & & & & & 2 \end{pmatrix}.$$

Therefore there is a subspace of TM_p of dimension λ where f_{**} is negative definite, and a subspace V of dimension $n-\lambda$ where f_{**} is positive definite. If there were a subspace of TM_p of dimension greater than λ on which f_{**} were negative definite then this subspace would intersect V, which is clearly impossible. Therefore λ is the index of f_{**}.

We now show that a suitable coordinate system $(y^1,\dots,y^n)$ exists. Obviously we can assume that p is the origin of $\mathbf{R}^n$ and that $f(p) = f(0) = 0$. By 2.1 we can write

$$f(x_1,\dots,x_n) = \sum_{j=1}^{n} x_j g_j(x_1,\dots,x_n)$$

for $(x_1,\dots,x_n)$ in some neighborhood of 0. Since 0 is assumed to be a critical point:

$$g_j(0) = \frac{\partial f}{\partial x^j}(0) = 0\,.$$

Therefore, applying 2.1 to the g_j we have

$$g_j(x_1,\ldots,x_n) = \sum_{i=1}^{n} x_i h_{ij}(x_1,\ldots,x_n)$$

for certain smooth functions h_{ij}. It follows that

$$f(x_1,\ldots,x_n) = \sum_{i,j=1}^{n} x_i x_j h_{ij}(x_1,\ldots,x_n) .$$

We can assume that $h_{ij} = h_{ji}$, since we can write $\bar{h}_{ij} = \frac{1}{2}(h_{ij} + h_{ji})$, and then have $\bar{h}_{ij} = \bar{h}_{ji}$ and $f = \Sigma\, x_i x_j \bar{h}_{ij}$. Moreover the matrix $(\bar{h}_{ij}(0))$ is equal to $\left(\frac{1}{2}\frac{\partial^2 f}{\partial x^i\, \partial x^j}(0)\right)$, and hence is non-singular.

There is a non-singular transformation of the coordinate functions which gives us the desired expression for f, in a perhaps smaller neighborhood of 0. To see this we just imitate the usual diagonalization proof for quadratic forms. (See for example, Birkhoff and MacLane, "A survey of modern algebra," p. 271.) The key step can be described as follows.

Suppose by induction that there exist coordinates $u_1,\ldots,u_n$ in a neighborhood U_1 of 0 so that

$$f = \pm (u_1)^2 \pm \cdots \pm (u_{r-1})^2 + \sum_{i,j\geq r} u_i u_j H_{ij}(u_1,\ldots,u_n)$$

throughout U_1; where the matrices $(H_{ij}(u_1,\ldots,u_n))$ are symmetric. After a linear change in the last $n-r+1$ coordinates we may assume that $H_{rr}(0) \neq 0$. Let $g(u_1,\ldots,u_n)$ denote the square root of $|H_{rr}(u_1,\ldots,u_n)|$. This will be a smooth, non-zero function of $u_1,\ldots,u_n$ throughout some smaller neighborhood $U_2 \subset U_1$ of 0. Now introduce new coordinates $v_1,\ldots,v_n$ by

$$v_i = u_i \qquad \text{for } i \neq r$$

$$v_r(u_1,\ldots,u_n) = g(u_1,\ldots,u_n)\Big[u_r + \sum_{i> r} u_i H_{ir}(u_1,\ldots,u_n)/H_{rr}(u_1,\ldots,u_n)\Big].$$

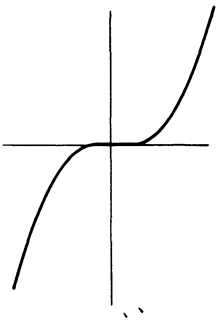
It follows from the inverse function theorem that $v_1,\ldots,v_n$ will serve as coordinate functions within some sufficiently small neighborhood U_3 of 0. It is easily verified that f can be expressed as

$$f = \sum_{i\leq r} \pm (v_i)^2 + \sum_{i,j>r} v_i v_j H'_{ij}(v_1,\ldots,v_n)$$

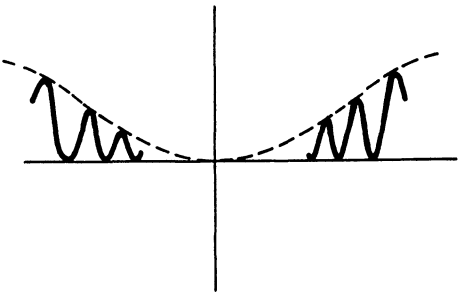
throughout U_3. This completes the induction; and proves Lemma 2.2.

COROLLARY 2.3 Non-degenerate critical points are isolated.

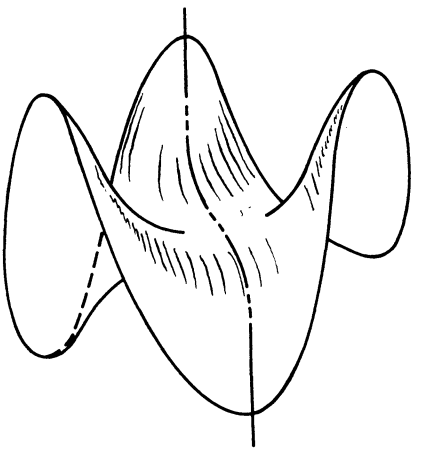
Examples of degenerate critical points (for functions on $\mathbf{R}$ and $\mathbf{R}^2$) are given below, together with pictures of their graphs.



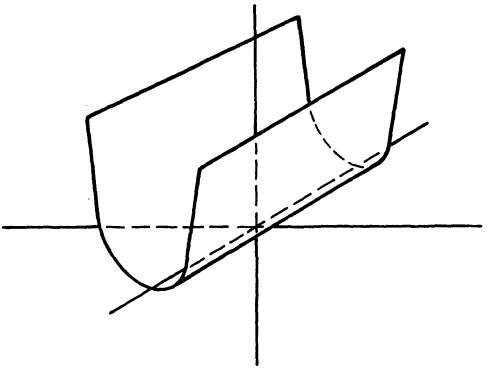
(a) $f(x) = x^3$. The origin is a degenerate critical point.



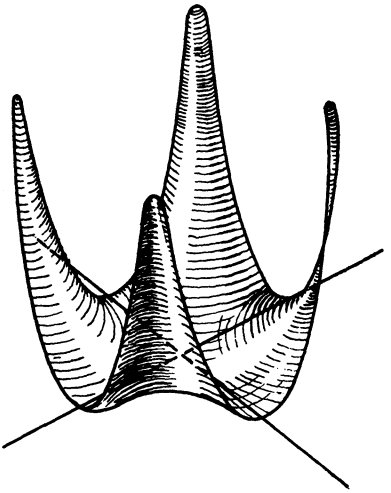
(b) $F(x) = e^{-1/x^2}\sin^2(1/x)$. The origin is a degenerate, and non-isolated, critical point.



(c) $f(x,y) = x^3 - 3xy^2$ = Real part of $(x + iy)^3$.
(0.0) is a degenerate critical point (a "monkey saddle").



(d) $f(x,y) = x^2$. The set of critical points, all of which are degenerate, is the x axis, which is a sub-manifold of $\mathbf{R}^2$.



(e) $f(x,y) = x^2y^2$. The set of critical points, all of which are degenerate, consists of the union of the x and y axis, which is not even a sub-manifold of $\mathbf{R}^2$.

We conclude this section with a discussion of 1-parameter groups of diffeomorphisms. The reader is referred to K. Nomizu, "Lie Groups and Differential Geometry," for more details.

A 1-parameter group of diffeomorphisms of a manifold M is a C^∞ map

$$\varphi : \mathbf{R} \times M \to M$$

such that

1) for each $t \in \mathbf{R}$ the map $\varphi_t : M \to M$ defined by $\varphi_t(q) = \varphi(t,q)$ is a diffeomorphism of M onto itself,

2) for all $t,s \in \mathbf{R}$ we have $\varphi_{t+s} = \varphi_t \circ \varphi_s$

Given a 1-parameter group φ of diffeomorphisms of M we define a vector field X on M as follows. For every smooth real valued function f let

$$X_q(f) = \lim_{h \to 0} \frac{f(\varphi_h(q)) - f(q)}{h} .$$

This vector field X is said to <u>generate</u> the group φ.

LEMMA 2.4. A smooth vector field on M which vanishes outside of a compact set $K \subset M$ generates a unique 1-parameter group of diffeomorphisms of M.

PROOF: Given any smooth curve

$$t \to c(t) \in M$$

it is convenient to define the <u>velocity vector</u>

$$\frac{dc}{dt} \in TM_{c(t)}$$

by the identity $\frac{dc}{dt}(f) = \lim_{h \to 0} \frac{f(c(t+h)) - f(c(t))}{h}$. (Compare §8.) Now let φ be a 1-parameter group of diffeomorphisms, generated by the vector field X. Then for each fixed q the curve

$$t \to \varphi_t(q)$$

satisfies the differential equation

$$\frac{d\varphi_t(q)}{dt} = X_{\varphi_t(q)} ,$$

with initial condition $\varphi_0(q) = q$. This is true since

$$\frac{d\varphi_t(q)}{dt}(f) = \lim_{h \to 0} \frac{f(\varphi_{t+h}(q)) - f(\varphi_t(q))}{h} = \lim_{h \to 0} \frac{f(\varphi_h(p)) - f(p)}{h} = X_p(f) ,$$

where $p = \varphi_t(q)$. But it is well known that such a differential equation, locally, has a unique solution which depends smoothly on the initial condition. (Compare Graves, "The Theory of Functions of Real Variables," p. 166. Note that, in terms of local coordinates $u^1,\ldots,u^n$, the differential equation takes on the more familiar form: $\frac{du^i}{dt} = x^i(u^1,\ldots,u^n)$, $i = 1,\ldots,n$.)

Thus for each point of M there exists a neighborhood U and a number $\varepsilon > 0$ so that the differential equation

$$\frac{d\varphi_t(q)}{dt} = X_{\varphi_t(q)}, \qquad \varphi_0(q) = q$$

has a unique smooth solution for $q \in U$, $|t| < \varepsilon$.

The compact set K can be covered by a finite number of such neighborhoods U. Let $\varepsilon_0 > 0$ denote the smallest of the corresponding numbers ε. Setting $\varphi_t(q) = q$ for $q \notin K$, it follows that this differential equation has a unique solution $\varphi_t(q)$ for $|t| < \varepsilon_0$ and for all $q \in M$. This solution is smooth as a function of both variables. Furthermore, it is clear that $\varphi_{t+s} = \varphi_t \circ \varphi_s$ providing that $|t|, |s|, |t+s| < \varepsilon_0$. Therefore each such φ_t is a diffeomorphism.

It only remains to define φ_t for $|t| \geq \varepsilon_0$. Any number t can be expressed as a multiple of $\varepsilon_0/2$ plus a remainder r with $|r| < \varepsilon_0/2$. If $t = k(\varepsilon_0/2) + r$ with $k \geq 0$, set

$$\varphi_t = \varphi_{\varepsilon_0/2} \circ \varphi_{\varepsilon_0/2} \circ \cdots \circ \varphi_{\varepsilon_0/2} \circ \varphi_r$$

where the transformation $\varphi_{\varepsilon_0/2}$ is iterated k times. If $k < 0$ it is only necessary to replace $\varphi_{\varepsilon_0/2}$ by $\varphi_{-\varepsilon_0/2}$ iterated $-k$ times. Thus φ_t is defined for all values of t. It is not difficult to verify that φ_t is well defined, smooth, and satisfies the condition $\varphi_{t+s} = \varphi_t \circ \varphi_s$. This completes the proof of Lemma 2.4

REMARK: The hypothesis that X vanishes outside of a compact set cannot be omitted. For example let M be the open unit interval $(0,1) \subset \mathbf{R}$, and let X be the standard vector field $\frac{d}{dt}$ on M. Then X does not generate any 1-parameter group of diffeomorphisms of M.

§3. Homotopy Type in Terms of Critical Values.

Throughout this section, if f is a real valued function on a manifold M, we let

$$M^a = f^{-1}(-\infty,a] = \{p \in M : f(p) \leq a\} .$$

THEOREM 3.1. Let f be a smooth real valued function on a manifold M. Let $a < b$ and suppose that the set $f^{-1}[a,b]$, consisting of all $p \in M$ with $a \leq f(p) \leq b$, is compact, and contains no critical points of f. Then M^a is diffeomorphic to M^b. Furthermore, M^a is a deformation retract of M^b, so that the inclusion map $M^a \to M^b$ is a homotopy equivalence.

The idea of the proof is to push M^b down to M^a along the orthogonal trajectories of the hypersurfaces f = constant. (Compare Diagram 2.)

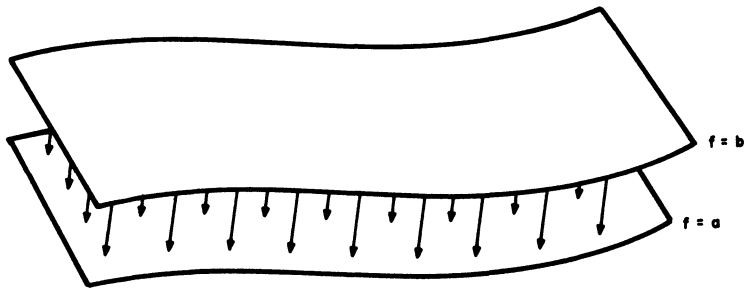


Diagram 2.

Choose a Riemannian metric on M; and let $\langle X,Y \rangle$ denote the inner product of two tangent vectors, as determined by this metric. The gradient of f is the vector field grad f on M which is characterized by the identity*

$$\langle X, \text{grad } f \rangle = X(f)$$

(= directional derivative of f along X) for any vector field X. This vector field grad f vanishes precisely at the critical points of f. If

---

* In classical notation, in terms of local coordinates $u^1,\ldots,u^n$, the gradient has components $\sum_j g^{ij} \frac{\partial f}{\partial u^j}$.

$c: \mathbf{R} \to M$ is a curve with velocity vector $\frac{dc}{dt}$ note the identity

$$\left\langle \frac{dc}{dt} , \text{grad } f \right\rangle = \frac{d(f \circ c)}{dt} .$$

Let $\rho: M \to \mathbf{R}$ be a smooth function which is equal to $1/\langle \text{grad } f, \text{grad } f \rangle$ throughout the compact set $f^{-1}[a,b]$; and which vanishes outside of a compact neighborhood of this set. Then the vector field X, defined by

$$X_q = \rho(q)\ (\text{grad } f)_q$$

satisfies the conditions of Lemma 2.4. Hence X generates a 1-parameter group of diffeomorphisms

$$\varphi_t: M \to M.$$

For fixed $q \in M$ consider the function $t \to f(\varphi_t(q))$. If $\varphi_t(q)$ lies in the set $f^{-1}[a,b]$, then

$$\frac{df(\varphi_t(q))}{dt} = \left\langle \frac{d\varphi_t(q)}{dt} , \text{grad } f \right\rangle = \langle X, \text{grad } f \rangle = +1.$$

Thus the correspondence

$$t \to f(\varphi_t(q))$$

is linear with derivative $+1$ as long as $f(\varphi_t(q))$ lies between a and b.

Now consider the diffeomorphism $\varphi_{b-a}: M \to M$. Clearly this carries M^a diffeomorphically onto M^b. This proves the first half of 3.1.

Define a 1-parameter family of maps

$$r_t: M^b \to M^b$$

by

$$r_t(q) = \begin{cases} q & \text{if } f(q) \leq a \\ \varphi_{t(a-f(q))}(q) & \text{if } a \leq f(q) \leq b . \end{cases}$$

Then r_0 is the identity, and r_1 is a retraction from M^b to M^a. Hence M^a is a deformation retract of M^b. This completes the proof.

REMARK: The condition that $f^{-1}[a,b]$ is compact cannot be omitted. For example Diagram 3 indicates a situation in which this set is not compact. The manifold M does not contain the point p. Clearly M^a is not a deformation retract of M^b.

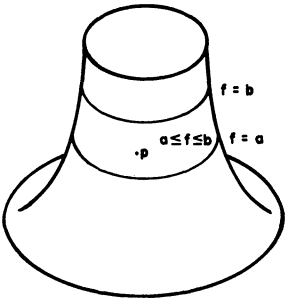


Diagram 3.

THEOREM 3.2. Let $f: M \to \mathbf{R}$ be a smooth function, and let p be a non-degenerate critical point with index λ. Setting $f(p) = c$, suppose that $f^{-1}[c-\varepsilon, c+\varepsilon]$ is compact, and contains no critical point of f other then p, for some $\varepsilon > 0$. Then, for all sufficiently small ε, the set $M^{c+\varepsilon}$ has the homotopy type of $M^{c-\varepsilon}$ with a λ-cell attached.

The idea of the proof of this theorem is indicated in Diagram 4, for the special case of the height function on a torus. The region

$$M^{c-\varepsilon} = f^{-1}(-\infty, c-\varepsilon]$$

is heavily shaded. We will introduce a new function $F: M \to \mathbf{R}$ which coincides with the height function f except that $F < f$ in a small neighborhood of p. Thus the region $F^{-1}(-\infty, c-\varepsilon]$ will consist of $M^{c-\varepsilon}$ together with a region H near p. In Diagram 4, H is the horizontally shaded region.

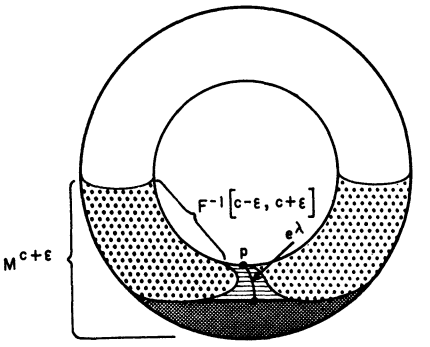


Diagram 4.

Choosing a suitable cell $e^\lambda \subset H$, a direct argument (i.e., pushing in along the horizontal lines) will show that $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $M^{c-\varepsilon} \cup H$. Finally, by applying 3.1 to the function F and the region $F^{-1}[c-\varepsilon, c+\varepsilon]$ we will see that $M^{c-\varepsilon} \cup H$ is a deformation retract of $M^{c+\varepsilon}$. This will complete the proof.

Choose a coordinate system $u^1, \ldots, u^n$ in a neighborhood U of p so that the identity

$$f = c - (u^1)^2 - \ldots - (u^\lambda)^2 + (u^{\lambda+1})^2 + \ldots + (u^n)^2$$

holds throughout U. Thus the critical point p will have coordinates

$$u^1(p) = \ldots = u^n(p) = 0 .$$

Choose $\varepsilon > 0$ sufficiently small so that

(1) The region $f^{-1}[c-\varepsilon, c+\varepsilon]$ is compact and contains no critical points other than p.

(2) The image of U under the diffeomorphic imbedding

$$(u^1, \ldots, u^n): \ U \longrightarrow \mathbf{R}^n$$

contains the closed ball.

$$\{(u^1, \ldots, u^n): \ \Sigma (u^i)^2 \leq 2\varepsilon\} .$$

Now define e^λ to be the set of points in U with

$$(u^1)^2 + \ldots + (u^\lambda)^2 \leq \varepsilon \quad \text{and} \quad u^{\lambda+1} = \ldots = u^n = 0.$$

The resulting situation is illustrated schematically in Diagram 5.

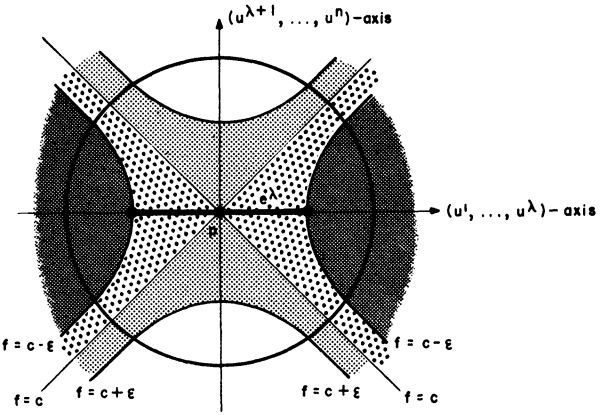


Diagram 5.

The coordinate lines represent the planes $u^{\lambda+1} = \ldots = u^n = 0$ and $u^1 = \ldots = u^\lambda = 0$ respectively; the circle represents the boundary of the ball of radius $\sqrt{2\varepsilon}$; and the hyperbolas represent the hypersurfaces $f^{-1}(c-\varepsilon)$ and $f^{-1}(c+\varepsilon)$. The region $M^{c-\varepsilon}$ is heavily shaded; the region $f^{-1}[c-\varepsilon,c]$ is heavily dotted; and the region $f^{-1}[c,c+\varepsilon]$ is lightly dotted. The horizontal dark line through p represents the cell e^λ.

Note that $e^\lambda \cap M^{c-\varepsilon}$ is precisely the boundary $\dot{e}^\lambda$, so that e^λ is attached to $M^{c-\varepsilon}$ as required. We must prove that $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $M^{c+\varepsilon}$.

Construct a new smooth function $F: M \longrightarrow \mathbf{R}$ as follows. Let

$$\mu : \mathbf{R} \longrightarrow \mathbf{R}$$

be a C^∞ function satisfying the conditions.

$$\mu(0) > \varepsilon$$
$$\mu(r) = 0 \quad \text{for} \quad r \geq 2\varepsilon$$
$$-1 < \mu'(r) \leq 0 \quad \text{for all} \quad r,$$

where $\mu'(r) = \frac{d\mu}{dr}$. Now let F coincide with f outside of the coordinate neighborhood U, and let

$$F = f - \mu((u^1)^2+\ldots+(u^\lambda)^2 + 2(u^{\lambda+1})^2+\ldots+2(u^n)^2)$$

within this coordinate neighborhood. It is easily verified that F is a well defined smooth function throughout M.

It is convenient to define two functions

$$\xi,\eta: U \longrightarrow [0,\infty)$$

by

$$\xi = (u^1)^2 + \ldots + (u^\lambda)^2$$
$$\eta = (u^{\lambda+1})^2 + \ldots + (u^n)^2$$

Then $f = c - \xi + \eta$; so that:

$$F(q) = c - \xi(q) + \eta(q) - \mu(\xi(q) + 2\eta(q))$$

for all $q \in U$.

ASSERTION 1. The region $F^{-1}(-\infty,c+\varepsilon]$ coincides with the region $M^{c+\varepsilon} = f^{-1}(-\infty,c+\varepsilon]$.

PROOF: Outside of the ellipsoid $\xi + 2\eta \leq 2\varepsilon$ the functions f and

F coincide. Within this ellipsoid we have

$$F \leq f = c-\xi+\eta \leq c+\tfrac{1}{2}\xi+\eta \leq c+\varepsilon \ .$$

This completes the proof.

ASSERTION 2. The critical points of F are the same as those of f.

PROOF: Note that

$$\frac{\partial F}{\partial \xi} = -1 - \mu'(\xi+2\eta) < 0$$

$$\frac{\partial F}{\partial \eta} = 1 - 2\mu'(\xi+2\eta) \geq 1 \ .$$

Since

$$dF = \frac{\partial F}{\partial \xi}\, d\xi + \frac{\partial F}{\partial \eta}\, d\eta$$

where the covectors $d\xi$ and $d\eta$ are simultaneously zero only at the origin, it follows that F has no critical points in U other than the origin.

Now consider the region $F^{-1}[c-\varepsilon, c+\varepsilon]$. By Assertion 1 together with the inequality $F \leq f$ we see that

$$F^{-1}[c-\varepsilon, c+\varepsilon] \subset f^{-1}[c-\varepsilon, c+\varepsilon] \ .$$

Therefore this region is compact. It can contain no critical points of F except possibly p. But

$$F(p) = c - \mu(0) < c - \varepsilon \ .$$

Hence $F^{-1}[c-\varepsilon, c+\varepsilon]$ contains no critical points. Together with 3.1 this proves the following.

ASSERTION 3. The region $F^{-1}(-\infty, c-\varepsilon]$ is a deformation retract of $M^{c+\varepsilon}$.

It will be convenient to denote this region $F^{-1}(-\infty, c-\varepsilon]$ by $M^{c-\varepsilon} \cup H$; where H denotes the closure of $F^{-1}(-\infty, c-\varepsilon] - M^{c-\varepsilon}$.

REMARK: In the terminology of Smale, the region $M^{c-\varepsilon} \cup H$ is described as $M^{c-\varepsilon}$ with a "handle" attached. It follows from Theorem 3.1 that the manifold-with-boundary $M^{c-\varepsilon} \cup H$ is diffeomorphic to $M^{c+\varepsilon}$. This fact is important in Smale's theory of differentiable manifolds. (Compare S. Smale, Generalized Poincaré's conjecture in dimensions greater than four, Annals of Mathematics, Vol. 74 (1961), pp. 391-406.)

Now consider the cell e^λ consisting of all points q with

$$\xi(q) \leq \varepsilon, \quad \eta(q) = 0.$$

Note that e^λ is contained in the "handle" H. In fact, since $\frac{\partial F}{\partial \xi} < 0$, we have

$$F(q) \leq F(p) < c-\varepsilon$$

but $f(q) \geq c-\varepsilon$ for $q \in e^\lambda$.

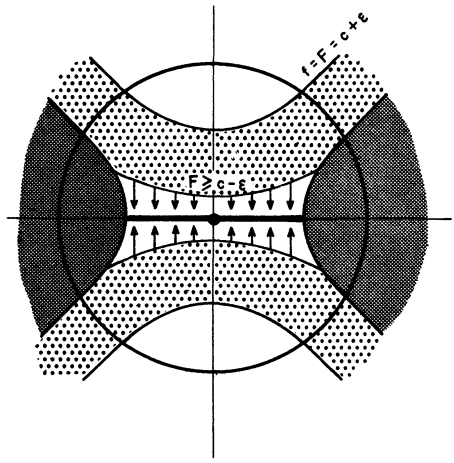


Diagram 6.

The present situation is illustrated in Diagram 6. The region $M^{c-\varepsilon}$ is heavily shaded; the handle H is shaded with vertical arrows; and the region $F^{-1}[c-\varepsilon, c+\varepsilon]$ is dotted.

ASSERTION 4. $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $M^{c-\varepsilon} \cup H$.

PROOF: A deformation retraction $r_t: M^{c-\varepsilon} \cup H \to M^{c-\varepsilon} \cup H$ is indicated schematically by the vertical arrows in Diagram 6. More precisely let r_t be the identity outside of U; and define r_t within U as follows. It in necessary to distinguish three cases as indicated in Diagram 7.

CASE 1. Within the region $\xi \leq \varepsilon$ let r_t correspond to the transformation

$$(u^1, \ldots, u^n) \to (u^1, \ldots, u^\lambda, tu^{\lambda+1}, \ldots, tu^n) .$$

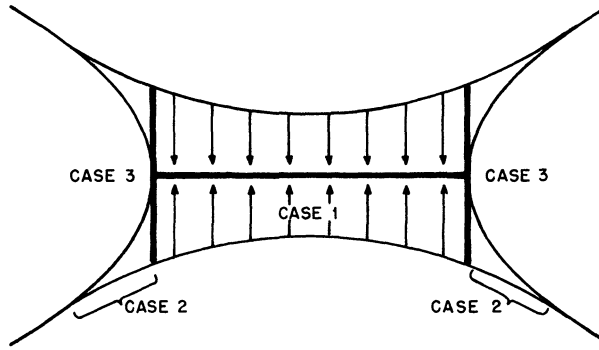


Diagram 7.

Thus r_1 is the identity and r_0 maps the entire region into e^λ. The fact that each r_t maps $F^{-1}(-\infty,c-\varepsilon]$ into itself, follows from the inequality $\frac{\partial F}{\partial \eta} > 0$.

CASE 2. Within the region $\varepsilon \leq \xi \leq \eta + \varepsilon$ let r_t correspond to the transformation

$$(u^1,\ldots,u^n) \to (u^1,\ldots,u^\lambda,s_t u^{\lambda+1},\ldots,s_t u^n)$$

where the number $s_t \in [0,1]$ is defined by

$$s_t = t + (1-t)((\xi-\varepsilon)/\eta)^{1/2} \quad .$$

Thus r_1 is again the identity, and r_0 maps the entire region into the hypersurface $f^{-1}(c-\varepsilon)$. The reader should verify that the functions $s_t u^i$ remain continuous as $\xi \to \varepsilon$, $\eta \to 0$. Note that this definition coincides with that of Case 1 when $\xi = \varepsilon$.

CASE 3. Within the region $\eta + \varepsilon \leq \xi$ (i.e., within $M^{c-\varepsilon}$) let r_t be the identity. This coincides with the preceeding definition when $\xi = \eta + \varepsilon$.

This completes the proof that $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of $F^{-1}(-\infty,c+\varepsilon]$. Together with Assertion 3, it completes the proof of Theorem 3.2.

REMARK 3.3. More generally suppose that there are k non-degenerate critical points $p_1,\ldots,p_k$ with indices $\lambda_1,\ldots,\lambda_k$ in $f^{-1}(c)$. Then a similar proof shows that $M^{c+\varepsilon}$ has the homotopy type of $M^{c-\varepsilon} \cup e^{\lambda_1} \cup \ldots \cup e^{\lambda_k}$.

REMARK 3.4. A simple modification of the proof of 3.2 shows that the set M^c is also a deformation retract of $M^{c+\varepsilon}$. In fact M^c is a deformation retract of $F^{-1}(-\infty,c]$, which is a deformation retract of $M^{c+\varepsilon}$. (Compare Diagram 8.) Combining this fact with 3.2 we see easily that $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of M^c.

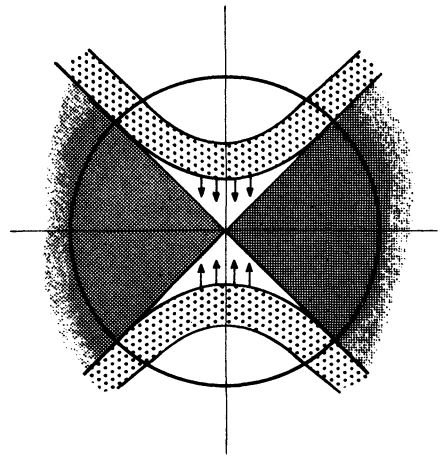


Diagram 8: M^c is heavily shaded, and $F^{-1}[c,c+\varepsilon]$ is dotted.

THEOREM 3.5. If f is a differentiable function on a manifold M with no degenerate critical points, and if each M^a is compact, then M has the homotopy type of a CW-complex, with one cell of dimension λ for each critical point of index λ.

(For the definition of CW-complex see J. H. C. Whitehead, Combinatorial Homotopy I, Bulletin of the American Mathematical Society, Vol. 55, (1949), pp. 213-245.)

The proof will be based on two lemmas concerning a topological space X with a cell attached.

LEMMA 3.6. (Whitehead) Let φ_0 and φ_1 be homotopic maps from the sphere $\dot{e}^\lambda$ to X. Then the identity map of X extends to a homotopy equivalence

$$k : X \underset{\varphi_0}{\cup} e^\lambda \to X \underset{\varphi_1}{\cup} e^\lambda .$$

PROOF: Define k by the formulas

$$\begin{aligned} k(x) &= x && \text{for } x \in X \\ k(tu) &= 2tu && \text{for } 0 \le t \le \tfrac{1}{2}, \quad u \in \dot{e}^\lambda \\ k(tu) &= \varphi_{2-2t}(u) && \text{for } \tfrac{1}{2} \le t \le 1, \quad u \in \dot{e}^\lambda . \end{aligned}$$

Here φ_t denotes the homotopy between φ_0 and φ_1; and tu denotes the product of the scalar t with the unit vector u. A corresponding map

$$\ell: X \cup_{\varphi_1} e^\lambda \to X \cup_{\varphi_0} e^\lambda$$

is defined by similar formulas. It is now not difficult to verify that the compositions $k\ell$ and ℓk are homotopic to the respective identity maps. Thus k is a homotopy equivalence.

For further details the reader is referred to Lemma 5 of J. H. C. Whitehead, On Simply Connected 4-Dimensional Polyhedra, Commentarii Math. Helvetici, Vol. 22 (1949), pp. 48-92.

LEMMA 3.7. Let $\varphi: \dot{e}^\lambda \to X$ be an attaching map. Any homotopy equivalence $f: X \to Y$ extends to a homotopy equivalence

$$F : X \cup_\varphi e^\lambda \to Y \cup_{f\varphi} e^\lambda .$$

PROOF: (Following an unpublished paper by P. Hilton.) Define F by the conditions

$$\begin{cases} F|X = f \\ F|e^\lambda = \text{identity} . \end{cases}$$

Let $g: Y \to X$ be a homotopy inverse to f and define

$$G: Y \underset{f\varphi}{\cup} e^\lambda \to X \underset{gf\varphi}{\cup} e^\lambda$$

by the corresponding conditions $G|Y = g$, $G|e^\lambda = \text{identity}$.

Since $gf\varphi$ is homotopic to φ, it follows from 3.6 that there is a homotopy equivalence

$$k: X \underset{gf\varphi}{\cup} e^\lambda \to X \underset{\varphi}{\cup} e^\lambda .$$

We will first prove that the composition

$$kGF: X \underset{\varphi}{\cup} e^\lambda \to X \underset{\varphi}{\cup} e^\lambda$$

is homotopic to the identity map.

Let h_t be a homotopy between gf and the identity. Using the specific definitions of k, G, and F, note that

$$\begin{aligned} kGF(x) &= gf(x) && \text{for } x \in X, \\ kGF(tu) &= 2tu && \text{for } 0 \le t \le \tfrac{1}{2}, \quad u \in \dot{e}^\lambda, \\ kGF(tu) &= h_{2-2t}\varphi(u) && \text{for } \tfrac{1}{2} \le t \le 1, \quad u \in \dot{e}^\lambda. \end{aligned}$$

The required homotopy

$$q_\tau : X \underset{\varphi}{\cup} e^\lambda \to X \underset{\varphi}{\cup} e^\lambda$$

is now defined by the formula

$$\begin{aligned} q_\tau(x) &= h_\tau(x) && \text{for } x \in X, \\ q_\tau(tu) &= \frac{2}{1+\tau}\, tu && \text{for } 0 \le t \le \tfrac{1+\tau}{2}, \quad u \in \dot{e}^\lambda, \\ q_\tau(tu) &= h_{2-2t+\tau}\varphi(u) && \text{for } \tfrac{1+\tau}{2} \le t \le 1, \quad u \in \dot{e}^\lambda. \end{aligned}$$

Therefore F has a left homotopy inverse.

The proof that F is a homotopy equivalence will now be purely formal, based on the following.

ASSERTION. If a map F has a left homotopy inverse L and a right homotopy inverse R, then F is a homotopy equivalence; and R (or L) is a 2-sided homotopy inverse.

PROOF: The relations

$$LF \simeq \text{identity}, \qquad FR \simeq \text{identity},$$

imply that

$$L \simeq L(FR) = (LF)R \simeq R.$$

Consequently

$$RF \simeq LF \simeq \text{identity},$$

which proves that R is a 2-sided inverse.

The proof of Lemma 3.7 can now be completed as follows. The relation

$$kGF \simeq \text{identity}$$

asserts that F has a left homotopy inverse; and a similar proof shows that G has a left homotopy inverse.

Step 1. Since $k(GF) \simeq$ identity, and k is known to have a left inverse, it follows that $(GF)k \simeq$ identity.

Step 2. Since $G(Fk) \simeq$ identity, and G is known to have a left inverse, it follows that $(Fk)G \simeq$ identity.

Step 3. Since $F(kG) \simeq$ identity, and F has kG as left inverse also, it follows that F is a homotopy equivalence. This completes the proof of 3.7.

PROOF OF THEOREM 3.5. Let $c_1 < c_2 < c_3 < \dots$ be the critical values of $f: M \to \mathbf{R}$. The sequence $\{c_i\}$ has no cluster point since each M^a is compact. The set M^a is vacuous for $a < c_1$. Suppose $a \neq c_1, c_2, c_3, \dots$ and that M^a is of the homotopy type of a CW-complex. Let c be the smallest $c_i > a$. By Theorems 3.1, 3.2, and 3.3, $M^{c+\varepsilon}$ has the homotopy type of $M^{c-\varepsilon} \cup_{\varphi_1} e^{\lambda_1} \cup \dots \cup_{\varphi_{j(c)}} e^{\lambda_{j(c)}}$ for certain maps $\varphi_1, \dots, \varphi_{j(c)}$ when ε is small enough, and there is a homotopy equivalence $h: M^{c-\varepsilon} \to M^a$. We have assumed that there is a homotopy equivalence $h': M^a \to K$, where K is a CW-complex.

Then each $h' \circ h \circ \varphi_j$ is homotopic by cellular approximation to a map

$$\psi_j: \dot{e}^{\lambda_j} \to (\lambda_j - 1) \text{ - skeleton of } K.$$

Then $K \cup_{\psi_1} e^{\lambda_1} \cup \dots \cup_{\psi_{j(c)}} e^{\lambda_{j(c)}}$ is a CW-complex, and has the same homotopy type as $M^{c+\varepsilon}$, by Lemmas 3.6, 3.7.

By induction it follows that each M^{a_i} has the homotopy type of a CW-complex. If M is compact this completes the proof. If M is not compact, but all critical points lie in one of the compact sets M^a, then a proof similar to that of Theorem 3.1 shows that the set M^a is a deformation retract of M, so the proof is again complete.

If there are infinitely many critical points then the above construction gives us an infinite sequence of homotopy equivalences

$$\begin{array}{ccccccc} M^{a_1} & \subset & M^{a_2} & \subset & M^{a_3} & \subset & \dots \\ \downarrow & & \downarrow & & \downarrow & & \\ K_1 & \subset & K_2 & \subset & K_3 & \subset & \dots \end{array},$$

each extending the previous one. Let K denote the union of the K_i in the direct limit topology, i.e., the finest possible compatible topology, and

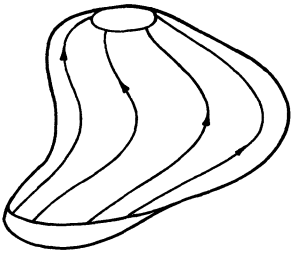
let $g: M \to K$ be the limit map. Then g induces isomorphisms of homotopy groups in all dimensions. We need only apply Theorem 1 of Combinatorial homotopy I to conclude that g is a homotopy equivalence. [Whitehead's theorem states that if M and K are both dominated by CW-complexes, then any map $M \to K$ which induces isomorphisms of homotopy groups is a homotopy equivalence. Certainly K is dominated by itself. To prove that M is dominated by a CW-complex it is only necessary to consider M as a retract of tubular neighborhood in some Euclidean space.] This completes the proof of Theorem 3.5.

REMARK. We have also proved that each M^a has the homotopy type of a finite CW-complex, with one cell of dimension λ for each critical point of index λ in M^a. This is true even if a is a critical value. (Compare Remark 3.4.)

§4. Examples.

As an application of the theorems of §3 we shall prove:

THEOREM 4.1 (Reeb). If M is a compact manifold and f is a differentiable function on M with only two critical points, both of which are non-degenerate, then M is homeomorphic to a sphere.



PROOF: This follows from Theorem 3.1 together with the Lemma of Morse (§2.2). The two critical points must be the minimum and maximum points. Say that $f(p) = 0$ is the mimimum and $f(q) = 1$ is the maximum. If ε is small enough then the sets $M^{\varepsilon} = f^{-1}[0,\varepsilon]$ and $f^{-1}[1-\varepsilon,1]$ are closed n-cells by §2.2. But M^{ε} is homeomorphic to $M^{1-\varepsilon}$ by §3.1. Thus M is the union of two closed n-cells, $M^{1-\varepsilon}$ and $f^{-1}[1-\varepsilon,1]$, matched along their common boundary. It is now easy to construct a homeomorphism between M and S^n.

REMARK 1. The theorem remains true even if the critical points are degenerate. However, the proof is more difficult. (Compare Milnor, Differential topology, in "Lectures on Modern Mathematics II," ed. by T. L. Saaty (Wiley, 1964), pp. 165-183; Theorem 1'; or R. Rosen, A weak form of the star conjecture for manifolds, Abstract 570-28, Notices Amer. Math Soc., Vol. 7 (1960), p. 380; Lemma 1.)

REMARK 2. It is not true that M must be diffeomorphic to S^n with its usual differentiable structure.(Compare: Milnor, On manifolds homeomorphic to the 7-sphere, Annals of Mathematics, Vol. 64 (1956), pp. 399-405. In this paper a 7-sphere with a non-standard differentiable structure is proved to be topologically S^7 by finding a function on it with two non-

degenerate critical points.)

As another application of the previous theorems we note that if an n-manifold has a non-degenerate function on it with only three critical points then they have index 0, n and n/2 (by Poincaré duality), and the manifold has the homotopy type of an n/2-sphere with an n-cell attached. See J. Eells and N. Kuiper, Manifolds which are like projective planes, Inst. des Hautes Etudes Sci., Publ. Math. 14, 1962. Such a function exists for example on the real or complex projective plane.

Let $\mathbf{C}P_n$ be complex projective n-space. We will think of $\mathbf{C}P_n$ as equivalence classes of (n+1)-tuples $(z_0,\dots,z_n)$ of complex numbers, with $\Sigma|z_j|^2 = 1$. Denote the equivalence class of $(z_0,\dots,z_n)$ by $(z_0:z_1:\dots:z_n)$.

Define a real valued function f on $\mathbf{C}P_n$ by the identity

$$f(z_0:z_1:\dots:z_n) = \sum c_j|z_j|^2$$

where $c_0,c_1,\dots,c_n$ are distinct real constants.

In order to determine the critical points of f, consider the following local coordinate system. Let U_0 be the set of $(z_0:z_1:\dots:z_n)$ with $z_0 \neq 0$, and set

$$|z_0|\,\frac{z_j}{z_0} = x_j + iy_j \ .$$

Then

$$x_1,y_1,\dots,x_n,y_n\colon\ U_0 \to \mathbf{R}$$

are the required coordinate functions, mapping U_0 diffeomorphically onto the open unit ball in $\mathbf{R}^{2n}$. Clearly

$$|z_j|^2 = x_j^2 + y_j^2 \qquad |z_0|^2 = 1 - \Sigma\left(x_j^2 + y_j^2\right)$$

so that

$$f = c_0 + \sum_{j=1}^{n}\left(c_j - c_0\right)\left(x_j^2 + y_j^2\right)$$

throughout the coordinate neighborhood U_0. Thus the only critical point of f within U_0 lies at the center point

$$p_0 = (1:0:0:\dots:0)$$

of the coordinate system. At this point f is non-degenerate; and has index equal to twice the number of j with $c_j < c_0$.

Similarly one can consider other coordinate systems centered at the points

$$p_1 = (0:1:0:\dots:0),\dots,p_n = (0:0:\dots:0:1) \ .$$

It follows that $p_0, p_1, \ldots, p_n$ are the only critical points of f. The index of f at p_k is equal to twice the number of j with $c_j < c_k$. Thus every possible even index between 0 and $2n$ occurs exactly once. By Theorem 3.5:

$\mathbf{C}P_n$ has the homotopy type of a CW-complex of the form

$$e^0 \cup e^2 \cup e^4 \cup \ldots \cup e^{2n} .$$

It follows that the integral homology groups of $\mathbf{C}P_n$ are given by

$$H_i(\mathbf{C}P_n; \mathbf{Z}) = \begin{cases} \mathbf{Z} & \text{for } i = 0,2,4,\ldots,2n \\ 0 & \text{for other values of } i . \end{cases}$$

§5. The Morse Inequalities.

In Morse's original treatment of this subject, Theorem 3.5 was not available. The relationship between the topology of M and the critical points of a real valued function on M were described instead in terms of a collection of inequalities. This section will describe this original point of view.

DEFINITION: Let S be a function from certain pairs of spaces to the integers. S is subadditive if whenever $X \supset Y \supset Z$ we have $S(X,Z) \leq S(X,Y) + S(Y,Z)$. If equality holds, S is called additive.

As an example, given any field F as coefficient group, let

$$R_\lambda(X,Y) = \lambda\text{th Betti number of } (X,Y)$$
$$= \text{rank over } F \text{ of } H_\lambda(X,Y;F) \ ,$$

for any pair (X,Y) such that this rank is finite. R_λ is subadditive, as is easily seen by examining the following portion of the exact sequence for (X,Y,Z):

$$\ldots \to H_\lambda(Y,Z) \to H_\lambda(X,Z) \to H_\lambda(X,Y) \to \ldots$$

The Euler characteristic $\chi(X,Y)$ is additive, where $\chi(X,Y) = \Sigma\,(-1)^\lambda R_\lambda(X,Y)$.

LEMMA 5.1. Let S be subadditive and let $X_0 \subset \ldots \subset X_n$. Then $S(X_n,X_0) \leq \sum_{i=1}^{n} S(X_i,X_{i-1})$. If S is additive then equality holds.

PROOF: Induction on n. For $n = 1$, equality holds trivially and the case $n = 2$ is the definition of [sub] additivity.

If the result is true for $n - 1$, then $S(X_{n-1},X_0) \leq \sum_{1}^{n-1} S(X_i,X_{i-1})$. Therefore $S(X_n,X_0) \leq S(X_{n-1},X_0) + S(X_n,X_{n-1}) \leq \sum_{1}^{n} S(X_i,X_{i-1})$ and the result is true for n.

Let $S(X,\emptyset) = S(X)$. Taking $X_0 = \emptyset$ in Lemma 5.1, we have

$$S(X_n) \leq \sum_{1}^{n} S(X_i,X_{i-1}) \tag{1}$$

with equality if S is additive.

Let M be a compact manifold and f a differentiable function on M with isolated, non-degenerate, critical points. Let $a_1 < \ldots < a_k$ be such that M^{a_i} contains exactly i critical points, and $M^{a_k} = M$. Then

$$H_*(M^{a_i}, M^{a_{i-1}}) = H_*(M^{a_{i-1}} \cup e^{\lambda_i}, M^{a_{i-1}})$$

where λ_i is the index of the critical point,

$$= H_*(e^{\lambda_i}, \dot{e}^{\lambda_i}) \quad \text{by excision,}$$

$$= \begin{cases} \text{coefficient group in dimension } \lambda_i \\ 0 \quad \text{otherwise.} \end{cases}$$

Applying (1) to $\emptyset = M^{a_0} \subset \ldots \subset M^{a_k} = M$ with $S = R_\lambda$ we have

$$R_\lambda(M) \leq \sum_{i=1}^{k} R_\lambda(M^{a_i}, M^{a_{i-1}}) = C_\lambda;$$

where C_λ denotes the number of critical points of index λ. Applying this formula to the case $S = \chi$ we have

$$\chi(M) = \sum_{i=1}^{k} \chi(M^{a_i}, M^{a_{i-1}}) = C_0 - C_1 + C_2 - + \ldots \pm C_n \ .$$

Thus we have proven:

THEOREM 5.2 (Weak Morse Inequalities). If C_λ denotes the number of critical points of index λ on the compact manifold M then

$$R_\lambda(M) \leq C_\lambda \ , \quad \text{and} \tag{2}$$

$$\Sigma\, (-1)^\lambda R_\lambda(M) = \Sigma\, (-1)^\lambda C_\lambda \ . \tag{3}$$

Slightly sharper inequalities can be proven by the following argument.

LEMMA 5.3. The function S_λ is subadditive, where

$$S_\lambda(X,Y) = R_\lambda(X,Y) - R_{\lambda-1}(X,Y) + R_{\lambda-2}(X,Y) - + \ldots \pm R_0(X,Y) \ .$$

PROOF: Given an exact sequence

$$\xrightarrow{h} A \xrightarrow{i} B \xrightarrow{j} C \xrightarrow{k} \ldots \ \ldots \rightarrow D \rightarrow 0$$

of vector spaces note that the rank of the homomorphism h plus the rank of i is equal to the rank of A. Therefore,

$$\begin{aligned} \text{rank } h &= \text{rank } A - \text{rank } i \\ &= \text{rank } A - \text{rank } B + \text{rank } j \\ &= \text{rank } A - \text{rank } B + \text{rank } C - \text{rank } k \\ &\quad \cdots \\ &= \text{rank } A - \text{rank } B + \text{rank } C - + \ldots \pm \text{rank } D \ . \end{aligned}$$

Hence the last expression is ≥ 0. Now consider the homology exact sequence of a triple $X \supset Y \supset Z$. Applying this computation to the homomorphism

$$H_{\lambda+1}(X,Y) \xrightarrow{\partial} H_\lambda(Y,Z)$$

we see that

$$\text{rank } \partial = R_\lambda(Y,Z) - R_\lambda(X,Z) + R_\lambda(X,Y) - R_{\lambda-1}(Y,Z) + \ldots \geq 0$$

Collecting terms, this means that

$$S_\lambda(Y,Z) - S_\lambda(X,Z) + S_\lambda(X,Y) \geq 0 \ ,$$

which completes the proof.

Applying this subadditive function S_λ to the spaces

$$\emptyset \subset M^{a_1} \subset M^{a_2} \subset \ldots \subset M^{a_k}$$

we obtain the Morse inequalities:

$$S_\lambda(M) \leq \sum_{i=1}^{k} S_\lambda(M^{a_i}, M^{a_{i-1}}) = C_\lambda - C_{\lambda-1} + - \ldots \pm C_0$$

or

$$(4_\lambda) \qquad R_\lambda(M) - R_{\lambda-1}(M) + - \ldots \pm R_0(M) \leq C_\lambda - C_{\lambda-1} + - \ldots \pm C_0 .$$

These inequalities are definitely sharper than the previous ones. In fact, adding (4_λ) and $(4_{\lambda-1})$, one obtains (2_λ); and comparing (4_λ) with $(4_{\lambda-1})$ for $\lambda > n$ one obtains the equality (3).

As an illustration of the use of the Morse inequalities, suppose that $C_{\lambda+1} = 0$. Then $R_{\lambda+1}$ must also be zero. Comparing the inequalities (4_λ) and $(4_{\lambda+1})$, we see that

$$R_\lambda - R_{\lambda-1} + - \ldots \pm R_0 = C_\lambda - C_{\lambda-1} + - \ldots \pm C_0 \ .$$

Now suppose that $C_{\lambda-1}$ is also zero. Then $R_{\lambda-1} = 0$, and a similar argument shows that

$$R_{\lambda-2} - R_{\lambda-3} + - \ldots \pm R_0 = C_{\lambda-2} - C_{\lambda-3} + - \ldots \pm C_0 \ .$$

Subtracting this from the equality above we obtain the following:

COROLLARY 5.4. If $C_{\lambda+1} = C_{\lambda-1} = 0$ then $R_\lambda = C_\lambda$ and $R_{\lambda+1} = R_{\lambda-1} = 0$.

(Of course this would also follow from Theorem 3.5.) Note that this corollary enables us to find the homology groups of complex projective space (see §4) without making use of Theorem 3.5.

§6. Manifolds in Euclidean Space.

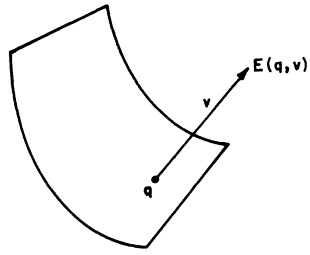
Although we have so far considered, on a manifold, only functions which have no degenerate critical points, we have not yet even shown that such functions always exist. In this section we will construct many functions with no degenerate critical points, on any manifold embedded in $\mathbf{R}^n$. In fact, if for fixed $p \in \mathbf{R}^n$ define the function $L_p\colon M \to \mathbf{R}$ by $L_p(q) = \|p-q\|^2$. It will turn out that for almost all p, the function L_p has only non-degenerate critical points.

Let $M \subset \mathbf{R}^n$ be a manifold of dimension $k < n$, differentiably embedded in $\mathbf{R}^n$. Let $N \subset M \times \mathbf{R}^n$ be defined by

$$N = \{(q,v)\colon q \in M,\ v \text{ perpendicular to } M \text{ at } q\}.$$

It is not difficult to show that N is an n-dimensional manifold differentiably embedded in $\mathbf{R}^{2n}$. (N is the total space of the normal vector bundle of M.)

Let $E\colon N \to \mathbf{R}^n$ be $E(q,v) = q + v$. (E is the "endpoint" map.)



DEFINITION. $e \in \mathbf{R}^n$ is a <u>focal point of</u> (M,q) <u>with multiplicity</u> μ if $e = q + v$ where $(q,v) \in N$ and the Jacobian of E at (q,v) has nullity $\mu > 0$. The point e will be called a <u>focal point</u> of M if e is a focal point of (M,q) for some $q \in M$.

Intuitively, a focal point of M is a point in $\mathbf{R}^n$ where nearby normals intersect.

We will use the following theorem, which we will not prove.

THEOREM 6.1 (Sard). If M_1 and M_2 are differentiable manifolds having a countable basis, of the same dimension, and $f: M_1 \to M_2$ is of class C^1, then the image of the set of critical points has measure 0 in M_2.

A critical point of f is a point where the Jacobian of f is singular. For a proof see de Rham, "Variétés Différentiables," Hermann, Paris, 1955, p. 10.

COROLLARY 6.2. For almost all $x \in \mathbf{R}^n$, the point x is not a focal point of M.

PROOF: We have just seen that N is an n-manifold. The point x is a focal point iff x is in the image of the set of critical points of $E: N \to \mathbf{R}^n$. Therefore the set of focal points has measure 0.

For a better understanding of the concept of focal point, it is convenient to introduce the "second fundamental form" of a manifold in Euclidean space. We will not attempt to give an invariant definition; but will make use of a fixed local coordinate system.

Let $u^1,\ldots,u^k$ be coordinates for a region of the manifold $M \subset \mathbf{R}^n$. Then the inclusion map from M to $\mathbf{R}^n$ determines n smooth functions

$$x_1(u^1,\ldots,u^k),\ldots,x_n(u^1,\ldots,u^k) \quad .$$

These functions will be written briefly as $\vec{x}(u^1,\ldots,u^k)$ where $\vec{x} = (x_1,\ldots,x_n)$. To be consistent the point $q \in M \subset \mathbf{R}^n$ will now be denoted by $\vec{q}$.

The first fundamental form associated with the coordinate system is defined to be the symmetric matrix of real valued functions

$$\left(g_{ij} \right) = \left(\frac{\partial \vec{x}}{\partial u_i} \cdot \frac{\partial \vec{x}}{\partial u_j} \right) \quad .$$

The second fundamental form on the other hand, is a symmetric matrix $\left(\vec{\ell}_{ij}\right)$ of vector valued functions.

It is defined as follows. The vector $\frac{\partial^2 \vec{x}}{\partial u^i \partial u^j}$ at a point of M can be expressed as the sum of a vector tangent to M and a vector normal to M. Define $\vec{\ell}_{ij}$ to be the normal component of $\frac{\partial^2 \vec{x}}{\partial u^i \partial u^j}$. Given any unit vector $\vec{v}$ which is normal to M at $\vec{q}$ the matrix

$$\left(\vec{v} \cdot \frac{\partial^2 \vec{x}}{\partial u^i \partial u^j} \right) = \left(\vec{v} \cdot \vec{\ell}_{ij} \right)$$

can be called the "second fundamental form of M at $\vec{q}$ in the direction $\vec{v}$."

It will simplify the discussion to assume that the coordinates have been chosen to that g_{ij}, evaluated at $\vec{q}$, is the identity matrix. Then the eigenvalues of the matrix $(\vec{v} \cdot \vec{\ell}_{ij})$ are called the <u>principal curvatures</u> $K_1,\dots,K_k$ of M at $\vec{q}$ in the normal direction $\vec{v}$. The reciprocals $K_1^{-1},\dots,K_k^{-1}$ of these principal curvatures are called the <u>principal radii of curvature</u>. Of course it may happen that the matrix $(\vec{v} \cdot \vec{\ell}_{ij})$ is singular. In this case one or more of the K_i will be zero; and hence the corresponding radii K_i^{-1} will not be defined.

Now consider the normal line ℓ consisting of all $\vec{q} + t\vec{v}$, where $\vec{v}$ is a fixed unit vector orthogonal to M at $\vec{q}$.

LEMMA 6.3. The focal points of $(M,\vec{q})$ along ℓ are precisely the points $\vec{q} + K_i^{-1}\vec{v}$, where $1 \leq i \leq k$, $K_i \neq 0$. Thus there are at most, k focal points of $(M,\vec{q})$ along ℓ, each being counted with its proper multiplicity.

PROOF: Choose $n-k$ vector fields $\vec{w}_1(u^1,\dots,u^k),\dots,\vec{w}_{n-k}(u^1,\dots,u^k)$ along the manifold so that $\vec{w}_1,\dots,\vec{w}_{n-k}$ are unit vectors which are orthogonal to each other and to M. We can introduce coordinates $(u^1,\dots,u^k, t^1,\dots,t^{n-k})$ on the manifold $N \subset M \times \mathbf{R}^n$ as follows. Let $(u^1,\dots,u^k,t^1,\dots t^{n-k})$ correspond to the point

$$\left(\vec{x}(u^1,\dots,u^k),\ \sum_{\alpha=1}^{n-k} t^\alpha \vec{w}_\alpha(u^1,\dots,u^k)\right) \in N \ .$$

Then the function

$$E\colon N \to \mathbf{R}^n$$

gives rise to the correspondence

$$(u^1,\dots,u^k,t^1,\dots,t^{n-k}) \xrightarrow{\vec{e}} \vec{x}(u^1,\dots,u^k) + \sum t^\alpha \vec{w}_\alpha(u^1,\dots,u^k) \ ,$$

with partial derivatives

$$\begin{cases} \dfrac{\partial \vec{e}}{\partial u^i} = \dfrac{\partial \vec{x}}{\partial u^i} + \displaystyle\sum_\alpha t^\alpha \frac{\partial \vec{w}_\alpha}{\partial u^i} \\[2ex] \dfrac{\partial \vec{e}}{\partial t^\beta} = \vec{w}_\beta \ . \end{cases}$$

Taking the inner products of these n-vectors with the linearly independent

vectors $\frac{\partial \vec{x}}{\partial u^1}, \ldots, \frac{\partial \vec{x}}{\partial u^k}, \vec{w}_1, \ldots, \vec{w}_{n-k}$ we will obtain an $n \times n$ matrix whose rank equals the rank of the Jacobian of E at the corresponding point.

This $n \times n$ matrix clearly has the following form

$$\begin{pmatrix} \left(\frac{\partial \vec{x}}{\partial u^i} \cdot \frac{\partial \vec{x}}{\partial u^j} + \sum_\alpha t^\alpha \frac{\partial \vec{w}_\alpha}{\partial u^i} \cdot \frac{\partial \vec{x}}{\partial u^j} \right) & \left(\sum_\alpha t^\alpha \frac{\partial \vec{w}_\alpha}{\partial u^i} \cdot \vec{w}_\beta \right) \\ \bigcirc & \text{identity matrix} \end{pmatrix}$$

Thus the nullity is equal to the nullity of the upper left hand block. Using the identity

$$0 = \frac{\partial}{\partial u^i} \left(\vec{w}_\alpha \cdot \frac{\partial \vec{x}}{\partial u^j} \right) = \frac{\partial \vec{w}_\alpha}{\partial u^i} \cdot \frac{\partial \vec{x}}{\partial u^j} + \vec{w}_\alpha \cdot \frac{\partial^2 \vec{x}}{\partial u^i \partial u^j}$$

we see that this upper left hand block is just the matrix

$$\left(g_{ij} - \sum_\alpha t^\alpha \vec{w}_\alpha \cdot \mathcal{U}_{ij} \right)$$

Thus:

ASSERTION 6.4. $\vec{q} + t\vec{v}$ is a focal point of $(M, \vec{q})$ with multiplicity μ if and only if the matrix

(*) $$(g_{ij} - t\vec{v} \cdot \mathcal{U}_{ij})$$

is singular, with nullity μ.

Now suppose that (g_{ij}) is the identity matrix. Then (*) is singular if and only if $\frac{1}{t}$ is an eigenvalue of the matrix $(\vec{v} \cdot \mathcal{U}_{ij})$. Furthermore the multiplicity μ is equal to the multiplicity of $\frac{1}{t}$ as eigenvalue. This completes the proof of Lemma 6.3.

Now for fixed $\vec{p} \in \mathbf{R}^n$ let us study the function

$$L_{\vec{p}} = f: M \to \mathbf{R}$$

where

$$f\left(\vec{x}(u^1, \ldots, u^k)\right) = \|\vec{x}(u^1, \ldots, u^k) - \vec{p}\|^2 = \vec{x} \cdot \vec{x} - 2\vec{x} \cdot \vec{p} + \vec{p} \cdot \vec{p} .$$

We have

$$\frac{\partial f}{\partial u^i} = 2 \frac{\partial \vec{x}}{\partial u^i} \cdot (\vec{x} - \vec{p}) .$$

Thus f has a critical point at $\vec{q}$ if and only if $\vec{q} - \vec{p}$ is normal to M at $\vec{q}$.

The second partial derivatives at a critical point are given by

$$\frac{\partial^2 f}{\partial u^i \partial u^j} = 2\left(\frac{\partial \vec{x}}{\partial u^i} \cdot \frac{\partial \vec{x}}{\partial u^j} + \frac{\partial^2 \vec{x}}{\partial u^i \partial u^j} \cdot (\vec{x} - \vec{p}) \right) .$$

Setting $\vec{p} = \vec{x} + t\vec{v}$, as in the proof of Lemma 6.3, this becomes

$$\frac{\partial^2 f}{\partial u^i \partial u^j} = 2(g_{ij} - t\vec{v} \cdot \vec{\gamma}_{ij}) .$$

Therefore:

LEMMA 6.5. The point $\vec{q} \in M$ is a degenerate critical point of $f = L_{\vec{p}}$ if and only if $\vec{p}$ is a focal point of $(M,\vec{q})$. The nullity of $\vec{q}$ as critical point is equal to the multiplicity of $\vec{p}$ as focal point.

Combining this result with Corollary 6.2 to Sard's theorem, we immediately obtain:

THEOREM 6.6. For almost all $p \in \mathbf{R}^n$ (all but a set of measure 0) the function

$$L_p : M \to \mathbf{R}$$

has no degenerate critical points.

This theorem has several interesting consequences.

COROLLARY 6.7. On any manifold M there exists a differentiable function, with no degenerate critical points, for which each M^a is compact.

PROOF: This follows from Theorem 6.6 and the fact that an n-dimensional manifold M can be embedded differentiably as a closed subset of $\mathbf{R}^{2n+1}$ (see Whitney, Geometric Integration Theory, p. 113).

APPLICATION 1. A differentiable manifold has the homotopy type of a CW-complex. This follows from the above corollary and Theorem 3.5.

APPLICATION 2. On a compact manifold M there is a vector field X such that the sum of the indices of the critical points of X equals $\chi(M)$, the Euler characteristic of M. This can be seen as follows: for any differentiable function f on M we have $\chi(M) = \Sigma (-1)^\lambda C_\lambda$ where C_λ is the number of critical points with index λ. But $(-1)^\lambda$ is the index of the vector field grad f at a point where f has index λ.

It follows that the sum of the indices of any vector field on M is equal to $\chi(M)$ because this sum is a topological invariant (see Steenrod, "The Topology of Fibre Bundles," §39.7).

The preceding corollary can be sharpened as follows. Let $k \geq 0$ be an integer and let $K \subset M$ be a compact set.

> COROLLARY 6.8. Any bounded smooth function $f: M \to \mathbf{R}$ can be uniformly approximated by a smooth function g which has no degenerate critical points. Furthermore g can be chosen so that the i-th derivatives of g on the compact set K uniformly approximate the corresponding derivatives of f, for $i \leq k$.

(Compare M. Morse, The critical points of a function of n variables, Transactions of the American Mathematical Society, Vol. 33 (1931), pp. 71-91.)

PROOF: Choose some imbedding $h: M \to \mathbf{R}^n$ of M as a bounded subset of some euclidean space so that the first coordinate h_1 is precisely the given function f. Let c be a large number. Choose a point

$$p = (-c+\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$$

close to $(-c,0,\ldots,0) \in \mathbf{R}^n$ so that the function $L_p: M \to \mathbf{R}$ is non-degenerate; and set

$$g(x) = \frac{L_p(x) - c^2}{2c} .$$

Clearly g is non-degenerate. A short computation shows that

$$g(x) = f(x) + \sum_1^n h_i(x)^2/2c - \sum_1^n \varepsilon_i h_i(x)/c + \sum_1^n \varepsilon_i^2/2c - \varepsilon_1 .$$

Clearly, if c is large and the ε_i are small, then g will approximate f as required.

The above theory can also be used to describe the index of the function

$$L_p: M \to \mathbf{R}$$

at a critical point.

> LEMMA 6.9. (Index theorem for L_p.) The index of L_p at a non-degenerate critical point $q \in M$ is equal to the number of focal points of (M,q) which lie on the segment from q to p; each focal point being counted with its multiplicity.

An analogous statement in Part III (the Morse Index Theorem) will be of fundamental importance.

PROOF: The index of the matrix

$$\left(\frac{\partial^2 L_p}{\partial u^i \partial u^j}\right) = 2(g_{ij} - t\vec{v} \cdot \vec{\Gamma}_{ij})$$

is equal to the number of negative eigenvalues. Assuming that (g_{ij}) is the identity matrix, this is equal to the number of eigenvalues of $(\vec{v} \cdot \vec{\Gamma}_{ij})$ which are $\geq \frac{1}{t}$. Comparing this statement with 6.3, the conclusion follows.

§7. The Lefschetz Theorem on Hyperplane Sections.

As an application of the ideas which have been developed, we will prove some results concerning the topology of algebraic varieties. These were originally proved by Lefschetz, using quite different arguments. The present version is due to Andreotti and Frankel*.

THEOREM 7.1. If $M \subset \mathbf{C}^n$ is a non-singular affine algebraic variety in complex n-space with real dimension $2k$, then

$$H_i(M;\mathbf{Z}) = 0 \quad \text{for} \quad i > k.$$

This is a consequence of the stronger:

THEOREM 7.2. A complex analytic manifold M of complex dimension k, bianalytically embedded as a closed subset of $\mathbf{C}^n$ has the homotopy type of a k-dimensional CW-complex.

The proof will be broken up into several steps. First consider a quadratic form in k complex variables

$$Q(z^1,\ldots,z^k) = \sum b_{hj}\, z^h z^j \quad .$$

If we substitute $x^h + iy^h$ for z^h, and then take the real part of Q we obtain a real quadratic form in $2k$ real variables:

$$Q'(x^1,\ldots,x^k,y^1,\ldots,y^k) = \text{real part of } \sum b_{hj}(x^h+iy^h)(x^j+iy^j) \quad .$$

ASSERTION 1. If e is an eigenvalue of Q' with multiplicity μ, then $-e$ is also an eigenvalue with the same multiplicity μ.

PROOF. The identity $Q(iz^1,\ldots,iz^k) = -Q(z^1,\ldots,z^k)$ shows that the quadratic form Q' can be transformed into $-Q'$ by an orthogonal change of variables. Assertion 1 clearly follows.

---

* See S. Lefschetz, "L'analysis situs et la géométrie algébrique," Paris, 1924; and A. Andreotti and T. Frankel, The Lefschetz theorem on hyperplane sections, Annals of Mathematics, Vol. 69 (1959), pp. 713-717.

Now consider a complex manifold M which is bianalytically imbedded as a subset of $\mathbf{C}^n$. Let q be a point of M.

ASSERTION 2. The focal points of (M,q) along any normal line ℓ occur in pairs which are situated symmetrically about q.

In other words if $q + tv$ is a focal point, then $q - tv$ is a focal point with the same multiplicity.

PROOF. Choose complex coordinates $z^1,\dots,z^k$ for M in a neighborhood of q so that $z^1(q) = \dots = z^k(q) = 0$. The inclusion map $M \to \mathbf{C}^n$ determines n complex analytic functions

$$w_\alpha = w_\alpha(z^1,\dots,z^k), \qquad \alpha = 1,\dots,n.$$

Let v be a fixed unit vector which is orthogonal to M at q. Consider the Hermitian inner product

$$\sum w_\alpha \bar{v}_\alpha = \sum w_\alpha(z^1,\dots,z^k)\bar{v}_\alpha$$

of w and v. This can be expanded as a complex power series

$$\sum w_\alpha(z^1,\dots,z^k)\bar{v}_\alpha = \text{constant} + Q(z^1,\dots,z^k) + \text{higher terms},$$

where Q denotes a homogeneous quadratic function. (The linear terms vanish since v is orthogonal to M.)

Now substitute $x^h + iy^h$ for z^h so as to obtain a real coordinate system for M; and consider the real inner product

$$w \cdot v = \text{real part of } \sum w_\alpha \bar{v}_\alpha \ .$$

This function has the real power series expansion

$$w \cdot v = \text{constant} + Q'(x^1,\dots,x^k,y^1,\dots,y^k) + \text{higher terms}.$$

Clearly the quadratic terms Q' determine the second fundamental form of M at q in the normal direction v. By Assertion 1 the eigenvalues of Q' occur in equal and opposite pairs. Hence the focal points of (M,q) along the line through q and $q + v$ also occur in symmetric pairs. This proves Assertion 2.

We are now ready to prove 7.2. Choose a point $p \in \mathbf{C}^n$ so that the squared-distance function

$$L_p : M \to \mathbf{R}$$

has no degenerate critical points. Since M is a closed subset of $\mathbf{C}^n$, it is clear that each set

$$M^a = L_p^{-1}[0,a]$$

is compact. Now consider the index of L_p at a critical point q. According to 6.9, this index is equal to the number of focal points of (M,q) which lie on the line segment from p to q. But there are at most $2k$ focal points along the full line through p and q; and these are distributed symmetrically about q. Hence at most k of them can lie between p and q.

Thus the index of L_p at q is $\leq k$. It follows that M has the homotopy type of a CW-complex of dimension $\leq k$; which completes the proof of 7.2.

COROLLARY 7.3 (Lefschetz). Let V be an algebraic variety of complex dimension k which lies in the complex projective space $\mathbf{C}P_n$. Let P be a hyperplane in $\mathbf{C}P_n$ which contains the singular points (if any) of V. Then the inclusion map

$$V \cap P \to V$$

induces isomorphisms of homology groups in dimensions less than $k-1$. Furthermore, the induced homomorphism

$$H_{k-1}(V \cap P;\mathbf{Z}) \to H_{k-1}(V;\mathbf{Z})$$

is onto.

PROOF. Using the exact sequence of the pair $(V,V \cap P)$ it is clearly sufficient to show that $H_r(V,V \cap P;\mathbf{Z}) = 0$ for $r \leq k-1$. But the Lefschetz duality theorem asserts that

$$H_r(V,V \cap P;\mathbf{Z}) \cong H^{2k-r}(V - (V \cap P);\mathbf{Z}) \quad .$$

But $V - (V \cap P)$ is a non-singular algebraic variety in the affine space $\mathbf{C}P_n - P$. Hence it follows from 7.2 that the last group is zero for $r \leq k-1$.

This result can be sharpened as follows:

THEOREM 7.4 (Lefschetz). Under the hypothesis of the preceding corollary, the relative homotopy group $\pi_r(V,V \cap P)$ is zero for $r < k$.

PROOF. The proof will be based on the hypothesis that some neighborhood U of $V \cap P$ can be deformed into $V \cap P$ within V. This can be proved, for example, using the theorem that algebraic varieties can be triangulated.

In place of the function $L_p: V - V \cap P \to \mathbf{R}$ we will use $f: V \to \mathbf{R}$ where

$$f(x) = \begin{cases} 0 & \text{for } x \in V \cap P \ , \\ 1/L_p(x) & \text{for } x \notin P. \end{cases}$$

Since the critical points of L_p have index $\leq k$ it follows that the critical points of f have index $\geq 2k - k = k$. The function f has no degenerate critical points with $\varepsilon \leq f < \infty$. Therefore V has the homotopy type of $V^\varepsilon = f^{-1}[0,\varepsilon]$ with finitely many cells of dimension $\geq k$ attached.

Choose ε small enough so that $V^\varepsilon \subset U$. Let I^r denote the unit r-cube. Then every map of the pair $(I^r, \dot{I}^r)$ into $(V, V \cap P)$ can be deformed into a map

$$(I^r, \dot{I}^r) \to (V^\varepsilon, V \cap P) \subset (U, V \cap P) \ ,$$

since $r < k$, and hence can be deformed into $V \cap P$. This completes the proof.

PART II

A RAPID COURSE IN RIEMANNIAN GEOMETRY

§8. Covariant Differentiation

The object of Part II will be to give a rapid outline of some basic concepts of Riemannian geometry which will be needed later. For more information the reader should consult Nomizu, "Lie groups and differential geometry. Math. Soc. Japan, 1956; Helgason, "Differential geometry and symmetric spaces," Academic Press, 1962; Sternberg, "Lectures on differential geometry," Prentice-Hall, 1964; or Laugwitz, "Differential and Riemannian geometry," Academic Press, 1965.

Let M be a smooth manifold.

DEFINITION. An affine connection at a point $p \in M$ is a function which assigns to each tangent vector $X_p \in TM_p$ and to each vector field Y a new tangent vector

$$X_p \vdash Y \in TM_p$$

called the covariant derivative* of Y in the direction X_p. This is required to be bilinear as a function of X_p and Y. Furthermore, if

$$f: M \to \mathbf{R}$$

is a real valued function, and if fY denotes the vector field

$$(fY)_q = f(q)Y_q$$

then $\vdash$ is required to satisfy the identity

$$X_p \vdash (fY) = (X_p f)Y_p + f(p)X_p \vdash Y \ .$$

---

* Note that our $X \vdash Y$ coincides with Nomizu's $\nabla_X Y$. The notation is intended to suggest that the differential operator X acts on the vector field Y.

(As usual, $X_p f$ denotes the directional derivative of f in the direction X_p.)

A global affine connection (or briefly a connection) on M is a function which assigns to each $p \in M$ an affine connection $\vdash_p$ at p, satisfying the following smoothness condition.

1) If X and Y are smooth vector fields on M then the vector field $X \vdash Y$, defined by the identity

$$(X \vdash Y)_p = X_p \vdash_p Y \quad ,$$

must also be smooth.

Note that:

(2) $X \vdash Y$ is bilinear as a function of X and Y.

(3) $(fX) \vdash Y = f(X \vdash Y)$,

(4) $(X \vdash (fY) = (Xf)Y + f(X \vdash Y)$.

Conditions (1), (2), (3), (4) can be taken as the definition of a connection.

In terms of local coordinates $u^1,\ldots,u^n$ defined on a coordinate neighborhood $U \subset M$, the connection $\vdash$ is determined by n^3 smooth real valued functions Γ_{ij}^k on U, as follows. Let ∂_k denote the vector field $\frac{\partial}{\partial u^k}$ on U. Then any vector field X on U can be expressed uniquely as

$$X = \sum_{k=1} x^k \partial_k$$

where the x^k are real valued functions on U. In particular the vector field $\partial_i \vdash \partial_j$ can be expressed as

$$\partial_i \vdash \partial_j = \sum_k \Gamma_{ij}^k \partial_k \tag{5}$$

These functions Γ_{ij}^k determine the connection completely on U. In fact given vector fields $X = \sum x^i \partial_i$ and $Y = \sum y^j \partial_j$ one can expand $X \vdash Y$ by the rules (2), (3), (4); yielding the formula

$$X \vdash Y = \sum_k \Big(\sum_i x^i y^k_{,i} \Big) \partial_k \tag{6}$$

where the symbol $y^k_{,i}$ stands for the real valued function

$$y^k_{,i} = \partial_i y^k + \sum_j \Gamma^k_{ij}\, y^j \quad .$$

Conversely, given any smooth real valued functions Γ^k_{ij} on U, one can define $X \vdash Y$ by the formula (6). The result clearly satisfies the conditions (1), (2), (3), (4), (5).

Using the connection $\vdash$ one can define the covariant derivative of a vector field along a curve in M. First some definitions.

A <u>parametrized curve</u> in M is a smooth function c from the real numbers to M. A <u>vector field</u> V <u>along</u> the curve c is a function which assigns to each $t \in \mathbf{R}$ a tangent vector

$$V_t \in TM_{c(t)} \quad .$$

This is required to be smooth in the following sense: For any smooth function f on M the correspondence

$$t \to V_t f$$

should define a smooth function on $\mathbf{R}$.

As an example the <u>velocity</u> vector field $\frac{dc}{dt}$ of the curve is the vector field along c which is defined by the rule

$$\frac{dc}{dt} = c_* \frac{d}{dt} \quad .$$

Here $\frac{d}{dt}$ denotes the standard vector field on the real numbers, and

$$c_*: \; T\mathbf{R}_t \to TM_{c(t)}$$

denotes the homomorphism of tangent spaces induced by the map c. (Compare Diagram 9.)

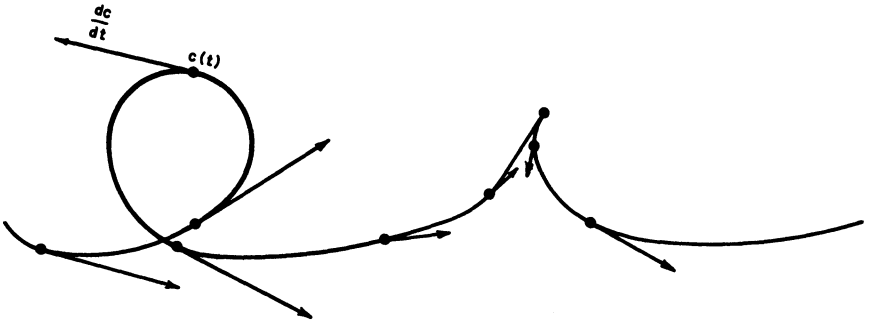


Diagram 9

Now suppose that M is provided with an affine connection. Then any vector field V along c determines a new vector field $\frac{DV}{dt}$ along c called the covariant derivative of V. The operation

$$V \to \frac{DV}{dt}$$

is characterized by the following three axioms.

a) $\frac{D(V+W)}{dt} = \frac{DV}{dt} + \frac{DW}{dt}$.

b) If f is a smooth real valued function on $\mathbf{R}$ then

$$\frac{D(fV)}{dt} = \frac{df}{dt} V + f \frac{DV}{dt} .$$

c) If V is induced by a vector field Y on M, that is if $V_t = Y_{c(t)}$ for each t, then $\frac{DV}{dt}$ is equal to $\frac{dc}{dt} \vdash Y$ (= the covariant derivative of Y in the direction of the velocity vector of c)

LEMMA 8.1. There is one and only one operation $V \to \frac{DV}{dt}$ which satisfies these three conditions.

PROOF: Choose a local coordinate system for M, and let $u^1(t), \ldots, u^n(t)$ denote the coordinates of the point $c(t)$. The vector field V can be expressed uniquely in the form

$$V = \sum v^j \partial_j$$

where $v^1, \ldots, v^n$ are real valued functions on $\mathbf{R}$ (or an appropriate open subset of $\mathbf{R}$), and $\partial_1, \ldots, \partial_n$ are the standard vector fields on the coordinate neighborhood. It follows from (a), (b), and (c) that

$$\frac{DV}{dt} = \sum_j \left(\frac{dv^j}{dt} \partial_j + v^j \frac{dc}{dt} \vdash \partial_j \right)$$

$$= \sum_k \left(\frac{dv^k}{dt} + \sum_{i,j} \frac{du^i}{dt} \Gamma_{ij}^k v^j \right) \partial_k .$$

Conversely, defining $\frac{DV}{dt}$ by this formula, it is not difficult to verify that conditions (a), (b), and (c) are satisfied.

A vector field V along c is said to be a parallel vector field if the covariant derivative $\frac{DV}{dt}$ is identically zero.

LEMMA 8.2. Given a curve c and a tangent vector V_0 at the point $c(0)$, there is one and only one parallel vector field V along c which extends V_0.

PROOF. The differential equations

$$\frac{dv^k}{dt} + \sum_{i,j} \frac{du^i}{dt}\, \Gamma_{ij}^k\, v^j = 0$$

have solutions $v^k(t)$ which are uniquely determined by the initial values $v^k(0)$. Since these equations are linear, the solutions can be defined for all relevant values of t. (Compare Graves, "The Theory of Functions of Real Variables," p. 152.)

The vector V_t is said to be obtained from V_0 by <u>parallel translation</u> along c.

Now suppose that M is a Riemannian manifold. The inner product of two vectors X_p, Y_p will be denoted by $\langle X_p, Y_p \rangle$.

DEFINITION. A connection $\vdash$ on M is <u>compatible</u> with the Riemannian metric if parallel translation preserves inner products. In other words, for any parametrized curve c and any pair P, P' of parallel vector fields along c, the inner product $\langle P, P' \rangle$ should be constant.

LEMMA 8.3. Suppose that the connection is compatible with the metric. Let V, W be any two vector fields along c. Then

$$\frac{d}{dt}\langle V, W \rangle = \langle \frac{DV}{dt}, W \rangle + \langle V, \frac{DW}{dt} \rangle .$$

PROOF: Choose parallel vector fields $P_1, \ldots, P_n$ along c which are orthonormal at one point of c and hence at every point of c. Then the given fields V and W can be expressed as $\sum v^i P_i$ and $\sum w^j P_j$ respectively (where $v^i = \langle V, P_i \rangle$ is a real valued function on $\mathbf{R}$). It follows that $\langle V, W \rangle = \sum v^i w^i$ and that

$$\frac{DV}{dt} = \sum \frac{dv^i}{dt} P_i, \qquad \frac{DW}{dt} = \sum \frac{dw^j}{dt} P_j .$$

Therefore

$$\langle \frac{DV}{dt}, W \rangle + \langle V, \frac{DW}{dt} \rangle = \sum \left(\frac{dv^i}{dt} w^i + v^i \frac{dw^i}{dt} \right) = \frac{d}{dt}\langle V, W \rangle ,$$

which completes the proof.

COROLLARY 8.4. For any vector fields Y,Y' on M and any vector $X_p \in TM_p$:

$$X_p \langle Y,Y' \rangle = \langle X_p \vdash Y, Y'_p \rangle + \langle Y_p, X_p \vdash Y' \rangle .$$

PROOF. Choose a curve c whose velocity vector at $t = 0$ is X_p; and apply 8.3.

DEFINITION 8.5. A connection $\vdash$ is called symmetric if it satisfies the identity*

$$(X \vdash Y) - (Y \vdash X) = [X,Y] .$$

(As usual, $[X,Y]$ denotes the poison bracket $[X,Y]f = X(Yf) - Y(Xf)$ of two vector fields.) Applying this identity to the case $X = \partial_i$, $Y = \partial_j$, since $[\partial_i,\partial_j] = 0$ one obtains the relation

$$\Gamma_{ij}^k - \Gamma_{ji}^k = 0.$$

Conversely if $\Gamma_{ij}^k = \Gamma_{ji}^k$ then using formula (6) it is not difficult to verify that the connection $\vdash$ is symmetric throughout the coordinate neighborhood.

LEMMA 8.6. (Fundamental lemma of Riemannian geometry.) A Riemannian manifold possesses one and only one symmetric connection which is compatible with its metric.

(Compare Nomizu p. 76, Laugwitz p. 95.)

PROOF of uniqueness. Applying 8.4 to the vector fields $\partial_i,\partial_j,\partial_k$, and setting $\langle \partial_j,\partial_k \rangle = g_{jk}$ one obtains the identity

$$\partial_i g_{jk} = \langle \partial_i \vdash \partial_j, \partial_k \rangle + \langle \partial_j, \partial_i \vdash \partial_k \rangle .$$

Permuting $i,j,$ and k this gives three linear equations relating the

---

* The following reformulation may (or may not) seem more intuitive. Define The "covariant second derivative" of a real valued function f along two vectors X_p,Y_p to be the expression

$$X_p(Yf) - (X_p \vdash Y)f$$

where Y denotes any vector field extending Y_p. It can be verified that this expression does not depend on the choice of Y. (Compare the proof of Lemma 9.1 below.) Then the connection is symmetric if this second derivative is symmetric as a function of X_p and Y_p.

three quantities

$$\langle \partial_i \vdash \partial_j, \partial_k \rangle, \quad \langle \partial_j \vdash \partial_k, \partial_i \rangle, \quad \text{and} \quad \langle \partial_k \vdash \partial_i, \partial_j \rangle \ .$$

(There are only three such quantities since $\partial_i \vdash \partial_j = \partial_j \vdash \partial_i$.) These equations can be solved uniquely; yielding the first Christoffel identity

$$\langle \partial_i \vdash \partial_j, \partial_k \rangle = \tfrac{1}{2} (\partial_i g_{jk} + \partial_j g_{ik} - \partial_k g_{ij}) \ .$$

The left hand side of this identity is equal to $\sum_\ell \Gamma_{ij}^\ell \, g_{\ell k}$. Multiplying by the inverse $(g^{k\ell})$ of the matrix $(g_{\ell k})$ this yields the second Christoffel identity

$$\Gamma_{ij}^\ell = \sum_k \tfrac{1}{2} \left(\partial_i \, g_{jk} + \partial_j \, g_{ik} - \partial_k \, g_{ij} \right) g^{k\ell} \ .$$

Thus the connection is uniquely determined by the metric.

Conversely, defining Γ_{ij}^ℓ by this formula, one can verify that the resulting connection is symmetric and compatible with the metric. This completes the proof.

An alternative characterization of symmetry will be very useful later. Consider a "parametrized surface" in M: that is a smooth function

$$s: \mathbf{R}^2 \to M \ .$$

By a vector field V along s is meant a function which assigns to each $(x,y) \in \mathbf{R}^2$ a tangent vector

$$V_{(x,y)} \in TM_{s(x,y)} \ .$$

As examples, the two standard vector fields $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ give rise to vector fields $s_* \frac{\partial}{\partial x}$ and $s_* \frac{\partial}{\partial y}$ along s. These will be denoted briefly by $\frac{\partial s}{\partial x}$ and $\frac{\partial s}{\partial y}$; and called the "velocity vector fields" of s.

For any smooth vector field V along s the covariant derivatives $\frac{DV}{\partial x}$ and $\frac{DV}{\partial y}$ are new vector fields, constructed as follows. For each fixed y_0, restricting V to the curve

$$x \to s(x, y_0)$$

one obtains a vector field along this curve. Its covariant derivative with respect to x is defined to be $\left(\frac{DV}{\partial x} \right)_{(x,y_0)}$. This defines $\frac{DV}{\partial x}$ along the entire parametrized surface s.

As examples, we can form the two covariant derivatives of the two

vector fields $\frac{\partial s}{\partial x}$ and $\frac{\partial s}{\partial y}$. The derivatives $\frac{D}{\partial x}\frac{\partial s}{\partial x}$ and $\frac{D}{\partial y}\frac{\partial s}{\partial y}$ are simply the acceleration vectors of suitable coordinate curves. However, the mixed derivatives $\frac{D}{\partial x}\frac{\partial s}{\partial y}$ and $\frac{D}{\partial y}\frac{\partial s}{\partial x}$ cannot be described so simply.

LEMMA 8.7. If the connection is symmetric then $\frac{D}{\partial x}\frac{\partial s}{\partial y} = \frac{D}{\partial y}\frac{\partial s}{\partial x}$

PROOF. Express both sides in terms of a local coordinate system, and compute.

§9. The Curvature Tensor

The curvature tensor R of an affine connection $\vdash$ measures the extent to which the second covariant derivative $\partial_i \vdash (\partial_j \vdash Z)$ is symmetric in i and j. Given vector fields X,Y,Z define a new vector field $R(X,Y)Z$ by the identity*

$$R(X,Y)Z = -X \vdash (Y \vdash Z) + Y \vdash (X \vdash Z) + [X,Y] \vdash Z .$$

LEMMA 9.1. The value of $R(X,Y)Z$ at a point $p \in M$ depends only on the vectors X_p, Y_p, Z_p at this point p and not on their values at nearby points. Furthermore the correspondence

$$X_p, Y_p, Z_p \to R(X_p, Y_p)Z_p$$

from $TM_p \times TM_p \times TM_p$ to TM_p is tri-linear.

Briefly, this lemma can be expressed by saying that R is a "tensor."

PROOF: Clearly $R(X,Y)Z$ is a tri-linear function of $X,Y,$ and Z. If X is replaced by a multiple fX then the three terms $-X \vdash (Y \vdash Z)$, $Y \vdash (X \vdash Z)$, $[X,Y] \vdash Z$ are replaced respectively by

i) $-fX \vdash (Y \vdash Z)$,

ii) $(Yf)(X \vdash Z) + fY \vdash (X \vdash Z)$,

iii) $-(Yf)(X \vdash Z) + f[X,Y] \vdash Z$.

Adding these three terms one obtains the identity

$$R(fX,Y)Z = fR(X,Y)Z .$$

Corresponding identities for Y and Z are easily obtained by similar computations.

Now suppose that $X = \sum x^i \partial_i$, $Y = \sum y^i \partial_j$, and $Z = \sum z^k \partial_k$. Then

$$\begin{aligned} R(X,Y)Z &= \sum R(x^i \partial_i, y^j \partial_j)(z^k \partial_k) \\ &= \sum x^i y^j z^k R(\partial_i, \partial_j)\partial_k . \end{aligned}$$

---

* Nomizu gives R the opposite sign. Our sign convention has the advantage that (in the Riemannian case) the inner product $\langle R(\partial_h, \partial_i)\partial_j, \partial_k \rangle$ coincides with the classical symbol R_{hijk}.

Evaluating this expression at p one obtains the formula

$$(R(X,Y)Z)_p = \sum x^i(p)y^j(p)z^k(p)\left(R(\partial_i,\partial_j)\partial_k\right)_p$$

which depends only on the values of the functions x^i, y^j, z^k at p, and not on their values at nearby points. This completes the proof.

Now consider a parametrized surface

$$s: \mathbf{R}^2 \longrightarrow M .$$

Given any vector field V along s, one can apply the two covariant differentiation operators $\frac{D}{\partial x}$ and $\frac{D}{\partial y}$ to V. In general these operators will not commute with each other.

LEMMA 9.2. $\frac{D}{\partial y}\frac{D}{\partial x}V - \frac{D}{\partial x}\frac{D}{\partial y}V = R\left(\frac{\partial s}{\partial x},\frac{\partial s}{\partial y}\right)V$.

PROOF: Express both sides in terms of a local coordinate system, and compute, making use of the identity

$$\partial_j \vdash (\partial_i \vdash \partial_k) - \partial_i \vdash (\partial_j \vdash \partial_k) = R(\partial_i,\partial_j)\partial_k .$$

[It is interesting to ask whether one can construct a vector field P along s which is parallel, in the sense that

$$\frac{D}{\partial x}P = \frac{D}{\partial y}P = 0,$$

and which has a given value $P_{(0,0)}$ at the origin. In general no such vector field exists. However, if the curvature tensor happens to be zero then P can be constructed as follows. Let $P_{(x,0)}$ be a parallel vector field along the x-axis, satisfying the given initial condition. For each fixed x_0 let $P_{(x_0,y)}$ be a parallel vector field along the curve

$$y \longrightarrow s(x_0,y) ,$$

having the right value for $y = 0$. This defines P everywhere along s. Clearly $\frac{D}{\partial y}P$ is identically zero; and $\frac{D}{\partial x}P$ is zero along the x-axis. Now the identity

$$\frac{D}{\partial y}\frac{D}{\partial x}P - \frac{D}{\partial x}\frac{D}{\partial y}P = R\left(\frac{\partial s}{\partial x},\frac{\partial s}{\partial y}\right)P = 0$$

implies that $\frac{D}{\partial y}\frac{D}{\partial x}P = 0$. In other words, the vector field $\frac{D}{\partial x}P$ is parallel along the curves

$$y \longrightarrow s(x_0,y) .$$

Since $\left(\frac{D}{\partial x} P\right)_{(x_0,0)} = 0$, this implies that $\frac{D}{\partial x} P$ is identically zero; and completes the proof that P is parallel along s.]

Henceforth we will assume that M is a Riemannian manifold, provided with the unique symmetric connection which is compatible with its metric. In conclusion we will prove that the tensor R satisfies four symmetry relations.

LEMMA 9.3. The curvature tensor of a Riemannian manifold satisfies:

(1) $R(X,Y)Z + R(Y,X)Z = 0$

(2) $R(X,Y)Z + R(Y,Z)X + R(Z,X)Y = 0$

(3) $\langle R(X,Y)Z,W\rangle + \langle R(X,Y)W,Z\rangle = 0$

(4) $\langle R(X,Y)Z,W\rangle = \langle R(Z,W)X,Y\rangle$.

PROOF: The skew-symmetry relation (1) follows immediately from the definition of R.

Since all three terms of (2) are tensors, it is sufficient to prove (2) when the bracket products $[X,Y]$, $[X,Z]$ and $[Y,Z]$ are all zero. Under this hypothesis we must verify the identity

$$\begin{aligned} &- X \vdash (Y \vdash Z) + Y \vdash (X \vdash Z) \\ &- Y \vdash (Z \vdash X) + Z \vdash (Y \vdash X) \\ &- Z \vdash (X \vdash Y) + X \vdash (Z \vdash Y) = 0 . \end{aligned}$$

But the symmetry of the connection implies that

$$Y \vdash Z - Z \vdash Y = [Y,Z] = 0 .$$

Thus the upper left term cancels the lower right term. Similarly the remaining terms cancel in pairs. This proves (2).

To prove (3) we must show that the expression $\langle R(X,Y)Z,W\rangle$ is skew-symmetric in Z and W. This is clearly equivalent to the assertion that

$$\langle R(X,Y)Z,Z\rangle = 0$$

for all X,Y,Z. Again we may assume that $[X,Y] = 0$, so that $\langle R(X,Y)Z,Z\rangle$ is equal to

$$\langle - X \vdash (Y \vdash Z) + Y \vdash (X \vdash Z), Z\rangle .$$

In other words we must prove that the expression

$$\langle Y \vdash (X \vdash Z), Z \rangle$$

is symmetric in X and Y.

Since $[X,Y] = 0$ the expression $YX\langle Z,Z\rangle$ is symmetric in X and Y. Since the connection is compatible with the metric, we have

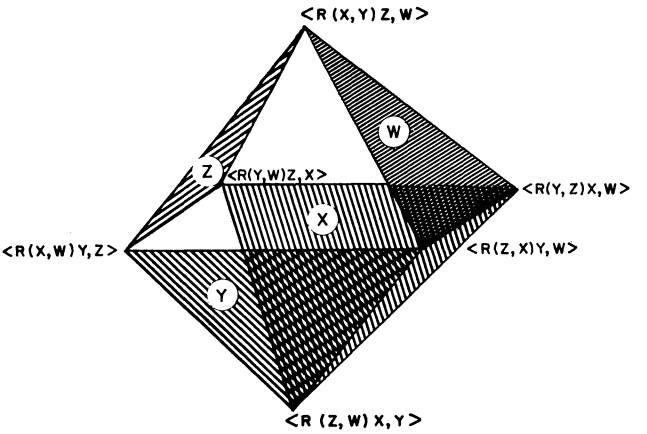
$$X\langle Z,Z\rangle = 2\langle X \vdash Z, Z\rangle$$

hence

$$YX\langle Z,Z\rangle = 2\langle Y \vdash (X \vdash Z), Z\rangle + 2\langle X \vdash Z, Y \vdash Z\rangle .$$

But the right hand term is clearly symmetric in X and Y. Therefore $\langle Y \vdash (X \vdash Z), Z\rangle$ is symmetric in X and Y; which proves property (3).

Property (4) may be proved from (1), (2), and (3) as follows.



Formula (2) asserts that the sum of the quantities at the vertices of shaded triangle W is zero. Similarly (making use of (1) and (3)) the sum of the vertices of each of the other shaded triangles is zero. Adding these identities for the top two shaded triangles, and subtracting the identities for the bottom ones, this means that twice the top vertex minus twice the bottom vertex is zero. This proves (4), and completes the proof.

§10. Geodesics and Completeness

Let M be a connected Riemannian manifold.

DEFINITION. A parametrized path

$$\gamma: \ I \to M,$$

where I denotes any interval of real numbers, is called a geodesic if the acceleration vector field $\frac{D}{dt}\frac{d\gamma}{dt}$ is identically zero. Thus the velocity vector field $\frac{d\gamma}{dt}$ must be parallel along γ. If γ is a geodesic, then the identity

$$\frac{d}{dt}\left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = 2\left\langle \frac{D}{dt}\frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle = 0$$

shows that the length $\left\| \frac{d\gamma}{dt} \right\| = \left\langle \frac{d\gamma}{dt}, \frac{d\gamma}{dt} \right\rangle^{1/2}$ of the velocity vector is constant along γ. Introducing the arc-length function

$$s(t) = \int \left\| \frac{d\gamma}{dt} \right\| dt + \text{constant}$$

This statement can be rephrased as follows: The parameter t along a geodesic is a linear function of the arc-length. The parameter t is actually equal to the arc-length if and only if $\left\| \frac{d\gamma}{dt} \right\| = 1$

In terms of a local coordinate system with coordinates $u^1,\ldots,u^n$ a curve $t \to \gamma(t) \in M$ determines n smooth functions $u^1(t),\ldots,u^n(t)$. The equation $\frac{D}{dt}\frac{d\gamma}{dt}$ for a geodesic then takes the form

$$\frac{d^2u^k}{dt^2} + \sum_{i,j=1}^{n} \Gamma_{ij}^{k}\,(u^1,\ldots,u^n)\,\frac{du^i}{dt}\frac{du^j}{dt} = 0$$

The existence of geodesics depends, therefore, on the solutions of a certain system of second order differential equations.

More generally consider any system of equations of the form

$$\frac{d^2\vec{u}}{dt^2} = \vec{F}\left(\vec{u}, \frac{d\vec{u}}{dt}\right)$$

Here $\vec{u}$ stands for $(u^1,\ldots,u^n)$ and $\vec{F}$ stands for an n-tuple of C^∞ functions, all defined throughout some neighborhood U of a point

$$(\vec{u}_1, \vec{v}_1) \in \mathbf{R}^{2n}$$

EXISTENCE AND UNIQUENESS THEOREM 10.1. There exists a neighborhood W of the point $(\vec{u}_1,\vec{v}_1)$ and a number $\varepsilon > 0$ so that, for each $(\vec{u}_0,\vec{v}_0) \in W$ the differential equation

$$\frac{d^2\vec{u}}{dt^2} = \vec{F}(\vec{u}, \frac{d\vec{u}}{dt})$$

has a unique solution $t \to \vec{u}(t)$ which is defined for $|t| < \varepsilon$, and satisfies the initial conditions

$$\vec{u}(0) = \vec{u}_0, \quad \frac{d\vec{u}}{dt}(0) = \vec{v}_0 .$$

Furthermore, the solution depends smoothly on the initial conditions. In other words, the correspondence

$$(\vec{u}_0,\vec{v}_0,t) \to \vec{u}(t)$$

from $W \times (-\varepsilon,\varepsilon)$ to $\mathbf{R}^n$ is a C^∞ function of all $2n+1$ variables.

PROOF: Introducing the new variables $v^i = \frac{du^i}{dt}$ this system of n second order equations becomes a system of $2n$ first order equations:

$$\begin{cases} \frac{d\vec{u}}{dt} = \vec{v} , \\ \frac{d\vec{v}}{dt} = \vec{F}(\vec{u},\vec{v}) . \end{cases}$$

The assertion then follows from Graves, "Theory of Functions of Real Variables," p. 166. (Compare our §2.4.)

Applying this theorem to the differential equation for geodesics, one obtains the following.

LEMMA 10.2. For every point p_0 on a Riemannian manifold M there exists a neighborhood U of p_0 and a number $\varepsilon > 0$ so that: for each $p \in U$ and each tangent vector $v \in TM_p$ with length $< \varepsilon$ there is a unique geodesic

$$\gamma_v: (-2,2) \to M$$

satisfying the conditions

$$\gamma_v(0) = p, \quad \frac{d\gamma_v}{dt}(0) = v .$$

PROOF. If we were willing to replace the interval $(-2,2)$ by an arbitrarily small interval, then this statement would follow immediately from 10.1. To be more precise; there exists a neighborhood U of p_0 and

numbers $\varepsilon_1, \varepsilon_2 > 0$ so that: for each $p \in U$ and each $v \in TM_p$ with $\|v\| < \varepsilon_1$ there is a unique geodesic

$$\gamma_v: \ (-2\varepsilon_2, 2\varepsilon_2) \ \to M$$

satisfying the required initial conditions.

To obtain the sharper statement it is only necessary to observe that the differential equation for geodesics has the following homogeneity property. Let c be any constant. If the parametrized curve

$$t \ \to \gamma(t)$$

is a geodesic, then the parametrized curve

$$t \ \to \gamma(ct)$$

will also be a geodesic.

Now suppose that ε is smaller than $\varepsilon_1\varepsilon_2$. Then if $\|v\| < \varepsilon$ and $|t| < 2$ note that

$$\|v/\varepsilon_2\| < \varepsilon_1 \quad \text{and} \quad |\varepsilon_2 t| < 2\varepsilon_2 \ .$$

Hence we can define $\gamma_v(t)$ to be $\gamma_{v/\varepsilon_2}(\varepsilon_2 t)$. This proves 10.2.

This following notation will be convenient. Let $v \in TM_q$ be a tangent vector, and suppose that there exists a geodesic

$$\gamma: [0,1] \ \to M$$

satisfying the conditions

$$\gamma(0) = q, \quad \frac{d\gamma}{dt}(0) = v.$$

Then the point $\gamma(1) \in M$ will be denoted by $\exp_q(v)$ and called the exponential* of the tangent vector v. The geodesic γ can thus be described by the formula

$$\gamma(t) = \exp_q(tv) \ .$$

---

* The historical motivation for this terminology is the following. If M is the group of all $n \times n$ unitary matrices then the tangent space TM_I at the identity can be identified with the space of $n \times n$ skew-Hermitian matrices. The function

$$\exp_I: \ TM_I \ \to M$$

as defined above is then given by the exponential power series

$$\exp_I(A) = I + A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \ldots \ .$$

Lemma 10.2 says that $\exp_q(v)$ is defined providing that $\|v\|$ is small enough. In general, $\exp_q(v)$ is not defined for large vectors v. However, if defined at all, $\exp_q(v)$ is always uniquely defined.

DEFINITION. The manifold M is geodesically complete if $\exp_q(v)$ is defined for all $q \in M$ and all vectors $v \in TM_q$. This is clearly equivalent to the following requirement:

For every geodesic segment $\gamma_0\colon [a,b] \to M$ it should be possible to extend γ_0 to an infinite geodesic

$$\gamma\colon \mathbf{R} \to M$$

We will return to a study of completeness after proving some local results.

Let TM be the tangent manifold of M, consisting of all pairs (p,v) with $p \in M$, $v \in TM_p$. We give TM the following C^∞ structure: if $(u^1,\dots,u^n)$ is a coordinate system in an open set $U \subset M$ then every tangent vector at $q \in U$ can be expressed uniquely as $t^1\partial_1 + \dots + t^n\partial_n$, where $\partial_i = \frac{\partial}{\partial u^i}\big|_q$. Then the functions $u^1,\dots,u^n,t^1,\dots,t^n$ constitute a coordinate system on the open set $TU \subset TM$.

Lemma 10.2 says that for each $p \in M$ the map

$$(q,v) \to \exp_q(v)$$

is defined throughout a neighborhood V of the point $(p,0) \in TM$. Furthermore this map is differentiable throughout V.

Now consider the smooth function $F\colon V \to M \times M$ defined by $F(q,v) = (q, \exp_q(v))$. We claim that the Jacobian of F at the point $(p,0)$ is non-singular. In fact, denoting the induced coordinates on $U \times U \subset M \times M$ by $(u_1^1,\dots,u_1^n,u_2^1,\dots,u_2^n)$, we have

$$F_*\Big(\frac{\partial}{\partial u^I}\Big) = \frac{\partial}{\partial u_1^I} + \frac{\partial}{\partial u_2^I}$$

$$F_*\Big(\frac{\partial}{\partial t^J}\Big) = \frac{\partial}{\partial u_2^J}$$

Thus the Jacobian matrix of F at $(p,0)$ has the form $\begin{pmatrix} I & I \\ 0 & I \end{pmatrix}$, and hence is non-singular.

It follows from the implicit function theorem that F maps some neighborhood V' of $(p,0) \in TM$ diffeomorphically onto some neighborhood

of $(p,p) \in M \times M$. We may assume that the first neighborhood V' consists of all pairs (q,v) such that q belongs to a given neighborhood U' of p and such that $\|v\| < \varepsilon$. Choose a smaller neighborhood W of p so that $F(V') \supset W \times W$. Then we have proven the following.

LEMMA 10.3. For each $p \in M$ there exists a neighborhood W and a number $\varepsilon > 0$ so that:

(1) Any two points of W are joined by a unique geodesic in M of length $< \varepsilon$.

(2) This geodesic depends smoothly upon the two points. (I.e., if $t \to \exp_{q_1}(tv)$, $0 \leq t \leq 1$, is the geodesic joining q_1 and q_2, then the pair $(q_1,v) \in TM$ depends differentiably on (q_1,q_2).)

(3) For each $q \in W$ the map $\exp_q$ maps the open ε-ball in TM_q diffeomorphically onto an open set $U_q \supset W$.

REMARK. With more care it would be possible to choose W so that the geodesic joining any two of its points lies completely within W. Compare J. H. C. Whitehead, Convex regions in the geometry of paths, Quarterly Journal of Mathematics (Oxford) Vol. 3, (1932), pp. 33-42.

Now let us study the relationship between geodesics and arc-length.

THEOREM 10.4. Let W and ε be as in Lemma 10.3. Let

$$\gamma: [0,1] \to M$$

be the geodesic of length $< \varepsilon$ joining two points of W, and let

$$\omega: [0,1] \to M$$

be any other piecewise smooth path joining the same two points. Then

$$\int_0^1 \|\frac{d\gamma}{dt}\| dt \leq \int_0^1 \|\frac{d\omega}{dt}\| dt \quad ,$$

where equality can hold only if the point set $\omega([0,1])$ coincides with $\gamma([0,1])$.

Thus γ is the shortest path joining its end points.

The proof will be based on two lemmas. Let $q = \gamma(0)$ and let U_q be as in 10.3.

LEMMA 10.5. In U_q, the geodesics through q are the orthogonal trajectories of hypersurfaces

$$\left\{ \exp_q(v) : v \in TM_q, \quad \|v\| = \text{constant} \right\} .$$

PROOF: Let $t \to v(t)$ denote any curve in TM_q with $\|v(t)\| = 1$. We must show that the corresponding curves

$$t \to \exp_q(r_0 v(t))$$

in U_q, where $0 < r_0 < \varepsilon$, are orthogonal to the radial geodesics

$$r \to \exp_q(rv(t_0)) .$$

In terms of the parametrized surface f given by

$$f(r,t) = \exp_q(rv(t)), \quad 0 \leq r < \varepsilon ,$$

we must prove that

$$\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = 0$$

for all (r,t).

Now

$$\frac{\partial}{\partial r} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle = \left\langle \frac{D}{\partial r} \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle + \left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial r} \frac{\partial f}{\partial t} \right\rangle .$$

The first expression on the right is zero since the curves

$$r \to f(r,t)$$

are geodesics. The second expression is equal to

$$\left\langle \frac{\partial f}{\partial r}, \frac{D}{\partial t} \frac{\partial f}{\partial r} \right\rangle = \frac{1}{2} \frac{\partial}{\partial t} \left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial r} \right\rangle = 0 ,$$

since $\left\|\frac{\partial f}{\partial r}\right\| = \|v(t)\| = 1$. Therefore the quantity $\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle$ is independent of r. But for $r = 0$ we have

$$f(0,t) = \exp_q(0) = q$$

hence $\frac{\partial f}{\partial t}(0,t) = 0$. Therefore $\left\langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \right\rangle$ is identically zero, which completes the proof.

Now consider any piecewise smooth curve

$$\omega: [a,b] \to U_q - \{q\} .$$

Each point $\omega(t)$ can be expressed uniquely in the form $\exp_q(r(t)v(t))$ with $0 < r(t) < \varepsilon$, and $\|v(t)\| = 1$, $v(t) \in TM_q$.

LEMMA 10.6. The length $\int_a^b \left\|\frac{d\omega}{dt}\right\| dt$ is greater than or equal to $|r(b) - r(a)|$, where equality holds only if the function $r(t)$ is monotone, and the function $v(t)$ is constant.

Thus the shortest path joining two concentric spherical shells around q is a radial geodesic.

PROOF: Let $f(r,t) = \exp_q(rv(t))$, so that $\omega(t) = f(r(t),t)$. Then

$$\frac{d\omega}{dt} = \frac{\partial f}{\partial r} r'(t) + \frac{\partial f}{\partial t} .$$

Since the two vectors on the right are mutually orthogonal, and since $\|\frac{\partial f}{\partial r}\| = 1$, this gives

$$\|\frac{d\omega}{dt}\|^2 = |r'(t)|^2 + \|\frac{\partial f}{\partial t}\|^2 \geq |r'(t)|^2$$

where equality holds only if $\frac{\partial f}{\partial t} = 0$; hence only if $\frac{dv}{dt} = 0$. Thus

$$\int_a^b \|\frac{d\omega}{dt}\| dt \geq \int_a^b |r'(t)| dt \geq |r(b) - r(a)|$$

where equality holds only if $r(t)$ is monotone and $v(t)$ is constant. This completes the proof.

The proof of Theorem 10.4 is now straightforward. Consider any piecewise smooth path ω from q to a point

$$q' = \exp_q(rv) \in U_q ;$$

where $0 < r < \varepsilon$, $\|v\| = 1$. Then for any $\delta > 0$ the path ω must contain a segment joining the spherical shell of radius δ to the spherical shell of radius r, and lying between these two shells. The length of this segment will be $\geq r - \delta$; hence letting δ tend to 0 the length of ω will be $\geq r$. If $\omega([0,1])$ does not coincide with $\gamma([0,1])$, then we easily obtain a strict inequality. This completes the proof of 10.4.

An important consequence of Theorem 10.4 is the following.

COROLLARY 10.7. Suppose that a path $\omega: [0,\ell] \to M$, parametrized by arc-length, has length less than or equal to the length of any other path from $\omega(0)$ to $\omega(\ell)$. Then ω is a geodesic.

PROOF: Consider any segment of ω lying within an open set W, as above, and having length $< \varepsilon$. This segment must be a geodesic by Theorem 10.4. Hence the entire path ω is a geodesic.

DEFINITION. A geodesic $\gamma: [a,b] \to M$ will be called <u>minimal</u> if

its length is less than or equal to the length of any other piecewise smooth path joining its endpoints.

Theorem 10.4 asserts that any sufficiently small segment of a geodesic is minimal. On the other hand a long geodesic may not be minimal. For example we will see shortly that a great circle arc on the unit sphere is a geodesic. If such an arc has length greater than π, it is certainly not minimal.

In general, minimal geodesics are not unique. For example two antipodal points on a unit sphere are joined by infinitely many minimal geodesics. However, the following assertion is true.

Define the distance $\rho(p,q)$ between two points $p,q \in M$ to be the greatest lower bound for the arc-lengths of piecewise smooth paths joining these points. This clearly makes M into a metric space. It follows easily from 10.4 that this metric is compatible with the usual topology of M.

> COROLLARY 10.8. Given a compact set $K \subset M$ there exists a number $\delta > 0$ so that any two points of K with distance less than δ are joined by a unique geodesic of length less than δ. Furthermore this geodesic is minimal; and depends differentiably on its endpoints.

PROOF. Cover K by open sets W_α, as in 10.3, and let δ be small enough so that any two points in K with distance less than δ lie in a common W_α. This completes the proof.

Recall that the manifold M is geodesically complete if every geodesic segment can be extended indifinitely.

> THEOREM 10.9 (Hopf and Rinow*). If M is geodesically complete, then any two points can be joined by a minimal geodesic.

PROOF. Given $p,q \in M$ with distance $r > 0$, choose a neighborhood U_p as in Lemma 10.3. Let $S \subset U_p$ denote a spherical shell of radius $\delta < \varepsilon$

---

* Compare p. 341 of G. de Rham, Sur la réductibilité d'un espace de Riemann, Commentarii Math. Helvetici, Vol. 26 (1952); as well as H. Hopf and W. Rinow, Ueber den Begriff der vollständigen differentialgeometrischen Fläche, Commentarii, Vol. 3 (1931), pp. 209-225.

about p. Since S is compact, there exists a point

$$p_0 = \exp_p(\delta v), \qquad \|v\| = 1,$$

on S for which the distance to q is minimized. We will prove that

$$\exp_p(rv) = q.$$

This implies that the geodesic segment $t \to \gamma(t) = \exp_p(tv)$, $0 \leq t \leq r$, is actually a minimal geodesic from p to q.

The proof will amount to showing that a point which moves along the geodesic γ must get closer and closer to q. In fact for each $t \in [\delta, r]$ we will prove that

$$(1_t) \qquad \rho(\gamma(t), q) = r - t \ .$$

This identity, for $t = r$, will complete the proof.

First we will show that the equality (1_δ) is true. Since every path from p to q must pass through S, we have

$$\rho(p,q) = \operatorname*{Min}_{s \in S} (\rho(p,s) + \rho(s,q)) = \delta + \rho(p_0,q) \ .$$

Therefore $\rho(p_0,q) = r - \delta$. Since $p_0 = \gamma(\delta)$, this proves (1_δ).

Let $t_0 \in [\delta, r]$ denote the supremum of those numbers t for which (1_t) is true. Then by continuity the equality (1_{t_0}) is true also. If $t_0 < r$ we will obtain a contradiction. Let S' denote a small spherical shell of radius δ' about the point $\gamma(t_0)$; and let $p_0' \in S'$ be a point of S' with minimum distance from q. (Compare Diagram 10.) Then

$$\rho(\gamma(t_0), q) = \operatorname*{Min}_{s \in S'} (\rho(\gamma(t_0), s) + \rho(s,q)) = \delta' + \rho(p_0', q) \ ,$$

hence

$$(2) \qquad \rho(p_0', q) = (r - t_0) - \delta' \ .$$

We claim that p_0' is equal to $\gamma(t_0 + \delta')$. In fact the triangle inequality states that

$$\rho(p, p_0') \geq \rho(p,q) - \rho(p_0', q) = t_0 + \delta'$$

(making use of (2)). But a path of length precisely $t_0 + \delta'$ from p to p_0' is obtained by following γ from p to $\gamma(t_0)$, and then following a minimal geodesic from $\gamma(t_0)$ to p_0'. Since this broken geodesic has minimal length, it follows from Corollary 10.7 that it is an (unbroken)

geodesic, and hence coincides with γ.

Thus $\gamma(t_0 + \delta') = p_0'$. Now the equality (2) becomes

$(1_{t_0+\delta'})$ $$\rho(\gamma(t_0 + \delta'),q) = r - (t_0 + \delta') \quad .$$

This contradicts the definition of t_0; and completes the proof.

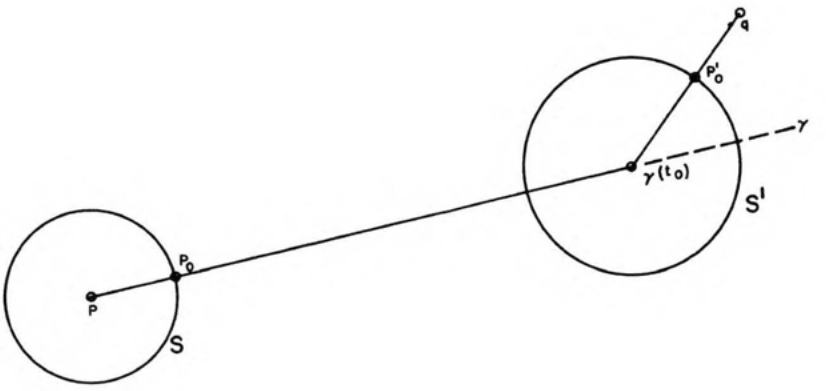


Diagram 10.

As a consequence one has the following.

COROLLARY 10.10. If M is geodesically complete then every bounded subset of M has compact closure. Consequently M is complete as a metric space (i.e., every Cauchy sequence converges).

PROOF. If $X \subset M$ has diameter d then for any $p \in X$ the map $\exp_p: TM_p \to M$ maps the disk of radius d in TM_p onto a compact subset of M which (making use of Theorem 10.9) contains X. Hence the closure of X is compact.

Conversely, if M is complete as a metric space, then it is not difficult, using Lemma 10.3, to prove that M is geodesically complete. For details the reader is referred to Hopf and Rinow. Henceforth we will not distinguish between geodesic completeness and metric completeness, but will refer simply to a complete Riemannian manifold.

FAMILIAR EXAMPLES OF GEODESICS. In Euclidean n-space, $\mathbf{R}^n$, with the usual coordinate system $x_1,\ldots,x_n$ and the usual Riemannian metric $dx_1 \otimes dx_1 + \ldots + dx_n \otimes dx_n$ we have $\Gamma_{ij}^k = 0$ and the equations for a geodesic γ, given by $t \to (x_1(t),\ldots,x_n(t)$ become

$$\frac{d^2x_i}{dt^2} = 0 \quad ,$$

whose solutions are the straight lines. This could also have been seen as follows: it is easy to show that the formula for arc length

$$\int \Big(\sum_{i=1}^{n} \Big(\frac{dx_i}{dt}\Big)^2\Big)^{\frac{1}{2}} dt$$

coincides with the usual definition of arc length as the least upper bound of the lengths of inscribed polygons; from this definition it is clear that straight lines have minimal length, and are therefore geodesics.

The geodesics on S^n are precisely the great circles, that is, the intersections of S^n with the planes through the center of S^n.

PROOF. Reflection through a plane E^2 is an isometry $I\colon S^n \to S^n$ whose fixed point set is $C = S^n \cap E^2$. Let x and y be two points of C with a unique geodesic C' of minimal length between them. Then, since I is an isometry, the curve $I(C')$ is a geodesic of the same length as C' between $I(x) = x$ and $I(y) = y$. Therefore $C' = I(C')$. This implies that $C' \subset C$.

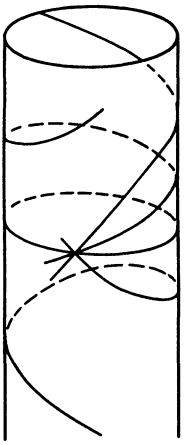
Finally, since there is a great circle through any point of S^n in any given direction, these are all the geodesics.

Antipodal points on the sphere have a continium of geodesics of minimal length between them. All other pairs of points have a unique geodesic of minimal length between them, but an infinite family of non-minimal geodesics, depending on how many times the geodesic goes around the sphere and in which direction it starts.

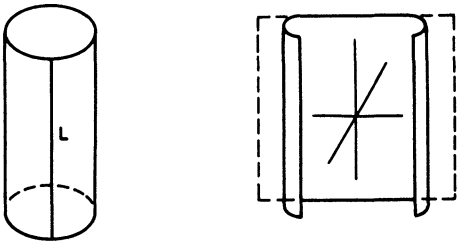
By the same reasoning every meridian line on a surface of revolution is a geodesic.

The geodesics on a right circular cylinder Z are the generating lines, the circles cut by planes perpendicular to the generating lines, and

the helices on Z.



PROOF: If L is a generating line of Z then we can set up an isometry $I: Z - L \to \mathbf{R}^2$ by rolling Z onto $\mathbf{R}^2$:



The geodesics on Z are just the images under I^{-1} of the straight lines in $\mathbf{R}^2$. Two points on Z have infinitely many geodesics between them.

PART III

THE CALCULUS OF VARIATIONS APPLIED TO GEODESICS

§11. The Path Space of a Smooth Manifold.

Let M be a smooth manifold and let p and q be two (not necessarily distinct) points of M. By a piecewise smooth path from p to q will be meant a map $\omega\colon [0,1] \to M$ such that

1) there exists a subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$ of $[0,1]$ so that each $\omega|[t_{i-1},t_i]$ is differentiable of class C^∞;

2) $\omega(0) = p$ and $\omega(1) = q$.

The set of all piecewise smooth paths from p to q in M will be denoted by $\Omega(M;p,q)$, or briefly by $\Omega(M)$ or Ω.

Later (in §16) Ω will be given the structure of a topological space, but for the moment this will not be necessary. We will think of Ω as being something like an "infinite dimensional manifold." To start the analogy we make the following definition.

By the tangent space of Ω at a path ω will be meant the vector space consisting of all piecewise smooth vector fields W along ω for which $W(0) = 0$ and $W(1) = 0$. The notation $T\Omega_\omega$ will be used for this vector space.

If F is a real valued function on Ω it is natural to ask what

$$F_*\colon T\Omega_\omega \to T\mathbf{R}_{F(\omega)} ,$$

the induced map on the tangent space, should mean. When F is a function which is smooth in the usual sense, on a smooth manifold M, we can define $F_*\colon TM_p \to T\mathbf{R}_{F(p)}$ as follows. Given $X \in TM_p$ choose a smooth path $u \to \alpha(u)$ in M, which is defined for $-\varepsilon < u < \varepsilon$, so that

$$\alpha(0) = p, \quad \frac{d\alpha}{du}(0) = X .$$

Then $F_*(X)$ is equal to $\frac{d(F(\alpha(u))}{du}\Big|_{u=0}$, multiplied by the basis vector $\left(\frac{d}{dt}\right)_{F(p)} \in T\mathbf{R}_{F(p)}$.

In order to carry out an analogous construction for $F: \Omega \to \mathbf{R}$, the following concept is needed.

DEFINITION. A variation of ω (keeping endpoints fixed) is a function

$$\bar{\alpha}: (-\varepsilon,\varepsilon) \to \Omega,$$

for some $\varepsilon > 0$, such that

1) $\bar{\alpha}(0) = \omega$

2) there is a subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$ of $[0,1]$ so that the map

$$\alpha: (-\varepsilon,\varepsilon) \times [0,1] \to M$$

defined by $\alpha(u,t) = \bar{\alpha}(u)(t)$ is C^∞ on each strip $(-\varepsilon,\varepsilon) \times [t_{i-1},t_i]$, $i = 1,\ldots,k$.

Since each $\bar{\alpha}(u)$ belongs to $\Omega = \Omega(M;p,q)$, note that:

3) $\alpha(u,0) = p$, $\alpha(u,1) = q$ for all $u \in (-\varepsilon,\varepsilon)$.

We will use either α or $\bar{\alpha}$ to refer to the variation. More generally if, in the above definition, $(-\varepsilon,\varepsilon)$ is replaced by a neighborhood U of 0 in $\mathbf{R}^n$, then α (or $\bar{\alpha}$) is called an n-parameter variation of ω.

Now $\bar{\alpha}$ may be considered as a "smooth path" in Ω. Its "velocity vector" $\frac{d\bar{\alpha}}{du}(0) \in T\Omega_\omega$ is defined to be the vector field W along ω given by

$$W_t = \frac{d\bar{\alpha}}{du}(0)_t = \frac{\partial\alpha}{\partial u}(0,t) .$$

Clearly $W \in T\Omega_\omega$. This vector field W is also called the variation vector field associated with the variation α.

Given any $W \in T\Omega_\omega$ note that there exists a variation $\bar{\alpha}: (-\varepsilon,\varepsilon) \to \Omega$ which satisfies the conditions $\bar{\alpha}(0) = \omega$, $\frac{d\bar{\alpha}}{du}(0) = W$. In fact one can set

$$\bar{\alpha}(u)(t) = \exp_{\omega(t)}\left(u\, W_t\right) .$$

By analogy with the definition given above, if F is a real valued

function on Ω, we attempt to define

$$F_*: \ T\Omega_\omega \to T\mathbf{R}_{F(\omega)}$$

as follows. Given $W \in T\Omega_\omega$ choose a variation $\bar{\alpha}: \ (-\varepsilon,\varepsilon) \to \Omega$ with

$$\bar{\alpha}(0) = \omega, \qquad \frac{d\bar{\alpha}}{du}(0) = W$$

and set $F_*(W)$ equal to $\left.\frac{d(F(\bar{\alpha}(u))}{du}\right|_{u=0}$ multiplied by the tangent vector $\left(\frac{d}{dt}\right)_{F(\omega)}$. Of course without hypothesis on F there is no guarantee that this derivative will exist, or will be independent of the choice of $\bar{\alpha}$. We will not investigate what conditions F must satisfy in order for F_* to have these properties. We have indicated how F_* might be defined only to motivate the following.

DEFINITION. A path ω is a <u>critical path</u> for a function $F: \ \Omega \to \mathbf{R}$ if and only if $\left.\frac{dF(\bar{\alpha}(u))}{du}\right|_{u=0}$ is zero for every variation $\bar{\alpha}$ of ω.

EXAMPLE. If F takes on its minimum at a path ω_0, and if the derivatives $\frac{dF(\bar{\alpha}(u))}{du}$ are all defined, then clearly ω_0 is a critical path.

§12. The Energy of a Path.

Suppose now that M is a Riemannian manifold. The length of a vector $v \in TM_p$ will be denoted by $\|v\| = \langle v,v \rangle^{\frac{1}{2}}$. For $\omega \in \Omega$ define the energy of ω from a to b (where $0 \leq a < b \leq 1$) as

$$E_a^b(\omega) = \int_a^b \left\|\frac{d\omega}{dt}\right\|^2 dt \quad .$$

We will write E for E_0^1.

This can be compared with the arc-length from a to b given by

$$L_a^b(\omega) = \int_a^b \left\|\frac{d\omega}{dt}\right\| dt$$

as follows. Applying Schwarz's inequality

$$\left(\int_a^b fg\,dt\right)^2 \leq \left(\int_a^b f^2 dt\right)\left(\int_a^b g^2 dt\right)$$

with $f(t) = 1$ and $g(t) = \left\|\frac{d\omega}{dt}\right\|$ we see that

$$\left(L_a^b\right)^2 \leq (b - a)E_a^b \quad ,$$

where equality holds if and only if g is constant; that is if and only if the parameter t is proportional to arc-length.

Now suppose that there exists a minimal geodesic γ from $p = \omega(0)$ to $q = \omega(1)$. Then

$$E(\gamma) = L(\gamma)^2 \leq L(\omega)^2 \leq E(\omega) \quad .$$

Here the equality $L(\gamma)^2 = L(\omega)^2$ can hold only if ω is also a minimal geodesic, possibly reparametrized. (Compare §10.7.) On the other hand the equality $L(\omega)^2 = E(\omega)$ can hold only if the parameter is proportional to arc-length along ω. This proves that $E(\gamma) < E(\omega)$ unless ω is also a minimal geodesic. In other words:

LEMMA 12.1. Let M be a complete Riemannian manifold and let $p,q \in M$ have distance d. Then the energy function

$$E: \Omega(M;p,q) \to \mathbf{R}$$

takes on its minimum d^2 precisely on the set of minimal geodesics from p to q.

We will now see which paths $\omega \in \Omega$ are critical paths for the energy function E.

Let $\bar{\alpha}: (-\varepsilon,\varepsilon) \to \Omega$ be a variation of ω, and let $W_t = \frac{\partial \alpha}{\partial u}(0,t)$ be the associated variation vector field. Furthermore, let:

$V_t = \frac{d\omega}{dt}$ = velocity vector of ω,

$A_t = \frac{D}{dt}\frac{d\omega}{dt}$ = acceleration vector of ω,

$\Delta_t V = V_{t+} - V_{t-}$ = discontinuity in the velocity vector at t, where $0 < t < 1$.

Of course $\Delta_t V = 0$ for all but a finite number of values of t.

THEOREM 12.2 (First variation formula). The derivative $\frac{1}{2}\frac{dE(\bar{\alpha}(u))}{du}\Big|_{u=0}$ is equal to

$$-\sum_t \langle W_t, \Delta_t V\rangle - \int_0^1 \langle W_t, A_t\rangle \, dt .$$

PROOF: According to Lemma 8.3, we have

$$\frac{\partial}{\partial u}\left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle = 2\left\langle \frac{D}{\partial u}\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle .$$

Therefore

$$\frac{dE(\bar{\alpha}(u))}{du} = \frac{d}{du}\int_0^1 \left\langle \frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle dt = 2\int_0^1 \left\langle \frac{D}{\partial u}\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial t}\right\rangle dt .$$

By Lemma 8.7 we can substitute $\frac{D}{\partial t}\frac{\partial \alpha}{\partial u}$ for $\frac{D}{\partial u}\frac{\partial \alpha}{\partial t}$ in this last formula.

Choose $0 = t_0 < t_1 < \ldots < t_k = 1$ so that α is differentiable on each strip $(-\varepsilon,\varepsilon) \times [t_{i-1}, t_i]$. Then we can "integrate by parts" on $[t_{i-1}, t_i]$, as follows. The identity

$$\frac{\partial}{\partial t}\left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle = \left\langle \frac{D}{\partial t}\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle + \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t}\right\rangle$$

implies that

$$\int_{t_{i-1}}^{t_i} \left\langle \frac{D}{\partial t}\frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle dt = \left\langle \frac{\partial \alpha}{\partial u}, \frac{\partial \alpha}{\partial t}\right\rangle \Big|_{t=t_{i-1}}^{t=t_i} - \int_{t_{i-1}}^{t_i} \left\langle \frac{\partial \alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t}\right\rangle dt .$$

Adding up the corresponding formulas for $i = 1,\dots,k$; and using the fact that $\frac{\partial\alpha}{\partial u} = 0$ for $t = 0$ or 1, this gives

$$\frac{1}{2}\frac{dE(\bar{\alpha}(u))}{du} = -\sum_{i=1}^{k-1} \left\langle \frac{\partial\alpha}{\partial u}, \Delta_{t_i} \frac{\partial\alpha}{\partial t} \right\rangle - \int_0^1 \left\langle \frac{\partial\alpha}{\partial u}, \frac{D}{\partial t}\frac{\partial\alpha}{\partial t} \right\rangle dt \quad .$$

Setting $u = 0$, we now obtain the required formula

$$\frac{1}{2}\frac{dE \circ \bar{\alpha}}{du}(0) = -\sum_t \langle W, \Delta_t V \rangle - \int_0^1 \langle W, A \rangle \, dt \quad .$$

This completes the proof.

Intuitively, the first term in the expression for $\frac{dE \circ \bar{\alpha}}{du}(0)$ shows that varying the path ω in the direction of decreasing "kink," tends to decrease E; see Diagram 11.

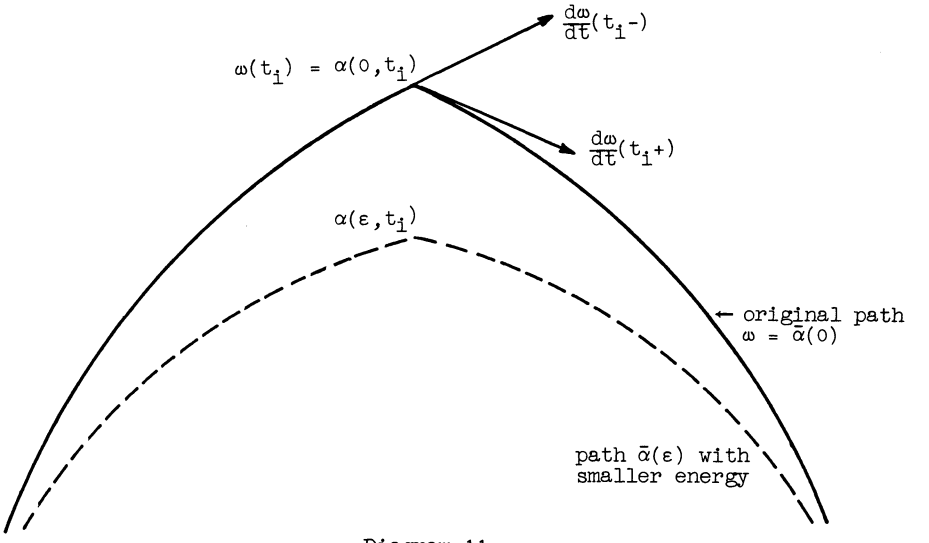


Diagram 11.

The second term shows that varying the curve in the direction of its acceleration vector $\frac{D}{dt}\left(\frac{d\omega}{dt}\right)$ tends to reduce E.

Recall that the path $\omega \in \Omega$ is called a <u>geodesic</u> if and only if ω is C^∞ on the whole interval $[0,1]$, and the acceleration vector $\frac{D}{dt}\left(\frac{d\omega}{dt}\right)$ of ω is identically zero along ω.

COROLLARY 12.3. The path ω is a critical point for the function E if and only if ω is a geodesic.

PROOF: Clearly a geodesic is a critical point. Let ω be a critical point. There is a variation of ω with $W(t) = f(t)A(t)$ where $f(t)$ is positive except that it vanishes at the t_i. Then

$$\frac{1}{2}\frac{dE}{du}(0) = -\int_0^1 f(t) \langle A(t), A(t) \rangle dt.$$

This is zero if and only if $A(t) = 0$ for all t. Hence each $\omega|[t_i, t_{i+1}]$ is a geodesic.

Now pick a variation such that $W(t_i) = \Delta_{t_i} V$. Then $\frac{1}{2}\frac{dE}{du}(0) = -\sum \langle \Delta_{t_i} V, \Delta_{t_i} V \rangle$. If this is zero then all $\Delta_t V$ are 0, and ω is differentiable of class C^1, even at the points t_i. Now it follows from the uniqueness theorem for differential equations that ω is C^∞ everywhere: thus ω is an unbroken geodesic.

§13. The Hessian of the Energy Function at a Critical Path.

Continuing with the analogy developed in the preceding section, we now wish to define a bilinear functional

$$E_{**}: \ T\Omega_\gamma \times T\Omega_\gamma \ \to \mathbf{R}$$

when γ is a critical point of the function E, i.e., a geodesic. This bilinear functional will be called the Hessian of E at γ.

If f is a real valued function on a manifold M with critical point p, then the Hessian

$$f_{**}: \ TM_p \times TM_p \ \to \mathbf{R}$$

can be defined as follows. Given $X_1, X_2 \in TM_p$ choose a smooth map $(u_1,u_2) \to \alpha(u_1,u_2)$ defined on a neighborhood of $(0,0)$ in $\mathbf{R}^2$, with values in M, so that

$$\alpha(0,0) = p, \quad \frac{\partial\alpha}{\partial u_1}(0,0) = X_1, \quad \frac{\partial\alpha}{\partial u_2}(0,0) = X_2 \ .$$

Then

$$f_{**}(X_1,X_2) = \frac{\partial^2 f(\alpha(u_1,u_2))}{\partial u_1 \, \partial u_2}\bigg|_{(0,0)} \ .$$

This suggests defining E_{**} as follows. Given vector fields $W_1, W_2 \in T\Omega_\gamma$ choose a 2-parameter variation

$$\alpha: \ U \times [0,1] \to M \ ,$$

where U is a neighborhood of $(0,0)$ in $\mathbf{R}^2$, so that

$$\alpha(0,0,t) = \gamma(t), \quad \frac{\partial\alpha}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial\alpha}{\partial u_2}(0,0,t) = W_2(t) \ .$$

(Compare §11.) Then the Hessian $E_{**}(W_1,W_2)$ will be defined to be the second partial derivative

$$\frac{\partial^2 E(\bar\alpha(u_1,u_2))}{\partial u_1 \, \partial u_2}\bigg|_{(0,0)} \ ;$$

where $\bar\alpha(u_1,u_2) \in \Omega$ denotes the path $\bar\alpha(u_1,u_2)(t) = \alpha(u_1,u_2,t)$. This second derivative will be written briefly as $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$.

The following theorem is needed to prove that E_{**} is well defined.

THEOREM 13.1 (Second variation formula). Let $\bar{\alpha}\colon U \to \Omega$ be a 2-parameter variation of the geodesic γ with variation vector fields

$$W_i = \frac{\partial \bar{\alpha}}{\partial u_i}(0,0) \in T\Omega_\gamma, \qquad i = 1,2 \ .$$

Then the second derivative $\frac{1}{2}\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$ of the energy function is equal to

$$-\sum_t \langle W_2(t), \Delta_t \frac{DW_1}{dt} \rangle - \int_0^1 \langle W_2, \frac{D^2 W_1}{dt^2} + R(V,W_1)V \rangle \, dt \ ;$$

where $V = \frac{d\gamma}{dt}$ denotes the velocity vector field and where

$$\Delta_t \frac{DW_1}{dt} = \frac{DW_1}{dt}(t^+) - \frac{DW_1}{dt}(t^-)$$

denotes the jump in $\frac{DW_1}{dt}$ at one of its finitely many points of discontinuity in the open unit interval.

PROOF: According to 12.2 we have

$$\frac{1}{2}\frac{\partial E}{\partial u_2} = -\sum_t \langle \frac{\partial \alpha}{\partial u_2}, \Delta_t \frac{\partial \alpha}{\partial t} \rangle - \int_0^1 \langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \rangle \, dt \ .$$

Therefore

$$\frac{1}{2}\frac{\partial^2 E}{\partial u_1 \partial u_2} = -\sum_t \langle \frac{D}{\partial u_1}\frac{\partial \alpha}{\partial u_2}, \Delta_t \frac{\partial \alpha}{\partial t} \rangle - \sum_t \langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1} \Delta_t \frac{\partial \alpha}{\partial t} \rangle$$
$$- \int_0^1 \langle \frac{D}{\partial u_1}\frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \rangle \, dt - \int_0^1 \langle \frac{\partial \alpha}{\partial u_2}, \frac{D}{\partial u_1}\frac{D}{\partial t}\frac{\partial \alpha}{\partial t} \rangle \, dt.$$

Let us evaluate this expression for $(u_1,u_2) = (0,0)$. Since $\gamma = \bar{\alpha}(0,0)$ is an unbroken geodesic, we have

$$\Delta_t \frac{\partial \alpha}{\partial t} = 0, \qquad \frac{D}{\partial t}\frac{\partial \alpha}{\partial t} = 0 \ ,$$

so that the first and third terms are zero.

Rearranging the second term, we obtain

$$\frac{1}{2}\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0) = -\sum \langle W_2, \Delta_t \frac{D}{dt} W_1 \rangle - \int_0^1 \langle W_2, \frac{D}{\partial u_1}\frac{D}{\partial t} V \rangle \, dt \ . \tag{13.2}$$

In order to interchange the two operators $\frac{D}{\partial u_1}$ and $\frac{D}{\partial t}$, we need to bring in the curvature formula,

$$\frac{D}{\partial u_1}\frac{D}{\partial t} V - \frac{D}{\partial t}\frac{D}{\partial u_1} V = R\Big(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u_1}\Big)V = R(V,W_1)V \ .$$

Together with the identity $\frac{D}{\partial u_1} V = \frac{D}{\partial t} \frac{\partial \alpha}{\partial u_1} = \frac{D}{dt} W_1$, this yields

$$\frac{D}{\partial u_1} \frac{D}{\partial t} V = \frac{D^2 W_1}{dt^2} + R(V, W_1)V \quad . \tag{13.3}$$

Substituting this expression into (13.2) this completes the proof of 13.1.

COROLLARY 13.4. The expression $E_{**}(W_1, W_2) = \frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$ is a well defined symmetric and bilinear function of W_1 and W_2.

PROOF: The second variation formula shows that $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$ depends only on the variation vector fields W_1 and W_2, so that $E_{**}(W_1, W_2)$ is well defined. This formula also shows that E_{**} is bilinear. The symmetry property

$$E_{**}(W_1, W_2) = E_{**}(W_2, W_1)$$

is not at all obvious from the second variation formula; but does follow immediately from the symmetry property $\frac{\partial^2 E}{\partial u_1 \partial u_2} = \frac{\partial^2 E}{\partial u_2 \partial u_1}$.

REMARK 13.5. The diagonal terms $E_{**}(W,W)$ of the bilinear pairing E_{**} can be described in terms of a 1-parameter variation of γ. In fact

$$E_{**}(W,W) = \frac{d^2 E \circ \bar{\alpha}}{du^2}(0) \quad ,$$

where $\bar{\alpha}$: $(-\varepsilon, \varepsilon) \to \Omega$ denotes any variation of γ with variation vector field $\frac{d\bar{\alpha}}{du}(0)$ equal to W. To prove this identity it is only necessary to introduce the two parameter variation

$$\bar{\beta}(u_1, u_2) = \bar{\alpha}(u_1 + u_2)$$

and to note that

$$\frac{\partial \bar{\beta}}{\partial u_i} = \frac{d\bar{\alpha}}{du} \quad , \quad \frac{\partial^2 E \circ \bar{\beta}}{\partial u_1 \, \partial u_2} = \frac{d^2 E \circ \bar{\alpha}}{du^2} \quad .$$

As an application of this remark, we have the following.

LEMMA 13.6. If γ is a minimal geodesic from p to q then the bilinear pairing E_{**} is positive semi-definite. Hence the index λ of E_{**} is zero.

PROOF: The inequality $E(\bar{\alpha}(u)) \geq E(\gamma) = E(\bar{\alpha}(0))$ implies that $\frac{d^2 E(\bar{\alpha}(u))}{du^2}$, evaluated at $u = 0$, is ≥ 0. Hence $E_{**}(W,W) \geq 0$ for all W.

§14. Jacobi Fields: The Null Space of E_{**}

A vector field J along a geodesic γ is called a Jacobi field if it satisfies the Jacobi differential equation

$$\frac{D^2J}{dt^2} + R(V,J)V = 0$$

where $V = \frac{d\gamma}{dt}$. This is a linear, second order differential equation. [It can be put in a more familiar form by choosing orthonormal parallel vector fields $P_1,\dots,P_n$ along γ. Then setting $J(t) = \Sigma f^i(t)P_i(t)$, the equation becomes

$$\frac{d^2f^i}{dt^2} + \sum_{j=1}^{n} a^i_j(t)f^j(t) = 0, \qquad i = 1,\dots, n;$$

where $a^i_j = \langle R(V,P_j)V,P_i \rangle$.] Thus the Jacobi equation has $2n$ linearly independent solutions, each of which can be defined throughout γ. The solutions are all C^∞-differentiable. A given Jacobi field J is completely determined by its initial conditions:

$$J(0), \frac{DJ}{dt}(0) \in TM_{\gamma(0)} .$$

Let $p = \gamma(a)$ and $q = \gamma(b)$ be two points on the geodesic γ, with $a \neq b$.

DEFINITION. p and q are conjugate* along γ if there exists a non-zero Jacobi field J along γ which vanishes for $t = a$ and $t = b$. The multiplicity of p and q as conjugate points is equal to the dimension of the vector space consisting of all such Jacobi fields.

Now let γ be a geodesic in $\Omega = \Omega(M;p,q)$. Recall that the null-space of the Hessian

$$E_{**}: \quad T\Omega_\gamma \times T\Omega_\gamma \longrightarrow \mathbf{R}$$

is the vector space consisting of those $W_1 \in T\Omega_\gamma$ such that $E_{**}(W_1,W_2) = 0$

---

* If γ has self-intersections then this definition becomes ambiguous. One should rather say that the parameter values a and b are conjugate with respect to γ.

for all W_2. The nullity ν of E_{**} is equal to the dimension of this null space. E_{**} is <u>degenerate</u> if $\nu > 0$.

> THEOREM 14.1. A vector field $W_1 \in T\Omega_\gamma$ belongs to the null space of E_{**} if and only if W_1 is a Jacobi field. Hence E_{**} is degenerate if and only if the end points p and q are conjugate along γ. The nullity of E_{**} is equal to the multiplicity of p and q as conjugate points.

PROOF: (Compare the proof of 12.3.) If J is a Jacobi field which vanishes at p and q, then J certainly belongs to $T\Omega_\gamma$. The second variation formula (§13.1) states that

$$-\tfrac{1}{2}E_{**}(J,W_2) = \sum_t \langle W_2(t),0\rangle + \int_0^1 \langle W_2,0\rangle \, dt = 0 .$$

Hence J belongs to the null space.

Conversely, suppose that W_1 belongs to the null space of E_{**}. Choose a subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$ of $[0,1]$ so that $W_1|[t_{i-1},t_i]$ is smooth for each i. Let $f\colon [0,1] \to [0,1]$ be a smooth function which vanishes for the parameter values $t_0,t_1,\ldots,t_k$ and is positive otherwise; and let

$$W_2(t) = f(t)\Big(\frac{D^2W_1}{dt^2} + R(V,W_1)V\Big)_t .$$

Then

$$-\tfrac{1}{2}E_{**}(W_1,W_2) = \sum 0 + \int_0^1 f(t)\left\|\frac{D^2W_1}{dt^2} + R(V,W_1)V\right\|^2 dt .$$

Since this is zero, it follows that $W_1|[t_{i-1},t_i]$ is a Jacobi field for each i.

Now let $W_2' \in T\Omega_\gamma$ be a field such that $W_2'(t_i) = \Delta_{t_i}\frac{DW_1}{dt}$ for $i = 1,2,\ldots,k-1$. Then

$$-\tfrac{1}{2}E_{**}(W_1,W_2') = \sum_{i=1}^{k-1} \left\|\Delta_{t_i}\frac{DW_1}{dt}\right\|^2 + \int_0^1 0 \, dt = 0$$

Hence $\frac{DW_1}{dt}$ has no jumps. But a solution W_1 of the Jacobi equation is completely determined by the vectors $W_1(t_i)$ and $\frac{DW_1}{dt}(t_i)$. Thus it follows that the k Jacobi fields $W_1|[t_{i-1},t_i]$, $i = 1,\ldots,k$, fit together to give a Jacobi field W_1 which is C^∞-differentiable throughout the

entire unit interval. This completes the proof of 14.1.

It follows that the nullity ν of E_{**} is always finite. For there are only finitely many linearly independent Jacobi fields along γ.

REMARK 14.2. Actually the nullity ν satisfies $0 \leq \nu < n$. Since the space of Jacobi fields which vanish for $t = 0$ has dimension precisely n, it is clear that $\nu \leq n$. We will construct one example of a Jacobi field which vanishes for $t = 0$, but not for $t = 1$. This will imply that $\nu < n$. In fact let $J_t = tV_t$ where $V = \frac{d\gamma}{dt}$ denotes the velocity vector field. Then

$$\frac{DJ}{dt} = 1 \cdot V + t \frac{DV}{dt} = V$$

(Since $\frac{DV}{dt} = 0$), hence $\frac{D^2J}{dt^2} = 0$. Furthermore $R(V,J)V = tR(V,V)V = 0$ since R is skew symmetric in the first two variables. Thus J satisfies the Jacobi equation. Since $J_0 = 0$, $J_1 \neq 0$, this completes the proof.

EXAMPLE 1. Suppose that M is "flat" in the sense that the curvature tensor is identically zero. Then the Jacobi equation becomes $\frac{D^2J}{dt^2} = 0$. Setting $J(t) = \Sigma f^i(t)P_i(t)$ where P_i are parallel, this becomes $\frac{d^2f^i}{dt^2} = 0$. Evidently a Jacobi field along γ can have at most one zero. Thus there are no conjugate points, and E_{**} is non-degenerate.

EXAMPLE 2. Suppose that p and q are antipodal points on the unit sphere S^n, and let γ be a great circle arc from p to q. Then we will see that p and q are conjugate with multiplicity $n-1$. Thus in this example the nullity ν of E_{**} takes its largest possible value. The proof will depend on the following discussion.

Let α be a 1-parameter variation of γ, not necessarily keeping the endpoints fixed, such that each $\bar{\alpha}(u)$ is a geodesic. That is, let

$$\alpha: (-\varepsilon,\varepsilon) \times [0,1] \to M$$

be a C^∞ map such that $\alpha(0,t) = \gamma(t)$, and such that each $\bar{\alpha}(u)$ [given by $\bar{\alpha}(u)(t) = \alpha(u,t)$] is a geodesic.

LEMMA 14.3. If α is such a variation of γ through geodesics, then the variation vector field $W(t) = \frac{\partial \alpha}{\partial u}(0,t)$ is a Jacobi field along γ.

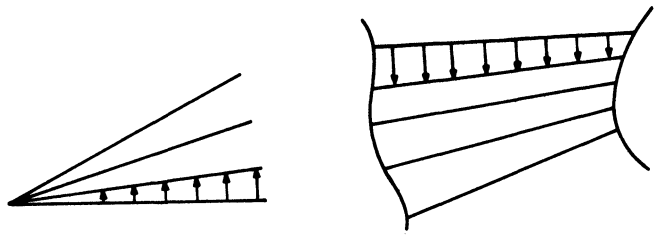
PROOF: If α is a variation of γ through geodesics, then $\frac{D}{dt}\frac{\partial \alpha}{\partial t}$ is identically zero. Hence

$$0 = \frac{D}{\partial u}\frac{D}{\partial t}\frac{\partial \alpha}{\partial t} = \frac{D}{\partial t}\frac{D}{\partial u}\frac{\partial \alpha}{\partial t} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t}$$

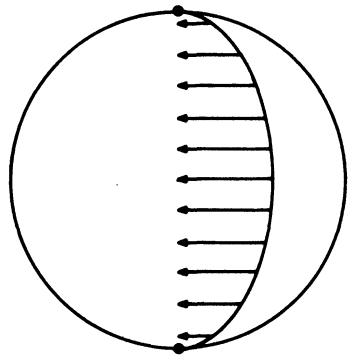
$$= \frac{D^2}{\partial t^2}\frac{\partial \alpha}{\partial u} + R\left(\frac{\partial \alpha}{\partial t}, \frac{\partial \alpha}{\partial u}\right)\frac{\partial \alpha}{\partial t} \quad .$$

(Compare §13.3.) Therefore the variation vector field $\frac{\partial \alpha}{\partial u}$ is a Jacobi field.

Thus one way of obtaining Jacobi fields is to move geodesics around.



Now let us return to the example of two antipodal points on a unit n-sphere. Rotating the sphere, keeping p and q fixed, the variation vector field along the geodesic γ will be a Jacobi field vanishing at p and q. Rotating in n-1 different directions one obtains n-1 linearly



independent Jacobi fields. Thus p and q are conjugate along γ with multiplicity n-1.

LEMMA 14.4. Every Jacobi field along a geodesic $\gamma\colon [0,1] \to M$ may be obtained by a variation of γ through geodesics.

PROOF: Choose a neighborhood U of $\gamma(0)$ so that any two points of U are joined by a unique minimal geodesic which depends differentiably on the endpoints. Suppose that $\gamma(t) \in U$ for $0 \leq t \leq \delta$. We will first construct a Jacobi field W along $\gamma|[0,\delta]$ with arbitrarily prescribed values at $t = 0$ and $t = \delta$. Choose a curve $a\colon (-\varepsilon,\varepsilon) \to U$ so that $a(0) = \gamma(0)$ and so that $\frac{da}{du}(0)$ is any prescribed vector in $TM_{\gamma(0)}$. Similarly choose $b\colon (-\varepsilon,\varepsilon) \to U$ with $b(0) = \gamma(\delta)$ and $\frac{db}{du}(0)$ arbitrary. Now define the variation

$$\alpha\colon (-\varepsilon,\varepsilon) \times [0,\delta] \to M$$

by letting $\bar{\alpha}(u)\colon [0,\delta] \to M$ be the unique minimal geodesic from $a(u)$ to $b(u)$. Then the formula $t \to \frac{\partial\alpha}{\partial u}(0,t)$ defines a Jacobi field with the given end conditions.

Any Jacobi field along $\gamma|[0,\delta]$ can be obtained in this way: If $\mathcal{J}(\gamma)$ denotes the vector space of all Jacobi fields W along γ, then the formula $W \to (W(0), W(\delta))$ defines a linear map

$$\ell\colon \mathcal{J}(\gamma) \to TM_{\gamma(0)} \times TM_{\gamma(\delta)} \ .$$

We have just shown that ℓ is onto. Since both vector spaces have the same dimension $2n$ it follows that ℓ is an isomorphism. I.e., a Jacobi field is determined by its values at $\gamma(0)$ and $\gamma(\delta)$. (More generally a Jacobi field is determined by its values at any two non-conjugate points.) Therefore the above construction yields all possible Jacobi fields along $\gamma|[0,\delta]$.

The restriction of $\bar{\alpha}(u)$ to the interval $[0,\delta]$ is not essential. If u is sufficiently small then, using the compactness of $[0,1]$, $\bar{\alpha}(u)$ can be extended to a geodesic defined over the entire unit interval $[0,1]$. This yields a variation through geodesics:

$$\alpha'\colon (-\varepsilon',\varepsilon') \times [0,1] \to M$$

with any given Jacobi field as variation vector.

REMARK 14.5. This argument shows that in any such neighborhood U the Jacobi fields along a geodesic segment in U are uniquely determined

by their values at the endpoints of the geodesic.

REMARK 14.6. The proof shows also, that there is a neighborhood $(-\delta,\delta)$ of 0 so that if $t \in (-\delta,\delta)$ then $\gamma(t)$ is not conjugate to $\gamma(0)$ along γ. We will see in §15.2 that the set of conjugate points to $\gamma(0)$ along the entire geodesic γ has no cluster points.

§15. The Index Theorem.

The index λ of the Hessian

$$E_{**}: \quad T\Omega_\gamma \times T\Omega_\gamma \longrightarrow \mathbf{R}$$

is defined to be the maximum dimension of a subspace of $T\Omega_\gamma$ on which E_{**} is negative definite. We will prove the following.

> THEOREM 15.1 (Morse). The index λ of E_{**} is equal to the number of points $\gamma(t)$, with $0 < t < 1$, such that $\gamma(t)$ is conjugate to $\gamma(0)$ along γ; each such conjugate point being counted with its multiplicity. This index λ is always finite*.

As an immediate consequence one has:

> COROLLARY 15.2. A geodesic segment $\gamma: [0,1] \to M$ can contain only finitely many points which are conjugate to $\gamma(0)$ along γ.

In order to prove 15.1 we will first make an estimate for λ by splitting the vector space $T\Omega_\gamma$ into two mutually orthogonal subspaces, on one of which E_{**} is positive definite.

Each point $\gamma(t)$ is contained in an open set U such that any two points of U are joined by a unique minimal geodesic which depends differentiably on the endpoints. (See §10.) Choose a subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$ of the unit interval which is sufficiently fine so that each segment $\gamma[t_{i-1}, t_i]$ lies within such an open set U; and so that each $\gamma|[t_{i-1}, t_i]$ is minimal.

Let $T\Omega_\gamma(t_0, t_1, t_2, \ldots, t_k) \subset T\Omega_\gamma$ be the vector space consisting of all vector fields W along γ such that

1) $W|[t_{i-1}, t_i]$ is a Jacobi field along $\gamma|[t_{i-1}, t_i]$ for each i;

2) W vanishes at the endpoints $t = 0$, $t = 1$.

Thus $T\Omega_\gamma(t_0, t_1, \ldots, t_k)$ is a finite dimensional vector space consisting of broken Jacobi fields along γ.

---

* For generalization of this result see: W. Ambrose, The index theorem in Riemannian geometry, Annals of Mathematics, Vol. 73 (1961), pp. 49-86.

Let $T' \subset T\Omega_\gamma$ be the vector space consisting of all vector fields $W \in T\Omega_\gamma$ for which $W(t_0) = 0$, $W(t_1) = 0$, $W(t_2) = 0, \ldots, W(t_k) = 0$.

LEMMA 15.3. The vector space $T\Omega_\gamma$ splits as the direct sum $T\Omega_\gamma(t_0,t_1,\ldots,t_k) \oplus T'$. These two subspaces are mutually perpendicular with respect to the inner product E_{**}. Furthermore, E_{**} restricted to T' is positive definite.

PROOF: Given any vector field $W \in T\Omega_\gamma$ let W_1 denote the unique "broken Jacobi field" in $T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ such that $W_1(t_i) = W(t_i)$ for $i = 0,1,\ldots,k$. It follows from §14.5 that W_1 exists and is unique. Clearly $W - W_1$ belongs to T'. Thus the two subspaces, $T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ and T' generate $T\Omega_\gamma$, and have only the zero vector field in common.

If W_1 belongs to $T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ and W_2 belongs to T', then the second variation formula (13.1) takes the form

$$\tfrac{1}{2}E_{**}(W_1,W_2) = -\sum_t \left\langle W_2(t), \Delta_t \frac{DW_1}{dt} \right\rangle - \int_0^1 \langle W_2, 0 \rangle \, dt = 0 .$$

Thus the two subspaces are mutually perpendicular with respect to E_{**}.

For any $W \in T\Omega_\gamma$ the Hessian $E_{**}(W,W)$ can be interpreted as the second derivative $\dfrac{d^2E \circ \bar{\alpha}}{du^2}(0)$; where $\bar{\alpha}\colon (-\varepsilon,\varepsilon) \to \Omega$ is any variation of γ with variation vector field $\dfrac{d\bar{\alpha}}{dt}(0)$ equal to W. (Compare 13.5.) If W belongs to T' then we may assume that $\bar{\alpha}$ is chosen so as to leave the points $\gamma(t_0),\gamma(t_1),\ldots,\gamma(t_k)$ fixed. In other words we may assume that $\bar{\alpha}(u)(t_i) = \gamma(t_i)$ for $i = 0,1,\ldots,k$.

Proof that $E_{**}(W,W) \geq 0$ for $W \in T'$. Each $\bar{\alpha}(u) \in \Omega$ is a piecewise smooth path from $\gamma(0)$ to $\gamma(t_1)$ to $\gamma(t_2)$ to ... to $\gamma(1)$. But each $\gamma|[t_{i-1},t_i]$ is a minimal geodesic, and therefore has smaller energy than any other path between its endpoints. This proves that

$$E(\bar{\alpha}(u)) \geq E(\gamma) = E(\bar{\alpha}(0)) .$$

Therefore the second derivative, evaluated at $u = 0$, must be ≥ 0.

Proof that $E_{**}(W,W) > 0$ for $W \in T'$, $W \neq 0$. Suppose that $E_{**}(W,W)$ were equal to 0. Then W would lie in the null space of E_{**}. In fact for any $W_1 \in T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ we have already seen that $E_{**}(W_1,W) = 0$. For any $W_2 \in T'$ the inequality

$$0 \leq E_{**}(W + c\,W_2,\ W + c\,W_2) = 2c\,E_{**}(W_2,W) + c^2\,E_{**}(W_2,W_2)$$

for all values of c implies that $E_{**}(W_2,W) = 0$. Thus W lies in the null space. But the null space of E_{**} consists of Jacobi fields. Since T' contains no Jacobi fields other than zero, this implies that $W = 0$.

Thus the quadratic form E_{**} is positive definite on T'. This completes the proof of 15.3.

An immediate consequence is the following:

LEMMA 15.4. The index (or the nullity) of E_{**} is equal to the index (or nullity) of E_{**} restricted to the space $T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ of broken Jacobi fields. In particular (since $T\Omega_\gamma(t_0,t_1,\ldots,t_k)$ is a finite dimensional vector space) the index λ is always finite.

The proof is straightforward.

Let γ_τ denote the restriction of γ to the interval $[0,\tau]$. Thus $\gamma_\tau\colon [0,\tau] \to M$ is a geodesic from $\gamma(0)$ to $\gamma(\tau)$. Let $\lambda(\tau)$ denote the index of the Hessian $\left(E_0^\tau\right)_{**}$ which is associated with this geodesic. Thus $\lambda(1)$ is the index which we are actually trying to compute. First note that:

ASSERTION (1). $\lambda(\tau)$ is a monotone function of τ.

For if $\tau < \tau'$ then there exists a $\lambda(\tau)$ dimensional space $\mathcal{V}$ of vector fields along γ_τ which vanish at $\gamma(0)$ and $\gamma(\tau)$ such that the Hessian $\left(E_0^\tau\right)_{**}$ is negative definite on this vector space. Each vector field in $\mathcal{V}$ extends to a vector field along $\gamma_{\tau'}$ which vanishes identically between $\gamma(\tau)$ and $\gamma(\tau')$. Thus we obtain a $\lambda(\tau)$ dimensional vector space of fields along $\gamma_{\tau'}$ on which $\left(E_0^{\tau'}\right)_{**}$ is negative definite. Hence $\lambda(\tau) \leq \lambda(\tau')$.

ASSERTION (2). $\lambda(\tau) = 0$ for small values of τ.

For if τ is sufficiently small then γ_τ is a minimal geodesic, hence $\lambda(\tau) = 0$ by Lemma 13.6.

Now let us examine the discontinuities of the function $\lambda(\tau)$. First note that $\lambda(\tau)$ is continuous from the left:

ASSERTION (3). For all sufficiently small $\varepsilon > 0$ we have $\lambda(\tau-\varepsilon) = \lambda(\tau)$.

PROOF. According to 15.3 the number $\lambda(1)$ can be interpreted as the index of a quadratic form on a finite dimensional vector space $T\Omega_\gamma(t_0,t_1,\dots,t_k)$. We may assume that the subdivision is chosen so that say $t_i < \tau < t_{i+1}$. Then the index $\lambda(\tau)$ can be interpreted as the index of a corresponding quadratic form H_τ on a corresponding vector space of broken Jacobi fields along γ_τ. This vector space is to be constructed using the subdivision $0 < t_1 < t_2 < \dots < t_i < \tau$ of $[0,\tau]$. Since a broken Jacobi field is uniquely determined by its values at the break points $\gamma(t_i)$, this vector space is isomorphic to the direct sum

$$\Sigma = TM_{\gamma(t_1)} \oplus TM_{\gamma(t_2)} \oplus \dots \oplus TM_{\gamma(t_i)} .$$

Note that this vector space Σ is independent of τ. Evidently the quadratic form H_τ on Σ varies continuously with τ.

Now H_τ is negative definite on a subspace $\mathcal{Y} \subset \Sigma$ of dimension $\lambda(\tau)$. For all τ' sufficiently close to τ it follows that $H_{\tau'}$ is negative definite on $\mathcal{Y}$. Therefore $\lambda(\tau') \geq \lambda(\tau)$. But if $\tau' = \tau - \varepsilon < \tau$ then we also have $\lambda(\tau-\varepsilon) \leq \lambda(\tau)$ by Assertion 1. Hence $\lambda(\tau-\varepsilon) = \lambda(\tau)$.

ASSERTION (4). Let ν be the nullity of the Hessian $\left(E_0^\tau \right)_{**}$. Then for all sufficiently small $\varepsilon > 0$ we have

$$\lambda(\tau+\varepsilon) = \lambda(\tau) + \nu .$$

Thus the function $\lambda(t)$ jumps by ν when the variable t passes a conjugate point of multiplicity ν; and is continuous otherwise. Clearly this assertion will complete the proof of the index theorem.

PROOF that $\lambda(\tau+\varepsilon) \leq \lambda(\tau) + \nu$. Let H_τ and Σ be as in the proof of Assertion 3. Since $\dim \Sigma = ni$ we see that H_τ is positive definite on some subspace $\mathcal{Y}' \subset \Sigma$ of dimension $ni - \lambda(\tau) - \nu$. For all τ' sufficiently close to τ, it follows that $H_{\tau'}$ is positive definite on $\mathcal{Y}'$. Hence

$$\lambda(\tau') \leq \dim \Sigma - \dim \mathcal{Y}' = \lambda(\tau) + \nu .$$

PROOF that $\lambda(\tau+\varepsilon) \geq \lambda(\tau) + \nu$. Let $W_1,\dots,W_{\lambda(\tau)}$ be $\lambda(\tau)$ vector fields along γ_τ, vanishing at the endpoints, such that the matrix

$$\left(\left(E_0^\tau \right)_{**}(W_i,W_j) \right)$$

is negative definite. Let $J_1,\ldots,J_\nu$ be ν linearly independent Jacobi fields along γ_τ, also vanishing at the endpoints. Note that the ν vectors

$$\frac{DJ_h}{dt}(\tau) \in TM_{\gamma(\tau)}$$

are linearly independent. Hence it is possible to choose ν vector fields $X_1,\ldots,X_\nu$ along $\gamma_{\tau+\varepsilon}$, vanishing at the endpoints of $\gamma\tau + \varepsilon$, so that

$$\left(< \frac{DJ_h}{dt}(\tau), X_k(\tau) >\right)$$

is equal to the $\nu \times \nu$ identity matrix. Extend the vector fields W_i and J_h over $\gamma_{\tau+\varepsilon}$ by setting these fields equal to 0 for $\tau \leq t \leq \tau + \varepsilon$.

Using the second variation formula we see easily that

$$\left(E_0^{\tau+\varepsilon}\right)_{**}\left(J_h, W_i\right) = 0$$

$$\left(E_0^{\tau+\varepsilon}\right)_{**}\left(J_h, X_k\right) = 2\delta_{hk} \quad \text{(Kronecker delta)}.$$

Now let c be a small number, and consider the $\lambda(\tau) + \nu$ vector fields

$$W_1,\ldots,W_{\lambda(\tau)},\ c^{-1}J_1 - c\,X_1,\ldots,\ c^{-1}J_\nu - c\,X_\nu$$

along $\gamma_{\tau+\varepsilon}$. We claim that these vector fields span a vector space of dimension $\lambda(\tau) + \nu$ on which the quadratic form $\left(E_0^{\tau+\varepsilon}\right)_{**}$ is negative definite. In fact the matrix of $\left(E_0^{\tau+\varepsilon}\right)_{**}$ with respect to this basis is

$$\begin{pmatrix} \left(\left(E_0^{\tau} \right)_{**}\left(W_i, W_j\right) \right) & c\,A \\ c\,A^t & -4\,I + c^2\,B \end{pmatrix}$$

where A and B are fixed matrices. If c is sufficiently small, this compound matrix is certainly negative definite. This proves Assertion (4).

The index theorem 15.1 clearly follows from the Assertions (2),(3), and (4).

§16. A Finite Dimensional Approximation to Ω^c.

Let M be a connected Riemannian manifold and let p and q be two (not necessarily distinct) points of M. The set $\Omega = \Omega(M;p,q)$ of piecewise C^∞ paths from p to q can be topologized as follows. Let ρ denote the topological metric on M coming from its Riemann metric. Given $\omega, \omega' \in \Omega$ with arc-lengths $s(t)$, $s'(t)$ respectively, define the distance $d(\omega,\omega')$ to be

$$\operatorname*{Max}_{0 \le t \le 1} \rho\big(\omega(t), \omega'(t)\big) + \left[\int_0^1 \left(\frac{ds}{dt} - \frac{ds'}{dt} \right)^2 dt \right]^{\frac{1}{2}} .$$

(The last term is added on so that the energy function,

$$E_a^b(\omega) = \int_a^b \left(\frac{ds}{dt} \right)^2 dt$$

will be a continuous function from Ω to the real numbers.) This metric induces the required topology on Ω.

Given $c > 0$ let Ω^c denote the closed subset $E^{-1}([0,c]) \subset \Omega$ and let $\operatorname{Int} \Omega^c$ denote the open subset $E^{-1}([0,c))$ (where $E = E_0^1 : \Omega \to \mathbf{R}$ is the energy function). We will study the topology of Ω^c by constructing a finite dimensional approximation to it.

Choose some subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$ of the unit interval. Let $\Omega(t_0,t_1,\ldots,t_k)$ be the subspace of Ω consisting of paths $\omega: [0,1] \to M$ such that

1) $\omega(0) = p$ and $\omega(1) = q$,

2) $\omega | [t_{i-1}, t_i]$ is a geodesic for each $i = 1,\ldots,k$.

Finally we define the subspaces

$$\Omega(t_0,t_1,\ldots,t_k)^c = \Omega^c \cap \Omega(t_0,t_1,\ldots,t_k)$$

$$\operatorname{Int} \Omega(t_0,t_1,\ldots,t_k)^c = (\operatorname{Int} \Omega^c) \cap \Omega(t_0,\ldots,t_k) .$$

LEMMA 16.1. Let M be a complete Riemannian manifold; and let c be a fixed positive number such that $\Omega^c \neq \emptyset$. Then for all sufficiently fine subdivisions $(t_0,t_1,\ldots,t_k)$ of $[0,1]$ the set $\operatorname{Int} \Omega(t_0,t_1,\ldots,t_k)^c$ can be given the structure of a smooth finite dimensional manifold in a natural way.

PROOF: Let S denote the ball

$$\{x \in M : \rho(x,p) \leq \sqrt{c}\} \quad .$$

Note that every path $\omega \in \Omega^c$ lies within this subset $S \subset M$. This follows from the inequality $L^2 \leq E \leq c$.

Since M is complete, S is a compact set. Hence by 10.8 there exists $\varepsilon > 0$ so that whenever $x, y \in S$ and $\rho(x,y) < \varepsilon$ there is a unique geodesic from x to y of length $< \varepsilon$; and so that this geodesic depends differentiably on x and y.

Choose the subdivision $(t_0, t_1, \ldots, t_k)$ of $[0,1]$ so that each difference $t_i - t_{i-1}$ is less than ε^2/c. Then for each broken geodesic

$$\omega \in \Omega(t_0, t_1, \ldots, t_k)^c$$

we have

$$\left(L_{t_{i-1}}^{t_i} \omega \right)^2 = (t_i - t_{i-1}) \left(E_{t_{i-1}}^{t_i} \omega \right) \leq (t_i - t_{i-1})(E\,\omega)$$

$$\leq (t_i - t_{i-1})c < \varepsilon^2 \quad .$$

Thus the geodesic $\omega | [t_{i-1}, t_i]$ is uniquely and differentiably determined by the two end points.

The broken geodesic ω is uniquely determined by the $(k-1)$-tuple

$$\omega(t_1), \omega(t_2), \ldots, \omega(t_{k-1}) \in M \times M \times \ldots \times M.$$

Evidently this correspondence

$$\omega \to \big(\omega(t_1), \ldots, \omega(t_{k-1})\big)$$

defines a homeomorphism between $\operatorname{Int} \Omega(t_0, t_1, \ldots, t_k)^c$ and a certain open subset of the $(k-1)$-fold product $M \times \ldots \times M$. Taking over the differentiable structure from this product, this completes the proof of 16.1.

To shorten the notation, let us denote this manifold $\operatorname{Int} \Omega(t_0, t_1, \ldots, t_k)^c$ of broken geodesics by B. Let

$$E' : B \to \mathbf{R}$$

denote the restriction to B of the energy function $E : \Omega \to \mathbf{R}$.

THEOREM 16.2. This function $E' : B \to \mathbf{R}$ is smooth. Furthermore, for each $a < c$ the set $B^a = (E')^{-1}[0,a]$ is compact, and is a deformation retract* of the corresponding set Ω^a. The critical points of E' are precisely the same as the critical points of E in $\operatorname{Int} \Omega^c$: namely the unbroken geodesics from p to q of length less than $\sqrt{c}$. The index [or the nullity] of the Hessian E'_{**} at each such critical point γ is equal to the index [or the nullity] of E_{**} at γ.

Thus the finite dimensional manifold B provides a faithful model for the infinite dimensional path space $\operatorname{Int} \Omega^c$. As an immediate consequence we have the following basic result.

THEOREM 16.3. Let M be a complete Riemannian manifold and let $p,q \in M$ be two points which are not conjugate along any geodesic of length $\leq \sqrt{a}$. Then Ω^a has the homotopy type of a finite CW-complex, with one cell of dimension λ for each geodesic in Ω^a at which E_{**} has index λ.

(In particular it is asserted that Ω^a contains only finitely many geodesics.)

PROOF. This follows from 16.2 together with §3.5.

PROOF of 16.2. Since the broken geodesic $\omega \in B$ depends smoothly on the $(k-1)$-tuple

$$\omega(t_1), \omega(t_2), \ldots, \omega(t_{k-1}) \in M \times \ldots \times M$$

it is clear that the energy $E'(\omega)$ also depends smoothly on this $(k-1)$-tuple. In fact we have the explicit formula

$$E'(\omega) = \sum_{i=1}^{k} \rho\big(\omega(t_{i-1}), \omega(t_i)\big)^2 / (t_i - t_{i-1}) \quad .$$

---

* Similarly B itself is a deformation retract of $\operatorname{Int} \Omega^c$.

For $a < c$ the set B^a is homeomorphic to the set of all $(k-1)$-tuples $(p_1,\dots,p_{k-1}) \in S \times S \times \dots \times S$ such that

$$\sum_{i=1}^{k} \rho(p_{i-1},p_i)^2/(t_i - t_{i-1}) \leq a \quad .$$

(Here it is to be understood that $p_0 = p$, $p_k = q$.) As a closed subset of a compact set, this is certainly compact.

A retraction $r\colon \operatorname{Int} \Omega^c \to B$ is defined as follows. Let $r(\omega)$ denote the unique broken geodesic in B such that each $r(\omega)|[t_{i-1},t_i]$ is a geodesic of length $< \varepsilon$ from $\omega(t_{i-1})$ to $\omega(t_i)$. The inequality

$$\rho\big(p,\omega(t)\big)^2 \leq (L\,\omega)^2 \leq E\,\omega < c$$

implies that $\omega[0,1] \subset S$. Hence the inequality

$$\rho\big(\omega(t_{i-1}),\omega(t_i)\big)^2 \leq (t_i - t_{i-1})\Big(E_{t_{i-1}}^{t_i}\,\omega\Big) < \frac{\varepsilon^2}{c} \cdot c = \varepsilon^2$$

implies that $r(\omega)$ can be so defined.

Clearly $E\big(r(\omega)\big) \leq E(\omega) < c$. This retraction r fits into a 1-parameter family of maps

$$r_u\colon \operatorname{Int} \Omega^c \to \operatorname{Int} \Omega^c$$

as follows. For $t_{i-1} \leq u \leq t_i$ let

$$\begin{cases} r_u(\omega)|[0,t_{i-1}] = r(\omega)|[0,t_{i-1}] \ , \\ r_u(\omega)|[t_{i-1},u] = \text{minimal geodesic from } \omega(t_{i-1}) \text{ to } \omega(u) \ , \\ r_u(\omega)|[u,1] = \omega|[u,1] \ . \end{cases}$$

Then r_0 is the identity map of $\operatorname{Int} \Omega^c$, and $r_1 = r$. It is easily verified that $r_u(\omega)$ is continuous as a function of both variables. This proves that B is a deformation retract of $\operatorname{Int} \Omega^c$.

Since $E\big(r_u(\omega)\big) \leq E(\omega)$ it is clear that each B^a is also a deformation retract of Ω^a.

Every geodesic is also a broken geodesic, so it is clear that every "critical point" of E in $\operatorname{Int} \Omega^c$ automatically lies in the submanifold B. Using the first variation formula (§12.2) it is clear that the critical points of E' are precisely the unbroken geodesics.

Consider the tangent space TB_γ to the manifold B at a geodesic γ. This will be identified with the space $T\Omega_\gamma(t_0,t_1,\dots,t_k)$ of broken

Jacobi fields along γ, as described in §15. This identification can be justified as follows. Let

$$\bar{\alpha}\colon\ (-\varepsilon,\varepsilon) \to B$$

be any variation of γ through broken geodesics. Then the corresponding variation vector field $\frac{\partial\alpha}{\partial u}(0,t)$ along γ is clearly a broken Jacobi field. (Compare §14.3)

Now the statement that the index (or the nullity) of E_{**} at γ is equal to the index (or nullity) of E'_{**} at γ is an immediate consequence of Lemma 15.4. This completes the proof of 16.2.

REMARK. As one consequence of this theorem we obtain an alternative proof of the existence of a minimal geodesic joining two given points p,q of a complete manifold. For if $\Omega^a(p,q)$ is non-vacuous, then the corresponding set B^a will be compact and non-vacuous. Hence the continuous function $E'\colon\ B^a \to \mathbf{R}$ will take on its minimum at some point $\gamma \in B^a$. This γ will be the required minimal geodesic.

§17. The Topology of the Full Path Space.

Let M be a Riemannian manifold with Riemann metric g, and let ρ be the induced topological metric. Let p and q be two (not necessarily distinct) points of M.

In homotopy theory one studies the space Ω^* of all continuous paths

$$\omega: [0,1] \to M$$

from p to q, in the compact open topology. This topology can also be described as that induced by the metric

$$d^*(\omega,\omega') = \operatorname*{Max}_t \rho\big(\omega(t),\omega'(t)\big) .$$

On the other hand we have been studying the space Ω of piecewise C^∞ paths from p to q with the metric

$$d(\omega,\omega') = d^*(\omega,\omega') + \Big[\int_0^1 \Big(\frac{ds}{dt} - \frac{ds'}{dt} \Big)^2 dt \Big]^{\frac{1}{2}} .$$

Since $d \geq d^*$ the natural map

$$i: \Omega \to \Omega^*$$

is continuous.

THEOREM 17.1. This natural map i is a homotopy equivalence between Ω and Ω^*.

[Added June 1968. The following proof is based on suggestions by W. B. Houston, Jr., who has pointed out that my original proof of 17.1 was incorrect. The original proof made use of an alleged homotopy inverse $\Omega^* \to \Omega$ which in fact was not even continuous.]

PROOF: We will use the fact that every point of M has an open neighborhood N which is "geodesically convex" in the sense that any two points of N are joined by a unique minimal geodesic which lies completely within N and depends differentiably on the endpoints. (This result is due to J. H. C. Whitehead. See for example Bishop and Crittenden, "Geometry of

manifolds," p. 246; Helgason, "Differential geometry and symmetric spaces," p. 53; or Hicks, "Notes on differential geometry," p. 134.)

Choose a covering of M by such geodesically convex open sets N_α. Subdividing the interval $[0,1]$ into 2^k equal subintervals $[(j-1)/2^k, j/2^k]$, let Ω_k^* denote the set of all continuous paths ω from p to q which satisfy the following condition: the image under ω of each subinterval $[(j-1)/2^k, j/2^k]$ should be contained in one of the sets N_α of the covering.

Clearly each Ω_k^* is an open subset of the space Ω^* of all paths from p to q, and clearly Ω^* is the union of the sequence of open subsets

$$\Omega_1^* \subset \Omega_2^* \subset \Omega_3^* \subset \cdots .$$

Similarly the corresponding sets

$$\Omega_k = i^{-1}(\Omega_k^*)$$

are open subsets of Ω with union equal to Ω.

We will first show that the natural map

$$(i|\Omega_k) : \Omega_k \to \Omega_k^*$$

is a homotopy equivalence. For each $\omega \in \Omega_k^*$ let $h(\omega) \in \Omega_k$ be the broken geodesic which coincides with ω for the parameter values $t = j/2^k$, $j = 0,1,2,\ldots,2^k$, and which is a minimal geodesic within each intermediate interval $[(j-1)/2^k, j/2^k]$. This construction defines a function

$$h : \Omega_k^* \to \Omega_k ,$$

and it is not difficult to verify that h is continuous.

Just as in the proof of 16.2 on page 91, it can be verified that the composition $(i|\Omega_k) \circ h$ is homotopic to the identity map of Ω_k^* and that the composition $h \circ (i|\Omega_k)$ is homotopic to the identity map of Ω_k. This proves that $i|\Omega_k$ is a homotopy equivalence.

To conclude the proof of 17.1 we appeal to the Appendix. Using Example 1 on page 149 note that the space Ω is the homotopy direct limit of the sequence of subsets Ω_k. Similarly note that Ω^* is the homotopy direct limit of the sequence of subsets Ω_k^*. Therefore, Theorem A (page 150) shows that $i : \Omega \to \Omega^*$ is a homotopy equivalence. This completes the proof.

It is known that the space Ω^* has the homotopy type of a CW-complex. (See Milnor, On spaces having the homotopy type of a CW-complex, Trans. Amer. Math. Soc., Vol. 90 (1959), pp. 272-280.) Therefore

COROLLARY 17.2. Ω has the homotopy type of a CW-complex.

This statement can be sharpened as follows.

THEOREM 17.3. (Fundamental theorem of Morse Theory.) Let M be a complete Riemannian manifold, and let $p,q \in M$ be two points which are not conjugate along any geodesic. Then $\Omega(M;p,q)$ (or $\Omega^*(M;p,q)$) has the homotopy type of a countable CW-complex which contains one cell of dimension λ for each geodesic from p to q of index λ.

The proof is analogous to that of 3.5. Choose a sequence $a_0 < a_1 < a_2 < \dots$ of real numbers which are not critical values of the energy function E, so that each interval (a_{i-1},a_i) contains precisely one critical value. Consider the sequence

$$\Omega^{a_0} \subset \Omega^{a_1} \subset \Omega^{a_2} \subset \dots ;$$

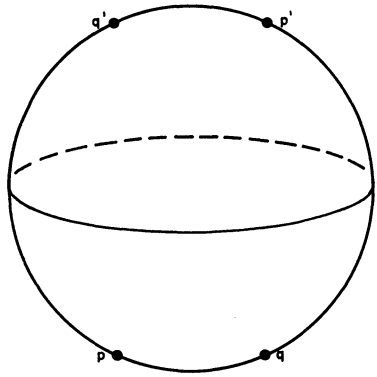
where we may assume that Ω^{a_0} is vacuous. It follows from 16.2 together with 3.3 and 3.7 that each Ω^{a_i} has the homotopy type of $\Omega^{a_{i-1}}$ with a finite number of cells attached: one λ-cell for each geodesic of index λ in $E^{-1}(a_{i-1},a_i)$. Now, just as in the proof of 3.5, one constructs a sequence $K_0 \subset K_1 \subset K_2 \subset \dots$ of CW-complexes with cells of the required description, and a sequence

$$\begin{array}{ccccccc} \Omega^{a_0} & \subset & \Omega^{a_1} & \subset & \Omega^{a_2} & \subset & \dots \\ \downarrow & & \downarrow & & \downarrow & & \\ K_0 & \subset & K_1 & \subset & K_2 & \subset & \dots \end{array}$$

of homotopy equivalences. Letting $f\colon \Omega \to K$ be the direct limit mapping, it is clear that f induces isomorphisms of homotopy groups in all dimensions. Since Ω is known to have the homotopy type of a CW-complex (17.2) it follows from Whitehead's theorem that f is a homotopy equivalence. This completes the proof. [For a different proof, not using 17.2, see p. 149.]

EXAMPLE. The path space of the sphere S^n. Suppose that p and q are two non-conjugate points on S^n. That is, suppose that $q \neq p,p'$ where p' denotes the antipode of p. Then there are denumerably many

geodesics $\gamma_0, \gamma_1, \gamma_2, \ldots$ from p to q, as follows. Let γ_0 denote the short great circle arc from p to q; let γ_1 denote the long great circle arc pq'p'q; let γ_2 denote the arc pqp'q'pq; and so on. The



subscript k denotes the number of times that p or p' occurs in the interior of γ_k.

The index $\lambda(\gamma_k) = \mu_1 + \ldots + \mu_k$ is equal to $k(n-1)$, since each of the points p or p' in the interior is conjugate to p with multiplicity n-1. Therefore we have:

> COROLLARY 17.4. The loop space $\Omega(S^n)$ has the homotopy type of a CW-complex with one cell each in the dimensions $0, n-1, 2(n-1), 3(n-1), \ldots$.

For $n > 2$ the homology of $\Omega(S^n)$ can be computed immediately from this information. Since $\Omega(S^n)$ has non-trivial homology in infinitely many dimensions, we can conclude:

> COROLLARY 17.5. Let M have the homotopy type of S^n, for $n > 2$. Then any two non-conjugate points of M are joined by infinitely many geodesics.

This follows since the homotopy type of $\Omega^*(M)$ (and hence of $\Omega(M)$) depends only on the homotopy type of M. There must be at least one geodesic in $\Omega(M)$ with index 0, at least one with index n-1, $2(n-1)$, $3(n-1)$, and so on.

REMARK. More generally if M is any complete manifold which is not contractible then any two non-conjugate points of M are joined by infinitely many geodesics. Compare p. 484 of J. P. Serre, Homologie singulière des espaces fibrés, Annals of Math. 54 (1951), pp. 425-505.

As another application of 17.4, one can give a proof of the Freudenthal suspension theorem. (Compare §22.3.)

§18. Existence of Non-Conjugate Points.

Theorem 17.3 gives a good description of the space $\Omega(M;p,q)$ providing that the points p and q are not conjugate to each other along any geodesic. This section will justify this result by showing that such non-conjugate points always exist.

Recall that a smooth map $f: N \to M$ between manifolds of the same dimension is critical at a point $x \in N$ if the induced map

$$f_*: TN_x \to TM_{f(x)}$$

of tangent spaces is not 1-1. We will apply this definition to the exponential map

$$\exp = \exp_p: TM_p \to M .$$

(We will assume that M is complete, so that exp is everywhere defined; although this assumption could easily be eliminated.)

> THEOREM 18.1. The point $\exp v$ is conjugate to p along the geodesic γ_v from p to $\exp v$ if and only if the mapping exp is critical at v.

PROOF: Suppose that exp is critical at $v \in TM_p$. Then $\exp_*(X) = 0$ for some non-zero $X \in T(TM_p)_v$, the tangent space at v to TM_p, considered as a manifold. Let $u \to v(u)$ be a path in TM_p such that $v(0) = v$ and $\frac{dv}{du}(0) = X$. Then the map α defined by $\alpha(u,t) = \exp tv(u)$ is a variation through geodesics of the geodesic γ_v given by $t \to \exp tv$. Therefore the vector field W given by $t \to \frac{\partial}{\partial u}(\exp tv(u))\Big|_{u=0}$ is a Jacobi field along γ_v. Obviously $W(0) = 0$. We also have

$$W(1) = \frac{\partial}{\partial u}(\exp v(u))\Big|_{u=0} = \exp_* \frac{dv(u)}{du}(0) = \exp_* X = 0.$$

But this field is not identically zero since

$$\frac{DW}{dt}(0) = \frac{D}{\partial u}\frac{\partial}{\partial t}(\exp tv(u))\Big|_{(0,0)} = \frac{D}{\partial u} v(u)\Big|_{u=0} \neq 0 .$$

So there is a non-trivial Jacobi field along γ_v from p to $\exp v$, vanishing at these points; hence p and $\exp v$ are conjugate along γ_v.

Now suppose that $\exp_*$ is non-singular at v. Choose n independent vectors $X_1,\ldots,X_n$ in $T(TM_p)_v$. Then $\exp_*(X_1),\ldots,\exp_*(X_n)$ are linearly independent. In TM_p choose paths $u \to v_1(u),\ldots,u \to v_n(u)$ with $v_i(0) = v$ and $\frac{dv_i(u)}{du}(0) = X_i$.

Then $\alpha_1,\ldots,\alpha_n$, constructed as above, provide n Jacobi fields $W_1,\ldots,W_n$ along γ_v, vanishing at p. Since the $W_i(1) = \exp_*(X_i)$ are independent, no non-trivial linear combination of the W_i can vanish at $\exp v$. Since n is the dimension of the space of Jacobi fields along γ_v, which vanish at p, clearly no non-trivial Jacobi field along γ_v vanishes at both p and $\exp v$. This completes the proof.

COROLLARY 18.2. Let $p \in M$. Then for almost all $q \in M$, p is not conjugate to q along any geodesic.

PROOF. This follows immediately from 18.1 together with Sard's theorem (§6.1).

§19. Some Relations Between Topology and Curvature.

This section will describe the behavior of geodesics in a manifold with "negative curvature" or with "positive curvature."

LEMMA 19.1. Suppose that $\langle R(A,B)A,B\rangle \leq 0$ for every pair of vectors A,B in the tangent space TM_p and for every $p \in M$. Then no two points of M are conjugate along any geodesic.

PROOF. Let γ be a geodesic with velocity vector field V; and let J be a Jacobi field along γ. Then

$$\frac{D^2J}{dt^2} + R(V,J)V = 0$$

so that

$$\left\langle \frac{D^2J}{dt^2}, J \right\rangle = - \langle R(V,J)V,J \rangle \geq 0.$$

Therefore

$$\frac{d}{dt}\left\langle \frac{DJ}{dt}, J \right\rangle = \left\langle \frac{D^2J}{dt^2}, J \right\rangle + \left\| \frac{DJ}{dt} \right\|^2 \geq 0 \quad .$$

Thus the function $\left\langle \frac{DJ}{dt}, J \right\rangle$ is monotonically increasing, and strictly so if $\frac{DJ}{dt} \neq 0$.

If J vanishes both at 0 and at $t_0 > 0$, then the function $\left\langle \frac{DJ}{dt}, J \right\rangle$ also vanishes at 0 and t_0, and hence must vanish identically throughout the interval $[0,t_0]$. This implies that

$$J(0) = \frac{DJ}{dt}(0) = 0,$$

so that J is identically zero. This completes the proof.

REMARK. If A and B are orthogonal unit vectors at p then the quantity $\langle R(A,B)A,B\rangle$ is called the sectional curvature determined by A and B. It is equal to the Gaussian curvature of the surface

$$(u_1,u_2) \rightarrow \exp_p(u_1A + u_2B)$$

spanned by the geodesics through p with velocity vectors in the subspace spanned by A and B. (See for example, Laugwitz "Differential-Geometrie," p. 101.)

[Intuitively the curvature of a manifold can be described in terms of "optics" within the manifold as follows. Suppose that we think of the geodesics as being the paths of light rays. Consider an observer at p looking in the direction of the unit vector U towards a point $q = \exp(rU)$. A small line segment at q with length L, pointed in a direction corresponding to the unit vector $W \in TM_p$, would appear to the observer as a line segment of length

$$L\Big(1 + \frac{r^2}{6} \langle R(U,W)U,W \rangle + (\text{terms involving higher powers of } r)\Big) .$$

Thus if sectional curvatures are negative then any object appears shorter than it really is. A small sphere of radius ε at q would appear to be an ellipsoid with principal radii $\varepsilon(1 + \frac{r^2}{6}K_1 + \ldots), \ldots, \varepsilon(1 + \frac{r^2}{6}K_n + \ldots)$ where $K_1, K_2, \ldots, K_n$ denote the eigenvalues of the linear transformation $W \to R(U,W)U$. Any small object of volume v would appear to have volume $v\big(1 + \frac{r^2}{6}(K_1 + K_2 + \ldots + K_n) + (\text{higher terms})\big)$ where $K_1 + \ldots + K_n$ is equal to the "Ricci curvature" $K(U,U)$, as defined later in this section.]

Here are some familiar examples of complete manifolds with curvature ≤ 0:

(1) The Euclidean space with curvature 0.

(2) The paraboloid $z = x^2 - y^2$, with curvature < 0.

(3) The hyperboloid of rotation $x^2 + y^2 - z^2 = 1$, with curvature < 0.

(4) The helicoid $x \cos z + y \sin z = 0$, with curvature < 0.

(REMARK. In all of these examples the curvature takes values arbitrarily close to 0. Cf. N. V. Efimov, Impossibility of a complete surface in 3-space whose Gaussian curvature has a negative upper bound, Soviet Math., Vol. 4 (1963), pp. 843-846.)

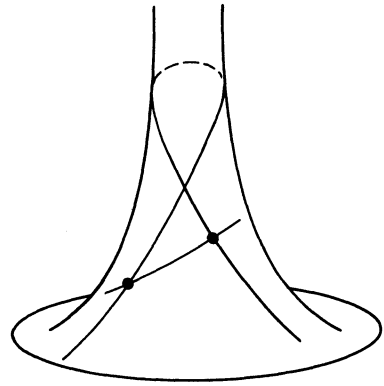
A famous example of a manifold with everywhere negative sectional curvature is the pseudo-sphere

$$z = -\sqrt{1 - x^2 - y^2} + \operatorname{sech}^{-1}\sqrt{x^2 + y^2}, \quad z > 0$$

with the Riemann metric induced from $\mathbf{R}^3$. Here the Gaussian curvature has the constant value -1.

No geodesic on this surface has conjugate points although two geodesics may intersect in more than one point. The pseudo-sphere gives a

non-Euclidean geometry, in which the sum of the angles of any triangle is $< \pi$ radians. This manifold is not complete. In fact a theorem of Hilbert



states that no complete surface of constant negative curvature can be imbedded in $\mathbf{R}^3$. (See Blaschke, "Differential Geometric I," 3rd edn., §96; or Efimov, ibid.)

However, there do exist Riemannian manifolds of constant negative curvature which are complete. (See for example Laugwitz, "Differential and Riemannian geometry," §12.6.2.) Such a manifold can even be compact; for example, a surface of genus ≥ 2. (Compare Hilbert and Cohn-Vossen, "Geometry and the imagination," p. 259.)

> THEOREM 19.2 (Cartan*). Suppose that M is a simply connected, complete Riemannian manifold, and that the sectional curvature $\langle R(A,B)A,B \rangle$ is everywhere ≤ 0. Then any two points of M are joined by a unique geodesic. Furthermore, M is diffeomorphic to the Euclidean space $\mathbf{R}^n$.

PROOF: Since there are no conjugate points, it follows from the index theorem that every geodesic from p to q has index $\lambda = 0$. Thus Theorem 17.3 asserts that the path space $\Omega(M;p,q)$ has the homotopy type of a 0-dimensional CW-complex, with one vertex for each geodesic.

The hypothesis that M is simply connected implies that $\Omega(M;p,q)$ is connected. Since a connected 0-dimensional CW-complex must consist of a single point, it follows that there is precisely one geodesic from p to q.

* See E. Cartan, "Lecons sur la Géométrie des Espaces de Riemann," Paris, 1926 and 1951.

Therefore, the exponential map $\exp_p\colon TM_p \to M$ is one-one and onto. But it follows from 18.1 that $\exp_p$ is non-critical everywhere; so that $\exp_p$ is locally a diffeomorphism. Combining these two facts, we see that $\exp_p$ is a global diffeomorphism. This completes the proof of 19.2.

More generally, suppose that M is not simply connected; but is complete and has sectional curvature ≤ 0. (For example M might be a flat torus $S^1 \times S^1$, or a compact surface of genus ≥ 2 with constant negative curvature.) Then Theorem 19.2 applies to the universal covering space $\tilde{M}$ of M. For it is clear that $\tilde{M}$ inherits a Riemannian metric from M which is geodesically complete, and has sectional curvature ≤ 0.

Given two points $p,q \in M$, it follows that each homotopy class of paths from p to q contains precisely one geodesic.

The fact that $\tilde{M}$ is contractible puts strong restrictions on the topology of M. For example:

> COROLLARY 19.3. If M is complete with $\langle R(A,B)A,B\rangle \leq 0$ then the homotopy groups $\pi_i(M)$ are zero for $i > 1$; and $\pi_1(M)$ contains no element of finite order other than the identity.

PROOF: Clearly $\pi_i(M) = \pi_i(\tilde{M}) = 0$ for $i > 1$. Since $\tilde{M}$ is contractible the cohomology group $H^k(M)$ can be identified with the cohomology group $H^k(\pi_1(M))$ of the group $\pi_1(M)$. (See for example pp. 200-202 of S. T. Hu "Homotopy Theory," Academic Press, 1959.) Now suppose that $\pi_1(M)$ contains a non-trivial finite cyclic subgroup G. Then for a suitable covering space $\hat{M}$ of M we have $\pi_1(\hat{M}) = G$; hence

$$H^k(G) = H^k(\hat{M}) = 0 \quad \text{for} \quad k > n \ .$$

But the cohomology groups of a finite cyclic group are non-trivial in arbitrarily high dimensions. This gives a contradiction; and completes the proof.

Now we will consider manifolds with "positive curvature." Instead of considering the sectional curvature, one can obtain sharper results in this case by considering the Ricci tensor (sometines called the "mean curvature tensor").

DEFINITION. The Ricci tensor at a point p of a Riemannian manifold M is a bilinear pairing

$$K: TM_p \times TM_p \to \mathbf{R}$$

defined as follows. Let $K(U_1,U_2)$ be the trace of the linear transformation

$$W \to R(U_1,W)U_2$$

from TM_p to TM_p. (In classical terminology the tensor K is obtained from R by contraction.) It follows easily from §9.3 that K is symmetric: $K(U_1,U_2) = K(U_2,U_1)$.

The Ricci tensor is related to sectional curvature as follows. Let $U_1,U_2,\ldots,U_n$ be an orthonormal basis for the tangent space TM_p.

ASSERTION. $K(U_n,U_n)$ is equal to the sum of the sectional curvatures $\langle R(U_n,U_i)U_n,U_i \rangle$ for $i = 1,2,\ldots,n-1$.

PROOF: By definition $K(U_n,U_n)$ is equal to the trace of the matrix $\left(\langle R(U_n,U_i)U_n,U_j \rangle \right)$. Since the n-th diagonal term of this matrix is zero, we obtain a sum of $n-1$ sectional curvatures, as asserted.

THEOREM 19.4 (Myers*). Suppose that the Ricci curvature K satisfies

$$K(U,U) \geq (n-1)/r^2$$

for every unit vector U at every point of M; where r is a positive constant. Then every geodesic on M of length $> \pi r$ contains conjugate points; and hence is not minimal.

PROOF: Let $\gamma: [0,1] \to M$ be a geodesic of length L. Choose parallel vector fields $P_1,\ldots,P_n$ along γ which are orthonormal at one point, and hence are orthonormal everywhere along γ. We may assume that P_n points along γ, so that

$$V = \frac{d\gamma}{dt} = L\,P_n, \quad \text{and} \quad \frac{DP_i}{dt} = 0 .$$

Let $W_i(t) = (\sin \pi t)\, P_i(t)$. Then

---

* See S. B. Myers, Riemann manifolds with positive mean curvature, Duke Math. Journal, Vol. 8 (1941), pp. 401-404.

$$\tfrac{1}{2}E_{**}(W_i,W_i) = -\int_0^1 \left\langle W_i, \frac{D^2W_i}{dt^2} + R(V,W_i)V \right\rangle dt$$

$$= \int_0^1 (\sin \pi t)^2 \left(\pi^2 - L^2 \left\langle R(P_n,P_i)P_n,P_i \right\rangle\right) dt.$$

Summing for $i = 1,\ldots,n-1$ we obtain

$$\frac{1}{2}\sum_1^{n-1} E_{**}(W_i,W_i) = \int_0^1 (\sin \pi t)^2\left((n-1)\pi^2 - L^2\, K(P_n,P_n)\right) dt \ .$$

Now if $K(P_n,P_n) \geq (n-1)/r^2$ and $L > \pi r$ then this expression is < 0. Hence $E_{**}(W_i,W_i) < 0$ for some i. This implies that the index of γ is positive, and hence, by the Index Theorem, that γ contains conjugate points.

It follows also that γ is not a minimal geodesic. In fact if $\bar{\alpha}$: $(-\varepsilon,\varepsilon) \to \Omega$ is a variation with variation vector field W_i then

$$\frac{dE(\bar{\alpha}(u))}{du} = 0, \qquad \frac{d^2E(\bar{\alpha}(u))}{du^2} < 0 \ ,$$

for $u = 0$. Hence $E(\bar{\alpha}(u)) < E(\gamma)$ for small values of $u \neq 0$. This completes the proof.

EXAMPLE. If M is a sphere of radius r then every sectional curvature is equal to $1/r^2$. Hence $K(U,U)$ takes the constant value $(n-1)/r^2$. It follows from 19.4 that every geodesic of length $> \pi r$ contains conjugate points: a best possible result.

COROLLARY 19.5. If M is complete, and $K(U,U) \geq (n-1)/r^2 > 0$ for all unit vectors U, then M is compact, with diameter $\leq \pi r$.

PROOF. If $p,q \in M$ let γ be a minimal geodesic from p to q. Then the length of γ must be $\leq \pi r$. Therefore, all points have distance $\leq \pi r$. Since closed bounded sets in a complete manifold are compact, it follows that M itself is compact.

This corollary applies also to the universal covering space $\tilde{M}$ of M. Since $\tilde{M}$ is compact, it follows that the fundamental group $\pi_1(M)$ is finite. This assertion can be sharpened as follows.

THEOREM 19.6. If M is a compact manifold, and if the Ricci tensor K of M is everywhere positive definite, then the path space $\Omega(M;p,q)$ has the homotopy type of a CW-complex having only finitely many cells in each dimension.

PROOF. Since the space consisting of all unit vectors U on M is compact, it follows that the continuous function $K(U,U) > 0$ takes on a minimum, which we can denote by $(n-1)/r^2 > 0$. Then every geodesic $\gamma \in \Omega(M;p,q)$ of length $> \pi r$ has index $\lambda \geq 1$.

More generally consider a geodesic γ of length $> k\pi r$. Then a similar argument shows that γ has index $\lambda \geq k$. In fact for each $i = 1,2,\ldots,k$ one can construct a vector field X_i along γ which vanishes outside of the interval $\left(\frac{i-1}{k}, \frac{i}{k}\right)$, and such that $E_{**}(X_i,X_i) < 0$. Clearly $E_{**}(X_i,X_j) = 0$ for $i \neq j$; so that $X_1,\ldots,X_k$ span a k-dimensional subspace of $T\Omega_\gamma$ on which E_{**} is negative definite.

Now suppose that the points p and q are not conjugate along any geodesic. Then according to § 16.3 there are only finitely many geodesics from p to q of length $\leq k\pi r$. Hence there are only finitely many geodesics with index $< k$. Together with §17.3, this completes the proof.

REMARK. I do not know whether or not this theorem remains true if M is allowed to be complete, but non-compact. The present proof certainly breaks down since, on a manifold such as the paraboloid $z = x^2 + y^2$, the curvature $K(U,U)$ will not be bounded away from zero.

It would be interesting to know which manifolds can carry a metric so that all sectional curvatures are positive. An instructive example is provided by the product $S^m \times S^k$ of two spheres; with $m,k \geq 2$. For this manifold the Ricci tensor is everywhere positive definite. However, the sectional curvatures in certain directions (corresponding to flat tori $S^1 \times S^1 \subset S^m \times S^k$) are zero. It is not known whether or not $S^m \times S^k$ can be remetrized so that all sectional curvatures are positive. The following partial result is known: If such a new metric exists, then it can not be invariant under the involution $(x,y) \to (-x,-y)$ of $S^m \times S^k$. This follows from a theorem of Synge. (See J. L. Synge, On the connectivity of spaces

of positive curvature, Quarterly Journal of Mathematics (Oxford), Vol. 7 (1936), pp. 316-320.

For other theorems relating topology and curvature, the following sources are useful.

K. Yano and S. Bochner, "Curvature and Betti Numbers," Annals Studies, No 32, Princeton, 1953.

S. S. Chern, On curvature and characteristic classes of a Riemann manifold, Abh. Math. Sem., Hamburg, Vol. 20 (1955), pp. 117-126.

M. Berger, Sur certaines variétés Riemanniennes à courbure positive, Comptes Rendus Acad. Sci., Paris, Vol. 247 (1958), pp. 1165-1168.

S. I. Goldberg, "Curvature and Homology," Academic Press, 1962.

PART IV.

APPLICATIONS TO LIE GROUPS AND SYMMETRIC SPACES

§20. Symmetric Spaces.

A symmetric space is a connected Riemannian manifold M such that, for each $p \in M$ there is an isometry $I_p: M \to M$ which leaves p fixed and reverses geodesics through p, i.e., if γ is a geodesic and $\gamma(0) = p$ then $I_p(\gamma(t)) = \gamma(-t)$.

LEMMA 20.1 Let γ be a geodesic in M, and let $p = \gamma(0)$ and $q = \gamma(c)$. Then $I_q I_p(\gamma(t)) = \gamma(t + 2c)$ (assuming $\gamma(t)$ and $\gamma(t + 2c)$ are defined). Moreover, $I_q I_p$ preserves parallel vector fields along γ.

PROOF: Let $\gamma'(t) = \gamma(t + c)$. Then γ' is a geodesic and $\gamma'(0) = q$. Therefore $I_q I_p(\gamma(t)) = I_q(\gamma(-t)) = I_q(\gamma'(-t - c)) = \gamma'(t + c) = \gamma(t + 2c)$.

If the vector field V is parallel along γ then $I_{p*}(V)$ is parallel (since I_p is an isometry) and $I_{p*}V(0) = -V(0)$; therefore $I_{p*}V(t) = -V(-t)$. Therefore $I_{q*}\, I_{p*}(V(t)) = V(t + 2c)$.

COROLLARY 20.2. M is complete.

Since 20.1 shows that geodesics can be indefinitely extended.

COROLLARY 20.3. I_p is unique.

Since any point is joined to p by a geodesic.

COROLLARY 20.4. If U,V and W are parallel vector fields along γ then $R(U,V)W$ is also a parallel field along γ.

PROOF. If X denotes a fourth parallel vector field along γ, note that the quantity $< R(U,V)W,X >$ is constant along γ. In fact,

given $p = \gamma(0)$, $q = \gamma(c)$, consider the isometry $T = I_{\gamma(c/2)}I_p$ which carries p to q. Then

$$\langle R(U_q,V_q)W_q,X_q\rangle = \langle R(T_*U_p,T_*V_p)T_*W_p,T_*X_p\rangle$$

by 20.1. Since T is an isometry, this quantity is equal to $\langle R(U_p,V_p)W_p,X_p\rangle$. Thus $\langle R(U,V)W,X\rangle$ is constant for every parallel vector field X. It clearly follows that $R(U,V)W$ is parallel.

Manifolds with the property of 20.4 are called <u>locally symmetric</u>. (A classical theorem, due to Cartan states that a complete, simply connected, locally symmetric manifold is actually symmetric.)

In any locally symmetric manifold the Jacobi differential equations have simple explicit solutions. Let γ: $\mathbf{R} \to M$ be a geodesic in a locally symmetric manifold. Let $V = \frac{d\gamma}{dt}(0)$ be the velocity vector at $p = \gamma(0)$. Define a linear transformation

$$K_V\colon\ TM_p \to TM_p$$

by* $K_V(W) = R(V,W)V$. Let $e_1,\ldots,e_n$ denote the eigenvalues of K_V.

THEOREM 20.5. The conjugate points to p along γ are the points $\gamma(\pi k/\sqrt{e_i})$ where k is any non-zero integer, and e_i is any positive eigenvalue of K_V. The multiplicity of $\gamma(t)$ as a conjugate point is equal to the number of e_i such that t is a multiple of $\pi/\sqrt{e_i}$.

PROOF: First observe that K_V is self-adjoint:

$$\langle K_V(W),W'\rangle = \langle W,K_V(W')\rangle .$$

This follows immediately from the symmetry relation

$$\langle R(V,W)V',W'\rangle = \langle R(V',W')V,W\rangle .$$

Therefore we may choose an orthonormal basis $U_1,\ldots,U_n$ for M_p so that

$$K_V(U_i) = e_iU_i ,$$

where $e_1,\ldots,e_n$ are the eigenvalues. Extend the U_i to vector fields along γ by parallel translation. Then since M is locally symmetric,

---

* K_V should not be confused with the Ricci tensor of §19.

the condition

$$R(V,U_i)V = e_iU_i$$

remains true everywhere along γ. Any vector field W along γ may be expressed uniquely as

$$W(t) = w_1(t)U_1(t) + \ldots + w_n(t)U_n(t) .$$

Then the Jacobi equation $\frac{D^2W}{dt^2} + K_V(W) = 0$ takes the form

$$\sum \frac{d^2w_i}{dt^2} U_i + \sum e_iw_iU_i = 0.$$

Since the U_i are everywhere linearly independent this is equivalent to the system of n equations

$$\frac{d^2w_i}{dt^2} + e_iw_i = 0 .$$

We are interested in solutions that vanish at $t = 0$. If $e_i > 0$ then

$$w_i(t) = c_i \sin(\sqrt{e_i}\, t), \text{ for some constant } c_i.$$

Then the zeros of $w_i(t)$ are at the multiples of $t = \pi/\sqrt{e_i}$.

If $e_i = 0$ then $w_i(t) = c_it$ and if $e_i < 0$ then $w_i(t) = c_i \sinh(\sqrt{|e_i|}\, t)$ for some constant c_i. Thus if $e_i \leq 0$, $w_i(t)$ vanishes only at $t = 0$. This completes the proof of 20.5.

§21. Lie Groups as Symmetric Spaces.

In this section we consider a Lie group G with a Riemannian metric which is invariant both under left translations

$$L_\tau: G \to G, \quad L_\tau(\sigma) = \tau\sigma$$

and right translation, $R_\tau(\sigma) = \sigma\tau$. If G is commutative such a metric certainly exists. If G is compact then such a metric can be constructed as follows: Let $\langle , \rangle$ be any Riemannian metric on G, and Let μ denote the Haar measure on G. Then μ is right and left invariant. Define a new inner product $\langle\langle , \rangle\rangle$ on G by

$$\langle\langle V,W \rangle\rangle = \int_{G\times G} \langle L_{\sigma *}R_{\tau *}(V),\ L_{\sigma_*}R_{\tau *}(W) \rangle \ d\mu(\sigma)\ d\mu(\tau) \ .$$

Then $\langle\langle , \rangle\rangle$ is left and right invariant.

LEMMA 21.1 If G is a Lie group with a left and right invariant metric, then G is a symmetric space. The reflection I_τ in any point $\tau \in G$ is given by the formula $I_\tau(\sigma) = \tau\sigma^{-1}\tau$.

PROOF: By hypothesis L_τ and R_τ are isometries. Define a map $I_e: G \to G$ by

$$I_e(\sigma) = \sigma^{-1} \ .$$

Then $I_{e*}: TG_e \to TG_e$ reverses the tangent space of e; so is certainly an isometry on this tangent space. Now the identity

$$I_e = R_{\sigma^{-1}} I_e L_{\sigma^{-1}}$$

shows that $I_{e*}: TG_\sigma \to TG_{\sigma^{-1}}$ is an isometry for any $\sigma \in G$. Since I_e reverses the tangent space at e, it reverses geodesics through e.

Finally, defining $I_\tau(\sigma) = \tau\sigma^{-1}\tau$, the identity $I_\tau = R_\tau I_e R_\tau^{-1}$ shows that each I_τ is an isometry which reverses geodesics through τ.

A 1-parameter subgroup of G is a C^∞ homomorphism of $\mathbf{R}$ into G. It is well known that a 1-parameter subgroup of G is determined by its tangent vector at e. (Compare Chevalley, "Theory of Lie Groups," Princeton, 1946.)

LEMMA 21.2. The geodesics γ in G with $\gamma(0) = e$ are precisely the one-parameter subgroups of G.

PROOF: Let $\gamma: \mathbf{R} \to G$ be a geodesic with $\gamma(0) = e$. By Lemma 20.1 the map $I_{\gamma(t)}I_e$ takes $\gamma(u)$ into $\gamma(u + 2t)$. Now $I_{\gamma(t)}I_e(\sigma) = \gamma(t)\sigma\,\gamma(t)$ so $\gamma(t)\gamma(u)\gamma(t) = \gamma(u + 2t)$. By induction it follows that $\gamma(nt) = \gamma(t)^n$ for any integer n. If t'/t'' is rational so that $t' = n't$ and $t'' = n''t$ for some t and some integers n' and n'' then $\gamma(t' + t'') = \gamma(t)^{n'+n''} = \gamma(t')\gamma(t'')$. By continuity γ is a homomorphism.

Now let $\gamma: \mathbf{R} \to G$ be a 1-parameter subgroup. Let γ' be the geodesic through e such that the tangent vector of γ' at e is the tangent vector of γ at e. We have just seen that γ' is a 1-parameter subgroup. Hence $\gamma' = \gamma$. This completes the proof.

A vector field X on a Lie group G is called <u>left invariant</u> if and only if $(L_a)_*(X_b) = X_{a\cdot b}$ for every a and b in G. If X and Y are left invariant then $[X,Y]$ is also. The Lie algebra $\mathfrak{g}$ of G is the vector space of all left invariant vector fields, made into an algebra by the bracket $[\ ]$.

$\mathfrak{g}$ is actually a Lie algebra because the Jacobi identity

$$[[X,Y],Z] + [[Y,Z],X] + [[Z,X],Y] = 0$$

holds for all (not necessarily left invariant) vector fields X,Y and Z.

THEOREM 21.3. Let G be a Lie group with a left and right invariant Riemannian metric. If X,Y,Z and W are left invariant vector fields on G then:

a) $\langle [X,Y],Z\rangle = \langle X,[Y,Z]\rangle$

b) $R(X,Y)Z = \frac{1}{4}[[X,Y],Z]$

c) $\langle R(X,Y)Z,W\rangle = \frac{1}{4}\langle [X,Y],[Z,W]\rangle$.

PROOF: As in §8 we will use the notation $X \vdash Y$ for the covariant derivative of Y in the direction X. For any left invariant X the identity

$$X \vdash X = 0$$

is satisfied, since the integral curves of X are left translates of 1-parameter subgroups, and therefore are geodesics.

Therefore

$$(X + Y) \vdash (X + Y) = (X \vdash X) + (X \vdash Y) + (Y \vdash X) + (Y \vdash Y)$$

is zero; hence

$$X \vdash Y + Y \vdash X = 0 .$$

On the other hand

$$X \vdash Y - Y \vdash X = [X,Y]$$

by §8.5. Adding these two equations we obtain:

d) $2X \vdash Y = [X,Y]$.

Now recall the identity

$$Y \langle X,Z \rangle = \langle Y \vdash X,Z \rangle + \langle X,Y \vdash Z \rangle .$$

(See §8.4.) The left side of this equation is zero, since $\langle X,Z \rangle$ is constant. Substituting formula (d) in this equation we obtain

$$0 = \langle [Y,X],Z \rangle + \langle X,[Y,Z] \rangle .$$

Finally, using the skew commutativity of $[Y,X]$, we obtain the required formula*

(a) $$\langle [X,Y],Z \rangle = \langle X,[Y,Z] \rangle .$$

By definition, $R(X,Y)Z$ is equal to

$$- X \vdash (Y \vdash Z) + Y \vdash (X \vdash Z) + [X,Y] \vdash Z.$$

Substituting formula (d), this becomes

$$- \tfrac{1}{4}[X,[Y,Z]] + \tfrac{1}{4}[Y,[X,Z]] + \tfrac{1}{2}[[X,Y],Z]$$

Using the Jacobi identity, this yields the required formula

(b) $$R(X,Y)Z = \tfrac{1}{4}[[X,Y],Z] .$$

The formula (c) follows from (a) and (b) .

---

* It follows that the tri-linear function $X,Y,Z \to \langle [X,Y],Z \rangle$ is skew-symmetric in all three variables. Thus one obtains a left invariant differential 3-form on G, representing an element of the de Rham cohomology group $H^3(G)$. In this way Cartan was able to prove that $H^3(G) \neq 0$ if G is a non-abelian compact connected Lie group. (See E. Cartan, "La Topologie des Espaces Représentatives des Groupes de Lie," Paris, Hermann, 1936.)

COROLLARY 21.4. The sectional curvature $\langle R(X,Y)X,Y\rangle = \frac{1}{4}\langle [X,Y],[X,Y]\rangle$ is always ≥ 0. Equality holds if and only if $[X,Y] = 0$.

Recall that the center $\mathfrak{c}$ of a Lie algebra $\mathfrak{g}$ is defined to be the set of $X \in \mathfrak{g}$ such that $[X,Y] = 0$ for all $Y \in \mathfrak{g}$.

COROLLARY 21.5. If G has a left and right invariant metric, and if the Lie algebra $\mathfrak{g}$ has trivial center, then G is compact, with finite fundamental group.

PROOF: This follows from Meyer's theorem (§19). Let X_1 be any unit vector in $\mathfrak{g}$ and extend to a orthonormal basis $X_1,\ldots,X_n$. The Ricci curvature

$$K(X_1,X_1) = \sum_{i=1}^{n} \langle R(X_1,X_i)X_1,X_i\rangle$$

must be strictly positive, since $[X_1,X_i] \neq 0$ for some i. Furthermore $K(X_1,X_1)$ is bounded away from zero, since the unit sphere in $\mathfrak{g}$ is compact. Therefore, by Corollary 19.5, the manifold G is compact.

This result can be sharpened slightly as follows.

COROLLARY 21.6. A simply connected Lie group G with left and right invariant metric splits as a Cartesian product $G' \times \mathbf{R}^k$ where G' is compact and $\mathbf{R}^k$ denotes the additive Lie group of some Euclidean space. Furthermore, the Lie algebra of G' has trivial center.

Conversely it is clear that any such product $G' \times \mathbf{R}^k$ possesses a left and right invariant metric.

PROOF. Let $\mathfrak{c}$ be the center of the Lie algebra $\mathfrak{g}$ and let

$$\mathfrak{g}' = \{X \in \mathfrak{g} : \langle X,C\rangle = 0 \text{ for all } C \in \mathfrak{c}\}$$

be the orthogonal complement of $\mathfrak{c}$. Then $\mathfrak{g}'$ is a Lie sub-algebra. For if $X,Y \in \mathfrak{g}'$ and $C \in \mathfrak{c}$ then

$$\langle [X,Y],C\rangle = \langle X,[Y,C]\rangle = 0;$$

hence $[X,Y] \in \mathfrak{g}'$. It follows that $\mathfrak{g}$ splits as a direct sum $\mathfrak{g}' \oplus \mathfrak{c}$ of Lie algebras. Hence G splits as a Cartesian product $G' \times G''$; where G' is compact by 21.5 and G'' is simply connected and abelian, hence isomorphic

to some $\mathbf{R}^k$. (See Chevalley, "Theory of Lie Groups.") This completes the proof.

> THEOREM 21.7 (Bott). Let G be a compact, simply connected Lie group. Then the loop space $\Omega(G)$ has the homotopy type of a CW-complex with no odd dimensional cells, and with only finitely many λ-cells for each even value of λ.

Thus the λ-th homology groups of $\Omega(G)$ is zero for λ odd, and is free abelian of finite rank for λ even.

REMARK 1. This CW-complex will always be infinite dimensional. As an example, if G is the group S^3 of unit quaternions, then we have seen that the homology group $H_i\Omega(S^3)$ is infinite cyclic for all even values of i.

REMARK 2. This theorem remains true even for a non-compact group. In fact any connected Lie group contains a compact subgroup as deformation retract. (See K. Iwasawa, On some types of topological groups, Annals of Mathematics 50 (1949), Theorem 6.)

PROOF of 21.7. Choose two points p and q in G which are not conjugate along any geodesic. By Theorem 17.3, $\Omega(G;p,q)$ has the homotopy type of a CW-complex with one cell of dimension λ for each geodesic from p to q of index λ. By §19.4 there are only finitely many λ-cells for each λ. Thus it only remains to prove that the index λ of a geodesic is always even.

Consider a geodesic γ starting at p with velocity vector

$$V = \frac{d\gamma}{dt}(0) \in TG_p \cong \mathfrak{g} \ .$$

According to §20.5 the conjugate points of p on γ are determined by the eigenvalues of the linear transformation

$$K_V: \ TG_p \to TG_p \ ,$$

defined by

$$K_V(W) = R(V,W)V = \tfrac{1}{4}[[V,W],V] \ .$$

Defining the adjoint homomorphism

$$\mathrm{Ad}\ V: \ \mathfrak{g} \to \mathfrak{g}$$

by

$$\text{Ad } V(W) = [V,W]$$

we have

$$K_V = -\tfrac{1}{4}\,(\text{Ad } V) \circ (\text{Ad } V) \quad .$$

The linear transformation Ad V is skew-symmetric; that is

$$\langle \text{Ad } V(W), W' \rangle = - \langle W, \text{Ad } V(W') \rangle \quad .$$

This follows immediately from the identity 21.3a. Therefore we can choose an orthonormal basis for $\mathfrak{G}$ so that the matrix of Ad V takes the form

$$\begin{pmatrix} 0 & a_1 & & & & \\ -a_1 & 0 & & & & \\ & & 0 & a_2 & & \\ & & -a_2 & 0 & & \\ & & & & \ddots & \\ & & & & & \ddots \end{pmatrix} .$$

It follows that the composite linear transformation $(\text{Ad } V)\circ(\text{Ad } V)$ has matrix

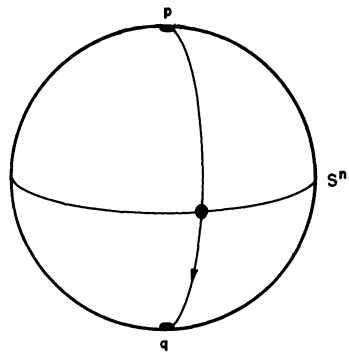
$$\begin{pmatrix} -a_1^2 & & & & & \\ & -a_1^2 & & & & \\ & & -a_2^2 & & & \\ & & & -a_2^2 & & \\ & & & & \ddots & \\ & & & & & \ddots \end{pmatrix} .$$

Therefore the non-zero eigenvalues of $K_V = -\frac{1}{4}(\text{Ad } V)^2$ are positive, and occur in pairs.

It follows from 20.5 that the conjugate points of p along γ also occur in pairs. In other words every conjugate point has even multiplicity. Together with the Index Theorem, this implies that the index λ of any geodesic from p to q is even. This completes the proof.

§22. Whole Manifolds of Minimal Geodesics.

So far we have used a path space $\Omega(M;p,q)$ based on two points $p,q \in M$ which are in "general position." However, Bott has pointed out that very useful results can be obtained by considering pairs p,q in some special position. As an example let M be the unit sphere S^{n+1}, and let p,q be antipodal points. Then there are infinitely many minimal geodesics from p to q. In fact the space Ω^{π^2} of minimal geodesics forms a smooth manifold of dimension n which can be identified with the equator $S^n \subset S^{n+1}$. We will see that this space of minimal geodesics provides a



fairly good approximation to the entire loop space $\Omega(S^{n+1})$.

Let M be a complete Riemannian manifold, and let $p,q \in M$ be two points with distance $\rho(p,q) = \sqrt{d}$.

> THEOREM 22.1. If the space Ω^d of minimal geodesics from p to q is a topological manifold, and if every non-minimal geodesic from p to q has index $\geq \lambda_0$, then the relative homotopy group $\pi_i(\Omega,\Omega^d)$ is zero for $0 \leq i < \lambda_0$.

It follows that the inclusion homomorphism

$$\pi_i(\Omega^d) \to \pi_i(\Omega)$$

is an isomorphism for $i \leq \lambda_0 - 2$. But it is well known that the homotopy group $\pi_i(\Omega)$ is isomorphic to $\pi_{i+1}(M)$ for all values of i. (Compare S. T. Hu, "Homotopy Theory," Academic Press, 1959, p. 111; together with §17.1.)

Thus we obtain:

COROLLARY 22.2. With the same hypotheses, $\pi_i(\Omega^d)$ is isomorphic to $\pi_{i+1}(M)$ for $0 \leq i \leq \lambda_0 - 2$.

Let us apply this corollary to the case of two antipodal points on the $(n+1)$-sphere. Evidently the hypotheses are satisfied with $\lambda_0 = 2n$. For any non-minimal geodesic must wind one and a half times around S^{n+1}; and contain two conjugate points, each of multiplicity n, in its interior. This proves the following.

COROLLARY 22.3. (The Freudenthal suspension theorem.) The homotopy group $\pi_i(S^n)$ is isomorphic to $\pi_{i+1}(S^{n+1})$ for $i \leq 2n-2$.

Theorem 22.1 also implies that the homology groups of the loop space Ω are isomorphic to those of Ω^d in dimensions $\leq \lambda_0 - 2$. This fact follows from 22.1 together with the relative Hurewicz theorem. (See for example Hu, p. 306. Compare also J. H. C. Whitehead, Combinatorial homotopy I, Theorem 2.)

The rest of §22 will be devoted to the proof of Theorem 22.1. The proof will be based on the following lemma, which asserts that the condition "all critical points have index $\geq \lambda_0$" remains true when a function is jiggled slightly.

Let K be a compact subset of the Euclidean space $\mathbf{R}^n$; let U be a neighborhood of K; and let

$$f\colon U \to \mathbf{R}$$

be a smooth function such that all critical points of f in K have index $\geq \lambda_0$.

LEMMA 22.4. If $g\colon U \to \mathbf{R}$ is any smooth function which is "close" to f, in the sense that

$$\left|\frac{\partial g}{\partial x_i} - \frac{\partial f}{\partial x_i}\right| < \varepsilon, \quad \left|\frac{\partial^2 g}{\partial x_i \partial x_j} - \frac{\partial^2 f}{\partial x_i \partial x_j}\right| < \varepsilon, \quad (i,j = 1,\ldots,n)$$

uniformly throughout K, for some sufficiently small constant ε, then all critical points of g in K have index $\geq \lambda_0$.

(Note that f is allowed to have degenerate critical points. In the application, g will be a nearby function without degenerate critical points.)

PROOF of 22.4. The first derivatives of g are roughly described by the single real valued function

$$k_g(x) = \sum_i \left|\frac{\partial g}{\partial x_i}\right| \geq 0$$

on U; which vanishes precisely at the critical points of g. The second derivatives of g can be roughly described by n continuous functions

$$e_g^1, \ldots, e_g^n: U \to \mathbf{R} ,$$

as follows. Let

$$e_g^1(x) \leq e_g^2(x) \leq \ldots \leq e_g^n(x)$$

denote the n eigenvalues of the matrix $\left(\frac{\partial^2 g}{\partial x_i \partial x_j}\right)$. Thus a critical point x of g has index $\geq \lambda$ if and only if the number $e_g^\lambda(x)$ is negative.

The continuity of the functions e_g^λ follows from the fact that the λ-th eigenvalue of a symmetric matrix depends continuously on the matrix*. This can be proved, for example, using the fact that the roots of a complex polynomial of degree n vary continuously with the coefficient of the polynomial. (Rouché's theorem.)

Let $m_g(x)$ denote the larger of the two numbers $k_g(x)$ and $-e_g^{\lambda_0}(x)$. Similarly let $m_f(x)$ denote the larger of the corresponding numbers $k_f(x)$ and $-e_f^{\lambda_0}(x)$. The hypothesis that all critical points of f in K have index $\geq \lambda_0$ implies that $-e_f^{\lambda_0}(x) > 0$ whenever $k_f(x) = 0$. In other words $m_f(x) > 0$ for all $x \in K$.

Let $\delta > 0$ denote the minimum of m_f on K. Now suppose that g is so close to f that

$$(*) \qquad \left|k_g(x) - k_f(x)\right| < \delta, \qquad \left|e_g^{\lambda_0}(x) - e_f^{\lambda_0}(x)\right| < \delta$$

for all $x \in K$. Then $m_g(x)$ will be positive for $x \in K$; hence every critical point of g in K will have index $\geq \lambda_0$.

---

* This statement can be sharpened as follows. Consider two $n \times n$ symmetric matrices. If corresponding entries of the two matrices differ by at most ε, then corresponding eigenvalues differ by at most $n\varepsilon$. This can be proved using Courant's minimax definition of the λ-th eigenvalue. (See §1 of Courant, Über die Abhängigkeit der Schwingungszahlen einer Membran..., Nachrichten, Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math. Phys. Klasse 1919, pp. 255-264.)

To complete the proof of 22.4, it is only necessary to show that the inequalities (*) will be satisfied providing that

$$\left|\frac{\partial g}{\partial x_i} - \frac{\partial f}{\partial x_i}\right| < \varepsilon \quad \text{and} \quad \left|\frac{\partial^2 g}{\partial x_i \partial x_j} - \frac{\partial^2 f}{\partial x_i \partial x_j}\right| < \varepsilon$$

for sufficiently small ε. This follows by a uniform continuity argument which will be left to the reader (or by the footnote above).

We will next prove an analogue of Theorem 22.1 for real valued functions on a manifold.

Let $f\colon M \to \mathbf{R}$ be a smooth real valued function with minimum 0, such that each $M^c = f^{-1}[0,c]$ is compact.

LEMMA 22.5. If the set M^0 of minimal points is a manifold, and if every critical point in $M - M^0$ has index $\geq \lambda_0$, then $\pi_r(M,M^0) = 0$ for $0 \leq r < \lambda_0$.

PROOF: First observe that M^0 is a retract of some neighborhood $U \subset M$. In fact Hanner has proved that any manifold M^0 is an absolute neighborhood retract. (See Theorem 3.3 of O. Hanner, Some theorems on absolute neighborhood retracts, Arkiv för Matematik, Vol. 1 (1950), pp. 389-408.) Replacing U by a smaller neighborhood if necessary, we may assume that each point of U is joined to the corresponding point of M^0 by a unique minimal geodesic. Thus U can be deformed into M^0 within M.

Let I^r denote the unit cube of dimension $r < \lambda_0$, and let

$$h\colon (I^r, \dot{I}^r) \to (M, M^0)$$

be any map. We must show that h is homotopic to a map h' with $h'(I^r) \subset M^0$.

Let c be the maximum of f on $h(I^r)$. Let $3\delta > 0$ be the minimum of f on the set $M - U$. (The function f has a minimum on $M - U$ since each subset $M^{\bar{c}} - U$ is compact.)

Now choose a smooth function

$$g\colon M^{c+2\delta} \to \mathbf{R}$$

which approximates f closely, but has no degenerate critical points. This is possible by §6.8. To be more precise the approximation should be so close that:

(1) $|f(x) - g(x)| < \delta$ for all $x \in M^{c+2\delta}$; and

(2) The index of g at each critical point which lies in the compact set $f^{-1}[\delta, c+2\delta]$ is $\geq \lambda_0$.

It follows from Lemma 22.4 that any g which approximates f sufficiently closely, the first and second derivatives also being approximated, will satisfy (2). In fact the compact set $f^{-1}[\delta, c+2\delta]$ can be covered by finitely many compact set K_i, each of which lies in a coordinate neighborhood. Lemma 22.4 can then be applied to each K_i.

The proof of 22.5 now proceeds as follows. The function g is smooth on the compact region $g^{-1}[2\delta, c+\delta] \subset f^{-1}[\delta, c+2\delta]$, and all critical points are non-degenerate, with index $\geq \lambda_0$. Hence the manifold $g^{-1}(-\infty, c+\delta]$ has the homotopy type of $g^{-1}(-\infty, 2\delta]$ with cells of dimension $\geq \lambda_0$ attached.

Now consider the map

$$h: \ I^r, \dot{I}^r \to M^c, M^0 \subset g^{-1}(-\infty, c+\delta], M^0 \ .$$

Since $r < \lambda_0$ it follows that h is homotopic within $g^{-1}(-\infty, c+\delta], M^0$ to a map

$$h': \ I^r, \dot{I}^r \to g^{-1}(-\infty, 2\delta], M^0 \ .$$

But this last pair is contained in (U, M^0); and U can be deformed into M^0 within M. It follows that h' is homotopic within (M, M^0) to a map $h'': \ I^r, \dot{I}^r \to M^0, M^0$. This completes the proof of 22.5.

The original theorem, 22.1, now can be proved as follows. Clearly it is sufficient to prove that

$$\pi_i(\text{Int } \Omega^c, \Omega^d) = 0$$

for arbitrarily large values of c. As in §16 the space $\text{Int } \Omega^c$ contains a smooth manifold $\text{Int } \Omega^c(t_0, t_1, \ldots, t_k)$ as deformation retract. The space Ω^d of minimal geodesics is contained in this smooth manifold.

The energy function $E: \Omega \to \mathbf{R}$, when restricted to $\text{Int } \Omega^c(t_0, t_1, \ldots, t_k)$, almost satisfies the hypothesis of 22.5. The only difficulty is that $E(\omega)$ ranges over the interval $d \leq E < c$, instead of the required interval $[0, \infty)$. To correct this, let

$$F: \ [d, c) \to [0, \infty)$$

be any diffeomorphism.

Then

$$F \circ E: \quad \text{Int } \Omega^c(t_0,t_1,\ldots,t_k) \to \mathbf{R}$$

satisfies the hypothesis of 22.5. Hence

$$\pi_i(\text{Int } \Omega^c(t_0,\ldots,t_k),\Omega^d) \cong \pi_i(\text{Int } \Omega^c,\Omega^d)$$

is zero for $i < \lambda_0$. This completes the proof.

§23. The Bott Periodicity Theorem for the Unitary Group.

First a review of well known facts concerning the unitary group. Let $\mathbf{C}^n$ be the space of n-tuples of complex numbers, with the usual Hermitian inner product. The unitary group $\mathbf{U}(n)$ is defined to be the group of all linear transformations $S: \mathbf{C}^n \to \mathbf{C}^n$ which preserve this inner product. Equivalently, using the matrix representation, $\mathbf{U}(n)$ is the group of all $n \times n$ complex matrices S such that $S\,S^* = I$; where S^* denotes the conjugate transpose of S.

For any $n \times n$ complex matrix A the exponential of A is defined by the convergent power series expansion

$$\exp A = I + A + \tfrac{1}{2!}A^2 + \tfrac{1}{3!}A^3 + \ldots \quad .$$

The following properties are easily verified:

(1) $\exp(A^*) = (\exp A)^*$; $\exp(TAT^{-1}) = T(\exp A)T^{-1}$.

(2) If A and B commute then

$\exp(A + B) = (\exp A)(\exp B)$. In particular:

(3) $(\exp A)(\exp -A) = I$

(4) The function exp maps a neighborhood of 0 in the space of $n \times n$ matrices diffeomorphically onto a neighborhood of I.

If A is skew-Hermitian (that is if $A + A^* = 0$), then it follows from (1) and (3) that $\exp A$ is unitary. Conversely if $\exp A$ is unitary, and A belongs to a sufficiently small neighborhood of 0, then it follows from (1), (3), and (4) that $A + A^* = 0$. From these facts one easily proves that:

(5) $\mathbf{U}(n)$ is a smooth submanifold of the space of $n \times n$ matrices;

(6) the tangent space $T\mathbf{U}(n)_I$ can be identified with the space of $n \times n$ skew-Hermitian matrices.

Therefore the Lie algebra $\mathfrak{g}$ of $\mathbf{U}(n)$ can also be identified with the space of skew-Hermitian matrices. For any tangent vector at I extends uniquely to a left invariant vector field on $\mathbf{U}(n)$. Computation shows that the bracket product of left invariant vector fields corresponds to the product $[A,B] = AB - BA$ of matrices.

Since $\mathbf{U}(n)$ is compact, it possesses a left and right invariant Riemannian metric. Note that the function

$$\exp: T\mathbf{U}(n)_I \to \mathbf{U}(n)$$

defined by exponentiation of matrices coincides with the function exp defined (as in §10) by following geodesics on the resulting Riemannian manifold. In fact for each skew-Hermitian matrix A the correspondence

$$t \to \exp(t\,A)$$

defines a 1-parameter subgroup of $\mathbf{U}(n)$ (by Assertion (2) above); and hence defines a geodesic.

A specific Riemannian metric on $\mathbf{U}(n)$ can be defined as follows. Given matrices $A,B \in \mathfrak{g}$ let $\langle A,B\rangle$ denote the real part of the complex number

$$\text{trace}\,(AB^*) = \sum_{i,j} A_{ij}\bar{B}_{ij} \ .$$

Clearly this inner product is positive definite on $\mathfrak{g}$.

This inner product on $\mathfrak{g}$ determines a unique left invariant Riemannian metric on $\mathbf{U}(n)$. To verify that the resulting metric is also right invariant, we must check that it is invariant under the adjoint action of $\mathbf{U}(n)$ on $\mathfrak{g}$.

DEFINITION of the adjoint action. Each $S \in \mathbf{U}(n)$ determines an inner automorphism

$$X \to S\,X\,S^{-1} = (L_S R_S^{-1})X$$

of the group $\mathbf{U}(n)$. The induced linear mapping

$$(L_S R_S^{-1})_* : T\mathbf{U}(n)_I \to T\mathbf{U}(n)_I$$

is called $\mathrm{Ad}(S)$. Thus $\mathrm{Ad}(S)$ is an automorphism of the Lie algebra of $\mathbf{U}(n)$. Using Assertion (1) above we obtain the explicit formula

$$\mathrm{Ad}(S)A = SAS^{-1} \ ,$$

for $A \in \mathfrak{g}$, $S \in \mathbf{U}(n)$.

The inner product $\langle A,B\rangle$ is invariant under each such automorphism $\mathrm{Ad}(S)$. In fact if $A_1 = \mathrm{Ad}(S)A$, $B_1 = \mathrm{Ad}(S)B$ then the identity

$$A_1 B_1^* = SAS^{-1}(SBS^{-1})^* = SAB^*S^{-1}$$

implies that

$$\text{trace } (A_1 B_1^*) = \text{trace } (SAB^*S^{-1}) = \text{trace } (AB^*) ;$$

and hence that

$$\langle A_1, B_1 \rangle = \langle A, B \rangle .$$

It follows that the corresponding left invariant metric on $\mathbf{U}(n)$ is also right invariant.

Given $A \in \mathfrak{g}$ we know by ordinary matrix theory that there exists $T \in \mathbf{U}(n)$ so that TAT^{-1} is in diagonal form

$$TAT^{-1} = \begin{pmatrix} ia_1 & & & \\ & ia_2 & & \\ & & \ddots & \\ & & & ia_n \end{pmatrix}$$

where the a_i's are real. Also, given any $S \in \mathbf{U}(n)$, there is a $T \in \mathbf{U}(n)$ such that

$$TST^{-1} = \begin{pmatrix} e^{ia_1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & e^{ia_n} \end{pmatrix}$$

where again the a_i's are real. Thus we see directly that exp: $\mathfrak{g} \to \mathbf{U}(n)$ is onto.

One may treat the special unitary group $\mathbf{SU}(n)$ in the same way. $\mathbf{SU}(n)$ is defined as the subgroup of $\mathbf{U}(n)$ consisting of matrices of determinant 1. If exp is regarded as the ordinary exponential map of matrices, it is easy to show, using the diagonal form, that

$$\det (\exp A) = e^{\text{trace } A} .$$

Using this equation, one may show that $\mathfrak{g}'$, the Lie algebra of $\mathbf{SU}(n)$ is the set of all matrices A such that $A + A^* = 0$ and trace $A = 0$.

In order to apply Morse theory to the topology of $\mathbf{U}(n)$ and $\mathbf{SU}(n)$, we begin by considering the set of all geodesics in $\mathbf{U}(n)$ from I to $-I$. In other words, we look for all $A \in T\mathbf{U}(n)_I = \mathfrak{g}$ such that $\exp A = -I$. Suppose A is such a matrix; if it is not already in diagonal form, let $T \in \mathbf{U}(n)$ be such that TAT^{-1} is in diagonal form. Then

$$\exp TAT^{-1} = T(\exp A)T^{-1} = T(-I)T^{-1} = -I$$

so that we may as well assume that A is already in diagonal form

$$A = \begin{pmatrix} ia_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & ia_n \end{pmatrix}$$

In this case,

$$\exp A = \begin{pmatrix} e^{ia_1} & & & \\ & \cdot & & \\ & & \cdot & \\ & & & e^{ia_n} \end{pmatrix}$$

so that $\exp A = -I$ if and only if A has the form

$$\begin{pmatrix} k_1 i\pi & & & \\ & k_2 i\pi & & \\ & & \cdot & \\ & & & k_n i\pi \end{pmatrix}$$

for some odd integers $k_1, \ldots, k_n$.

Since the length of the geodesic $t \to \exp tA$ from $t = 0$ to $t = 1$ is $|A| = \sqrt{\operatorname{tr} AA^*}$, the length of the geodesic determined by A is $\pi\sqrt{k_1^2 + \ldots + k_n^2}$. Thus A determines a minimal geodesic if and only if each k_i equals ± 1, and in that case, the length is $\pi\sqrt{n}$. Now, regarding such an A as a linear map of $\mathbf{C}^n$ to $\mathbf{C}^n$ observe that A is completely determined by specifying Eigen($i\pi$), the vector space consisting of all $v \in \mathbf{C}^n$ such that $Av = i\pi v$; and Eigen($-i\pi$), the space of all $v \in \mathbf{C}^n$ such that $Av = -i\pi v$. Since $\mathbf{C}^n$ splits as the orthogonal sum Eigen($i\pi$) $\oplus$ Eigen($-i\pi$), the matrix A is then completely determined by Eigen($i\pi$), which is an arbitrary subspace of $\mathbf{C}^n$. Thus the space of all minimal geodesics in $\mathbf{U}(n)$ from I to $-I$ may be identified with the space of all sub-vector-spaces of $\mathbf{C}^n$.

Unfortunately, this space is rather inconvenient to use since it has components of varying dimensions. This difficulty may be removed by replacing $\mathbf{U}(n)$ by $\mathbf{SU}(n)$ and setting $n = 2m$. In this case, all the above considerations remain valid. But the additional condition that $a_1 + \ldots + a_{2m} = 0$ with $a_i = \pm\pi$ restricts Eigen($i\pi$) to being an arbitrary m dimensional sub-vector-space of $\mathbf{C}^{2m}$. This proves the following:

LEMMA 23.1. The space of minimal geodesics from I to $-I$ in the special unitary group $\mathbf{SU}(2m)$ is homeomorphic to the complex Grassmann manifold $G_m(\mathbf{C}^{2m})$, consisting of all m dimensional vector subspaces of $\mathbf{C}^{2m}$.

We will prove the following result at the end of this section.

LEMMA 23.2. Every non-minimal geodesic from I to $-I$ in $\mathbf{SU}(2m)$ has index $\geq 2m+2$.

Combining these two lemmas with §22 we obtain:

THEOREM 23.3 (Bott). The inclusion map $G_m(\mathbf{C}^{2m}) \to \Omega(\mathbf{SU}(2m); I, -I)$ induces isomorphisms of homotopy groups in dimensions $\leq 2m$. Hence

$$\pi_i\, G_m(\mathbf{C}^{2m}) \cong \pi_{i+1}\mathbf{SU}(2m)$$

for $i \leq 2m$.

On the other hand using standard methods of homotopy theory one obtains somewhat different isomorphisms.

LEMMA 23.4. The group $\pi_i G_m(\mathbf{C}^{2m})$ is isomorphic to $\pi_{i-1}\,\mathbf{U}(m)$ for $i \leq 2m$. Furthermore,

$$\pi_{i-1}\,\mathbf{U}(m) \cong \pi_{i-1}\,\mathbf{U}(m+1) \cong \pi_{i-1}\,\mathbf{U}(m+2) \cong \cdots$$

for $i \leq 2m$; and

$$\pi_j\,\mathbf{U}(m) \cong \pi_j\,\mathbf{SU}(m)$$

for $j \neq 1$.

PROOF. First note that for each m there exists a fibration

$$\mathbf{U}(m) \to \mathbf{U}(m+1) \to S^{2m+1}\ .$$

From the homotopy exact sequence

$$\cdots \to \pi_i\, S^{2m+1} \to \pi_{i-1}\,\mathbf{U}(m) \to \pi_{i-1}\,\mathbf{U}(m+1) \to \pi_{i-1}\, S^{2m+1} \to \cdots$$

of this fibration we see that

$$\pi_{i-1}\,\mathbf{U}(m) \cong \pi_{i-1}\,\mathbf{U}(m+1) \quad \text{for } i \leq 2m.$$

(Compare Steenrod, "The Topology of Fibre Bundles," Princeton, 1951, p. 35 and p. 90.) It follows that the inclusion homomorphisms

$$\pi_{i-1}\,\mathbf{U}(m) \to \pi_{i-1}\,\mathbf{U}(m+1) \to \pi_{i-1}\,\mathbf{U}(m+2) \to \cdots$$

are all isomorphisms for $i \leq 2m$. These mutually isomorphic groups are

called the (i-1)-st stable homotopy group of the unitary group. They will be denoted briefly by $\pi_{i-1}\mathbf{U}$.

The same exact sequence shows that, for $i = 2m+1$, the homomorphism $\pi_{2m}\mathbf{U}(m) \to \pi_{2m}\mathbf{U}(m+1) \cong \pi_{2m}\mathbf{U}$ is onto.

The complex Stiefel manifold is defined to be the coset space $\mathbf{U}(2m)/\mathbf{U}(m)$. From the exact sequence of the fibration

$$\mathbf{U}(m) \to \mathbf{U}(2m) \to \mathbf{U}(2m)/\mathbf{U}(m)$$

we see that $\pi_i(\mathbf{U}(2m)/\mathbf{U}(m)) = 0$ for $i \leq 2m$.

The complex Grassmann manifold $G_m(\mathbf{C}^{2m})$ can be identified with the coset space $\mathbf{U}(2m)/\mathbf{U}(m) \times \mathbf{U}(m)$. (Compare Steenrod §7.) From the exact sequence of the fibration

$$\mathbf{U}(m) \to \mathbf{U}(2m)/\mathbf{U}(m) \to G_m(\mathbf{C}^{2m})$$

we see now that

$$\pi_i G_m(\mathbf{C}^{2m}) \xrightarrow{\cong} \pi_{i-1}\mathbf{U}(m)$$

for $i \leq 2m$.

Finally, from the exact sequence of the fibration $\mathbf{SU}(m) \to \mathbf{U}(m) \to S^1$ we see that $\pi_j\mathbf{SU}(m) \cong \pi_j\mathbf{U}(m)$ for $j \neq 1$. This completes the proof of Lemma 23.4.

Combining Lemma 23.4 with Theorem 23.3 we see that

$$\pi_{i-1}\mathbf{U} = \pi_{i-1}\mathbf{U}(m) \cong \pi_i G_m(\mathbf{C}^{2m}) \cong \pi_{i+1}\mathbf{SU}(2m) \cong \pi_{i+1}\mathbf{U}$$

for $1 \leq i \leq 2m$. Thus we obtain:

PERIODICITY THEOREM. $\pi_{i-1}\mathbf{U} \cong \pi_{i+1}\mathbf{U}$ for $i \geq 1$.

To evaluate these groups it is now sufficient to observe that $\mathbf{U}(1)$ is a circle; so that

$$\pi_0\mathbf{U} = \pi_0\mathbf{U}(1) = 0$$
$$\pi_1\mathbf{U} = \pi_1\mathbf{U}(1) \cong \mathbf{Z} \quad \text{(infinite cyclic)}.$$

As a check, since $\mathbf{SU}(2)$ is a 3-sphere, we have:

$$\pi_2\mathbf{U} = \pi_2\mathbf{SU}(2) = 0$$
$$\pi_3\mathbf{U} = \pi_3\mathbf{SU}(2) \cong \mathbf{Z}\ .$$

Thus we have proved the following result.

THEOREM 23.5 (Bott). The stable homotopy groups $\pi_i \mathbf{U}$ of the unitary groups are periodic with period 2. In fact the groups

$$\pi_0 \mathbf{U} \cong \pi_2 \mathbf{U} \cong \pi_4 \mathbf{U} \cong \dots$$

are zero, and the groups

$$\pi_1 \mathbf{U} \cong \pi_3 \mathbf{U} \cong \pi_5 \mathbf{U} \cong \dots$$

are infinite cyclic.

The rest of §23 will be concerned with the proof of Lemma 23.2. We must compute the index of any non-minimal geodesic from I to -I on $\mathbf{SU}(n)$, where n is even. Recall that the Lie algebra

$$\mathfrak{g}' = T(\mathbf{SU}(n))_I$$

consists of all $n \times n$ skew-Hermitian matrices with trace zero. A given matrix $A \in \mathfrak{g}'$ corresponds to a geodesic from I to -I if and only if the eigenvalues of A have the form $i\pi k_1, \dots, i\pi k_n$ where $k_1, \dots, k_n$ are odd integers with sum zero.

We must find the conjugate points to I along the geodesic

$$t \to \exp(tA) .$$

According to Theorem 20.5 these will be determined by the positive eigenvalues of the linear transformation

$$K_A: \mathfrak{g}' \to \mathfrak{g}'$$

where

$$K_A(W) = R(A,W)A = \tfrac{1}{4}[[A,W],A] .$$

(Compare §21.7.)

We may assume that A is the diagonal matrix

$$\begin{pmatrix} i\pi k_1 & & \\ & \ddots & \\ & & i\pi k_n \end{pmatrix}$$

with $k_1 \geq k_2 \geq \dots \geq k_n$. If $W = (w_{j\ell})$ then a short computation shows that

$$[A,W] = \left(i\pi(k_j - k_\ell)w_{j\ell}\right) ,$$

hence

$$[A,[A,W]] = \left(-\pi^2(k_j - k_\ell)^2 w_{j\ell}\right) ,$$

and

$$K_A(W) = \left(\tfrac{\pi^2}{4}(k_j - k_\ell)^2 \, w_{j\ell} \right) .$$

Now we find a basis for $\mathfrak{g}'$ consisting of eigenvectors of K_A, as follows:

1) For each $j < \ell$ the matrix $E_{j\ell}$ with $+1$ in the $(j\ell)$-th place, -1 in the (ℓj)-th place and zeros elsewhere, is in $\mathfrak{g}'$ and is an eigenvector corresponding to the eigenvalue $\frac{\pi^2}{4}(k_j - k_\ell)^2$.

2) Similarly for each $j < \ell$ the matrix $E'_{j\ell}$ with $+i$ in the $(j\ell)$-th place and $+i$ in the (ℓj)-th place is an eigenvector, also with eigenvalue $\frac{\pi^2}{4}(k_j - k_\ell)^2$.

3) Each diagonal matrix in $\mathfrak{g}'$ is an eigenvector with eigenvalue 0.

Thus the non-zero eigenvalues of K_A are the numbers $\frac{\pi^2}{4}(k_j - k_\ell)^2$ with $k_j > k_\ell$. Each such eigenvalue is to be counted twice.

Now consider the geodesic $\gamma(t) = \exp tA$. Each eigenvalue $e = \frac{\pi^2}{4}(k_j - k_\ell)^2 > 0$ gives rise to a series of conjugate points along γ corresponding to the values

$$t = \pi/\sqrt{e},\ 2\pi/\sqrt{e},\ 3\pi/\sqrt{e},\ \ldots .$$

(See §20.5.) Substituting in the formula for e, this gives

$$t = \frac{2}{k_j - k_\ell},\ \frac{4}{k_j - k_\ell},\ \frac{6}{k_j - k_\ell},\ \ldots .$$

The number of such values of t in the open interval $(0,1)$ is evidently equal to $\frac{k_j - k_\ell}{2} - 1$.

Now let us apply the Index Theorem. For each j,ℓ with $k_j > k_\ell$ we obtain two copies of the eigenvalue $\frac{\pi^2}{4}(k_j - k_\ell)^2$, and hence a contribution of

$$2\left(\frac{k_j - k_\ell}{2} - 1 \right)$$

to the index. Adding over all j,ℓ this gives the formula

$$\lambda = \sum_{k_j > k_\ell} (k_j - k_\ell - 2)$$

for the index of the geodesic γ.

As an example, if γ is a minimal geodesic, then all of the k_j

are equal to ± 1. Hence $\lambda = 0$, as was to be expected.

Now consider a non-minimal geodesic. Let $n = 2m$.

CASE 1. At least $m+1$ of the k_i's are (say) negative. In this case at least one of the positive k_i must be ≥ 3, and we have

$$\lambda \geq \sum_{1}^{m+1} (3 - (-1) - 2) = 2(m+1) .$$

CASE 2. m of the k_i are positive and m are negative but not all are ± 1. Then one is ≥ 3 and one is ≤ -3 so that

$$\lambda \geq \sum_{1}^{m-1} (3 - (-1) - 2) + \sum_{1}^{m-1} (1 - (-3) - 2) + (3 - (-3) - 2)$$

$$= 4m \geq 2(m+1) .$$

Thus in either case we have $\lambda \geq 2m+2$. This proves Lemma 23.2, and therefore completes the proof of the Theorem 23.3.

§24. The Periodicity Theorem for the Orthogonal Group.

This section will carry out an analogous study of the iterated loop space of the orthogonal group. However the treatment is rather sketchy, and many details are left out. The point of view in this section was suggested by the paper Clifford modules by M. Atiyah, R. Bott, and A. Shapiro, which relates the periodicity theorem with the structure of certain Clifford algebras. (See Topology, Vol. 3, Supplement 1 (1964), pp. 3-38.)

Consider the vector space $\mathbf{R}^n$ with the usual inner product. The orthogonal group $\mathbf{O}(n)$ consists of all linear maps

$$T : \mathbf{R}^n \to \mathbf{R}^n$$

which preserve this inner product. Alternatively $\mathbf{O}(n)$ consists of all real $n \times n$ matrices T such that $T\,T^* = I$. This group $\mathbf{O}(n)$ can be considered as a smooth subgroup of the unitary group $\mathbf{U}(n)$; and therefore inherits a right and left invariant Riemannian metric.

Now suppose that n is even.

DEFINITION. A complex structure J on $\mathbf{R}^n$ is a linear transformation $J : \mathbf{R}^n \to \mathbf{R}^n$, belonging to the orthogonal group, which satisfies the identity $J^2 = -I$. The space consisting of all such complex structures on $\mathbf{R}^n$ will be denoted by $\mathbf{\Omega}_1(n)$.

We will see presently (Lemma 24.4) that $\mathbf{\Omega}_1(n)$ is a smooth submanifold of the orthogonal group $\mathbf{O}(n)$.

REMARK. Given some fixed $J_1 \in \mathbf{\Omega}_1(n)$ let $\mathbf{U}(n/2)$ be the subgroup of $\mathbf{O}(n)$ consisting of all orthogonal transformations which commute with J_1. Then $\mathbf{\Omega}_1(n)$ can be identified with the quotient space $\mathbf{O}(n)/\mathbf{U}(n/2)$.

LEMMA 24.1. The space of minimal geodesics from I to $-I$ on $\mathbf{O}(n)$ is homeomorphic to the space $\mathbf{\Omega}_1(n)$ of complex structures on $\mathbf{R}^n$.

PROOF: The space $\mathbf{O}(n)$ can be identified with the group of $n \times n$ orthogonal matrices. Its tangent space $\mathfrak{g} = T\mathbf{O}(n)_I$ can be identified with the space of $n \times n$ skew-symmetric matrices. Any geodesic γ with

$\gamma(0) = I$ can be written uniquely as

$$\gamma(t) = \exp(\pi t A)$$

for some $A \in \mathfrak{g}$.

Let $n = 2m$. Since A is skew-symmetric, there exists an element $T \in \mathbf{O}(n)$ so that

$$TAT^{-1} = \begin{pmatrix} 0 & a_1 & & & & & \\ -a_1 & 0 & & & & & \\ & & 0 & a_2 & & & \\ & & -a_2 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & a_m \\ & & & & & -a_m & 0 \end{pmatrix}$$

with $a_1, a_2, \ldots, a_m \geq 0$. A short computation shows that $T(\exp \pi A)T^{-1}$ is equal to

$$\begin{pmatrix} \cos \pi a_1 & \sin \pi a_1 & 0 & 0 & \cdots \\ -\sin \pi a_1 & \cos \pi a_1 & 0 & 0 & \cdots \\ 0 & 0 & \cos \pi a_2 & \sin \pi a_2 & \cdots \\ 0 & 0 & -\sin \pi a_2 & \cos \pi a_2 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} .$$

Thus $\exp(\pi A)$ is equal to $-I$ if and only if $a_1, a_2, \ldots, a_m$ are odd integers.

The inner product $\langle A, A \rangle$ is easily seen to be $2(a_1^2 + a_2^2 + \ldots + a_m^2)$. Therefore the geodesic $\gamma(t) = \exp(\pi t A)$ from I to $-I$ is minimal if and only if $a_1 = a_2 = \ldots = a_m = 1$.

If γ is minimal then

$$A^2 = T^{-1} \begin{pmatrix} 0 & 1 & & & & \\ -1 & 0 & & & & \\ & & 0 & 1 & & \\ & & -1 & 0 & & \\ & & & & \ddots & \\ & & & & & \ddots \end{pmatrix}^2 T = -I \quad ,$$

hence A is a complex structure.

Conversely, let J be any complex structure. Since J is orthogonal we have

$$J J^* = I$$

where J^* denotes the transpose of J. Together with the identity $J\,J = -I$ this implies that $J^* = -J$. Thus J is skew-symmetric. Hence

$$T\,J\,T^{-1} = \begin{pmatrix} 0 & a_1 & & & \\ -a_1 & 0 & & & \\ & & \cdot & & \\ & & & \cdot & \end{pmatrix}$$

for some $a_1, a_2, \ldots, a_m \geq 0$ and some T. Now the identity $J^2 = -I$ implies that $a_1 = \ldots = a_m = 1$; and hence that $\exp \pi J = -I$. This completes the proof.

LEMMA 24.2. Any non-minimal geodesic from I to $-I$ in $\mathbf{O}(2m)$ has index $\geq 2m-2$.

The proof is similar to that of 23.2. Suppose that the geodesic has the form $t \to \exp(\pi t\, A)$ with

$$A = \begin{pmatrix} 0 & a_1 & & & & \\ -a_1 & 0 & & & & \\ & & 0 & a_2 & & \\ & & -a_2 & 0 & & \\ & & & & \cdot & \\ & & & & & \cdot \end{pmatrix}$$

where $a_1 \geq a_2 \geq \ldots \geq a_m > 0$ are odd integers. Computation shows that the non-zero eigenvalues of the linear transformation $K_A = -\frac{1}{4}(\mathrm{Ad}\, A)^2$ are

1) for each $i < j$ the number $(a_i + a_j)^2/4$, and

2) for each $i < j$ with $a_i \neq a_j$ the number $(a_i - a_j)^2/4$.

Each of these eigenvalues is to be counted twice. This leads to the formula

$$\lambda = \sum_{i<j} (a_i + a_j - 2) + \sum_{a_i > a_j} (a_i - a_j - 2) \quad .$$

For a minimal geodesic we have $a_1 = a_2 = \ldots = a_m = 1$ so that $\lambda = 0$, as expected. For a non-minimal geodesic we have $a_1 \geq 3$; so that

$$\lambda \geq \sum_{2}^{m} (3+1-2) + 0 = 2m - 2.$$

This completes the proof.

Now let us apply Theorem 22.1. The two lemmas above, together with

the statement that $\Omega_1(n)$ is a manifold imply the following.

THEOREM 24.3 (Bott). The inclusion map $\Omega_1(n) \to \Omega\, \mathbf{O}(n)$ induces isomorphisms of homotopy groups in dimensions $\leq n-4$. Hence

$$\pi_i\, \Omega_1(n) \cong \pi_{i+1}\, \mathbf{O}(n)$$

for $i \leq n-4$.

Now we will iterate this procedure, studying the space of geodesics from J to $-J$ in $\Omega_1(n)$; and so on. Assume that n is divisible by a high power of 2.

Let $J_1, \ldots, J_{k-1}$ be fixed complex structures on $\mathbf{R}^n$ which anti-commute *, in the sense that

$$J_r J_s + J_s J_r = 0$$

for $r \neq s$. Suppose that there exists at least one other complex structure J which anti-commutes with $J_1, \ldots, J_{k-1}$.

DEFINITION. Let $\Omega_k(n)$ denote the set of all complex structures J on $\mathbf{R}^n$ which anti-commute with the fixed structures $J_1, \ldots, J_{k-1}$.

Thus we have

$$\Omega_k(n) \subset \Omega_{k-1}(n) \subset \ldots \subset \Omega_1(n) \subset \mathbf{O}(n) \quad .$$

Clearly each $\Omega_k(n)$ is a compact set. To complete the definition it is natural to define $\Omega_0(n)$ to be $\mathbf{O}(n)$

LEMMA 24.4. Each $\Omega_k(n)$ is a smooth, totally geodesic** submanifold of $\mathbf{O}(n)$. The space of minimal geodesics from J_ℓ to $-J_\ell$ in $\Omega_\ell(n)$ is homeomorphic to $\Omega_{\ell+1}(n)$, for $0 \leq \ell < k$.

It follows that each component of $\Omega_k(n)$ is a symmetric space. For the isometric reflection of $\mathbf{O}(n)$ in a point of $\Omega_k(n)$ will automatically carry $\Omega_k(n)$ to itself.

---

* These structures make $\mathbf{R}^n$ into a module over a suitable Clifford algebra. However, the Clifford algebras will be suppressed in the following presentation.

** A submanifold of a Riemannian manifold is called totally geodesic if each geodesic in the submanifold is also a geodesic in larger manifold.

PROOF of 24.4. Any point in $\mathbf{O}(n)$ close to the identity can be expressed uniquely in the form $\exp A$, where A is a "small," skew-symmetric matrix. Hence any point in $\mathbf{O}(n)$ close to the complex structure J can be expressed uniquely as $J \exp A$; where again A is small and skew.

ASSERTION 1. $J \exp A$ is a complex structure if and only if A anti-commutes with J.

PROOF: If A anti-commutes with J, then $J^{-1}A\,J = -A$ hence

$$I = \exp(J^{-1}A\,J)\exp A = J^{-1}(\exp A)J \exp A .$$

Therefore $(J \exp A)^2 = -I$. Conversely if $(J \exp A)^2 = -I$ then the above computation shows that

$$\exp(J^{-1}A\,J)\exp A = I .$$

Since A is small, this implies that

$$J^{-1}A\,J = -A$$

so that A anti-commutes with J.

ASSERTION 2. $J \exp A$ anti-commutes with the complex structures $J_1,\ldots,J_{k-1}$ if and only if A commutes with $J_1,\ldots,J_{k-1}$.

The proof is similar and straightforward.

Note that Assertions 1 and 2 both put linear conditions on A. Thus a neighborhood of J in $\mathbf{\Omega}_k(n)$ consists of all points $J \exp A$ where A ranges over all small matrices in a linear subspace of the Lie algebra $\mathfrak{g}$. This clearly implies that $\mathbf{\Omega}_k(n)$ is a totally geodesic submanifold of $\mathbf{O}(n)$.

Now choose a specific point $J_k \in \mathbf{\Omega}_k(n)$, and assume that there exists a complex structure J which anti-commutes with $J_1,\ldots,J_k$. Setting $J = J_k A$ we see easily that A is also a complex structure which anti-commutes with J_k. However, A comutes with $J_1,\ldots,J_{k-1}$. Hence the formula

$$t \to J_k \exp(\pi t\, A)$$

defines a geodesic from J_k to $-J_k$ in $\mathbf{\Omega}_k(n)$. Since this geodesic is minimal in $\mathbf{O}(n)$, it is certainly minimal in $\mathbf{\Omega}_k(n)$.

Conversely, let γ be any minimal geodesic from J_k to $-J_k$ in $\Omega_k(n)$. Setting $\gamma(t) = J_k \exp(\pi t A)$, it follows from 24.1 that A is a complex structure, and from Assertions 1,2 that A commutes with $J_1,\dots,J_{k-1}$ and anti-commutes with J_k. It follows easily that J_kA belongs to $\Omega_{k+1}(n)$. This completes the proof of 24.4.

REMARK. The point $J_kA \in \Omega_{k+1}(n)$ which corresponds to a given geodesic γ has a very simple interpretation: it is the midpoint $\gamma(\frac{1}{2})$ of the geodesic.

In order to pass to a stable situation, note that $\Omega_k(n)$ can be imbedded in $\Omega_k(n+n')$ as follows. Choose fixed anti-commuting complex structures $J'_1,\dots,J'_k$ on $\mathbf{R}^{n'}$. Then each $J \in \Omega_k(n)$ determines a complex structure $J \oplus J'_k$ on $\mathbf{R}^n \oplus \mathbf{R}^{n'}$ which anti-commutes with $J_\alpha \oplus J'_\alpha$ for $\alpha = 1,\dots,k-1$.

DEFINITION. Let Ω_k denote the direct limit as $n \to \infty$ of the spaces $\Omega_k(n)$, with the direct limit topology. (I.e., the fine topology.) The space $\mathbf{O} = \Omega_0$ is called the <u>infinite orthogonal group</u>.

It is not difficult to see that the inclusions $\Omega_{k+1}(n) \to \Omega\,\Omega_k(n)$ give rise, in the limit, to inclusions $\Omega_{k+1} \to \Omega\,\Omega_k$.

THEOREM 24.5. For each $k \geq 0$ this limit map $\Omega_{k+1} \to \Omega\,\Omega_k$ is a homotopy equivalence. Thus we have isomorphisms

$$\pi_h \mathbf{O} \cong \pi_{h-1}\,\Omega_1 \cong \pi_{h-2}\,\Omega_2 \cong \dots \cong \pi_1\,\Omega_{h-1} .$$

The proof will be given presently.

Next we will give individual descriptions of the manifolds $\Omega_k(n)$ for $k = 0,1,2,\dots,8$.

$\Omega_0(n)$ is the orthogonal group.

$\Omega_1(n)$ is the set of all complex structures on $\mathbf{R}^n$.

Given a fixed complex structure J_1 we may think of $\mathbf{R}^n$ as being a vector space $\mathbf{C}^{n/2}$ over the complex numbers.

$\Omega_2(n)$ can be described as the set of "quaternionic structures" on the complex vector space $\mathbf{C}^{n/2}$. Given a fixed $J_2 \in \Omega_2(n)$ we may think of

$\mathbf{C}^{n/2}$ as being a vector space $\mathbf{H}^{n/4}$ over the quaternions $\mathbf{H}$. Let $\mathbf{Sp}(n/4)$ be the group of isometries of this vector space onto itself. Then $\Omega_2(n)$ can be identified with the quotient space $\mathbf{U}(n/2)/\mathbf{Sp}(n/4)$.

Before going further it will be convenient to set $n = 16r$.

LEMMA 24.6 - (3). The space $\Omega_3(16r)$ can be identified with the quaternionic Grassmann manifold consisting of all quaternionic subspaces of $\mathbf{H}^{4r}$.

PROOF: Any complex structure $J_3 \in \Omega_3(16r)$ determines a splitting of $\mathbf{H}^{4r} = \mathbf{R}^{16r}$ into two mutually orthogonal subspaces V_1 and V_2 as follows. Note that $J_1J_2J_3$ is an orthogonal transformation with square $J_1J_2J_3J_1J_2J_3$ equal to $+ I$. Hence the eigenvalues of $J_1J_2J_3$ are ± 1. Let $V_1 \subset \mathbf{R}^{16r}$ be the subspace on which $J_1J_2J_3$ equals $+ I$; and let V_2 be the orthogonal subspace on which it equals $-I$. Then clearly $\mathbf{R}^{16r} = V_1 \oplus V_2$. Since $J_1J_2J_3$ commutes with J_1 and J_2 it is clear that both V_1 and V_2 are closed under the action of J_1 and J_2.

Conversely, given the splitting $\mathbf{H}^{4r} = V_1 \oplus V_2$ into mutually orthogonal quaternionic subspaces, we can define $J_3 \in \Omega_3(16r)$ by the identities

$$\begin{cases} J_3|V_1 = -J_1J_2|V_1 \\ J_3|V_2 = J_1J_2|V_2 . \end{cases}$$

This proves Lemma 24.6 -(3).

The space $\Omega_3(16r)$ is awkward in that it contains components of varying dimension. It is convenient to restrict attention to the component of largest dimension: namely the space of $2r$-dimensional quaternionic subspaces of $\mathbf{H}^{4r}$. Henceforth, we will assume that J_3 has been chosen in this way, so that $\dim_{\mathbf{H}} V_1 = \dim_{\mathbf{H}} V_2 = 2r$.

LEMMA 24.6 - (4). The space $\Omega_4(16r)$ can be identified with the set of all quaternionic isometries from V_1 to V_2. Thus $\Omega_4(16r)$ is diffeomorphic to the symplectic group $\mathbf{Sp}(2r)$.

PROOF: Given $J_4 \in \Omega_4(16r)$ note that the product J_3J_4 anti-commutes with $J_1J_2J_3$. Hence J_3J_4 maps V_1 to V_2 (and V_2 to V_1).

Since J_3J_4 commutes with J_1 and J_2 we see that

$$J_3J_4|V_1 : V_1 \to V_2$$

is a quaternionic isomorphism. Conversely, given any such isomorphism $T : V_1 \to V_2$ it is easily seen that J_4 is uniquely determined by the identities:

$$\begin{cases} J_4|V_1 = J_3^{-1} T \\ J_4|V_2 = -T^{-1} J_3 \end{cases} .$$

This proves 24.6 - (4).

LEMMA 24.6 - (5). The space $\Omega_5(16r)$ can be identified with the set of all vector spaces $W \subset V_1$ such that
(1) W is closed under J_1 (i.e., W is a complex vector space) and
(2) V_1 splits as the orthogonal sum $W \oplus J_2 W$.

PROOF: Given $J_5 \in \Omega_5(16r)$ note that the transformation $J_1J_4J_5$ commutes with $J_1J_2J_3$ and has square $+ I$. Thus $J_1J_4J_5$ maps V_1 into itself; and determines a splitting of V_1 into two mutually orthogonal subspaces. Let $W \subset V_1$ be the subspace on which $J_1J_4J_5$ coincides with $+ I$. Since J_2 anti-commutes with $J_1J_4J_5$, it follows that $J_2W \subset V_1$ is precisely the orthogonal subspace, on which $J_1J_4J_5$ equals $-I$. Clearly $J_1W = W$.

Conversely, given the subspace W, it is not difficult to show that J_5 is uniquely determined.

REMARK. If $\mathbf{U}(2r) \subset \mathbf{Sp}(2r)$ denotes the group of quaternionic automorphisms of V_1 keeping W fixed, then the quotient space $\mathbf{Sp}(2r)/\mathbf{U}(2r)$ can be identified with $\Omega_5(16r)$.

LEMMA 24.6 - (6). The space $\Omega_6(16r)$ can be identified with the set of all real subspaces $X \subset W$ such that W splits as the orthogonal sum $X \oplus J_1X$.

PROOF. Given $J_6 \in \Omega_6(16r)$ note that the transformation $J_2J_4J_6$ commutes both with $J_1J_2J_3$ and with $J_1J_4J_5$. Hence $J_2J_4J_6$ maps W into itself. Since $(J_2J_4J_6)^2 = I$, it follows that $J_2J_4J_6$ determines a

splitting of W into two mutually orthogonal subspaces. Let $X \subset W$ be the subspace on which $J_2J_4J_6$ equals $+I$. Then J_1X will be the orthogonal subspace on which it equals $-I$.

Conversely, given $X \subset W$, it is not hard to see that J_6 is uniquely determined.

REMARK. If $\mathbf{O}(2r) \subset \mathbf{U}(2r)$ denotes the group of complex automorphisms of W keeping X fixed, then the quotient space $\mathbf{U}(2r)/\mathbf{O}(2r)$ can be identified with $\mathbf{\Omega}_6(16r)$.

LEMMA 24.6 - (7). The space $\mathbf{\Omega}_7(16r)$ can be identified with the real Grassmann manifold consisting of all real subspaces of $X \cong \mathbf{R}^{2r}$.

PROOF: Given J_7, anti-commuting with $J_1,\ldots,J_6$ note that $J_1J_6J_7$ commutes with $J_1J_2J_3$, with $J_1J_4J_5$, and with $J_2J_4J_6$; and has square $+I$. Thus $J_1J_6J_7$ determines a splitting of X into two mutually orthogonal subspaces: X_1 (where $J_1J_6J_7$ equals $+I$) and X_2 (where $J_1J_6J_7$ equals $-I$). Conversely, given $X_1 \subset X$ it can be shown that J_7 is uniquely determined.

This space $\mathbf{\Omega}_7(16r)$, like $\mathbf{\Omega}_3(16r)$, has components of varying dimension. Again we will restrict attention to the component of largest dimension, by assuming that

$$\dim X_1 = \dim X_2 = r.$$

Thus we obtain:

ASSERTION. The largest component of $\mathbf{\Omega}_7(16r)$ is diffeomorphic to the Grassmann manifold consisting of r-dimensional subspaces of R^{2r}.

LEMMA 24.6 - (8). The space $\mathbf{\Omega}_8(16r)$ can be identified with the set of all real isometries from X_1 to X_2.

PROOF. If $J_8 \in \mathbf{\Omega}_8(16r)$ then the orthogonal transformation J_7J_8 commutes with $J_1J_2J_3$, $J_1J_4J_5$, and $J_2J_4J_6$; but anti-commutes with $J_1J_6J_7$. Hence J_7J_8 maps X_1 isomorphically onto X_2. Clearly this isomorphism determines J_8 uniquely.

Thus we see that $\mathbf{\Omega}_8(16r)$ is diffeomorphic to the orthogonal

group* $\mathbf{O}(r)$.

Let us consider this diffeomorphism $\Omega_8(16r) \to \mathbf{O}(r)$, and pass to the limit as $r \to \infty$. It follows that Ω_8 is homeomorphic to the infinite orthogonal group $\mathbf{O}$. Combining this fact with Theorem 24.5, we obtain the following.

THEOREM 24.7 (Bott). The infinite orthogonal group $\mathbf{O}$ has the same homotopy type as its own 8-th loop space. Hence the homotopy group $\pi_i \mathbf{O}$ is isomorphic to $\pi_{i+8} \mathbf{O}$ for $i \geq 0$.

If $\mathbf{Sp} = \Omega_4$ denotes the infinite symplectic group, then the above argument also shows that $\mathbf{O}$ has the homotopy type of the 4-fold loop space $\Omega\Omega\Omega\Omega \mathbf{Sp}$, and that $\mathbf{Sp}$ has the homotopy type of the 4-fold loop space $\Omega\Omega\Omega\Omega \mathbf{O}$. The actual homotopy groups can be tabulated as follows.

i modulo 8	$\pi_i \mathbf{O}$	$\pi_i \mathbf{Sp}$
0	$\mathbf{Z}_2$	0
1	$\mathbf{Z}_2$	0
2	0	0
3	$\mathbf{Z}$	$\mathbf{Z}$
4	0	$\mathbf{Z}_2$
5	0	$\mathbf{Z}_2$
6	0	0
7	$\mathbf{Z}$	$\mathbf{Z}$

The verification that these groups are correct will be left to the reader. (Note that $\mathbf{Sp}(1)$ is a 3-sphere, and that $\mathbf{SO}(3)$ is a projective 3-space.)

The remainder of this section will be concerned with the proof of Theorem 24.5. It is first necessary to prove an algebraic lemma.

Consider a Euclidean vector space V with anti-commuting complex structures $J_1, \ldots, J_k$.

---

* For $k > 8$ it can be shown that $\Omega_k(16r)$ is diffeomorphic to $\Omega_{k-8}(r)$. In fact any additional complex structures $J_9, J_{10}, \ldots, J_k$ on $\mathbf{R}^{16r}$ give rise to anti-commuting complex structures J_8J_9, J_8J_{10}, $J_8J_{11}, \ldots, J_8J_k$ on X_1; and hence to an element of $\Omega_{k-8}(r)$. However, for our purposes it will be sufficient to stop with $k = 8$.

DEFINITION. V is a minimal $(J_1,\dots,J_k)$-space if no proper, non-trivial subspace is closed under the action of $J_1,\dots,$ and J_k. Two such minimal vector spaces are isomorphic if there is an isometry between them which commutes with the action of $J_1,\dots,J_k$.

LEMMA 24.8 (Bott and Shapiro). For $k \not\equiv 3 \pmod 4$, any two minimal $(J_1,\dots,J_k)$ vector spaces are isomorphic.

The proof of 24.8 follows that of 24.6. For $k = 0,1,$ or 2 a minimal space is just a 1-dimensional vector space over the reals, the complex numbers or the quaternions. Clearly any two such are isomorphic.

For $k = 3$ a minimal space is still a 1-dimensional vector space over the quaternions. However, there are two possibilities, according as J_3 is equal to $+J_1J_2$ or $-J_1J_2$. This gives two non-isomorphic minimal spaces, both with dimension equal to 4. Call these H and H'.

For $k = 4$ a minimal space must be isomorphic to $H \oplus H'$, with J_3J_4 mapping H to H'. The dimension is equal to 8.

For $k = 5,6$ we obtain the same minimal vector space $H \oplus H'$. The complex structures J_5,J_6 merely determine preferred complex or real subspaces. For $k = 7$ we again obtain the same space, but there are two possibilities, according as J_7 is equal to $+J_1J_2J_3J_4J_5J_6$ or to $-J_1J_2J_3J_4J_5J_6$. Thus in this case there are two non-isomorphic minimal vector spaces; call these L and L'.

For $k = 8$ a minimal vector space must be isomorphic to $L \oplus L'$, with J_7J_8 mapping L onto L'. The dimension is equal to 16.

For $k > 8$ it can be shown that the situation repeats more or less periodically. However, the cases $k \leq 8$ will suffice for our purposes.

Let m_k denote the dimension of a minimal $(J_1,\dots,J_k)$-vector space. From the above discussion we see that:

$$m_0 = 1,\ m_1 = 2,\ m_2 = m_3 = 4,$$
$$m_4 = m_5 = m_6 = m_7 = 8,\ m_8 = 16.$$

For $k > 8$ it can be shown that $m_k = 16m_{k-8}$.

REMARK. These numbers m_k are closely connected with the problem of constructing linearly independent vector fields on spheres. Suppose for example that $J_1,\dots,J_k$ are anti-commuting complex structures on a vector

space V of dimension rm_k. Here r can be any positive integer. Then for each unit vector $u \in V$ the k vectors $J_1u, J_2u, \ldots, J_ku$ are perpendicular to each other and to u. Thus we obtain k linearly independent vector fields on an (rm_k-1)-sphere. For example we obtain 3 vector fields on a $(4r-1)$-sphere; 7 vector fields on an $(8r-1)$-sphere; 8 vector fields on a $(16r-1)$-sphere; and so on. These results are due to Hurwitz and Radon. (Compare B. Eckmann, Gruppentheoretischer Beweis des Satzes von Hurwitz-Radon..., Commentarii Math. Helv. Vol. 15 (1943), pp. 358-366.) J. F. Adams has recently proved that these estimates are best possible.

PROOF of Theorem 24.5 for $k \not\equiv 2 \pmod 4$. We must study non-minimal geodesics from J to $-J$ in $\mathbf{\Omega}_k(n)$. Recall that the tangent space of $\mathbf{\Omega}_k(n)$ at J consists of all matrices JA where

1) A is skew

2) A anti-commutes with J

3) A commutes with $J_1, \ldots, J_{k-1}$.

Let T denote the vector space of all such matrices A. A given $A \in T$ corresponds to a geodesic $t \to J \exp(\pi tA)$ from J to $-J$ if and only if its eigenvalues are all odd multiples of i.

Each such $A \in T$ determines a self-adjoint transformation $K_A: T \to T$. Since $\mathbf{\Omega}_k(n)$ is a totally geodesic submanifold of $\mathbf{O}(n)$, we can compute K_A by the formula

$$K_AB = -\tfrac{1}{4}[A,[A,B]] = (-A^2B + 2ABA - BA^2)/4 ,$$

just as before. We must construct some non-zero eigenvalues of K_A so as to obtain a lower bound for the index of the corresponding geodesic

$$t \to J \exp(\pi t A) .$$

Split the vector space $\mathbf{R}^n$ as a direct sum $M_1 \oplus M_2 \oplus \ldots \oplus M_s$ of mutually orthogonal subspaces which are closed and minimal under the action of $J_1, \ldots, J_{k-1}$, J and A. Then the eigenvalues of A on M_h must be all equal, except for sign.* For otherwise M_h would split as a sum of

* We are dealing with the complex eigenvalues of a real, skew-symmetric transformation. Hence these eigenvalues are pure imaginary; and occur in conjugate pairs.

eigenspaces of A; and hence would not be minimal. Let $\pm ia_h$ be the two eigenvalues of $A|M_h$; where $a_1,\ldots,a_s$ are odd, positive integers.

Now note that $J' = a_h^{-1}JA|M_h$; is a complex structure on M_h which anti-commutes with $J_1,\ldots,J_{k-1}$, and J. Thus M_h is $(J_1,\ldots,J_{k-1},J,J')$-minimal. Hence the dimension of M_h is m_{k+1}. Since $k+1 \not\equiv 3 \pmod 4$ we see that $M_1,M_2,\ldots,M_s$ are mutually isomorphic.

For each pair h,j with $h \neq j$ we can construct an eigenvector $B: \mathbf{R}^n \to \mathbf{R}^n$ of the linear transformation $K_A: T \to T$ as follows. Let $B|M_\ell$ be zero for $\ell \neq h,j$. Let $B|M_h$ be an isometry from M_h to M_j which satisfies the conditions

$$BJ_\alpha = J_\alpha B \quad \text{for} \quad \alpha = 1,\ldots,k-1;$$
$$BJ = -JB \quad \text{and} \quad BJ' = +J'B \ .$$

In other words $B|M_h$ is an isomorphism from M_h to $\bar{M}_j$; where the bar indicates that we have changed the sign of J on M_j. Such an isomorphism exists by 24.8. Finally let $B|M_j$ be the negative adjoint of $B|M_h$.

Proof that B belongs to the vector space T. Since

$$\langle Bv,w\rangle = \langle v,-Bw\rangle \quad \text{for} \quad v \in M_h, \quad w \in M_j$$

it is clear that B is skew-symmetric. It is also clear that $B|M_h$ commutes with $J_1,\ldots,J_{k-1}$ and anti-commutes with J. It follows easily that the negative adjoint $B|M_j$ also commutes with $J_1,\ldots,J_{k-1}$ and anti-commutes with J. Thus $B \in T$.

We claim that B is an eigenvector of K_A corresponding to the eigenvalue $(a_h + a_j)^2/4$. For example if $v \in M_h$ then

$$\begin{aligned}(K_AB)v &= \tfrac{1}{4}(-A^2B + 2ABA - BA^2)v \\ &= \tfrac{1}{4}(a_j^2Bv + 2a_jBa_hv + Ba_h^2v) \\ &= \tfrac{1}{4}(a_j + a_h)^2\,Bv\ ;\end{aligned}$$

and a similar computation applies for $v \in M_j$.

Now let us count. The number of minimal spaces $M_h \subset \mathbf{R}^n$ is given by $s = n/m_{k+1}$. For at least one of these the integer a_h must be ≥ 3. For otherwise we would have a minimal geodesic. This proves the following (always for $k \not\equiv 2 \pmod 4$):

ASSERTION. K_A has at least $s-1$ eigenvalues which are $\geq (3+1)^2/4 = 4$. The integer $s = n/m_{k+1}$ tends to infinity with n.

Now consider the geodesic $t \to J \exp(\pi t A)$. Each eigenvalue e^2 of K_A gives rise to conjugate points along this geodesic for $t = e^{-1}, 2e^{-1}, 3e^{-1}, \ldots$ by 20.5. Thus if $e^2 \geq 4$ then one obtains at least one interior conjugate point. Applying the index theorem, this proves the following.

ASSERTION. The index of a non-minimal geodesic from J to $-J$ in $\Omega_k(n)$ is $\geq n/m_{k+1} - 1$.

It follows that the inclusion map

$$\Omega_{k+1}(n) \to \Omega\, \Omega_k(n)$$

induces isomorphisms of homotopy groups in dimensions $\leq n/m_{k+1} - 3$. This number tends to infinity with n. Therefore, passing to the direct limit as $n \to \infty$, it follows that the inclusion map $i : \Omega_{k+1} \to \Omega\, \Omega_k$ induces isomorphisms of homotopy groups in all dimensions. But it can be shown that both Ω_{k+1} and $\Omega\, \Omega_k$ have the homotopy type of a CW-complex. Therefore, by Whitehead's theorem, it follows that i is a homotopy equivalence. This completes the proof of 24.5 providing that $k \not\equiv 2 \pmod 4$.

PROOF of 24.5 for $k \equiv 2 \pmod 4$. The difficulty in this case may be ascribed to the fact that $\Omega_k(n)$ has an infinite cyclic fundamental group. Thus $\Omega\, \Omega_k(n)$ has infinitely many components, while the approximating subspace $\Omega_{k+1}(n)$ has only finitely many.

To describe the fundamental group $\pi_1 \Omega_k(n)$ we construct a map

$$f : \Omega_k(n) \to S^1 \subset \mathbf{C}$$

as follows. Let $J_1, \ldots, J_{k-1}$ be the fixed anti-commuting complex structure on $\mathbf{R}^n$. Make $\mathbf{R}^n$ into an $(n/2)$-dimensional complex vector space by defining

$$iv = J_1 J_2 \cdots J_{k-1} v$$

for $v \in \mathbf{R}^n$; where $i = \sqrt{-1} \in \mathbf{C}$. The condition $k \equiv 2 \pmod 4$ guarantees that $i^2 = -1$, and that $J_1, J_2, \ldots, J_{k-1}$ commute with i.

Choose a base point $J \in \Omega_k(n)$. For any $J' \in \Omega_k(n)$ note that the composition $J^{-1}J'$ commutes with i. Thus $J^{-1}J'$ is a unitary complex linear transformation, and has a well defined complex determinant which will be denoted by $f(J')$.

Now consider a geodesic

$$t \to J \exp(\pi t A)$$

from J to $-J$ in $\Omega_k(n)$. Since A commutes with $i = J_1 J_2 \dots J_{k-1}$ (compare Assertion 2 in the proof of 24.4) we may think of A also as a complex linear transformation. In fact A is skew-Hermitian; hence the trace of A is a pure imaginary number. Now

$$f(J \exp(\pi t A)) = \text{determinant } (\exp(\pi t A)) = e^{\pi t \text{ trace } A}$$

Thus f maps the given geodesic into a closed loop on S^1 which is completely determined by the trace of A. It follows that this trace is invariant under homotopy of the geodesic within the path space $\Omega(\Omega_k(n); J, -J)$.

The index λ of this geodesic can be estimated as follows. As before split $\mathbf{R}^n$ into an orthogonal sum $M_1 \oplus \dots \oplus M_r$ where each M_h is closed under the action of $J_1, \dots, J_{k-1}, J$, and A; and is minimal. Thus for each h, the complex linear transformation $A|M_h$ can have only one eigenvalue, say ia_h. For otherwise M_h would split into eigenspaces. Thus $A|M_h$ coincides with $a_h J_1 J_2 \dots J_{k-1}|M_h$. Since M_h is minimal under the action of $J_1, \dots, J_{k-1}$, and J; its complex dimension is $m_k/2$. Therefore the trace of A is equal to $i(a_1 + \dots + a_r)m_k/2$.

Now for each $h \neq j$ an eigenvector B of the linear transformation

$$B \to K_A B = (-A^2 B + 2ABA - BA^2)/4$$

can be constructed much as before. Since M_h and M_j are $(J_1, \dots, J_{k-1}, J)$-minimal it follows from 24.8 that there exists an isometry

$$B|M_h : M_h \to M_j$$

which commutes with $J_1, \dots, J_{k-1}$ and anti-commutes with J. Let $B|M_j$ be the negative adjoint of $B|M_h$; and let $B|M_\ell$ be zero for $\ell \neq h, j$. Then an easy computation shows that

$$K_A B = (a_h - a_j)^2 B/4 \quad .$$

Thus for each $a_h > a_j$ we obtain an eigenvalue $(a_h - a_j)^2/4$ for K_A. Since each such eigenvalue makes a contribution of $(a_h - a_j)/2 - 1$ towards the index λ, we obtain the inequality

$$2\lambda \geq \sum_{a_h > a_j} (a_h - a_j - 2) \quad .$$

Now let us restrict attention to some fixed component of $\Omega\,\Omega_k(n)$. That is let us look only at matrices A such that trace $A = icm_k/2$ where c is some constant integer.

Thus the integers $a_1,\ldots a_r$ satisfy

1) $a_1 \equiv a_2 \equiv \ldots \equiv a_r \equiv 1 \pmod 2$, (since $\exp(\pi A) = -I$),

2) $a_1 + \ldots + a_r = c$, and

3) $\operatorname{Max}_h |a_h| \geq 3$ (for a non-minimal geodesic).

Suppose for example that some a_h is equal to -3. Let p be the sum of the positive a_h and $-q$ the sum of the negative a_h. Thus

$$p - q = c, \quad p + q \geq r \ ,$$

hence $2p \geq r + c$. Now

$$2\lambda \geq \sum_{a_h > a_j} (a_h - a_j - 2) > \sum_{a_h > 0} (a_h - (-3) - 3) = p \ ,$$

hence $4\lambda \geq 2p \geq r + c$; where $r = n/m_k$ tends to infinity with n. It follows that the component of $\Omega\,\Omega_k(n)$ is approximated up to higher and higher dimensions by the corresponding component of $\Omega_{k+1}(n)$, as $n \to \infty$. Passing to the direct limit, we obtain a homotopy equivalence on each component. This completes the proof of 24.5.

APPENDIX. THE HOMOTOPY TYPE OF A MONOTONE UNION

The object of this appendix will be to give an alternative version for the final step in the proof of Theorem 17.3 (the fundamental theorem of Morse theory). Given the subsets $\Omega^{a_0} \subset \Omega^{a_1} \subset \Omega^{a_2} \subset \ldots$ of the path space $\Omega = \Omega(M;p,q)$, and given the information that each Ω^{a_i} has the homotopy type of a certain CW-complex, we wish to prove that the union Ω also has the homotopy type of a certain CW-complex.

More generally consider a topological space X and a sequence $X_0 \subset X_1 \subset X_2 \subset \ldots$ of subspaces. To what extent is the homotopy type of X determined by the homotopy types of the X_i?

It is convenient to consider the infinite union

$$X_\Sigma = X_0 \times [0,1] \cup X_1 \times [1,2] \cup X_2 \times [2,3] \cup \ldots .$$

This is to be topologized as a subset of $X \times \mathbf{R}$.

DEFINITION. We will say that X is the homotopy direct limit of the sequence $\{X_i\}$ if the projection map $p : X_\Sigma \to X$, defined by $p(x,\tau) = x$, is a homotopy equivalence.

EXAMPLE 1. Suppose that each point of X lies in the interior of some X_i, and that X is paracompact. Then using a partition of unity one can construct a map

$$f : X \to \mathbf{R}$$

so that $f(x) \geq i+1$ for $x \notin X_i$, and $f(x) \geq 0$ for all x. Now the correspondence $x \to (x,f(x))$ maps X homeomorphically onto a subset of X_Σ which is clearly a deformation retract. Therefore p is a homotopy equivalence; and X is a homotopy direct limit.

EXAMPLE 2. Let X be a CW-complex, and let the X_i be subcomplexes with union X. Since $p : X_\Sigma \to X$ induces isomorphisms of homotopy groups in all dimensions, it follows from Whitehead's theorem that X is a homotopy direct limit.

EXAMPLE 3. The unit interval $[0,1]$ is not the homotopy direct limit of the sequence of closed subsets $[0] \cup [1/i,1]$.

The main result of this appendix is the following.

THEOREM A. Suppose that X is the homotopy direct limit of $\{X_i\}$ and Y is the homotopy direct limit of $\{Y_i\}$. Let $f: X \to Y$ be a map which carries each X_i into Y_i by a homotopy equivalence. Then f itself is a homotopy equivalence.

Assuming Theorem A, the alternative proof of Theorem 17.3 can be given as follows. Recall that we had constructed a commutative diagram

$$\begin{array}{ccccccc} \Omega^{a_0} & \subset & \Omega^{a_1} & \subset & \Omega^{a_2} & \subset & \cdots \\ \downarrow & & \downarrow & & \downarrow & & \\ K_0 & \subset & K_1 & \subset & K_2 & \subset & \cdots \end{array}$$

of homotopy equivalences. Since $\Omega = \cup\, \Omega^{a_i}$ and $K = \cup\, K_i$ are homotopy direct limits (compare Examples 1 and 2 above), it follows that the limit mapping $\Omega \to K$ is also a homotopy equivalence.

PROOF of Theorem A. Define $f_\Sigma : X_\Sigma \to Y_\Sigma$ by $f_\Sigma(x,t) = (f(x),t)$. It is clearly sufficient to prove that f_Σ is a homotopy equivalence.

CASE 1. Suppose that $X_i = Y_i$ and that each map $f_i : X_i \to Y_i$ (obtained by restricting f) is homotopic to the identity. We must prove that f_Σ is a homotopy equivalence.

REMARK. Under these conditions it would be natural to conjecture that f_Σ must actually be homotopic to the identity. However counter-examples can be given.

For each n let

$$h_u^n : X_n \to X_n$$

be a one-parameter family of mappings, with $h_0^n = f_n$, $h_1^n =$ identity. Define the homotopy

$$h_u : X_\Sigma \to X_\Sigma$$

as follows (where it is always to be understood that $0 \le t \le 1$, and $n = 0,1,2,\ldots$).

$$h_u(x,n+t) = \begin{cases} (h_u^n(x),n+2t) & \text{for } 0 \le t \le \tfrac{1}{2} \\ (h^n_{(3-4t)u}(x),n+1) & \text{for } \tfrac{1}{2} \le t \le \tfrac{3}{4} \\ (h^{n+1}_{(4-3t)u}(x),n+1) & \text{for } \tfrac{3}{4} \le t \le 1 \end{cases} .$$

Taking $u = 0$ this defines a map $h_0 : X_\Sigma \to X_\Sigma$ which is clearly homotopic to f_Σ. The mapping $h_1 : X_\Sigma \to X_\Sigma$ on the other hand has the following properties:

$$h_1(x,n+t) = (x,n+2t) \quad \text{for } 0 \le t \le \tfrac{1}{2}$$

$$h_1(x,n+t) \in X_{n+1} \times [n+1] \quad \text{for } \tfrac{1}{2} \le t \le 1 .$$

We will show that any such map h_1 is a homotopy equivalence. In fact a homotopy inverse $g : X_\Sigma \to X_\Sigma$ can be defined by the formula

$$g(x,n+t) = \begin{cases} (x,n+2t) & 0 \le t \le \tfrac{1}{2} \\ h_1(x,n+\tfrac{3}{2} - t) & \tfrac{1}{2} \le t \le 1 \end{cases} .$$

This is well defined since

$$h_1(x,n+\tfrac{1}{2}) = h_1(x,n+1) = (x,n+1) .$$

Proof that the composition $h_1 g$ is homotopic to the identity map of X_Σ. Note that

$$h_1 g(x,n+t) = \begin{cases} (x,n+4t) & 0 \le t \le \tfrac{1}{4} \\ h_1(x,n+2t) & \tfrac{1}{4} \le t \le \tfrac{1}{2} \\ h_1(x,n+\tfrac{3}{2} - t) & \tfrac{1}{2} \le t \le 1 \end{cases} .$$

Define a homotopy $H_u : X_\Sigma \to X_\Sigma$ as follows. For $0 \le u \le \tfrac{1}{2}$ let

$$H_u(x,n+t) = \begin{cases} h_1 g(x,n+t) & \text{for } 0 \le t \le (1-u)/2 \\ & \text{and for } \tfrac{1}{2}+u \le t \le 1 \\ h_1(x,n+1-u) & \text{for } (1-u)/2 \le t \le \tfrac{1}{2} + u. \end{cases}$$

This is well defined since

$$h_1 g(x,n+(1-u)/2) = h_1 g(x,n+\tfrac{1}{2}+u) = h_1(x,n+1-u).$$

Now H_0 is equal to $h_1 g$ and $H_{\frac{1}{2}}$ is given by

$$H_{\frac{1}{2}}(x,n+t) = \begin{cases} (x,n+4t) & 0 \le t \le \tfrac{1}{4} \\ (x,n+1) & \tfrac{1}{4} \le t \le 1 \end{cases}.$$

Clearly this is homotopic to the identity.

Thus $h_1 g$ is homotopic to the identity; and a completely analogous argument shows that gh_1 is homotopic to the identity. This completes the proof in Case 1.

CASE 2. Now let X and Y be arbitrary. For each n let $g_n : Y_n \to X_n$ be a homotopy inverse to f_n. Note that the diagram

$$\begin{array}{ccc} Y_n & \xrightarrow{g_n} & X_n \\ \downarrow{\scriptstyle j_n} & & \downarrow{\scriptstyle i_n} \\ Y_{n+1} & \xrightarrow{g_{n+1}} & X_{n+1} \end{array}$$

(where i_n and j_n denote inclusion maps) is homotopy commutative. In fact

$$i_n g_n \sim g_{n+1} f_{n+1} i_n g_n = g_{n+1} j_n f_n g_n \sim g_{n+1} j_n .$$

Choose a specific homotopy $h_u^n : Y_n \to X_{n+1}$ with $h_0^n = i_n g_n$, $h_1^n = g_{n+1} j_n$; and define $G : Y_\Sigma \to X_\Sigma$ by the formula

$$G(y,n+t) = \begin{cases} (g_n(y),n+2t) & 0 \le t \le \tfrac{1}{2} \\ (h_{2t-1}^n(y),n+1) & \tfrac{1}{2} \le t \le 1 \end{cases}.$$

We will show that the composition $Gf_\Sigma : X_\Sigma \to X_\Sigma$ is a homotopy equivalence. Let X_Σ^n denote the subset of X_Σ consisting of all pairs (x,τ) with $\tau \le n$. (Thus $X_\Sigma^n = X_0 \times [0,1] \cup \ldots \cup X_{n-1} \times [n-1,n] \cup X_n \times [n]$.) The composition Gf_Σ carries X_Σ^n into itself by a mapping which is homotopic to the identity. In fact X_Σ^n contains $X_n \times [n]$ as deformation retract; and the mapping Gf_Σ restricted to $X_n \times [n]$ can be identified with $g_n f_n$, and hence is homotopic to the identity. Thus we can apply Case 1 to the sequence

$\{X_\Sigma^n\}$, and conclude that Gf_Σ is a homotopy equivalence.

This proves that f_Σ has a left homotopy inverse. A similar argument shows that $f_\Sigma G : Y_\Sigma \to Y_\Sigma$ is a homotopy equivalence, so that f_Σ has a right homotopy inverse. This proves that f_Σ is a homotopy equivalence (compare page 22) and completes the proof of Theorem A.

> COROLLARY. Suppose that X is the homotopy direct limit of $\{X_\lambda\}$. If each X_λ has the homotopy type of a CW-complex, then X itself has the homotopy type of a CW-complex.

The proof is not difficult.