MORSE THEORY

BY

J. Milnor

Based on lecture notes by

M. SPIVAK and R. WELLS

PRINCETON, NEW JERSEY

PRINCETON UNIVERSITY PRESS

1963
CONTENTS

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 RICCIANIAN 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 $E_{\ast}$...... 77
§15. The Index Theorem...................... 83
§16. A Finite Dimensional Approximation to $g^0$...... 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

Printed in the United States of America
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 ... 131

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.

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

1. If $a < b < f(p)$, then $\mathcal{A}^a$ is vacuous.
2. If $f(p) < a < f(q)$, then $\mathcal{A}^a$ is homeomorphic to a 2-cell.
3. If $f(q) < a < f(r)$, then $\mathcal{A}^a$ is homeomorphic to a cylinder:

4. If $f(r) < a < f(s)$, then $\mathcal{A}^a$ is homeomorphic to a compact manifold of genus one having a circle as boundary:
I. NON-DEGENERATE FUNCTIONS

(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:

\[(1) \rightarrow (2) \] 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:

Here "\( = \)" means "is of the same homotopy type as."

\[(2) \rightarrow (3) \] is the operation of attaching a 1-cell:

\[(3) \rightarrow (4) \] is again the operation of attaching a 1-cell:

\[(4) \rightarrow (5) \] 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

\[ e^k = \{ x \in \mathbb{R}^k : \|x\| < 1 \} \]

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

§1. INTRODUCTION

The boundary

\[ e^k \] will be denoted by \( S^k-1 \). If \( g: S^k-1 \rightarrow Y \) is a continuous map then

\[ Y \cup_{g} e^k \]

is obtained by first taking the topological sum (a disjoint union) of \( Y \) and \( e^k \), and then identifying each \( x \in S^{k-1} \) with \( \{ x \} \in Y \). To take care of the case \( k = 0 \) let \( e^0 \) be a point and let \( e^0 = S^{-1} \) 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 \( p, q, r, s \) 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.


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

§2. Definitions and Lemmas.

The words "smooth" and "differentiable" will be used interchangeably to mean differentiable of class $C^\infty$. The tangent space of a smooth manifold $M$ at a point $p$ will be denoted by $T_pM$. If $\mathbf{G}: M \to N$ is a smooth map with $\mathbf{G}(p) = q$, then the induced linear map of tangent spaces will be denoted by $\mathbf{G}_p: T_pM \to T_qN$.

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 $f^*: T_pM \to \mathbb{R}_f(p)$ is zero. If we choose a local coordinate system $(x^1, \ldots, x^n)$ in a neighborhood $U$ of $p$ this means that

$$f(x) = a_1 x_1 + \ldots + a_n x_n$$

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 $T_pM$, called the Hessian of $f$ at $p$. If $v, w \in T_pM$ then $v$ and $w$ have extensions $\mathbf{v}$ and $\mathbf{w}$ to vector fields. We let

$$f_*(v, w) = \mathbf{v}(\mathbf{w}(f)), \quad \forall \mathbf{v}, \mathbf{w}$$

is well-defined since $\mathbf{v}(\mathbf{w}(f))$ is independent of the extension $\mathbf{v}$ of $v$, while $\mathbf{v}(\mathbf{w}(f))$ is independent of $\mathbf{w}$.

If $(x_1, \ldots, x^n)$ is a local coordinate system and $v = \sum a_i \frac{\partial}{\partial x^i}(p)$ we can take $\mathbf{v} = \sum b_j \frac{\partial}{\partial x^j}$ where $b_j$ now denotes a constant function. Then

$$f_*(v, w) = \mathbf{v}(\mathbf{w}(f)) = \sum a_i b_j \frac{\partial^2 f}{\partial x^i \partial x^j}(p)$$

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 $\left( \frac{\partial}{\partial x^1}, \ldots, \frac{\partial}{\partial x^n} \right)$.

We can now talk about the index and the nullity of the bilinear functional $f_*$ on $T_pM$. 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 $T_pM$ has nullity equal to 0. The index of $f_*$ on $T_pM$ 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 $C^\infty$ function in a convex neighborhood $V$ of 0 in $\mathbb{R}^n$, with $f(0) = 0$. Then

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

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

**Proof:**

$$f(x_1, \ldots, x_n) = \int_0^1 df(x_1 t_1, \ldots, x_n t_1) dt = \sum_{i=1}^n \int_0^1 \frac{\partial f}{\partial x^i}(x_1 t_1, \ldots, x_n t_1) x_i dt .$$

Therefore we can let $g_i(x) \frac{\partial f}{\partial x^i}(x_1 t_1, \ldots, x_n t_1) dt$.
1. NON-DEGENERATE FUNCTIONS

Lemma 2.2 (Lemma of Morse). Let \( p \) be a non-degenerate critical point for \( f \). Then there is a local coordinate system \( (y^1, \ldots, y^n) \) in a neighborhood \( U \) of \( p \) with
\[
y^i(p) = 0 \quad \text{for all } i
\]
and such that the identity
\[
f = f(p) + (y^1)^2 - \cdots - (y^i)^2 + (y^{i+1})^2 + \cdots + (y^n)^2
\]
holds throughout \( U \), where \( \lambda \) is the index of \( f \) at \( p \).

Proof: We first show that if there is any such expression for \( f \), then \( \lambda \) must be the index of \( f \) at \( p \). For any coordinate system \( (z^1, \ldots, z^n) \), if
\[
f(q) = f(p) - (z^1(p))^2 - \cdots - (z^i(p))^2 + (z^{i+1}(p))^2 + \cdots + (z^n(p))^2
\]
then we have
\[
\frac{\partial^2 f}{\partial z^i \partial z^j}(p) = \begin{cases} \lambda^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}
\]
which shows that the matrix representing \( f \) with respect to the basis \( \frac{\partial}{\partial z^1}, \ldots, \frac{\partial}{\partial z^n} \) at \( p \) is
\[
\begin{pmatrix} \lambda^2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda^2 \end{pmatrix}
\]
Therefore there is a subspace of \( T_p \mathbb{R}^n \) of dimension \( \lambda \) where \( f \) is negative definite, and a subspace \( V \) of dimension \( n-\lambda \) where \( f \) is positive definite. If there were a subspace of \( T_p \mathbb{R}^n \) of dimension greater than \( \lambda \) on which \( f \) were negative definite then this subspace would intersect \( V \), which is clearly impossible. Therefore \( \lambda \) is the index of \( f \).

We now show that a suitable coordinate system \( (y^1, \ldots, y^n) \) exists. Obviously we can assume that \( p \) is the origin of \( \mathbb{R}^n \) and that \( f(p) = f(0) = 0 \). By 2.1 we can write
\[
f(x_1, \ldots, x_n) = \sum_{j=1}^{n-\lambda} x_j h_j(x_1, \ldots, x_n)
\]
for \( (x_1, \ldots, x_n) \) in some neighborhood of 0. Since 0 is assumed to be a critical point:
\[
\nabla f(0) = \frac{\partial f}{\partial x^i}(0) = 0
\]
for certain smooth functions \( h_j \). It follows that
\[
f(x_1, \ldots, x_n) = \sum_{j=1}^{n-\lambda} x_j h_j(x_1, \ldots, x_n)
\]
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
\[
f = \sum (u_1^2 + \cdots + u_r^2) + \sum_{i,j} u_i u_j h_{ij}(u_1, \ldots, u_r)
\]
throughout \( U_1 \), where the matrices \((h_{ij}(u_1, \ldots, u_r))\) are symmetric. After a linear change in the last \( n-r+1 \) coordinates we may assume that \( h_{rr}(0) \neq 0 \). Let \( g_1(u_1, \ldots, u_n) \) denote the square root of \( h_{rr}(u_1, \ldots, u_r) \). This will be a smooth, non-zero function of \( u_1, \ldots, u_r \) throughout some smaller neighborhood \( U_2 \subset U_1 \) of 0. Now introduce new coordinates \( v_1, \ldots, v_n \) by
\[
v_i = u_i \quad \text{for } i \neq r
\]
and \( v_r = g(u_1, \ldots, u_n) \). 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
\[
f = \sum_{i=1}^{r} (v_i^2) + \sum_{i,j} v_i v_j h_{ij}(v_1, \ldots, v_r)
\]
1. NON-DEGENERATE FUNCTIONS

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 $\mathbb{R}$ and $\mathbb{R}^2$) 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) $f(x, y) = x^3 - 3xy^2 = \text{Real part of } (x + iy)^3$. (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 $\mathbb{R}^2$.

(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 $\mathbb{R}^2$.

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 $C^\infty$ map

$$\varphi: \mathbb{R} \times M \to M$$
such that

1) for each \( t \in \mathbb{R} \) 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 \( t,s \in \mathbb{R} \) we have \( \varphi_{ts} = \varphi_t \circ \varphi_s \).

Given a 1-parameter group \( \varphi \) 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 \textit{generate} the group \( \varphi \).

\section*{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 \( M \).

\textbf{Proof:} Given any smooth curve \( t = c(t) \in M \).

It is convenient to define the \textit{velocity vector}
\[
\frac{dc}{dt} \in TM_c(t)
\]
by the identity \( \frac{d}{dt}(f) = \lim_{h \to 0} \frac{f(c(t+h)) - f(c(t))}{h} \). (Compare §4.1) Now let \( \varphi \) be a 1-parameter group of diffeomorphisms, generated by the vector field \( X \).

Then for each fixed \( q \) the curve
\[
t \mapsto \varphi_t(q)
\]
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:
\[
\frac{dc^i}{dt} = X^i(u^1, \ldots, u^n), \quad i = 1, \ldots, n.
\]

Thus for each point of \( M \) there exists a neighborhood \( U \) and a number \( \varepsilon > 0 \) so that the differential equation
\[
\frac{d\varphi_t(q)}{dt} = X_{\varphi_t(q)}, \quad \varphi_0(q) = q
\]
has a unique smooth solution for \( q \in U, \quad |t| < \varepsilon \).

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 \( \varepsilon \). 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 \( \varphi_t \) is a diffeomorphism.

It only remains to define \( \varphi_t \) 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 \( \varphi_t \) is defined for all values of \( t \). It is not difficult to verify that \( \varphi_t \) is well defined, smooth, and satisfies the condition \( \varphi_{t+s} = \varphi_t \circ \varphi_s \). This completes the proof of Lemma 2.4.

\textbf{Remark:} The hypothesis that \( X \) vanishes outside of a compact set cannot be omitted. For example let \( M \) be the open unit interval \( (0,1) \subset \mathbb{R} \), 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 \( a \leq f(p) < b \), 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 \( M^a \hookrightarrow M^b \) 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.)

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 \( \text{grad} f \) 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 \( \text{grad} f \) vanishes precisely at the critical points of \( f \).

In classical notation, in terms of local coordinates \( u^1, \ldots, u^n \), the gradient has components \( \frac{\partial f}{\partial u^i} \).

\[ \frac{dc}{dt} = \langle \text{grad} f, X \rangle. \]

Let \( \gamma: \mathbb{R} \to M \) be a smooth function which is equal to \( \frac{1}{2} \langle \text{grad} f, \text{grad} f \rangle \) 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

\[ X_q = \gamma(q) (\text{grad} f)_q \]

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

\[ \phi_t: M \to M. \]

For fixed \( q \in M \) consider the function \( \lambda = f(\phi_t(q)) \). If \( \phi_t(q) \)

lies in the set \( f^{-1}[a, b] \), then

\[ \frac{d\phi_t}{dt}(q) X = \langle \text{grad} f, X \rangle, \quad \text{grad} f \rangle = \langle \text{grad} f, \text{grad} f \rangle \]

Thus the correspondence

\[ t \to f(\phi_t(q)) \]

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

Now consider the diffeomorphism \( \phi_1: M \to M \). Clearly this carries \( M^a \) diffeomorphically onto \( M^b \). This proves the first half of §1.1.

Define a 1-parameter family of maps

\[ r_t: M^b \to M^b \]

by

\[ r_t(q) = \begin{cases} q & \text{if } f(q) \leq a \\ q_t(a-f(q)) & \text{if } a \leq f(q) \leq b \end{cases} \]

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 \( M^b \).
I. NON-DEGENERATE FUNCTIONS

THEOREM 3.2. Let \( f: M \rightarrow \mathbb{R} \) be a smooth function, and let \( p \) be a non-degenerate critical point with index \( \lambda \). Setting \( f(p) = c \), suppose that \( f^{-1}(c-e, c+e) \) is compact, and contains no critical point of \( f \) other than \( p \). Then, for all sufficiently small \( \varepsilon > 0 \). Then, for all points other than \( p \), the set \( M^0 \) has the homotopy type of \( M^0 \) with a \( k \)-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

\[
M^0 = f^{-1}(-\varepsilon, \varepsilon)
\]

is heavily shaded. We will introduce a new function \( F: M \rightarrow \mathbb{R} \) which coincides with the height function \( f \) except that \( F < f \) in a small neighborhood of \( p \). Thus the region \( F^{-1}(-\varepsilon, \varepsilon) \) will consist of \( M^0 \) together with a region \( H \) near \( p \). In Diagram 4, \( H \) is the horizontally shaded region.

Diagram 4.

Diagram 3.

§3. HOMOTOPY TYPE

Choosing a suitable cell \( e^k \subset H \), a direct argument (i.e., pushing in along the horizontal lines) will show that \( M^{0-} \cup e^k \) is a deformation retract of \( M^{0-} \cup H \). Finally, by applying 3.1 to the function \( F \) and the region \( F^{-1}[0-\varepsilon, 0+\varepsilon] \) we will see that \( M^{0-} \cup H \) is a deformation retract of \( M^{0+} \). 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

\[
u_i(p) = \ldots = u^n(p) = 0.
\]

Choose \( \varepsilon > 0 \) sufficiently small so that

1. The region \( F^{-1}[0-\varepsilon, 0+\varepsilon] \) is compact and contains no critical points other than \( p \).

2. The image of \( U \) under the diffeomorphic imbedding

\[
(u_1, \ldots, u^n): U \rightarrow \mathbb{R}^n
\]

contains the closed ball.

\[
\{(u_1, \ldots, u^n): \sum (u^k)^2 \leq 2\varepsilon\}.
\]

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

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

Construct a new smooth function \( F: M \longrightarrow \mathbb{R} \) as follows. Let

\[
\mu: \mathbb{R} \longrightarrow \mathbb{R}
\]

be a \( C^\infty \) function satisfying the conditions

\[
\begin{align*}
\mu(0) &> \varepsilon \\
\mu(r) = 0 & \text{ for } r \geq 2\varepsilon \\
-1 < \mu'(r) &\leq 0 \text{ for all } r,
\end{align*}
\]

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

\[
\begin{align*}
\xi &= (u^1)^2 + \ldots + (u^\lambda)^2 \\
\eta &= (u^{\lambda+1})^2 + \ldots + (u^N)^2
\end{align*}
\]

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 \( F^{-1}(-\varepsilon, c-\varepsilon] \) coincides with the region \( \mathcal{M}^{c-\varepsilon} = f^{-1}(-\varepsilon, c+\varepsilon] \).

**Proof:** Outside of the ellipsoid \( \xi + 2\eta \leq 2\varepsilon \) the functions \( F \) and \( f \) coincide. Within this ellipsoid we have

\[
F \leq f = c - \xi + \eta \leq c + \frac{1}{2}\xi + \frac{1}{2}\eta \leq c + \varepsilon
\]

This completes the proof.

**Assertion 2.** The critical points of \( F \) are the same as those of \( f \).

**Proof:** Note that

\[
\frac{dF}{dx} = -1 - \mu'(t + 2\eta) < 0
\]

and

\[
\frac{dF}{d\eta} = 1 - 2\mu'(t + 2\eta) > 1.
\]

Since

\[
\frac{dF}{dx} = -1 - \mu'(t + 2\eta) < 0
\]

and

\[
\frac{dF}{d\eta} = 1 - 2\mu'(t + 2\eta) > 1,
\]

where the covectors \( dx \) 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

\[
F^{-1}[c-\varepsilon, c+\varepsilon] \subseteq r^{-1}[c-\varepsilon, c+\varepsilon].
\]

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 \( F^{-1}(-\varepsilon, c-\varepsilon] \) is a deformation retract of \( \mathcal{M}^{c+\varepsilon} \).

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

**Remark:** In the terminology of Smale, the region \( \mathcal{M}^{c-\varepsilon} \cup H \) is described as \( \mathcal{M}^{c-\varepsilon} \) with a "handle" attached. It follows from Theorem 3.1 that the manifold-with-boundary \( \mathcal{M}^{c-\varepsilon} \cup H \) is diffeomorphic to \( \mathcal{M}^{c+\varepsilon} \). 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.)
1. NON-DEGENERATE FUNCTIONS

Now consider the cell $e^\lambda$ consisting of all points $q$ with
$$t(q) \leq \epsilon, \quad \eta(q) = 0.$$  

Note that $e^\lambda$ is contained in the "handle" $H$. In fact, since $\frac{\partial F}{\partial \eta} < 0$, we have
$$F(q) \leq F(p) < c-\epsilon$$
but $t(q) \geq c-\epsilon$ for $q \in e^\lambda$.

Diagram 6.

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

Assertion 4. $M^\epsilon - E \cup e^\lambda$ is a deformation retract of $M^\epsilon - E \cup H$.

Proof: A deformation retraction $r_t: M^\epsilon - E \cup H \rightarrow M^\epsilon - E \cup H$ 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 is necessary to distinguish three cases as indicated in Diagram 7.

CASE 1. Within the region $\xi \leq \epsilon$ let $r_t$ correspond to the transformation
$$(u_1, \ldots, u^n) \rightarrow (u_1, \ldots, u_1^n, t, u_{n+1}^{\lambda+1}, \ldots, u_n^{\lambda})$$

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

CASE 2. Within the region $\epsilon \leq \xi \leq \eta + \epsilon$ let $r_t$ correspond to the transformation
$$(u_1, \ldots, u^n) \rightarrow (u_1, \ldots, u_1^n, s_t, u_{n+1}^{\lambda+1}, \ldots, u_n^{\lambda})$$
where the number $s_t \in [0,1]$ is defined by
$$s_t = t + (1-t)(\frac{1-\xi}{\eta})^{1/2}.$$  

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

CASE 3. Within the region $\eta + \epsilon \leq \xi$ (i.e., within $M^\epsilon - E$) let $r_t$ be the identity. This coincides with the preceding definition when $\xi = \eta + \epsilon$.

This completes the proof that $M^\epsilon - E \cup e^\lambda$ is a deformation retract of $F^{-1}(c-\epsilon, c+\epsilon)$. 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 $M^\epsilon - E \cup e_{\lambda_1} \cup \ldots \cup e_{\lambda_k}$ has the homotopy type of $M^\epsilon - E \cup e_{\lambda_1} \cup \ldots \cup e_{\lambda_k}$.
**Remark 2.4.** A simple modification of the proof of 3.2 shows that the set $M^o$ is also a deformation retract of $M^{o\pm\epsilon}$. In fact $M^o$ is a deformation retract of $F^{-1}[c, c+\epsilon]$, which is a deformation retract of $M^{o\pm\epsilon}$.

(Compare Diagram 8.) Combining this fact with 3.2 we see easily that $M^{o-\epsilon} \cup e^\lambda$ is a deformation retract of $M^o$.

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

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


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

**Lemma 3.6.** (Whitehead) Let $\varphi_0$ and $\varphi_1$ be homotopic maps from the sphere $e^\lambda$ to $X$. Then the identity map of $X$ extends to a homotopy equivalence $k: X \cup e^\lambda \rightarrow X \cup e^\lambda$.

**Proof:** Define $k$ by the formulas

\[
\begin{align*}
k(x) &= x & \text{for } x \in X \\
k(tu) &= 2tu & \text{for } 0 \leq t \leq \frac{1}{2}, u \in e^\lambda \\
k(tu) &= \varphi_2 - 2t(1) & \text{for } \frac{1}{2} \leq t \leq 1, u \in e^\lambda
\end{align*}
\]

Here $\varphi_t$ denotes the homotopy between $\varphi_0$ and $\varphi_1$; and $tu$ denotes the product of the scalar $t$ with the unit vector $u$. A corresponding map

\[f: X \cup e^\lambda \rightarrow X \cup e^\lambda\]

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


**Lemma 3.7.** Let $\varphi: e^\lambda \rightarrow X$ be an attaching map. Any homotopy equivalence $\xi: X \rightarrow Y$ extends to a homotopy equivalence $F: X \cup e^\lambda \rightarrow Y \cup \varphi e^\lambda$.

**Proof:** (Following an unpublished paper by P. Hilton.) Define $F$ by the conditions

\[
\begin{align*}
\det & = f \\
F|e^\lambda & = \text{identity}.
\end{align*}
\]

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

\[G: Y \cup e^\lambda \rightarrow X \cup e^\lambda\]

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

Since $g f_\varphi$ is homotopic to $\varphi$, it follows from 3.6 that there is a homotopy equivalence $k: X \cup e^\lambda \rightarrow X \cup e^\lambda$.

We will first prove that the composition

\[k G: X \cup e^\lambda \rightarrow X \cup e^\lambda\]

is homotopic to the identity map.
I. NON-DEGENERATE FUNCTIONS

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

$$kGF(x) = gf(x) \quad \text{for } x \in X,$$

$$kGF(tu) = 2tu \quad \text{for } 0 \leq t \leq \frac{1}{2}, \quad u \in e^\lambda,$$

$$kGF(tu) = h_{2-2t} \phi(u) \quad \text{for } \frac{1}{2} \leq t \leq 1, \quad u \in e^\lambda.$$ 

The required homotopy

$$q_t : X \cup e^\lambda \to X \cup e^\lambda$$

is now defined by the formula

$$q_t(x) = h_t(x) \quad \text{for } x \in X,$$

$$q_t(tu) = \frac{2}{1-t} tu \quad \text{for } 0 \leq t \leq \frac{1}{2}, \quad u \in e^\lambda,$$

$$q_t(tu) = h_{2-2t+\phi}(u) \quad \text{for } \frac{1}{2} \leq t \leq 1, \quad u \in e^\lambda.$$ 

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

$$LF \cong \text{identity}, \quad RF \cong \text{identity},$$

imply that

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

Consequently

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

which proves that $R$ is a 2-sided inverse.

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

$$kGF \cong \text{identity}$$

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

Step 1. Since $X(GF) \cong \text{identity}$, and $k$ is known to have a left inverse, it follows that $(GF)k \cong \text{identity}$.

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

Step 3. Since $F(kG) \cong \text{identity}$, 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 $c_1 < c_2 < c_3 < \ldots$ be the critical values of $f : M \to \mathbb{R}$. The sequence $\{c_j\}$ has no cluster point since each $M^a$ is compact. The set $M^a$ is vacuous for $a < c_1$. Suppose $a \not\in c_1, c_2, c_3, \ldots$ and that $M^a$ is of the homotopy type of a CW-complex.

Let $\epsilon$ be the smallest $c_{j+1} - a$. By Theorems 3.1, 3.2, and 3.3, $M^{c+\epsilon}$ has the homotopy type of $M^{c-\epsilon} \cup e^{\lambda_1} \cup \ldots \cup e^{\lambda_j(c)}$ for certain maps $\varphi_1, \ldots, \varphi_j(c)$ when $\epsilon$ is small enough, and there is a homotopy equivalence $h : M^{c+\epsilon} \to M^{a}$. We have assumed that there is a homotopy equivalence $h' : M^a \to M$. Let $K$ be a CW-complex.

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

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

Then $K \cup e^{\lambda_1} \cup \ldots \cup e^{\lambda_j(c)}$ is a CW-complex, and has the same homotopy type as $M^{c+\epsilon}$, by Lemmas 3.6, 3.7.

By induction it follows that each $M^{a'}$ 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 $E$, 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

$$M_{a_1} \subset M_{a_2} \subset M_{a_3} \subset \ldots,$$

$$K_1 \subset K_2 \subset K_3 \subset \ldots,$$

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
I. NON-DEGENERATE FUNCTIONS

Let \( g: M \rightarrow K \) 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 \( M \rightarrow K \) 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 \( M^\alpha \) has the homotopy type of a finite CW-complex, with one cell of dimension \( \lambda \) for each critical point of index \( \lambda \) in \( M^\alpha \). This is true even if \( \alpha \) is a critical value. (Compare Remark 3.4.)

\[ \]

§4. EXAMPLES

94. 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 (92.2). The two critical points must be the minimum and maximum points. Say that \( f(p) = 0 \) is the minimum and \( f(q) = 1 \) is the maximum. If \( \varepsilon \) is small enough then the sets \( M^\varepsilon = f^{-1}[0, \varepsilon] \) and \( f^{-1}[1-\varepsilon, 1] \) are closed \( n \)-cells by 92.2. But \( M^\varepsilon \) is homeomorphic to \( M^1-\varepsilon \) by 93.1. 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, Sommes de variétés différentiables et structures différentiables des sphères, Bull. Soc. Math. de France, Vol. 87 (1959), pp. 439-444; Theorem 1(3); 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-
I. NON-DEGENERATE FUNCTIONS

degenerate critical points.)

As another application of the previous theorems we note that if an
n-manifold has a differentiable 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.
This situation is studied in forthcoming papers by J. Eells and N. Kuiper.

Such a function exists for example on the real or complex projective plane.

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

Define a real valued function $f$ on $\mathbb{C}P^n$ by the identity

$$f(z_0; z_1; \ldots; z_n) = c_i |z_i|$$

where $c_0, c_1, \ldots, c_n$ 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 $(z_0; z_1; \ldots; z_n)$
with $z_n \neq 0$, and set

$$|z_0|^2|z_n|^{-1} = x_j + iy_j$$

Then

$$\begin{align*}
x_1, y_1, \ldots, x_n, y_n : U_0 &\to \mathbb{R} \\
x_j &\neq 0
\end{align*}$$

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

$$|z_j|^2 = x_j^2 + y_j^2$$

so that

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

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

$$p = (1:0:0:\ldots:0)$$

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

$$p_1 = (0:1:0:0:0:\ldots), p_n = (0:0:0:0:0:0:1)$$

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:

$$\mathbb{C}P^n$$

has the homotopy type of a CW-complex of the form

$$e^0 \cup e^2 = e^1 \cup e^{2n}$$

It follows that the integral homology groups of $\mathbb{C}P^n$ are given by

$$H_i(\mathbb{C}P^n; \mathbb{Z}) = \begin{cases} \mathbb{Z} & \text{for } i = 0, 2, 4, \ldots, 2n \\ 0 & \text{for other values of } i \end{cases}$$
I. NON-DEGENERATE FUNCTIONS

§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 $\mathcal{S}$ be a function from certain pairs of spaces to the integers. $\mathcal{S}$ is subadditive if whenever $(X,Y) \subset Z$ we have $\mathcal{S}(X,Z) \leq \mathcal{S}(X,Y) + \mathcal{S}(Y,Z)$. If equality holds, $\mathcal{S}$ is called additive.

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

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

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

$$\cdots \rightarrow H_\lambda(Y,Z) \rightarrow H_\lambda(X,Z) \rightarrow H_\lambda(X,Y) \rightarrow \cdots$$

The Euler characteristic $x(X,Y)$ is additive, where

$$x(X,Y) = \sum (-1)^\lambda R_\lambda(X,Y).$$

LEMMA 5.1. Let $\mathcal{S}$ be subadditive and let $X_0 \subset \subset \cdots \subset X_n$.

Then $S(X_n, X_0) \leq \sum_{i=0}^{n-1} S(X_i, X_{i+1})$. If $\mathcal{S}$ 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 $S(X_{n-1}, X_0) \leq \sum S(X_i, X_{i+1})$. Therefore $S(X_n, X_0) \leq S(X_{n-1}, X_0) + S(X_n, X_{n-1}) \leq S(X_i, X_{i+1})$ and the result is true for $n$.

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

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

with equality if $\mathcal{S}$ is additive.

§5. The Morse Inequalities

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 $M^{a_1} = \ldots = M^{a_k} = M$. Then

$$H_\lambda(M^{a_1}, M^{a_1-1}) = H_\lambda(M^{a_1-1} \cup \cdots, M^{a_1-1})$$

where $\lambda$ is the index of the critical point.

Applying (1) to $\mathcal{S} = M^{a_1}$ we have

$$R_\lambda(M^{a_1}) \leq R_\lambda(M^{a_1-1}) \leq C;$$

where $C$ denotes the number of critical points of index $\lambda$. Applying this formula to the case $\mathcal{S} = x$ we have

$$x(M) = x(M^{a_1-1}) = C_0 + C_1 + C_2 + \ldots + C_n.$$

Thus we have proven:

THEOREM 5.2 (Weak Morse Inequalities). If $C_\lambda$ denotes the number of critical points of index $\lambda$ on the compact manifold $M$ then

$$R_\lambda(M) \leq C_\lambda,$$

and

$$\sum (-1)^\lambda R_\lambda(M) = \sum (-1)^\lambda C_\lambda.$$

Slightly sharper inequalities can be proven by the following argument.

LEMMA 5.3. The function $S_\lambda$ is subadditive, where

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

PROOF: Given an exact sequence

$$A \rightarrow B \rightarrow C \rightarrow \ldots \rightarrow D \rightarrow 0$$

of vector spaces note that the rank of the homomorphism $h$ plus the rank of $1$ is equal to the rank of $A$. Therefore,
Hence the last expression is \( \geq 0 \). Now consider the homology exact sequence or a triple \( X \supseteq Y \supseteq Z \). Applying this computation to the homomorphism

\[
H_{k+1}(X,Y) \xrightarrow{\partial} H_k(Y,Z)
\]

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 \( S_\lambda \) to the spaces

\[\emptyset \subset \mathbb{M}^1 \subset \mathbb{M}^2 \subset \ldots \subset \mathbb{M}^k\]

we obtain the Morse inequalities

\[
S_\lambda(M) \leq \sum_{l=1}^{\lambda} S_\lambda(M^l), \quad \lambda = C_{\lambda-1} + \cdots + C_0
\]

or

\[
(\lambda) \quad R_\lambda(M) - R_{\lambda-1}(M) + \cdots + R_0(M) \leq C_{\lambda-1} + \cdots + C_0.
\]

These inequalities are definitely sharper than the previous ones.

In fact, adding (4) and \( (C_{\lambda-1}) \), one obtains (2\( \lambda \)) and comparing \( (\lambda) \) with \( (C_{\lambda-1}) \) for \( \lambda < \alpha \) one obtains the equality (3).

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

\[
R_\lambda - R_{\lambda-1} + \cdots + R_0 = C_{\lambda} + C_{\lambda-1} + \cdots + C_0.
\]

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

\[
R_{\lambda-2} - R_{\lambda-3} + \cdots + R_0 = C_{\lambda-2} + C_{\lambda-3} + \cdots + C_0.
\]
56. 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 $\mathbb{R}^n$.

In fact, if for fixed $p \in \mathbb{R}^n$ define the function $L_p: M \to \mathbb{R}$ 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 $M \subset \mathbb{R}^n$ be a manifold of dimension $k < n$, differentiably embedded in $\mathbb{R}^n$. Let $N \subset M \times \mathbb{R}^n$ be defined by

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

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

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

**DEFINITION.** $e \in \mathbb{R}^n$ is a focal point of $(M,q)$ with multiplicity $\mu$ 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 focal point of $(M,q)$ for some $q \in M$.

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

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

**THEOREM 6.1 (Sard).** If $M$ 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 $M$.

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 $x \in \mathbb{R}^n$, 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 if and only if $x$ is in the image of the set of critical points of $E: N \to \mathbb{R}^n$. 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 $M \subset \mathbb{R}^n$. Then the inclusion map from $M$ to $\mathbb{R}^n$ determines $n$ smooth functions

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

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

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

$$\left( g_{ij} \right) = \left( \frac{\partial \xi}{\partial u^i} \cdot \frac{\partial \xi}{\partial u^j} \right).$$

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

It is defined as follows. The vector $\frac{\partial \xi}{\partial u^i}$ at a point of $M$ can be expressed as the sum of a vector tangent to $\mathbb{R}^n$ and a vector normal to $M$.

Define $\mathbf{I}_{ij}$ to be the normal component of $\frac{\partial ^2 \xi}{\partial u^i \partial u^j}$. Given any unit vector $\mathbf{v}$ which is normal to $M$ at $\xi$ the matrix $\frac{\partial \mathbf{v}}{\partial u^i}$.

$$\left( \mathbf{v} \cdot \frac{\partial \xi}{\partial u^i} \right) = \left( \mathbf{v} \cdot \mathbf{I}_{ij} \right).$$
can be called the "second fundamental form of \( M \) at \( \overrightarrow{q} \) in the direction \( \overrightarrow{T} \)."

It will simplify the discussion to assume that the coordinates have been chosen so that \( g_{ij} \) evaluated at \( \overrightarrow{q} \) is the identity matrix. Then the eigenvalues of the matrix \((\overrightarrow{T} \cdot \overrightarrow{T}_{ij})\) are called the principal curvatures \( K_1, \ldots, K_n \) of \( M \) at \( \overrightarrow{q} \) in the normal direction \( \overrightarrow{T} \). The reciprocals \( K_1^{-1}, \ldots, K_n^{-1} \) of these principal curvatures are called the principal radii of curvature. Of course it may happen that the matrix \((\overrightarrow{T} \cdot \overrightarrow{T}_{ij})\) 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 \( f \) consisting of all \( \overrightarrow{q} + t \overrightarrow{v} \), where \( \overrightarrow{v} \) is a fixed unit vector orthogonal to \( M \) at \( \overrightarrow{q} \).

**Lemma 6.3.** The focal points of \((M, \overrightarrow{q})\) along \( f \) are precisely the points \( \overrightarrow{q} + K_i^{-1} \overrightarrow{v} \), where \( 1 \leq i \leq k \). The focal points are not in \( f \) itself, each being counted with its proper multiplicity.

**Proof:** Choose n-k vector fields \( \overrightarrow{w}_1(u_1, \ldots, u_k, \ldots, u_{n-k}) \) along the manifold so that \( \overrightarrow{w}_1, \ldots, \overrightarrow{w}_{n-k} \) are unit vectors which are orthogonal to each other and to \( M \). We can introduce coordinates \((u_1, \ldots, u_k, t^{n-k})\) on the manifold \( N = M \times \mathbb{R}^k \) as follows. Let \((u_1, \ldots, u_k, t^{n-k})\) correspond to the point \((\overrightarrow{x}(u_1, \ldots, u_k), \sum_{\alpha=1}^{n-k} t^\alpha \overrightarrow{w}_\alpha(u_1, \ldots, u_k)) \in N \).

Then the function

\[
E: N \to \mathbb{R}^k
\]
gives rise to the correspondence

\[
(u_1, \ldots, u_k, t^{n-k}) \mapsto (x(u_1, \ldots, u_k) + \sum_{\alpha=1}^{n-k} t^\alpha w_\alpha(u_1, \ldots, u_k),
\]

with partial derivatives

\[
\begin{align*}
\frac{\partial x}{\partial u_i} &= \frac{\partial x}{\partial u_j} + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_i} \\
\frac{\partial x}{\partial u_i} &= \frac{\partial x}{\partial u_j} + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_i} \\
\frac{\partial x}{\partial t^\alpha} &= \overrightarrow{v}_\alpha .
\end{align*}
\]

Taking the inner products of these n-vectors with the linearly independent vectors \( \overrightarrow{x}, \overrightarrow{x}, \ldots, \overrightarrow{x}_k \) we will obtain an \( nxn \) matrix whose rank equals the rank of the Jacobian of \( E \) at the corresponding point.

The \( nxn \) matrix clearly has the following form

\[
\begin{pmatrix}
\frac{\partial x}{\partial u_1} \cdot \overrightarrow{v}_1 + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_1} & \ldots & \frac{\partial x}{\partial u_n} \cdot \overrightarrow{v}_1 + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial x}{\partial u_1} \cdot \overrightarrow{v}_n + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_1} & \ldots & \frac{\partial x}{\partial u_n} \cdot \overrightarrow{v}_n + \sum_\alpha t^\alpha \frac{\partial w_\alpha}{\partial u_n}
\end{pmatrix}
\]

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

\[
0 = \frac{\partial}{\partial u_i} \left( \overrightarrow{v}_\alpha \cdot \overrightarrow{v}_\beta \right) = \frac{\partial w_\alpha}{\partial u_i} \cdot \overrightarrow{v}_\beta + \overrightarrow{v}_\alpha \cdot \frac{\partial w_\beta}{\partial u_i}
\]

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

\[
\begin{pmatrix}
\varepsilon_{ij} & \sum_\alpha t^\alpha \overrightarrow{v}_\alpha \cdot \overrightarrow{T}_{ij}
\end{pmatrix}
\]

Thus:

**Assertion 6.4.** \( \overrightarrow{q} + t \overrightarrow{v} \) is a focal point of \((M, \overrightarrow{q})\) with multiplicity \( \mu \) if and only if the matrix

\[
\varepsilon_{ij} - t \overrightarrow{v}_\alpha \overrightarrow{T}_{ij}
\]

is singular, with multiplicity \( \mu \).

Now suppose that \((\overrightarrow{q})\) is the identity matrix. Then \((\overrightarrow{q})\) is singular if and only if \( \overrightarrow{v}_i \) is an eigenvalue of the matrix \((\overrightarrow{T} \cdot \overrightarrow{T}_{ij}) \). Furthermore the multiplicity \( \mu \) is equal to the multiplicity of \( \overrightarrow{v}_i \) as eigenvalue.

This completes the proof of Lemma 6.3.
I. NON-DEGENERATE FUNCTIONS

The second partial derivatives at a critical point are given by
\[ \frac{\partial^2 f}{\partial u_i \partial u_j} = \mp \left( \frac{\partial^2 f}{\partial u_i \partial u_j} + \frac{\partial^2 f}{\partial u_j \partial u_i} \right) \cdot (\nabla - \mathbf{p}) \]

Setting \( \overline{\mathbf{p}} = \overline{x} + t \overline{v} \), as in the proof of Lemma 6.3, this becomes
\[ \frac{\partial^2 f}{\partial u_i \partial u_j} = \pm (\overline{x} - \mathbf{p}) \]

Therefore:

**LEMMA 6.5.** The point \( \overline{q} \in \mathcal{M} \) is a degenerate critical point of \( f = \frac{\partial f}{\partial u_i} \) if and only if \( \overline{v} \) is a focal point of \( (\mathcal{M}, \nabla) \).

The nullity of \( \overline{q} \) as critical point is equal to the multiplicity of \( \overline{v} \) as focal point.

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

**THEOREM 6.6.** For almost all \( p \in \mathbb{R}^n \) (all but a set of measure 0) the function \( L_p : \mathcal{M} \rightarrow \mathbb{R} \)

has no degenerate critical points.

This theorem has several interesting consequences.

**COROLLARY 6.7.** On any manifold \( \mathcal{M} \) there exists a differentiable function, with no degenerate critical points, for which each \( \mathcal{M}^h \) is compact.

**PROOF:** This follows from Theorem 6.6 and the fact that an n-dimensional manifold \( \mathcal{M} \) can be embedded differentially as a closed subset of \( \mathbb{R}^{2n+1} \) (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 \( \mathcal{M} \) there is a vector field \( X \) such that the sum of the indices of the critical points of \( X \) equals \( \chi(\mathcal{M}) \), the Euler characteristic of \( \mathcal{M} \). This can be seen as follows: for any differentiable function \( f \) on \( \mathcal{M} \) we have \( \chi(\mathcal{M}) = \sum (-1)^{\lambda} \varepsilon_\lambda \) where \( \varepsilon_\lambda \) is the number of critical points with index \( \lambda \). But \( (-1)^{\lambda} \) is the index of the vector field \( \nabla f \) at a point where \( f \) has index \( \lambda \).

\[ \frac{\partial^2 f}{\partial u_i \partial u_j} = \pm (\overline{x} - \mathbf{p}) \]

\[ \frac{\partial^2 f}{\partial u_i \partial u_j} = \pm (\overline{x} - \mathbf{p}) \]

It follows that the sum of the indices of any vector field on \( \mathcal{M} \) is equal to \( \chi(\mathcal{M}) \) because this sum is a topological invariant (see Steenrod, The Topology of Fibre Bundles, p. 139.7).

The preceding corollary can be sharpened as follows. Let \( k \) be an integer and let \( \mathcal{K} \subset \mathcal{M} \) be a compact set.

**COROLLARY 6.8.** Any bounded smooth function \( f : \mathcal{M} \rightarrow \mathbb{R} \) 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 \( \mathcal{K} \) uniformly approximate the corresponding derivatives of \( f \), for \( 1 \leq i \leq k \).

(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 embedding \( h : \mathcal{M} \rightarrow \mathbb{R}^n \) of \( \mathcal{M} \) as a bounded subset of some euclidean space so that the first coordinate \( \mathbb{R}^1 \) 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 \( (-c, \varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n) \in \mathbb{R}^n \) so that the function \( L_p : \mathcal{M} \rightarrow \mathbb{R} \) is non-degenerate; and set
\[ g(x) = \frac{h_1(x) - c^\varepsilon}{\varepsilon} \]
Clearly \( g \) is non-degenerate. A short computation shows that
\[ g(x) = f(x) + \sum_{1 \leq i \leq n} \frac{\varepsilon_1 h_i(x)^2}{\varepsilon} - \sum_{1 \leq i \leq n} \varepsilon_1 h_i(x)/\varepsilon_1 \]
Clearly, if \( c \) is large and the \( \varepsilon_1 \) are small, then \( g \) will approximate \( f \) as required.

The above theory can also be used to describe the index of the function \( L_p : \mathcal{M} \rightarrow \mathbb{R} \)

at a critical point

**LEMMA 6.9.** (Index theorem for \( L_p \)). The index of \( L_p \) at a non-degenerate critical point \( q \in \mathcal{M} \) is equal to the number of focal points of \( (\mathcal{M}, q) \) which lie on the segment from \( q \) to \( p \), each focal point being counted with its multiplicity.
I. NON-DEGENERATE FUNCTIONS

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

PROOF: The index of the matrix

\[
\begin{pmatrix}
\frac{\partial^2 F}{\partial u_i \partial u_j} \\
\frac{\partial^2 F}{\partial u_i \partial u_j}
\end{pmatrix} = \sum (g_{ij} - \epsilon_i \cdot \epsilon_j)
\]

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 \( (\epsilon_i \cdot \epsilon_j) \) which are \( \geq \frac{1}{2} \). 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 \( M \subset \mathbb{C}^n \) is a non-singular affine algebraic variety in complex n-space with real dimension \( 2k \), then

\[ H_i(M; \mathbb{Z}) = 0 \quad \text{for} \quad i > k. \]

This is a consequence of the stronger:

**THEOREM 7.2.** A complex analytic manifold \( M \) of complex dimension \( k \), biaalytically embedded as a closed subset of \( \mathbb{C}^n \) 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

\[ Q(z_1, \ldots, z_k) = \sum b_{ij} z_i z_j. \]

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:

\[ Q'(x^1, \ldots, x^k, y^1, \ldots, y^k) = \text{real part of} \sum b_{ij}(x^h z^h)(x^j z^j). \]

**ASSERTION 1.** If \( e \) is an eigenvalue of \( Q' \) with multiplicity \( \mu \), then \(-e\) is also an eigenvalue with the same multiplicity \( \mu \).

PROOF: The identity \( Q(iz^1, \ldots, iz^k) = -\text{Re}(z_1, \ldots, z^k) \) shows that the quadratic form \( Q' \) can be transformed into \(-Q'\) by an orthogonal change of variables. Assertion 1 clearly follows.

Let $M$ be a complex manifold which is biholomorphically imbedded as a subset of $\mathbb{C}^n$. Let $q$ be a point of $M$.

**Assertion 1.** The focal points of $(M, q)$ along any normal line $\ell$ 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 $z^1, \ldots, z^n$ for $M$ in a neighborhood of $q$ so that $z^1(q) = \ldots = z^k(q) = 0$. The inclusion map $M \to \mathbb{C}^n$ determines $n$ complex analytic functions

$$v_\alpha = w_\alpha(z^1, \ldots, z^n), \quad \alpha = 1, \ldots, n.$$ 

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

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

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

$$\sum w_\alpha(z^1, \ldots, z^n) v_\alpha = \text{constant} + Q(z^1, \ldots, z^n) + \text{higher terms},$$

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

Now substitute $x^\alpha + iy^\alpha$ for $z^\alpha$ 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

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

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 $p \in \mathbb{C}^n$ so that the squared-distance function $\rho_p(v)$ is a homogeneous quadratic function. (The constant term is $0$.)

**Theorem 7.2** (Lefschetz). Let $V$ be an algebraic variety of complex dimension $k$ which lies in the complex projective space $\mathbb{C}P^n$. Let $P$ be a hyperplane in $\mathbb{C}P^n$ 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

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

is onto.

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

$$H_k(V, V \cap P; \mathbb{Z}) = H^{n-k}(V - (V \cap P); \mathbb{Z}).$$

But $V - (V \cap P)$ is a non-singular algebraic variety in the affine space $\mathbb{C}P^n - P$. Hence it follows from 7.2 that the last group is zero for $k \leq k-1$. This result can be sharpened as follows:

**Theorem 7.4** (Lefschetz). Under the hypothesis of the preceding corollary, the relative homotopy group $\pi_k(V, V \cap P)$ is zero for $k < k$. 
I. NON-DEGENERATE FUNCTIONS

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 \( L_p : V \to V \cap P \to \mathbb{R} \) we will use \( f: V \to \mathbb{R} \) 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 \( r_p \) 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 \( \mathbb{V} = f^{-1}[0, \varepsilon] \) with finitely many cells of dimension \( \geq k \) attached.

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

\[
(I^r, I^r) \to (\mathbb{V}, V \cap P) \subseteq (U, V \cap P),
\]

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." The Mathematical Society of Japan, 1956; Laugwitz, "Differential-geometric," Teubner 1960; or Helgason, "Differential geometry and symmetric spaces," Academic Press, 1962.

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 \( X_p \in \mathbb{T}_pM \) and to each vector field \( Y \) a new tangent vector

\[
X_p \vdash Y \in \mathbb{T}_pM
\]
called the covariant derivative* of \( Y \) in the direction \( X \). This is required to be bilinear as a function of \( X_p \) and \( Y \). Furthermore, if

\[
f: M \to \mathbb{R}
\]
is a real valued function, and if \( \nabla Y \) denotes the vector field

\[
(\nabla Y)_q = f(q)Y_q
\]
then \( \vdash \) is required to satisfy the identity

\[
X_p \vdash (\nabla Y) = (X_p f)Y_p + f(p)X_p \vdash Y.
\]

* 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 \( Y \).
11. RIEMANNIAN GEOMETRY

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 $\Gamma_{ij}$ at $p$, satisfying the following smoothness condition.

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

$$ (X \cdot Y)_p = X_p \cdot Y,$$

must also be smooth.

Note that:

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

3) $(fX) \cdot Y = f(X \cdot Y)$.

4) $(X + Y) \cdot Z = X \cdot Z + Y \cdot Z$.

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 $\cdot$ is determined by $\Gamma_{ij}^k$ smooth real valued functions $\Gamma_{ij}^k$ on $U$, as follows. Let $\partial_k$ denote the vector field $\frac{\partial}{\partial u^k}$ on $U$. Then any vector field $X$ on $U$ can be expressed uniquely as

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

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

$$ \partial_i \cdot \partial_j = \sum_k \Gamma_{ij}^k \partial_k.$$

These functions $\Gamma_{ij}^k$ 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 $X \cdot Y$ by the rules (2), (3), (4); yielding the formula

$$ X \cdot Y = \sum_k \left( \sum_i x^i y^k_{ij} \right) \partial_k,$$

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

$$ y^k_{ij} = \partial_j y^k + \sum_j \Gamma_{ij}^k y^j.$$

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

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

A parametrized curve in $M$ is a smooth function $c$ from the real numbers to $M$. A vector field $V$ along the curve $c$ is a function which assigns to each $t \in \mathbb{R}$ a tangent vector $V_t \in \mathcal{T}_c(t)$. 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 $\mathbb{R}$.

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

$$ \frac{dc}{dt} = c_x c.$$  

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

$$ c_x: \mathbb{R} \to \mathcal{T}_c(t)$$

denotes the homomorphism of tangent spaces induced by the map $c$. (Compare 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 \( V \rightarrow \frac{DV}{dt} \) is characterized by the following three axioms.

a) \( \frac{D(VW)}{dt} = \frac{DV}{dt} + \frac{DW}{dt} \).

b) If \( f \) is a smooth real valued function on \( \mathbb{R} \), then \( \frac{D(fV)}{dt} = \frac{d}{dt} f \cdot \frac{DV}{dt} \).

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 \( \frac{d}{dt} Y + Y \) (the covariant derivative of \( Y \) in the direction of the velocity vector of \( c \)).

**DEFINITION.** There is one and only one operation \( V \rightarrow \frac{DV}{dt} \) 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 \( \mathbb{R} \) (or an appropriate open subset of \( \mathbb{R} \)), and \( \partial_1, \ldots, \partial_n \) are the standard vector fields on the coordinate neighborhood. It follows from (a), (b), and (c) that

\[
\left( \frac{DV}{dt} \right)^j = v^i \frac{d}{dt} \partial_i + \partial_j
\]

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.

## §8. COVARIANT DIFFERENTIATION

**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

\[
\frac{dv^k}{dt} + \sum \frac{du^i}{dt} r^k_{ij} v^j = 0
\]

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 **parallel translation** 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 \( \mathcal{X} \) on \( M \) is compatible 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 \( \langle P_p, P'_p \rangle \) 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 \( F_1, \ldots, F_n \) 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 \( \sum_v v^i F_i \) and \( \sum_v v^i F_i \) respectively (where \( v^i = \langle F_i, P_i \rangle \) is a real valued function on \( \mathbb{R} \)). It follows that \( \langle V, W \rangle = \sum v^i w^i \) and that

\[
\frac{DV}{dt} = \sum \frac{dv^i}{dt} F_i, \quad \frac{DW}{dt} = \sum \frac{dw^i}{dt} F_i.
\]

Therefore

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

which completes the proof.
COROLLARY 8.4. For any vector fields \( Y, Y' \) on \( M \) and any vector \( X_p \in \mathcal{T}_pM: \)
\[
X_p \langle Y, Y' \rangle = \langle X_p + Y, Y' \rangle - \langle Y, X_p + Y' \rangle.
\]

PROOF. Choose a curve \( c \) whose velocity vector at \( t = 0 \) is \( X_p \) and apply 8.3.

DEFINITION 8.5. A connection \( \nabla \) is called symmetric if it satisfies the identity*
\[
\langle X \nabla Y - (Y \nabla X) \rangle = [X,Y],
\]
(As usual, \( [X,Y] \) denotes the poison bracket \( [X,Y] = X(Yf) - Y(Xf) \) 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
\[
\Gamma^k_{ij} - \Gamma^k_{ji} = 0.
\]
Conversely, if \( \Gamma^k_{ij} - \Gamma^k_{ji} = 0 \) then using formula (6) it is not difficult to verify that the connection \( \nabla \) 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 \( \langle \partial_i \nabla \partial_j, \partial_k \rangle = g_{jk} \) one obtains the identity
\[
\partial_i g_{jk} = \langle \partial_i \nabla \partial_j, \partial_k \rangle + \langle \partial_j \nabla \partial_i, \partial_k \rangle + \langle \partial_k \nabla \partial_i, \partial_j \rangle.
\]
Permuting \( i,j,k \) this gives three linear equations relating the three quantities
\[
\langle \partial_i \nabla \partial_j, \partial_k \rangle, \quad \langle \partial_j \nabla \partial_i, \partial_k \rangle, \quad \text{and} \quad \langle \partial_k \nabla \partial_i, \partial_j \rangle.
\]
(There are only three such quantities since \( \partial_i \nabla \partial_j = \partial_j \nabla \partial_i \).) These equations can be solved uniquely; yielding the first Christoffel identity
\[
\langle \partial_i \nabla \partial_j, \partial_k \rangle = g^{kl} (\partial_k g_{jk} + \partial_j g_{ik} - \partial_i g_{jk}) \partial_k.
\]
The left hand side of this identity is equal to \( \sum \Gamma^l_{ij} g_{lk} \). Multiplying by the inverse \( (g^{kf}) \) of the matrix \( (g_{ik}) \) this yields the second Christoffel identity
\[
\Gamma^f_{ij} = \sum_k g^{lf} (\partial_k g_{jk} + \partial_j g_{ik} - \partial_i g_{jk}) g^{kf}.
\]
Thus the connection is uniquely determined by the metric.

Conversely, defining \( \Gamma^f_{ij} \) 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
\[
s: \mathbb{R}^2 \to M.
\]
By a vector field \( V \) along \( s \) is meant a function which assigns to each \( (x,y) \in \mathbb{R}^2 \) a tangent vector
\[
V(x,y) \in \mathcal{T}_{s(x,y)}M.
\]
As examples, the two standard vector fields \( \partial_x \) and \( \partial_y \) give rise to vector fields \( s_* \partial_x \) and \( s_* \partial_y \) along \( s \). These will be denoted briefly by \( \partial_x \) and \( \partial_y \); and called the "velocity vector fields" of \( s \).

For any smooth vector field \( V \) along \( s \) the covariant derivatives \( \frac{\partial V}{\partial x} \) and \( \frac{\partial V}{\partial y} \) are new vector fields, constructed as follows. For each fixed \( y_0 \), restricting \( V \) to the curve
\[
x \mapsto s(x,y_0)
\]
one obtains a vector field along this curve. Its covariant derivative with respect to \( x \) is defined to be \( \left( \frac{\partial V}{\partial x} \right)(x,y_0) \) . This defines \( \frac{\partial V}{\partial x} \) 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 \( \frac{D}{\partial x} \frac{\partial s}{\partial y} \) and \( \frac{D}{\partial y} \frac{\partial s}{\partial x} \) 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 \( \nabla \) measures the extent to which the second covariant derivative \( \nabla_1 (\nabla_2 f) \) is symmetric in 1 and 2. Given vector fields \( X, Y, Z \) define a new vector field \( R(X, Y)Z \) by the identity

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

**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

\[
X_p, Y_p, Z_p \rightarrow R(X_p, Y_p)Z_p
\]

from \( T_p M \times T_p M \times T_p M \) to \( T_p M \) 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 \( -X \nabla (Y \nabla Z), \)
\( Y \nabla (X \nabla Z), \) \( [X, Y] \nabla Z \) are replaced respectively by

1. \( -fX \nabla (Y \nabla Z), \)
2. \( (Yf)(X \nabla Z) + fY \nabla (X \nabla Z), \)
3. \( (Yf)(X \nabla Z) + f[X, Y] \nabla Z. \)

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 \( X = \sum x^i \partial_i, \ Y = \sum y^i \partial_i, \) and \( Z = \sum z^k \partial_k. \)

Then

\[
R(X, Y)Z = \sum R(x^i \partial_i, y^j \partial_j)(z^k \partial_k)
\]

\[
= \sum x^i y^j z^k R(\partial_i, \partial_j) \partial_k.
\]

* Nominal gives \( R \) the opposite sign. Our sign convention has the advantage that (in the Riemannian case) the inner product coincides with the classical symbol \( * \).
Evaluating this expression at \( p \) one obtains the formula
\[
(R(X,Y)Z)_p = \sum x^j(p)y^k(p)\left(\delta_{ij}\delta_{kj}\right)_p
\]
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
\[ s: \mathbb{R}^2 \to M. \]

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

**Lemma 9.2.** \( \frac{\partial}{\partial y} \frac{\partial}{\partial x} V - \frac{\partial}{\partial x} \frac{\partial}{\partial y} V = R\left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right) V \).

**Proof:** Express both sides in terms of a local coordinate system, and compute, making use of the identity
\[
\partial_j \left( \partial_i \alpha_k - \partial_i \alpha_k \right) - \partial_i \left( \partial_j \alpha_k - \partial_j \alpha_k \right) = R(\partial_i, \partial_j) \alpha_k.
\]

It is interesting to ask whether one can construct a vector field \( P \) along \( s \) which is parallel, in the sense that
\[
\frac{\partial}{\partial x} P = \frac{\partial}{\partial y} P = 0,
\]
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
\[ y = s(x_0, y), \]
having the right value for \( y = 0 \). This defines \( P \) everywhere along \( s \).

Clearly \( \frac{\partial}{\partial y} P \) is identically zero, and \( \frac{\partial}{\partial x} P \) is zero along the \( x \)-axis. Now the identity
\[
\frac{\partial}{\partial y} \frac{\partial}{\partial x} P - \frac{\partial}{\partial x} \frac{\partial}{\partial y} P = R\left( \frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right) P = 0
\]
implies that \( \frac{\partial}{\partial y} \frac{\partial}{\partial x} P = 0 \). In other words, the vector field \( \frac{\partial}{\partial x} P \) is parallel along the curves
\[ y = s(x_0, y). \]

§9. The curvature tensor

Since \( \left( \frac{\partial}{\partial x} P \right)(x_0,0) = 0 \), this implies that \( \frac{\partial}{\partial x} P \) 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. \( \langle R(X,Y)Z,W \rangle + \langle R(Y,Z)W,X \rangle = 0 \)
4. \( \langle R(X,Y)Z,W \rangle - \langle R(Z,X)W,Y \rangle = 0 \).

**Proof:** For 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
\[
- X \wedge (Y \wedge Z) + Y \wedge (X \wedge Z) + Z \wedge (Y \wedge X) - Z \wedge (X \wedge Y) + X \wedge (-Z \wedge Y) = 0.
\]

But the symmetry of the connection implies that
\[
Y \wedge Z - Z \wedge Y = [Y,Z] = 0.
\]

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
\[
\langle R(X,Y)Z,W \rangle = 0
\]
for all \( X,Y,Z \). Again we may assume that \([X,Y] = 0 \), so that \( \langle R(X,Y)Z,W \rangle \) is equal to
\[
\langle -X \wedge (Y \wedge Z) + Y \wedge (X \wedge Z), W \rangle.
\]
In other words we must prove that the expression
\[ \langle Y + (X + Z), Z \rangle \]
is symmetric in \( X \) and \( Y \).

Since \( [X,Y] = 0 \) the expression \( \langle X \wedge Z, Z \rangle \) is symmetric in \( X \) and \( Y \). Since the connection is compatible with the metric, we have
\[ X \wedge Z, Z = 2 \langle X + Z, Y + Z \rangle \]
hence
\[ \langle X \wedge Z, Z \rangle = 2 \langle Y + (X + Z), Z \rangle + 2 \langle X + Z, Y + Z \rangle . \]
But the right hand term is clearly symmetric in \( X \) and \( Y \). Therefore
\[ \langle Y + (X + Z), Z \rangle \]
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.

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

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

\[
d\overline{u} = f(\overline{u}, \overline{v})
\]

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

\[
\overline{u}(0) = \overline{u}_0, \quad \frac{d\overline{u}}{dt}(0) = \overline{v}_0.
\]

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

\[
(\overline{u}_0, \overline{v}_0, t) \rightarrow \overline{u}(t)
\]

from \( W \times (\varepsilon, \varepsilon) \) to \( R^n \) is a \( C^\infty \) function of all \( 2n + 1 \) variables.

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

\[
\begin{align*}
\frac{d\overline{u}}{dt} &= \overline{v}, \\
\frac{d\overline{v}}{dt} &= f(\overline{u}, \overline{v}).
\end{align*}
\]

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

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 \( v \in T_p M \) with length \( < \varepsilon \), there is a unique geodesic

\[
\gamma_v: (-\varepsilon, \varepsilon) \rightarrow M
\]

satisfying the conditions

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

**PROOF:** If we were willing to replace the interval \((-\varepsilon, \varepsilon)\) 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 \( v \in T_p M \) with

\[
|v| < \varepsilon_1
\]

there is a unique geodesic

\[
\gamma_v: (-2\varepsilon_2, 2\varepsilon_2) \rightarrow M
\]

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 \( \varepsilon \) be any constant. If the parametrized curve

\[
t \rightarrow \gamma(t)
\]

is a geodesic, then the parametrized curve

\[
t \rightarrow \gamma(ct)
\]

will also be a geodesic.

Now suppose that \( \varepsilon \) is smaller than \( \varepsilon_1 \varepsilon_2 \). Then if \( |v| < \varepsilon \) and

\[
|v/\varepsilon_2| < \varepsilon_1 \quad \text{and} \quad |\varepsilon_2| < 2\varepsilon_2,
\]

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 \( v \in T_p M \) be a tangent vector, and suppose that there exists a geodesic

\[
\gamma: [0,1] \rightarrow M
\]

satisfying the conditions

\[
\gamma(0) = p, \quad \frac{d\gamma}{dt}(0) = v.
\]

Then the point \( \gamma(t) \in M \) will be denoted by \( \exp_p(v) \) and called the exponential of the tangent vector \( v \). The geodesic \( \gamma \) can thus be described by the formula

\[
\gamma(t) = \exp_p(tv).
\]

* 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 \( T_M \) at the identity can be identified with the space of \( n \times n \) skew-Hermitian matrices. The function

\[
\exp: T_M \rightarrow M
\]

as defined above is then given by the exponential power series

\[
\exp(A) = I + A + \frac{1}{2} A^2 + \frac{1}{3!} A^3 + \ldots.
\]
II. RIEMANNIAN GEOMETRY

58

Lemma 10.2 says that \( \exp_q(v) \) is defined proving 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 \( \exp(v) \) 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 \( \gamma_0: [a,b] \to M \) it should be possible to extend \( \gamma_0 \) to an infinite geodesic \( \gamma: \mathbb{R} \to M \).

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 \( C^\infty \) structure: if \((u^1,\ldots,u^n) \) is a coordinate system in an open set \( U \subset M \) then every tangent vector at \( q \in U \) can be expressed uniquely as \( t^1(\partial_1)+\cdots+t^m(\partial_n) \), where \( \partial_i = \frac{\partial}{\partial u^i} \mid q \). Then the functions \( u^1,\ldots,u^n,t^1,\ldots,t^n \) 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 \( F: V \to \mathbb{R} \times M \) 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 \((u^1,\ldots,u^n,u^1_1,\ldots,u^n_2)\), we have

\[
\begin{align*}
F_*\left(\frac{\partial}{\partial u^i}\right) &= \frac{\partial}{\partial u^i} + \frac{\partial}{\partial u^i_1} + \frac{\partial}{\partial u^i_2} \\
F_*\left(\frac{\partial}{\partial t^i}\right) &= \frac{\partial}{\partial t^i}
\end{align*}
\]

Thus the Jacobian matrix of \( F \) at \((p,0)\) has the form \(\begin{pmatrix} I & I \end{pmatrix} \), 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 \( F(V') \subset W \times W \). 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., \( q = \exp_q(tv), \varepsilon_t = \exp_q(tv), 0 \leq t \leq 1 \), is the geodesic joining \( q_1 \) and \( q_2 \), then the pair \((q_1,v)\) \( TM \) depends differentiably on \((q_1,q_2)\).
3. For each \( q \in W \) the map \( \exp_q \) maps the open \( \varepsilon \)-ball in \( TM_q \) diffeomorphically onto an open set \( U_q \subset W \).

**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 arclength.

**Theorem 10.4.** Let \( W \) and \( \varepsilon \) 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

\[
\int_0^1 \frac{\|\gamma'\|^2}{\varepsilon} \, dt \leq \int_0^1 \frac{\|\omega'\|^2}{\varepsilon} \, dt
\]

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

Thus \( \gamma \) 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

\[
\{ \text{exp}_q(v) : v \in TM_q, \quad \|v\| = \text{constant} \}.
\]

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

\[ t \mapsto \text{exp}_q(r_0 v(t)) \]

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

\[ r \mapsto \text{exp}_q(r v(t_0)) \]

In terms of the parametrized surface $f$ given by

\[ f(r, t) = \exp_q(r v(t)), \quad 0 \leq r < \varepsilon \]

we must prove that

\[ \langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \rangle = 0 \]

for all $(r, t)$.

Now

\[ \frac{\partial}{\partial r} \langle \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \rangle = -\langle \frac{\partial}{\partial r} \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \rangle - \langle \frac{\partial}{\partial r} \frac{\partial f}{\partial t}, \frac{\partial f}{\partial t} \rangle = -\langle \frac{\partial^2 f}{\partial r^2}, \frac{\partial f}{\partial t} \rangle - \langle \frac{\partial f}{\partial t}, \frac{\partial f}{\partial t} \rangle = 0. \]

The first expression on the right is zero since the curves

\[ r \mapsto f(r, t) \]

are geodesics. The second expression is equal to

\[ -\langle \frac{\partial}{\partial r} \frac{\partial f}{\partial r}, \frac{\partial f}{\partial t} \rangle = -\langle \frac{\partial^2 f}{\partial r \partial t}, \frac{\partial f}{\partial t} \rangle = 0, \]

since $\|\frac{\partial f}{\partial t}\| = \|v(t)\| = 1$. Therefore the quantity $\langle \frac{\partial^2 f}{\partial r \partial t}, \frac{\partial f}{\partial t} \rangle$ is independent of $r$. But for $r = 0$ we have

\[ f(0, t) = \exp_q(0) = q \]

define $\frac{\partial f}{\partial t}(0, t) = 0$. Therefore $\langle \frac{\partial^2 f}{\partial r \partial t}, \frac{\partial f}{\partial t} \rangle$ is identically zero, which completes the proof.

Now consider any piecewise smooth curve

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

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

LEMMA 10.6. The length

\[ \int_a^b \sqrt{\frac{\text{d} s}{\text{d} t}} \text{dt} \]

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 $f(r, t) = \exp_q(r v(t))$, so that $\omega(t) = f(r(t), t)$

Then

\[ \frac{\text{d} s}{\text{d} t} = \frac{\partial f}{\partial r} \frac{\partial f}{\partial t} = \frac{\partial r}{\partial t} v + r \frac{\partial v}{\partial t}. \]

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

\[ \frac{\text{d} s}{\text{d} t} \geq |r'(t)| \quad \text{for all} \quad t \in [a, b]. \]

This completes the proof.

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

\[ q' = \exp_q(rv) \in U_q; \]

where $0 < r < \varepsilon$, $\|v\| = 1$. Then for any $s > 0$ the path $\omega$ must contain a segment joining the spherical shell of radius $s$ to the spherical shell of radius $r$, and lying between these two shells. The length of this segment will be $\geq r - s$; hence letting $s$ tend to $0$ the length of $\omega$ 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 $\omega : [0, 1] \to M$, parametrized by arc-length, has length less than or equal to the length of any other path from $\omega(0)$ to $\omega(1)$. Then $\omega$ is a geodesic.

PROOF: Consider any segment of $\omega$ 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 $\omega$ is a geodesic.

DEFINITION. A geodesic $\gamma : [a, b] \to M$ will be called minimal 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 $\pi$, 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 $d(p,q)$ 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 $\delta$ are joined by a unique geodesic of length less than $\delta$. Furthermore this geodesic is minimal: and depends differentiably on its endpoints.

**Proof:** Cover $K$ by open sets $U_j$ as in 10.3. and let $\delta$ be small enough so that any two points in $K$ with distance less than $\delta$ lie in a common $U_j$. This completes the proof.

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

**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 $\delta$. In fact the triangle inequality states that $d(p,p') \geq d(p,q) - d(q,p') = r + \delta'$

We claim that $p'$ is equal to $r(t_0 + \delta')$. In fact the triangle inequality states that $d(p,p') \geq d(p,q) - d(q,p') = r + \delta'$

(making use of (2)). But a path of length precisely $t_0 + \delta'$ from $p$ to $p'$ is obtained by following $\gamma$ from $p$ to $\gamma(t_0)$, and then following a minimal geodesic from $\gamma(t_0)$ to $p'$. Since this broken geodesic has minimal length, it follows from Corollary 10.7 that it is an (unbroken)
and hence coincides with $7$.

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

$$(t_0 + \delta') \gamma(t_0 + \delta') = r - (t_0 + \delta')t_0'$$

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 $exp_p: TM_p \rightarrow M$ 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.

**§10. GEODESICS AND COMPLETEINESS**

**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 $R^k_{ij} = 0$ and the equations for a geodesic $7$, given by $t \rightarrow (x_1(t), \ldots, x_n(t))$ become

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

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

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

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 $E^k$ is an isometry $I: S^n \rightarrow S^n$ whose fixed point set is $C = S^n \cap E^k$. 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 $I(t)$ 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 continuum 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 $I: Z - L \rightarrow \mathbb{R}^2$ by rolling $Z$ onto $\mathbb{R}^2$:

![Diagram of a helix being rolled into a plane]

The geodesics on $Z$ are just the images under $I$ of the straight lines in $\mathbb{R}^2$. 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 $\omega: [0,1] \rightarrow M$ 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 $C^\infty$;

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 $\mathcal{P}(M;p,q)$, or briefly by $\mathcal{P}(M)$ or $\mathcal{P}$.

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

By the tangent space of $\mathcal{P}$ at a path $\omega$ will be meant the vector space consisting of all piecewise smooth vector fields $W$ along $\omega$ for which $W(0) = 0$ and $W(1) = 0$. The notation $T_{\omega}\mathcal{P}$ will be used for this vector space.

If $F$ is a real valued function on $\mathcal{P}$ it is natural to ask what

$$F_*: T_{\omega}\mathcal{P} \rightarrow TR_F(\omega),$$

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 $\mathcal{M}$, we can define $F_*: T_M \mathcal{M} \rightarrow TR_F(\mathcal{M})$ as follows. Given $X \in T_M \mathcal{M}$ choose a smooth path $u \rightarrow \sigma(u)$ in $\mathcal{M}$ which is defined for $-\epsilon < u < \epsilon$, so that $\sigma(0) = p$, $\frac{d\sigma(0)}{du} = X$.
Then $F_\epsilon(X)$ is equal to $\frac{d(F(\alpha(u)))}{du}\bigg|_{u=0}$, multiplied by the basis vector $\left(\frac{d}{du}F(p)\right) \in T_{F(p)}$. 

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

**DEFINITION.** A *variation* of $\alpha$ (keeping endpoints fixed) is a function $6: (-\epsilon, \epsilon) \rightarrow \Omega$, for some $\epsilon > 0$, such that

1) $\alpha(0) = \alpha$

2) there is a subdivision $0 = t_0 < t_1 < \ldots < t_k = 1$

of $[0,1]$ so that the map $a: (-\epsilon, \epsilon) \times [0,1] \rightarrow \mathbb{R}^n$ defined by $a(u,t) = \alpha(u)(t)$ is $C^\infty$ on each strip $(-\epsilon, \epsilon) \times [t_{i-1}, t_i]$, $i = 1, \ldots, k$.

Since each $\alpha(u)$ belongs to $\alpha \in \mathbb{W}(\Sigma; p, q)$, note that:

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

We will use either $a$ or $\alpha$ to refer to the variation. More generally, if, in the above definition, $(-\epsilon, \epsilon)$ is replaced by a neighborhood $U$ of 0 in $\mathbb{R}^n$, then $a$ (or $6$) is called an *n-parameter variation* of $\alpha$.

Now 6 may be considered as a "smooth path" in $\Omega$. Its "velocity vector" $\frac{d\alpha}{du}(0) \in T_{\alpha}$ is defined to be the vector field $W$ along $\alpha$ given by

$$W = \frac{d\alpha}{du}(0) = \frac{d\alpha}{du}(0,t).$$

Clearly $W \in T_{\alpha}$. This vector field $W$ is also called the *variation vector field* associated with the variation $a$.

Given any $W \in T_{\alpha}$, note that there exists a variation $\tilde{\alpha}: (-\epsilon, \epsilon) \rightarrow \Omega$ which satisfies the conditions $\tilde{\alpha}(0) = \alpha$, $\frac{d\tilde{\alpha}}{du}(0) = W$.

In fact one can set

$$\tilde{\alpha}(u)(t) = \exp_{\alpha(t)}(uW).$$

By analogy with the definition given above, if $F$ is a real valued function on $\Omega$, we attempt to define

$$F_\epsilon: T_{\alpha} \rightarrow T_{F(\alpha)}$$
as follows. Given $W \in T_{\alpha}$ choose a variation $\tilde{\alpha}: (-\epsilon, \epsilon) \rightarrow \Omega$ with

$$\tilde{\alpha}(0) = \alpha, \quad \frac{d\tilde{\alpha}}{du}(0) = W$$

and set $F_\epsilon(W)$ equal to $\left.\frac{d(F(\tilde{\alpha}(u)))}{du}\right|_{u=0}$ multiplied by the tangent vector $\left(\frac{d}{du}F(\alpha)\right)$. Of course without hypothesis on $F$ there is no guarantee that this derivative will exist, or it will be independent of the choice of $\tilde{\alpha}$.

We will not investigate what conditions $F$ must satisfy in order for $F_\epsilon$ to have these properties. We have indicated how $F_\epsilon$ might be defined only to motivate the following.

**DEFINITION.** A path $\alpha$ is a *critical path* for a function $F: \Omega \rightarrow \mathbb{R}$ if and only if $\frac{d(F(\tilde{\alpha}(u)))}{du}$ is zero for every variation $\tilde{\alpha}$ of $\alpha$.

**EXAMPLE.** If $F$ takes on its minimum at a path $\alpha_0$, and if the derivatives $\frac{d(F(\tilde{\alpha}(u)))}{du}$ are all defined, then clearly $\alpha_0$ 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 $E$ of $\omega$ from $a$ to $b$ (where $0 \leq a < b \leq 1$) as

$$E^b_a(\omega) = \int_a^b \|dv\| dt.$$ 

We will write $E$ for $E^1_0(\omega)$.

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

$$L^b_a(\omega) = \int_a^b \|d\omega\| dt,$$

as follows. Applying Schwarz’s inequality

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

with $f(t) = 1$ and $g(t) = \|d\omega\|$ we see that

$$\left( L^b_a(\omega) \right)^2 \leq (b - a) E^b_a(\omega),$$

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

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

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

Here the equality $L(\gamma)^2 = L(\omega)^2$ can hold only if $w$ 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 $w$. This proves that $E(\gamma) < E(\omega)$ unless $\omega$ 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

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

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

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

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

$$\Delta_t v = V_{t^+} - V_{t^-} = \text{discontinuity in the velocity vector at } t,$$

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

$$1 \frac{dE(\omega(t))}{du}$$

is equal to

$$\int \langle W_t, A_t \rangle dt.$$

**Proof:** According to Lemma 8.3, we have

$$\frac{d}{du} \int \langle \omega(t), \omega(t) \rangle = 2 \langle \frac{d}{du} \omega(t), \omega(t) \rangle.$$

Therefore

$$\frac{dE(\omega(t))}{du} = \frac{d}{du} \int_0^1 \langle \frac{d}{du} \omega(t), \omega(t) \rangle dt = 2 \int_0^1 \langle \frac{d}{du} \omega(t), \omega(t) \rangle dt.$$

By Lemma 8.5 we can substitute $\frac{D}{dt} \frac{d\omega}{dt}$ for $\frac{D}{dt} \frac{d\omega}{du}$ in this last formula.

Choose $0 < t_0 < t_1 < \ldots < t_k = 1$ so that $\delta$ 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

$$\frac{d}{dt} \langle \delta, \omega \rangle = \langle \frac{D}{dt} \frac{d\omega}{dt}, \omega \rangle + \langle \omega, \frac{D}{dt} \frac{d\omega}{dt} \rangle$$

implies that

$$\int_{t_{i-1}}^{t_i} \langle \frac{D}{dt} \frac{d\omega}{dt}, \omega \rangle dt = \langle \frac{d}{dt} \frac{d\omega}{dt}, \omega \rangle \bigg|_{t=t_i}^{t=t_{i-1}} - \int_{t_{i-1}}^{t_i} \langle \omega, \frac{D}{dt} \frac{d\omega}{dt} \rangle dt.$$
111. CALCULUS OF VARIATIONS

Adding up the corresponding formulas for $i = 1, \ldots, k$; and using the fact that $\frac{\partial^2}{\partial u^2} = 0$ for $t = 0$ or 1, this gives

$$\frac{1}{\tau} \frac{d}{du} \left( \frac{d}{dt} \right) = - \sum_{i=1}^{k-1} \left< \frac{\partial^2}{\partial u^2}, \epsilon_i \right> + \int_0^\tau \left< \frac{\partial^2}{\partial u^2}, \frac{d}{dt} \epsilon_i \right> \, dt .$$

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

$$\frac{1}{\tau} \frac{d}{du} \left( \frac{d}{dt} \right) = - \sum_{i=1}^{k-1} \left< \frac{\partial^2}{\partial u^2}, \epsilon_i \right> + \int_0^\tau \left< \frac{\partial^2}{\partial u^2}, A(t) \right> \, dt .$$

This completes the proof.

Intuitively, the first term in the expression for $\frac{d}{du} \left( \frac{d}{dt} \right)$ shows that varying the path $\omega$ in the direction of decreasing "kink," tends to decrease $E$; see Diagram 11.

![Diagram 11](attachment:diagram11.png)

The second term shows that varying the curve in the direction of its acceleration vector $\frac{d}{dt} \left( \frac{d}{du} \right)$ tends to reduce $E$.

Recall that the path $\omega \in \Omega$ is called a geodesic if and only if $\omega$ is $C^\infty$ on the whole interval $[0,1]$ and the acceleration vector $\frac{d}{dt} \left( \frac{d}{du} \right)$ of $\omega$ is identically zero along $\omega$.

COROLLARY 12.3. The path $\omega$ is a critical point for the function $E$ if and only if $\omega$ is a geodesic.

PROOF: Clearly a geodesic is a critical point. Let $\omega$ be a critical point. There is a variation of $\omega$ with $W(t) = f(t)A(t)$ where $f(t)$ is positive except that it vanishes at the $t_i$. Then

$$\frac{1}{\tau} \frac{d}{du} \left( \frac{d}{dt} \right) = - \int_0^\tau f(t) < A(t), A(t) > \, dt .$$

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

$$\frac{1}{\tau} \frac{d}{du} \left( \frac{d}{dt} \right) = - \sum \left< \Delta t_i V, \Delta t_i V > > . \right.$$ If this is zero then all $\Delta t_i V$ are 0, and $\omega$ is differentiable of class $C^1$, even at the points $t_i$. Now it follows from the uniqueness theorem for differential equations that $\omega$ is $C^\infty$ everywhere: thus $\omega$ is an unbroken geodesic.

Continuing with the analogy developed in the preceding section, we now wish to define a bilinear functional

$$E_{**}: T_0^* x T_0^* \to \mathbb{R}$$

when $\gamma$ is a critical point of the function $E$, i.e., a geodesic. This bilinear functional will be called the Hessian of $E$ at $\gamma$.

If $f$ is a real valued function on a manifold $M$ with critical point $p$, then the Hessian

$$f_{**}: T_p^* x T_p^* \to \mathbb{R}$$

can be defined as follows. Given $X_1, X_2 \in T_p^*$, choose a smooth map $(u_1, u_2) \to \sigma(u_1, u_2)$ defined on a neighborhood of $(0,0)$ in $\mathbb{R}^2$, with values in $M$, so that

$$\sigma(0,0) = p, \quad \frac{\partial \sigma}{\partial u_1}(0,0) = X_1, \quad \frac{\partial \sigma}{\partial u_2}(0,0) = X_2.$$ 

Then

$$f_{**}(X_1, X_2) = \frac{\partial^2 f(\sigma(u_1, u_2))}{\partial u_1 \partial u_2} \bigg|_{(0,0)}.$$ 

This suggests defining $E_{**}$ as follows. Given vector fields $W_1, W_2 \in T_0^*$, choose a 2-parameter variation

$$a: U \times [0,1] \to M,$$

where $U$ is a neighborhood of $(0,0)$ in $\mathbb{R}^2$, so that

$$\sigma(0,0,t) = \gamma(t), \quad \frac{\partial \sigma}{\partial u_1}(0,0,t) = W_1(t), \quad \frac{\partial \sigma}{\partial u_2}(0,0,t) = W_2(t).$$

(Compare § 11.) Then the Hessian $E_{**}(W_1, W_2)$ will be defined to be the second partial derivative

$$\frac{\partial^2 E(\sigma(u_1, u_2))}{\partial u_1 \partial u_2} \bigg|_{(0,0)}$$

where $\sigma(u_1, u_2) \in \Omega$ denotes the path $\sigma(u_1, u_2)(t) = \sigma(u_1, u_2, t)$. 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 $\bar{\sigma}: U \to \Omega$ be a 2-parameter variation of the geodesic $\gamma$ with variation vector fields

$$\bar{W}_1 = \frac{\partial \bar{\sigma}}{\partial u_1}(0,0) \in T_0^*, \quad i = 1, 2.$$ 

Then the second derivative $\frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0)$ of the energy function is equal to

$$<W_2(t), \frac{D}{dt}W_1> - \int_0^1 <W_2, \frac{D}{dt}W_1 + R(V, W_1)V> \, dt$$

where $V = \frac{d\bar{\sigma}}{dt}$ denotes the velocity vector field and where

$$A_t \frac{D}{dt} = \frac{D}{dt}(t') - \frac{D}{dt}(t^-)$$

denotes the jump in $\frac{D}{dt}$ at one of its finitely many points of discontinuity in the open unit interval.

**Proof:** According to 12.2 we have

$$\frac{1}{\gamma} \frac{\partial^2 E}{\partial u_1 \partial u_2} = - \int_0^1 \frac{d}{dt} <\frac{\partial \sigma}{\partial u_1}, A_t \frac{D}{dt} > - \int_0^1 <\frac{\partial \sigma}{\partial u_1}, D \frac{\partial}{\partial t} > \, dt.$$ 

Therefore

$$\frac{1}{\gamma} \frac{\partial^2 E}{\partial u_1 \partial u_2} = - \int_0^1 <\frac{D}{dt} \frac{\partial \sigma}{\partial u_1}, A_t \frac{D}{dt} > - \int_0^1 <\frac{D}{dt} \frac{\partial \sigma}{\partial u_1}, D \frac{\partial}{\partial t} > \, dt - \int_0^1 <\frac{\partial \sigma}{\partial u_1}, D \frac{\partial}{\partial t} > \, dt.$$ 

Let us evaluate this expression for $(u_1, u_2) = (0,0)$. Since $\gamma = \bar{\sigma}(0,0)$ is an unbroken geodesic, we have

$$A_t \frac{\partial}{\partial t} = 0, \quad \frac{D}{dt} \frac{\partial}{\partial t} = 0,$$

so that the first and third terms are zero.

Rearranging the second term, we obtain

$$\frac{1}{\gamma} \frac{\partial^2 E}{\partial u_1 \partial u_2}(0,0) = - \int_0^1 <W_2, A_t \frac{D}{dt} W_1> - \int_0^1 <W_2, D \frac{\partial}{\partial t} V> \, dt.$$ 

In order to interchange the two operators $\frac{D}{dt}$ and $\frac{D}{dt}$, we need to bring in the curvature formula

$$\frac{D}{dt} V - \frac{D}{dt} \frac{D}{dt} V = R(\frac{\partial}{\partial u_1}, \frac{\partial}{\partial u_2}) V - R(V, W_1)V.$$
Together with the identity \( \frac{D}{dt}W_1 = \frac{d}{dt}u_1 = \frac{D}{dt} W_1 \), this yields

\[
\frac{D}{dt}W_1 \frac{D}{dt}V = \frac{D^2 W_1}{dt^2} + R(V,W_1)V.
\]

Substituting this expression into (13.2) this completes the proof of 13.1.

**COROLLARY 13.4.** The expression \( E_{*}(W_1,W_2) = \frac{d^2 E}{du_1^2}W_1^2 \) 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 \( \frac{d}{dt}u_1 \frac{d}{dt}u_2 = \frac{d^2 E}{du_1^2}u_1 u_2 \).

**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 \( \gamma \). In fact

\[
E_{*}(W,W) = \frac{d^2 E}{du_1^2}W_1^2(0,0),
\]

where \( \delta : (-\varepsilon,\varepsilon) \to N \) denotes any variation of \( \gamma \) with variation vector field \( \frac{d}{dt}W(0) \) equal to \( W \). To prove this identity it is only necessary to introduce the two parameter variation

\[
\delta(u_1,u_2) = \delta(u_1) + u_2
\]

and to note that

\[
\frac{d}{dt}u_1 \frac{d}{dt}u_2 = \frac{d^2 E}{du_1^2}u_1 u_2,
\]

\[
\frac{d^2 E}{du_1^2}u_1 u_2 = \frac{d^2 E}{du_1^2}u_1
\]

As an application of this remark, we have the following.

**LEMMA 13.6.** If \( \gamma \) is a minimal geodesic from \( p \) to \( q \) then the bilinear pairing \( E_{*} \) is positive semi-definite. Hence the index \( \lambda \) of \( E_{*} \) is zero.

**PROOF:** The inequality \( B(\delta(u)) \geq E(\gamma) = E(\delta(0)) \) implies that \( \frac{d^2 E(\delta(u))}{du_1^2} \), evaluated at \( u = 0 \), is \( \geq 0 \). Hence \( E_{*}(W,W) \geq 0 \) for all \( W \).

§14. Jacobi Fields: The Null Space of \( E_{*} \)

A vector field \( J \) along a geodesic \( \gamma \) is called a Jacobi field if it satisfies the Jacobi differential equation

\[
\frac{d^2 J}{dt^2} + R(V,J)V = 0
\]

where \( V = \frac{d}{dt} \). This is a linear, second order differential equation. It can be put in a more familiar form by choosing orthonormal parallel vector fields \( P_1, \ldots, P_n \) along \( \gamma \). Then setting \( J(t) = \sum P_i(t)J_i(t) \), the equation becomes

\[
a_i^2(t)J_i(t) = 0, \quad i = 1, \ldots, n;
\]

where \( a_i^2(t) = \langle R(V,P_i)V,P_i \rangle \). Thus the Jacobi equation has \( n \) linearly independent solutions, each of which can be defined throughout \( \gamma \). The solutions are all \( \mathcal{C}^\infty \)-differentiable. A given Jacobi field \( J \) is completely determined by its initial conditions:

\[
J(0), \frac{dJ}{dt}(0) \in T_{\gamma}(0).
\]

Let \( p = \gamma(a) \) and \( q = \gamma(b) \) be two points on the geodesic \( \gamma \), with \( a \neq b \).

**DEFINITION.** \( p \) and \( q \) are conjugate along \( \gamma \) if there exists a non-zero Jacobi field \( J \) along \( \gamma \) 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 \( \gamma \) be a geodesic in \( \mathcal{M} = \mathcal{M}(M_1,p,q) \). Recall that the null space of the Hessian

\[
E_{**} : T_{\gamma}^2 \times T_{\gamma} \to \mathbb{R}
\]

is the vector space consisting of those \( W_1 \in T_{\gamma} \) such that \( E_{**}(W_1,W_2) = 0 \). If \( \gamma \) has self-intersections then this definition becomes ambiguous.

One should rather say that the parameter values \( a \) and \( b \) are conjugate with respect to \( \gamma \).
III. CALCULUS OF VARIATIONS

for all \( W_2 \). The nullity \( v \) of \( E \) is equal to the dimension of this null space. \( E \) is \textbf{degenerate} if \( v > 0 \).

**THEOREM 14.1.** A vector field \( W_1 \in T_0 \gamma \) 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 \( \gamma \). The nullity \( v \) 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 \( T_0 \gamma \). The second variation formula (§13.1) states that

\[
-\frac{d^2}{dt^2}e(J,W) = \sum_i \langle W_i(t),0 \rangle + \int_0^1 \langle W_2,0 \rangle \, dt = 0.
\]

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 \( f: [0,1] \to [0,1] \) be a smooth function which vanishes for the parameter values \( t_0, t_1, \ldots, t_k \) and is positive otherwise; and let

\[
W_2(t) = f(t) \left( \frac{d^2}{dt^2}W_1 + R(V,W_1)V \right)_t.
\]

Then

\[
\frac{d^2}{dt^2}e(W_1,W_2) = \sum_i \frac{d}{dt} \frac{d^2}{dt^2}W_1 + \int_0^1 \langle W_2,0 \rangle \, dt.
\]

Since this is zero, it follows that \( W_1|[t_{i-1},t_i] \) is a Jacobi field for each \( i \).

Now let \( W_1 \in T_0 \gamma \) be a field such that \( W_1(t_i) = A_{t_i} W_1 \) for \( i = 1,2,\ldots,k-1 \). Then

\[
-\frac{d^2}{dt^2}e(W_1,W_2) = \sum_{i=1}^{k-1} \| A_{t_i} W_1 \|^2 + \int_0^1 0 \, dt = 0.
\]

Hence \( \frac{dW_1}{dt} \) has no jumps. But a solution \( W_1 \) of the Jacobi equation is completely determined by the vectors \( W_1(t_1) \) and \( \frac{dW_1}{dt}(t_1) \). 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 \( C^\infty \)-differentiable throughout the entire unit interval. This completes the proof of 14.1.

It follows that the nullity \( v \) of \( E \) is always finite. For there are only finitely many linearly independent Jacobi fields along \( \gamma \).

**REMARK 14.2.** Actually the nullity \( v \) satisfies \( 0 \leq v < n \). Since the space of Jacobi fields which vanish for \( t = 0 \) has dimension precisely \( n \), it is clear that \( v \leq n \). We will construct one example of a Jacobi field which vanishes for \( t = 0 \), but not for \( t = 1 \). This will imply that \( v < n \). In fact let \( J_t = tW_t \) where \( V = \frac{dV}{dt} \) denotes the velocity vector field. Then

\[
\frac{dV}{dt} = 1 \cdot V + t \frac{dV}{dt} = V
\]

(Since \( \frac{dV}{dt} = 0 \), hence \( \frac{d^2V}{dt^2} = 0 \)). 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 \( \gamma \) is "flat" in the sense that the curvature tensor is identically zero. Then the Jacobi equation becomes \( \frac{d^2}{dt^2}W_1 = 0 \). Setting \( J(t) = E \int^t(t)P_1(t) \) where \( P_1 \) are parallel, this becomes \( \frac{d^2}{dt^2}J_1 = 0 \). Evidently a Jacobi field along \( \gamma \) 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 \( S^2 \), and let \( \gamma \) 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 \( v \) of \( E \) takes its largest possible value. The proof will depend on the following discussion.

Let \( a \) be a 1-parameter variation of \( \gamma \), not necessarily keeping the endpoints fixed, such that each \( \tilde{a}(u) \) is a geodesic. That is, let

\[
\alpha: (-\varepsilon,\varepsilon) \times [0,1] \to \mathbb{M}
\]

be a \( C^\infty \) map such that \( \alpha(0,t) = \gamma(t) \), and such that each \( \dot{\tilde{a}}(u) \) [given by \( \tilde{a}(u)(t) = \alpha(u,t) \)] is a geodesic.
III. CALCULUS OF VARIATIONS

**Lemma 14.3.** If $a$ is such a variation of $\gamma$ through geodesics, then the variation vector field $W(t) = \frac{\partial a}{\partial t}(0, t)$ is a Jacobi field along $\gamma$.

**Proof:** If $a$ is a variation of $\gamma$ through geodesics, then $\frac{D a}{\partial t}$ is identically zero. Hence

$$\frac{\partial}{\partial t} = \frac{D}{\partial t} \frac{\partial a}{\partial t} + R\left(\frac{\partial a}{\partial t}, \frac{\partial a}{\partial t}\right) \frac{\partial a}{\partial t}$$

(Compare §13.3.) Therefore the variation vector field $\frac{\partial a}{\partial t}$ 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 $\gamma$ 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 $\gamma$ with multiplicity $n$.

**§14. JACOBI FIELDS**

**Lemma 14.4.** Every Jacobi field along a geodesic $\gamma: [0, 1] \to M$ may be obtained by a variation of $\gamma$ 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 $0 \leq t \leq 6$. We will first construct a Jacobi field $W$ along $\gamma|[0, 5]$ with arbitrarily prescribed values at $t = 0$ and $t = 6$. Choose a curve $a: (-\varepsilon, \varepsilon) \to U$ so that $a(0) = \gamma(0)$ and so that $\frac{\partial a}{\partial t}(0)$ is any prescribed vector in $T\gamma(0)$.

Similarly choose $b$: $(-\varepsilon, \varepsilon) \to U$ with $b(0) = \gamma(5)$ and $\frac{\partial b}{\partial t}(0)$ arbitrary. Now define the variation

$$\alpha: (-\varepsilon, \varepsilon) \times [0, 5] \to M$$

by letting $\tilde{a}(u) = [0, 5] \to M$ be the unique minimal geodesic from $a(u)$ to $b(u)$. Then the formula $t \to \frac{\partial a}{\partial t}(0, t)$ defines a Jacobi field with the given end conditions.

Any Jacobi field along $\gamma|[0, 5]$ can be obtained in this way: If $\gamma(\gamma)$ denotes the vector space of all Jacobi fields $W$ along $\gamma$, then the formula $W = (W(0), W(5))$ defines a linear map

$$f: \gamma(\gamma) \to TM(0) \times TM(5)$$

We have just shown that $f$ is onto. Since both vector spaces have the same dimension $2n$, it follows that $f$ is an isomorphism. I.e., a Jacobi field is determined by its values at $\gamma(0)$ and $\gamma(5)$. (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 $\gamma|[0, 5]$.

The restriction of $\tilde{a}(u)$ to the interval $[0, 5]$ is not essential. If $u$ is sufficiently small then, using the compactness of $[0, 1]$, $\tilde{a}(u)$ can be extended to a geodesic defined over the entire unit interval $[0, 1]$. This yields a variation through geodesics:

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

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 (-δ, δ) of 0 so that if t ∈ (-δ, δ) then γ(t) is not conjugate to γ(0) along γ. We will see in §15.2 that the set of conjugate points to γ(0) along the entire geodesic γ has no cluster points.

**Theorem 15.1 (Morse).** The index $\lambda$ of E is equal to the number of points $γ(t)$, with $0 < t < 1$, such that $γ(t)$ is conjugate to $γ(0)$ along γ; each such conjugate point being counted with its multiplicity. This index $\lambda$ is always finite.

As an immediate consequence one has:

**Corollary 15.2.** A geodesic segment $γ: [0, 1] \to M$ can contain only finitely many points which are conjugate to $γ(0)$ along γ.

In order to prove 15.1 we will first make an estimate for $\lambda$ by splitting the vector space $T\gamma$ into two mutually orthogonal subspaces, on one of which $E_+$ is positive definite.

Each point $γ(t)$ 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 $γ[t_{i-1}, t_i]$ lies within such an open set $U$; and so that each $γ[t_{i-1}, t_i]$ is minimal.

Let $T\gamma_{γ}(t_0, t_1, t_2, \ldots, t_k) \subseteq T\gamma$ 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 $γ[t_{i-1}, t_i]$ for each $i$;
2) $W$ vanishes at the endpoints $t = 0, t = 1$.

Thus $T\gamma_{γ}(t_0, t_1, t_2, \ldots, t_k)$ is a finite dimensional vector space consisting of broken Jacobi fields along $γ$.

Let \( T' \subset T_0 \) be the vector space consisting of all vector fields \( W \in T_0 \) for which \( W(t_0) = 0 \). Then \( W(t_2) = 0, \ldots, W(t_k) = 0 \).

**Lemma 15.3.** The vector space \( T_0 \) splits as the direct sum \( T_0 = T_0'(t_0, t_1, \ldots, t_k) \oplus T' \). These two subspaces are mutually perpendicular with respect to the inner product \( E_{\ast\ast} \). Furthermore, \( E \) restricted to \( T' \) is positive definite.

**Proof:** Given any vector field \( W \in T_0 \), let \( W_1 \) denote the unique "broken Jacobi field" in \( T_0 \) such that \( W(t_1) = W_1(t_1) \) for \( t_1 = 0, 1, \ldots, k \). It follows from §14.5 that \( W_1 \) exists and is unique. Clearly \( W - W_1 \) belongs to \( T' \). Thus the two subspaces, \( T_0 = T_0'(t_0, t_1, \ldots, t_k) \) and \( T' \), generate \( T_0 \), and have only the zero vector field in common.

If \( W_1 \), belongs to \( T_0'(t_0, t_1, \ldots, t_k) \) and \( W_2 \) belongs to \( T' \), then the second variation formula (13.1) takes the form

\[
\sum I_{E_{\ast\ast}}(W_1, W_2) = -\int_0^t \left< W_2(t), \frac{d^2 W_1}{dt^2} \right> - \int_0^1 \left< W_2(0), \right> \, dt = 0
\]

Thus the two subspaces are mutually perpendicular with respect to \( E_{\ast\ast} \).

For any \( W_1 \in T_0 \), the Hessian \( E_{\ast\ast}(W, W) \) can be interpreted as the second derivative \( \frac{d^2 W_1}{dt^2} \) at \( 0 \), where \( \delta: (-\varepsilon, \varepsilon) \rightarrow \mathbb{R} \) is any variation of \( y \) with variation vector field \( \frac{d\delta}{dt} \) equal to \( W \). (Compare 13.5.) If \( W \) belongs to \( T' \) then we may assume that \( \delta \) is chosen so as to leave the points \( y(t_0), y(t_1), \ldots, y(t_k) \) fixed. In other words we may assume that \( \delta(u)(t_1) = y(t_1) \) for \( t_1 = 0, 1, \ldots, k \).

Proof that \( E_{\ast\ast}(W, W) \geq 0 \) for \( W \in T' \). Each \( \delta(u) \in \mathbb{R} \) is a piecewise smooth path from \( y(0) \) to \( y(t_1) \) to \( y(t_2) \) to \( \ldots \) to \( y(t_k) \). But each \( y(t_{i-1}, t_i) \) is a minimal geodesic, and therefore has smaller energy than any other path between its endpoints. This proves that

\[
E(\delta(u)) \geq E(y) = E(\delta(0))
\]

Therefore the second derivative, evaluated at \( u = 0 \), must be \( \geq 0 \).

Proof that \( E_{\ast\ast}(W, W) > 0 \) for \( W \in T' \), \( W \neq 0 \). Suppose that \( E_{\ast\ast}(W, W) = 0 \). Then \( W \) would lie in the null space of \( E_{\ast\ast} \). In fact for any \( W_1 \in T_0 (t_0, t_1, \ldots, t_k) \) we have already seen that \( E_{\ast\ast}(W_1, W) = 0 \). For any \( W_2 \in T' \), the inequality

\[
0 \leq E_{\ast\ast}(W + c W_2, W + c W_2) = c E_{\ast\ast}(W_2, W_2) + c^2 E_{\ast\ast}(W_2, W_2)
\]

for all values of \( c \) implies that \( E_{\ast\ast}(W_2, W_2) = 0 \). 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 \( T_0 = T_0'(t_0, t_1, \ldots, t_k) \) of broken Jacobi fields. In particular (since \( T_0'(t_0, t_1, \ldots, t_k) \) is a finite dimensional vector space) the index \( \lambda \) is always finite.

The proof is straightforward.

Let \( \gamma_T \) denote the restriction of \( \gamma \) to the interval \([0, T]\). Thus \( \gamma_T: [0, T] \rightarrow M \) is a geodesic from \( \gamma(0) \) to \( \gamma(T) \). Let \( \lambda(T) \) denote the index of the Hessian \( \left( E^T \right)_0 \), which is associated with this geodesic. Thus \( \lambda(T) \) is the index which we are actually trying to compute. First note that:

**Assertion (1).** \( \lambda(T) \) is a monotone function of \( T \).

For if \( T < T' \) then there exists a \( \lambda(T) \) dimensional space \( \mathcal{V} \) of vector fields along \( \gamma_T \) which vanish at \( \gamma(0) \) and \( \gamma(T) \) such that the Hessian \( \left( E^T \right)_0 \) is negative definite on this vector space. Each vector field in \( \mathcal{V} \) extends to a vector field along \( \gamma_T \) which vanishes identically between \( \gamma(T) \) and \( \gamma(T') \). Thus we obtain a \( \lambda(T) \) dimensional vector space of fields along \( \gamma_T \) on which \( \left( E^T \right)_0 \) is negative definite. Hence \( \lambda(T) \leq \lambda(T') \).

**Assertion (2).** \( \lambda(T) = 0 \) for small values of \( T \).

For if \( T \) is sufficiently small then \( \gamma_T \) is a minimal geodesic, hence \( \lambda(T) = 0 \) by Lemma 13.6.

Now let us examine the discontinuities of the function \( \lambda(T) \). First note that \( \lambda(T) \) is continuous from the left:

**Assertion (3).** For all sufficiently small \( \varepsilon > 0 \) we have \( \lambda(T + \varepsilon) = \lambda(T) \).

§15. **The Index Theorem**

\[
0 \leq E_{\ast\ast}(W + c W_2, W + c W_2) = c E_{\ast\ast}(W_2, W_2) + c^2 E_{\ast\ast}(W_2, W_2)
\]

III. CALCULUS OF VARIATIONS

PROOF. According to 15.3 the number \( \lambda(t) \) can be interpreted as the index of a quadratic form on a finite-dimensional vector space \( T_M(t_0, t_1, \ldots, t_k) \). We may assume that the subdivision is chosen so that say \( t_4 < t < t_{4+1} \). Then the index \( \lambda(t) \) can be interpreted as the index of a corresponding quadratic form \( H_t \) on a corresponding vector space of broken Jacobi fields along \( \gamma_t \). This vector space is to be constructed using the subdivision \( 0 < t_1 < t_2 < \ldots < t_4 < t \) of \([0, t]\). 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

\[
\Sigma = T_M(t_1) \oplus T_M(t_2) \oplus \ldots \oplus T_M(t_4).
\]

Note that this vector space \( \Sigma \) is independent of \( t \). Evidently the quadratic form \( H_t \) on \( \Sigma \) varies continuously with \( t \).

Now \( H_t \) is negative definite on a subspace \( \mathcal{K} \subset \Sigma \) of dimension \( \lambda(t) \). For all \( t' \) sufficiently close to \( t \) it follows that \( H_{t'} \) is negative definite on \( \mathcal{K} \). Therefore \( \lambda(t') \geq \lambda(t) \). But if \( t' = t - \varepsilon < t \) then we also have \( \lambda(t'; \varepsilon) \leq \lambda(t) \) by Assertion 1. Hence \( \lambda(t'; \varepsilon) = \lambda(t) \).

ASSWRTION (5). Let \( \gamma \) be the nullity of the Hessian \( (E^T_0)_{**} \).

Then for all sufficiently small \( \varepsilon > 0 \) we have

\[
\lambda(t'; \varepsilon) = \lambda(t) + \gamma \varepsilon.
\]

Thus the function \( \lambda(t) \) jumps by \( \gamma \) when the variable \( t \) passes a conjugate point of multiplicity \( \gamma \), and is continuous otherwise. Clearly this assertion will complete the proof of the index theorem.

PROOF that \( \lambda(t'; \varepsilon) \leq \lambda(t) + \gamma \varepsilon \). Let \( H \) and \( \Sigma \) be as in the proof of Assertion 3. Since \( \dim \Sigma - \dim \mathcal{K} \) we see that \( H \) is positive definite on some subspace \( \mathcal{K} \subset \Sigma \) of dimension \( \dim \Sigma - \dim \mathcal{K} \). For all \( t' \) sufficiently close to \( t \), it follows that \( H_{t'} \) is positive definite on \( \mathcal{K} \). Hence

\[
\lambda(t') \leq \dim \Sigma - \dim \mathcal{K} \leq \lambda(t) + \gamma \varepsilon.
\]

PROOF that \( \lambda(t'; \varepsilon) \geq \lambda(t) + \gamma \varepsilon \). Let \( W_1, \ldots, W_{H(t)} \) be \( \lambda(t) \) vector fields along \( \gamma_t \), vanishing at the endpoints, such that the matrix

\[
\left( \begin{array}{cc} (E^T_0)_{**} & (W_1, W_2) \\ (W_1, W_2) & cA \\
\end{array} \right)
\]

is negative definite. Let \( J_1, \ldots, J_p \) be \( \lambda \) linearly independent Jacobi fields along \( \gamma_t \), also vanishing at the endpoints. Note that the \( \lambda \) vectors

\[
\frac{d}{dt}(J) \in TM(t)
\]

are linearly independent. Hence it is possible to choose \( \lambda \) vector fields \( X_1, \ldots, X_{H(t)} \) along \( \gamma_t \) so that the matrix

\[
\left( \begin{array}{cc} \left( \begin{array}{cc} (E^T_0)_{**} & (J, X) \\ (J, X) & cA \\
\end{array} \right) \end{array} \right)
\]

is equal to the \( \lambda \times \lambda \) identity matrix. Extend the vector fields \( W_1, \ldots, W_{H(t)} \) and \( J_1, \ldots, J_p \) over \( \gamma_t + \varepsilon \) by setting these fields equal to 0 for \( t \leq t \leq t + \varepsilon \).

Using the second variation formula we see easily that

\[
\left( \begin{array}{cc} (E^T_0)_{**} & (J, W) \\ (J, W) & cA \\
\end{array} \right) = 0
\]

where

\[
\left( \begin{array}{cc} (E^T_0)_{**} & (J, X) \\ (J, X) & cA \\
\end{array} \right) = 2 \delta_{hk} \text{ (Kronecker delta)}
\]

Now let \( c \) be a small number, and consider the \( \lambda + \gamma \) vector fields

\[
W_1, \ldots, W_{H(t)}, c^{-1}J_1 - cX_1, \ldots, c^{-1}J_p - cX_p
\]

along \( \gamma_{t + \varepsilon} \). We claim that these vector fields span a vector space of dimension \( \lambda(t) + \gamma \) on which the quadratic form \( (E^T_0)_{**} \) is negative definite. In fact the matrix of \( (E^T_0)_{**} \) with respect to this basis is

\[
\left( \begin{array}{cc} \left( \begin{array}{cc} (E^T_0)_{**} & (W_1, W_p) \\ (W_1, W_p) & cA \\
\end{array} \right) & cA \\
\end{array} \right)
\]

where \( A \) and \( B \) are fixed matrices. If \( c \) is sufficiently small, this compound matrix is certainly negative definite. This proves Assertion (5).

The index theorem 15.1 clearly follows from the Assertions (2), (3), and (4).
§16. A Finite Dimensional Approximation to $\mathcal{O}^C$.

Let $M$ be a connected Riemannian manifold and let $p$ and $q$ be two (not necessarily distinct) points of $M$. The set $\mathcal{O} = \mathcal{O}(M;p,q)$ of piecewise $C^0$ paths from $p$ to $q$ can be topologized as follows. Let $\rho$ denote the topological metric on $M$ coming from its Riemannian metric. Given $\omega, \omega' \in \mathcal{O}$ with arc-lengths $s(t)$, $s'(t)$ respectively, define the distance $d(\omega,\omega')$ to be

$$\max_{0 \leq t \leq 1} \rho(s(t), s'(t)) + \left( \int_0^1 \left( \frac{ds}{dt} - \frac{ds'}{dt} \right)^2 dt \right)^{1/2}.$$  

(The last term is added on so that the energy function,

$$E^b_\mathcal{O}(\omega) = \int_a^b \left( \frac{ds}{dt} \right)^2 dt$$

will be a continuous function from $\mathcal{O}$ to the real numbers.) This metric induces the required topology on $\mathcal{O}$.

Given $c > 0$ let $\mathcal{O}^c$ denote the closed subset $E^{-1}(0,c) \subseteq \mathcal{O}$ and let $\text{Int} \mathcal{O}^c$ denote the open subset $E^{-1}(0,c)$ (where $E = E^1_\mathcal{O}$ $\mathcal{O} \to \mathbb{R}$ is the energy function). We will study the topology of $\mathcal{O}^c$ by constructing a finite dimensional approximation to it.

Choose some subdivision $0 = t_0 \leq t_1 \leq \ldots \leq t_k = 1$ of the unit interval. Let $\mathcal{O}(t_0, t_1, \ldots, t_k)$ be the subspace of $\mathcal{O}$ 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

$$\mathcal{O}(t_0, t_1, \ldots, t_k)^c = \mathcal{O}^c \cap \mathcal{O}(t_0, t_1, \ldots, t_k)$$

$$\text{Int} \mathcal{O}(t_0, t_1, \ldots, t_k)^c = \text{Int} \mathcal{O}^c \cap \mathcal{O}(t_0, t_1, \ldots, t_k).$$

**Lemma 16.1.** Let $M$ be a complete Riemannian manifold, and let $c$ be a fixed positive number such that $\mathcal{O}^c \neq \emptyset$. Then for all sufficiently fine subdivisions $(t_0, t_1, \ldots, t_k)$ of $[0,1]$ the set $\text{Int} \mathcal{O}(t_0, t_1, \ldots, t_k)^c$ can be given the structure of a smooth finite dimensional manifold in a natural way.

**Proof:** Let $S$ denote the ball

$$\{ x \in M : \rho(x,p) < \sqrt{c} \}$$

Note that every path $\omega \in \mathcal{O}^c$ lies within this subset $S \subseteq M$. 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 $\varepsilon^2/c$. Then for each broken geodesic $\omega \in \mathcal{O}(t_0, t_1, \ldots, t_k)^c$ we have

$$\left( \int_{t_{i-1}}^{t_i} \omega \right)^2 \lesssim (t_i - t_{i-1}) \left( E^1 \omega \right) \lesssim (t_i - t_{i-1}) \varepsilon < \varepsilon^2.$$ 

Thus the geodesic $\omega|[t_{i-1}, t_i]$ is uniquely and differentiably determined by the two end points.

The broken geodesic $\omega$ 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

$$\omega \mapsto (\omega(t_1), \ldots, \omega(t_{k-1}))$$

defines a homeomorphism between $\text{Int} \mathcal{O}(t_0, t_1, \ldots, t_k)^c$ and a certain open subset of the $(k-1)$-fold product $M \times M \times M$. Taking over the differentiable structure from this product, this completes the proof of 16.1.

To shorten the notation, let us denote this manifold

$$\text{Int} \mathcal{O}(t_0, t_1, \ldots, t_k)^c$$

denote the restriction to $B$ of the energy function $E : \mathcal{O} \to \mathbb{R}$. 

THEOREM 16.2. This function $E': \mathbb{R} \to \mathbb{R}$ is smooth. Furthermore, for each $a < c$ the set $E^a = (E')^{-1}(0, a)$ is compact, and is a deformation retract in the corresponding set $\mathbb{R}^2$. The critical points of $E'$ are precisely the same as the critical points of $E$ in $\text{Int} \, \mathbb{R}^2$; namely the unbroken geodesics from $p$ to $q$ of length less than $\sqrt{c}$. The index (or the nullity) of the Hessian $E''_{\ast \ast}$ at each such critical point $y$ is equal to the index (or the nullity) of $E_{\ast \ast}$ at $y$.

Thus the finite dimensional manifold $B$ provides a faithful model for the infinite dimensional path space $\text{Int} \, \mathbb{R}^C$. 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 $< \sqrt{c}$. Then $\mathbb{R}^a$ has the homotopy type of a finite CW-complex, with one cell of dimension $x$ for each geodesic in $\mathbb{R}^a$ at which $E_x$ has index $x$.

(In particular it is asserted that $\mathbb{R}^a$ 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 $\omega(t_1), \omega(t_2), \ldots, \omega(t_{k-1}) \in M \times \cdots \times M$ it is clear that the energy $E'(\omega)$ also depends smoothly on this $(k-1)$-tuple. In fact we have the explicit formula

$$E'(\omega) = \sum_{i=1}^{k} \rho(\omega(t_{i-1}), \omega(t_i))^2/(t_i - t_{i-1}) .$$

For $a < c$ the set $\mathbb{R}^a$ is homeomorphic to the set of all $(k-1)$-tuples $(p_1, \ldots, p_{k-1}) \in S \times S \times \cdots \times S$ such that

$$\rho(p_1, p_2)/(t_1 - t_{k-1}) \leq a.$$

(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 $r: \text{Int} \, \mathbb{R}^C \to B$ 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

$$\rho(p, \omega(t))^2 \leq (E)^2 \leq E < c$$

implies that $\omega[0, 1] \subset S$. Hence the inequality

$$\rho(\omega(t_{i-1}), \omega(t_i))^2 \leq (t_i - t_{i-1}) \left( \sum_{i=1}^{k} \rho(\omega(t_{i-1}), \omega(t_i))^2 \right) \leq \varepsilon^2$$

implies that $r(\omega)$ can be so defined.

Clearly $B(r(\omega)) \leq B(\omega) < r$. This retraction $r$ fits into a 1-parameter family of maps

$$r_u: \text{Int} \, \mathbb{R}^C \to \text{Int} \, \mathbb{R}^C$$

as follows. For $t_{i-1} \leq u \leq t_i$, let

$$r_u(\omega)[[0, t_{i-1}]] = r(\omega)[[0, t_{i-1}]],$$

$$r_u(\omega)[[t_{i-1}, u]] = \text{minimal geodesic from } \omega(t_{i-1}) \text{ to } \omega(u),$$

$$r_u(\omega)[[u, 1]] = \omega[[u, 1]].$$

Then $r_0$ is the identity map of $\text{Int} \, \mathbb{R}^C$, 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 $\mathbb{R}^a$.

Since $B(r_u(\omega)) \leq B(\omega)$ it is clear that each $\mathbb{R}^a$ is also a deformation retract of $\mathbb{R}^a$.

Every geodesic is also a broken geodesic, so it is clear that every "critical point" of $E$ in $\mathbb{R}^C$ 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 $T \mathbb{R}^a$ to the manifold $B$ at a geodesic $Y$. This will be identified with the space $T_{\mathbb{R}^a}(t_0, t_1, \ldots, t_{k-1})$ of broken

* Similarly $B$ itself is a deformation retract of $\text{Int} \, \mathbb{R}^C$. 

§16. AN APPROXIMATION TO $\mathbb{R}^C$

For $a < c$ the set $\mathbb{R}^a$ is homeomorphic to the set of all $(k-1)$-tuples $(p_1, \ldots, p_{k-1}) \in S \times S \times \cdots \times S$ such that

$$\rho(p_1, p_2)/(t_1 - t_{k-1}) \leq a .$$

(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.
III. CALCULUS OF VARIATIONS

Jacobi fields along \( \gamma \), as described in §15. This identification can be justified as follows. Let

\[ \tilde{\gamma}_1 : (-\varepsilon, \varepsilon) \to B \]

be any variation of \( \gamma \) through broken geodesics. Then the corresponding variation vector field \( \frac{d\gamma}{dt}(0, t) \) along \( \gamma \) is clearly a broken Jacobi field. (Compare §14.3.)

Now the statement that the index (or the nullity) of \( E^*_\gamma \) at \( \gamma \) is equal to the index (or nullity) of \( E^*_\gamma \) at \( \gamma \) 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 \( \tilde{\gamma}^N[p, q] \) is non-vacuous, then the corresponding set \( B^\tilde{\gamma} \) will be compact and non-vacuous. Hence the continuous function \( E^\tilde{\gamma} : B^\tilde{\gamma} \to R \) will take on its minimum at some point \( \gamma \in B^\tilde{\gamma} \). This \( \gamma \) will be the required minimal geodesic.

§17. THE FULL PATH SPACE

§17. The Topology of the Full Path Space.

Let \( M \) be a Riemannian manifold with Riemann metric \( g \), and let \( \sigma \) be the induced topological metric. Let \( p \) and \( q \) be two (not necessarily distinct) points of \( M \).

In homotopy theory one studies the space \( \tilde{\alpha}^* \) of all continuous paths

\[ \alpha : [0, 1] \to M \]

from \( p \) to \( q \), in the compact open topology. This topology can also be described as that induced by the metric

\[ d^*(\alpha, \alpha') = \max_t \sigma(\alpha(t), \alpha'(t)) \]

On the other hand we have been studying the space \( \alpha \) of piecewise \( C^\infty \) paths from \( p \) to \( q \) with the metric

\[ d(\alpha, \alpha') = d^*(\alpha, \alpha') + \left[ \int_0^1 \left( \frac{ds}{dt} - \frac{da}{dt} \right)^2 dt \right]^{1/2} \]

Since \( d \geq d^* \) the natural map

\[ \iota : \alpha \to \tilde{\alpha}^* \]

is continuous.

THEOREM 17.1. This natural map \( \iota \) is a homotopy equivalence between \( \alpha \) and \( \tilde{\alpha}^* \).

PROOF. We will first define a continuous function \( A : \alpha \to (0, 1) \) such that \( |t-t'| < \varepsilon A(\alpha) \) implies that \( \alpha(t) \) and \( \alpha(t') \) are joined by a unique minimal geodesic which varies differentiably with the end points.

Let \( f : M \to [0, \infty) \) be a continuous function such that \( f^{-1}(0, r) \) is compact for every \( r \in [0, \infty) \). Let \( \varepsilon_1(r) \leq \varepsilon_2 \) be the largest real number such that any two points with distance \( \varepsilon_1(r) \) in the set \( f^{-1}(0, r) \) are joined by a unique minimal geodesic which varies differentiably with the endpoints. Since \( \varepsilon_1 \) is a positive, monotone decreasing function, we can choose

\[ \varepsilon_2 : [0, \infty) \to R \]
III. CALCULUS OF VARIATIONS

so that \( \varepsilon \) is continuous and

\[ 0 < \varepsilon_2(r) < \varepsilon_1(r) \]

Now define

\[ \varepsilon: \mathbb{R}^+ \rightarrow \mathbb{R} \]

by

\[ \varepsilon(s) = \varepsilon_2 \left( \max_t f(s(t)) \right) \]

Thus \( \varepsilon \) is continuous, and any two points of \( s([0,1]) \) of distance \( \leq \varepsilon(s) \) are joined by a unique minimal geodesic.

Define a continuous function

\[ F: \mathbb{R}^+ \times [0,1] \rightarrow \mathbb{R} \]

by

\[ F(s,a) = (a-1)\varepsilon(s) + \max_{|t-t'| \leq a} s(s(t),s(t')) \]

Then \( F \), considered as a function of \( s \), is strictly monotone, and

\[ F(s,0) < 0 \leq F(s,1) \]

Hence for each \( s \in \mathbb{R}^+ \) there is a unique \( s \in (0,1] \) with \( F(s,a) = 0 \).

Set \( s = 2a(s) \). Since \( F \) is continuous, a straightforward epsilon, delta argument shows that the function \( A: \mathbb{R}^+ \rightarrow [0,1] \) is continuous. If \( |t-t'| \leq a = 2a(s) \) then \( s(s(t),s(t')) \leq (1-a)\varepsilon(s) \leq \varepsilon(s) \) hence \( s(t) \) and \( s(t') \) are joined by a unique minimal geodesic.

Now define \( h: \mathbb{R}^+ \rightarrow \mathbb{R} \) as follows. Let \( h(s) \) be the unique path such that

1) \( h(s) \) coincides with \( s \) for \( t = 0 \), \( A(s), 2A(s), ... kA(s) \), and for \( t = 1 \); where \( k \) is the largest integer in \( 1/A(s) \);

2) \( h(s) \) is a geodesic in each intermediate interval.

Clearly \( h \) is continuous.

Now an argument similar to that in §16 shows that the composition

\[ h \cdot h: \mathbb{R}^+ \rightarrow \mathbb{R} \]

is homotopic to the identity map of \( \mathbb{R} \). Similarly the composition

\[ 1 \cdot h: \mathbb{R}^+ \rightarrow \mathbb{R} \]

is homotopic to the identity. This completes the proof of 17.1.

§17. THE FULL PATH SPACE

It is known that the space \( \mathbb{H}^* \) 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. \( \mathbb{H} \) 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 \( \mathbb{H}(M,p,q) \) (or \( \mathbb{H}(M,\mathbb{H},q) \)) has the homotopy type of a countable \( CW \)-complex which contains one cell of dimension \( \lambda \) for each geodesic from \( p \) to \( q \) of index \( \lambda \).

The proof is analogous to that of 3.5. Choose a sequence \( a_0 < a_1 < a_2 < ... \) 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

\[ a_0 \subset a_1 \subset a_2 \subset \ldots \]

where we may assume that \( a_0 \) is vacuous. It follows from 16.2 together with 3.3 and 3.7 that each \( a_i \) has the homotopy type of \( a_{i-1} \) with a finite number of cells attached: one \( \lambda \)-cell for each geodesic of index \( \lambda \) in \( E^{-1}(a_{i-1}, a_i) \). Now, just as in the proof of 3.5, one constructs a sequence \( K_0 \subset K_1 \subset K_2 \subset \ldots \) of \( CW \)-complexes with cells of the required description, and a sequence

\[ \mathbb{H}_0 \subset \mathbb{H}_1 \subset \mathbb{H}_2 \subset \ldots \]

of homotopy equivalences. Letting \( f: \mathbb{H} \rightarrow K \) be the direct limit mapping, it is clear that \( f \) induces isomorphisms of homotopy groups in all dimensions. Since \( a_0 \) 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 17.1, see p. 149.]

EXAMPLE. The path space of the sphere \( S^3 \). Suppose that \( p \) and \( q \) are two non-conjugate points on \( S^3 \). 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 \( \gamma_0 \) denote the short great circle arc from \( p \) to \( q \); let \( \gamma_1 \) denote the long great circle arc \( pq'p'q \); let \( \gamma_2 \) denote the arc \( pp'p'qq \); and so on. The subscript \( k \) denotes the number of times that \( p \) or \( p' \) occurs in the interior of \( \gamma_k \).

The index \( \lambda(\gamma_k) = u_1 + \ldots + u_k \) 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:

\[ \text{COROLLARY 17.4.} \quad \text{The loop space } \pi(S^n) \text{ has the homotopy type of a } \text{CW-complex with one cell each in the dimensions} \]
\[ 0, n-1, 2(n-1), 3(n-1), \ldots. \]

For \( n > 2 \) the homology of \( \pi(S^n) \) can be computed immediately from this information. Since \( \pi(S^n) \) has non-trivial homology in infinitely many dimensions, we can conclude:

\[ \text{COROLLARY 17.5.} \quad \text{Let } M \text{ have the homotopy type of } S^n, \]
\[ \text{for } n > 2. \text{ Then any two non-conjugate points of } M \text{ are joined by infinitely many geodesics.} \]

This follows since the homotopy type of \( \pi^*(M) \) (and hence of \( \pi(M) \)) depends only on the homotopy type of \( M \). There must be at least one geodesic in \( \pi(M) \) with index \( 0 \), at least one with index \( n-1, 3(n-1), 5(n-1), \ldots \).
§18. Existence of Non-Conjugate Points.

Theorem 17.3 gives a good description of the space \( n(M;p,q) \) 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 \( f: \mathcal{N} \to M \) between manifolds of the same dimension is critical at a point \( x \in \mathcal{N} \) if the induced map

\[
T_xf: TM_x \to TM_{f(x)}
\]

is 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 \( \gamma_v \) 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 \( \exp_p(X) = 0 \) 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 \( a \) defined by \( \alpha(u,t) = \exp tv(u) \) is a variation through geodesics of the geodesic \( \gamma_v \) given by \( t \to \exp tv \).

Therefore the vector field \( W \) given by \( t \to \frac{\partial}{\partial t} \exp tv(u) \bigg|_{u=0} \) is a Jacobi field along \( \gamma_v \). Obvioulsy \( W(0) = 0 \). We also have

\[
W(1) = \frac{\partial}{\partial t} \exp v(u) \bigg|_{u=0} = \exp_p \frac{dv(u)}{du}(0) + \exp_p X = 0.
\]

But this field is not identically zero since

\[
\frac{\partial W}{\partial t}(0) = \frac{D}{du} \frac{\partial}{\partial t} \exp tv(u) \bigg|_{(0,0)} = D_{v(u)} v(u) \bigg|_{u=0} \neq 0.
\]

So there is a non-trivial Jacobi field along \( \gamma_v \) from \( p \) to \( \exp v \), vanishing at these points; hence \( p \) and \( \exp v \) are conjugate along \( \gamma_v \).

Now suppose that \( \exp \) is non-singular at \( v \). Choose \( n \) independent vectors \( X_1, \ldots, X_n \) in \( T(TM_p)_v \). Then \( \exp_p(X_1), \ldots, \exp_p(X_n) \) span \( TM_p \) with \( v_1(0) = v \) and \( \frac{dv_1}{du}(0) = X_1 \).

Then \( \alpha_1, \ldots, \alpha_n \), constructed as above, provide \( n \) Jacobi fields \( W_1, \ldots, W_n \) along \( \gamma_v \), vanishing at \( p \). Since the \( W_1 = \exp_p(X_1) \) are independent, no non-trivial linear combination of the \( W_1 \) can vanish at \( \exp v \). Since \( n \) is the dimension of the space of Jacobi fields along \( \gamma_v \), which vanish at \( p \), clearly no non-trivial Jacobi field along \( \gamma_v \) 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 (96.1).
§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 $\langle R(A, B)A, B \rangle < 0$ for every pair of vectors $A, B$ in the tangent space $T_M$, and for every $p \in M$. Then no two points of $M$ are conjugate along any geodesic.

**Proof.** Let $\gamma$ be a geodesic with velocity vector field $V$, and let $J$ be a Jacobi field along $\gamma$. Then

$$\frac{d}{dt} \left( \frac{d^2 J}{dt^2} + R(V, J)V \right) = 0$$

so that

$$\langle \frac{d^2 J}{dt^2}, J \rangle = - \langle R(V, J)V, J \rangle \geq 0.$$  

Therefore

$$\frac{d}{dt} \langle \frac{dJ}{dt}, J \rangle = \langle \frac{d^2 J}{dt^2}, J \rangle + \frac{d}{dt} \|J\|^2 \geq 0.$$  

Thus the function $\langle \frac{dJ}{dt}, J \rangle$ 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 $\langle \frac{dJ}{dt}, J \rangle$ 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

$$(u_1, u_2) \rightarrow \exp_p(u_1 A + u_2 B)$$

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 $q = \exp(\mu U)$. A small line segment at $q$ with length $L$, pointed in a direction corresponding to the unit vector $W \in T_q$ would appear to the observer as a line segment of length

$$L(1 + \frac{2}{R} \langle R(U, W)U, W \rangle + \text{terms involving higher powers of } r).$$

Thus if sectional curvatures are negative then any object appears shorter than it really is. A small sphere of radius $r$ at $q$ would appear to be an ellipsoid with principal radii $r(1 + \frac{2}{R} K_1 + \ldots), \ldots, r(1 + \frac{2}{R} K_n + \ldots)$ where $K_1, K_2, \ldots, K_n$ denote the eigenvalues of the linear transformation $W \rightarrow R(U, W)U$. Any small object of volume $v$ would appear to have volume $v(1 + \frac{2}{R}(K_1 + K_2 + \ldots + K_n) + \text{higher terms})$ where $K_1 + \ldots + K_n$ is equal to the "Ricci curvature" $R(U, U)$, as defined later in this section.

Here are some familiar examples of complete manifolds with curvature $\leq 0$:

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. It is not known whether or not there exists a complete surface in 3-space with curvature negative and bounded away from 0.)

A famous example of a manifold with everywhere negative sectional curvature is the pseudo-sphere

$$z = -\sqrt{1 - x^2 - y^2} + \text{sech}^{-1} \sqrt{x^2 + y^2}, \quad z > 0$$

with the Riemann metric induced from $R^3$. 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
III. CALCULUS OF VARIATIONS

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 embedded in \( \mathbb{R}^3 \). (See e.g. Willmore, "Differential Geometry," p. 137.)

However, there do exist Riemannian manifolds of constant negative curvature which are complete. (See for example Lang, "Differential-Geometrie," pp. 114-117.) Such a manifold can even be compact, for example, a surface of genus \( \geq 2 \). (Compare Hilbert and Cohn-Vossen, "Anschauliche Geometrie," p. 228.)

**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 \( \leq 0 \). Then any two points of \( M \) are joined by a unique geodesic. Furthermore, \( M \) is diffeomorphic to the Euclidean space \( \mathbb{R}^n \).

**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 \( \pi_1(M; p, q) \) 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 \( \pi_1(M; p, q) \) 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 \( q \).


Therefore, the exponential map \( \exp_p : TM \to M \) 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 \( \leq 0 \). (For example \( M \) might be a flat torus \( S^1 \times S^1 \), or a compact surface of genus \( \geq 2 \) 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 \( \leq 0 \).

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 \( \pi_1(M) = \pi_1(\tilde{M}) = 0 \) 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 EP. 200-202 of S. T. Hu "Homotopy Theory," Academic Press, 1959.) Now suppose that \( \pi_n(M) \) contains a non-trivial finite cyclic subgroup \( G \). Then for a suitable covering space \( \tilde{M} \) of \( M \) we have \( \pi_n(\tilde{M}) = G \); hence

\[
H^k(G) = H^k(\pi_1(M)) = 0 \quad \text{for } k > n.
\]

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 (sometimes called the "mean curvature tensor").
111. CALCULUS OF VARIATIONS

**DEFINITION.** The Ricci tensor at a point \( p \) of a Riemannian manifold \( M \) is a bilinear pairing

\[
K: TM_p \times TM_p \to \mathbb{R}
\]

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 99.3 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.** \( K(U_i, U_i) \) is equal to the sum of the sectional curvatures \( \langle R(U_i, U_j)U_i, U_j \rangle \) for \( i = 1, 2, \ldots, n-1 \).

**PROOF:** By definition \( K(U_i, U_i) \) is equal to the trace of the matrix

\[
(\langle R(U_i, U_j)U_i, U_j \rangle) \text{ since the } n \text{-th diagonal term of this matrix is zero, we obtain a sum of } n-1 \text{ sectional curvatures, as asserted.}
\]

**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 \( F_1, \ldots, F_n \) along \( \gamma \) which are orthonormal at one point, and hence are orthonormal everywhere along \( \gamma \). We may assume that \( F_1 \) points along \( \gamma \), so that

\[
V = \frac{d\gamma}{dt} = L F_n, \quad \text{and} \quad \frac{dF_1}{dt} = 0.
\]

Let \( W_1(t) = (\sin \pi t) F_1(t) \). Then

\[
\frac{dW_1}{dt}(t) = (\cos \pi t) F_1(t).
\]

Summing for \( i = 1, \ldots, n-1 \) we obtain

\[
E_{\ast\ast}(W_1, W_1) = (\sin \pi t)^2 (n-1) \pi^2 - L^2 K(P_n, P_n)
\]

for every unit vector \( W \) at every point of \( M \). Hence

\[
\frac{d^2 E(\bar{g}(u))}{du^2} < 0, \quad \frac{dE(\bar{g}(u))}{du} < 0.
\]

for \( u = 0 \). Hence \( E(\bar{g}(u)) < E(\gamma) \) 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 \( \gamma \) be a minimal geodesic from \( p \) to \( q \). Then the length of \( \gamma \) 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.

---

III. CALCULUS OF VARIATIONS

THEOREM 19.6. If \( M \) is a compact manifold, and if the Ricci tensor \( \mathbf{K} \) of \( M \) is everywhere positive definite, then the path space \( \mathcal{A}(M;p,q) \) has the homotopy type of a \( \text{CW} \)-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 \( \mathbf{K}(U;U) > 0 \) takes on a minimum, which we can denote by \( (n-1)/\pi^2 > 0 \). Then every geodesic \( \gamma \in \mathcal{A}(M;p,q) \) of length \( > \pi r \) has index \( \lambda > 1 \).

More generally consider a geodesic \( \gamma \) of length \( > k\pi r \). Then a similar argument shows that \( \gamma \) has index \( k \geq k \). In fact for each \( \ell = 1,2,\ldots,k \) one can construct a vector field \( X_\ell \) along \( \gamma \) which vanishes outside of the interval \( \left( \frac{\ell-1}{k}, \frac{\ell}{k} \right) \), and such that \( E_{\ell \ell}(X_\ell, X_\ell) < 0 \). Clearly \( E_{\ell \ell}(X_\ell, X_j) = 0 \) for \( \ell \neq j \); so that \( X_1, \ldots, X_k \) span a \( k \)-dimensional subspace of \( T\gamma \) on which \( E_k \) 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 with index \( \leq k \). Hence there are only finitely many geodesics with index \( \leq k \). 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 \( \mathbf{K}(U;U) \) 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 cannot be invariant under the involution \( (x,y) \mapsto (-x,-y) \) of \( S^m \times S^k \). This follows from a theorem of Synge. (See J. L. Synge, On the connectivity or 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.


PART IV

APPLICATIONS TO LIE GROUPS AND SYMMETRIC SPACES


A symmetric space is a connected Riemannian manifold $M$ such that, for each $p \in M$ there is an isometry $I_p: M \rightarrow M$ which leaves $p$ fixed and reverses geodesics through $p$, i.e., if $\gamma$ is a geodesic and $\gamma(0) = p$ then $I_p(\gamma(t)) = \gamma(-t)$.

**Lemma 20.1** Let $\gamma$ 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 $\gamma$.

**Proof:** Let $\gamma'(t) = \gamma(t + c)$. Then $\gamma'$ 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 $\gamma$ 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 $\gamma$ then $R(U, V)W$ is also a parallel field along $\gamma$.

**Proof.** If $X$ denotes a fourth parallel vector field along $\gamma$, note that the quantity $\langle R(U, V)W, X \rangle$ is constant along $\gamma$. In fact,
IV. APPLICATIONS

given \( p = \gamma(0) \), \( q = \gamma(c) \), consider the isometry \( T = I_{\gamma(c/2)}I_p \) which carries \( p \) to \( q \). Then

\[
<R(U_pV_p)_pX_p> = <R(T_pU_pV_p)_pX_p>
\]

by 20.1. Since \( T \) is an isometry, this quantity is equal to

\[
<R(U_pV_p)_pX_p> = <R(U_pV_p)_pX_p>.
\]

Thus \( R(U,V)W,X \) 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 locally symmetric. (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 \( \gamma: R \to M \) be a geodesic in a locally symmetric manifold. Let \( V = \frac{d}{dt}(\gamma(0)) \) be the velocity vector at \( p = \gamma(0) \).

Define a linear transformation

\[
K_V: TM_p \to TM_p
\]

by \( K_V(W) = R(V,W)V \). Let \( \epsilon_1, \ldots, \epsilon_n \) denote the eigenvalues of \( K_V \).

**THEOREM 20.5.** The conjugate points to \( p \) along \( \gamma \) are the points \( \gamma(tk/\sqrt{\epsilon_i}) \) where \( k \) is any non-zero integer, and \( \epsilon_i > 0 \) is any positive eigenvalue of \( K_V \). The multiplicity of \( \gamma(t) \) as a conjugate point is equal to the number of \( \epsilon_i \) such \( t \) is a multiple of \( \pi/\sqrt{\epsilon_i} \).

**PROOF:** First observe that \( K_V \) is self-adjoint:

\[
<K_V(W),W'> = <K_V(W'),W>.
\]

This follows immediately from the symmetry relation

\[
\]

Therefore we may choose an orthonormal basis \( U_1, \ldots, U_n \) for \( M_p \) so that

\[
K_V(U_i) = \epsilon_i U_i,
\]

where \( \epsilon_1, \ldots, \epsilon_n \) are the eigenvalues. Extend the \( U_i \) to vector fields along \( \gamma \) by parallel translation. Then since \( M \) is locally symmetric,

the condition

\[
R(V,U_i)V = \epsilon_i U_i
\]

remains true everywhere along \( \gamma \). Any vector field \( W \) along \( \gamma \) may be expressed uniquely as

\[
W(t) = \epsilon_1(t)U_1(t) + \ldots + \epsilon_n(t)U_n(t).
\]

Then the Jacobi equation \( \frac{d^2}{dt^2} + \sum \epsilon_i \frac{d}{dt} U_i = 0 \) takes the form

\[
\sum \epsilon_i \frac{d}{dt} U_i + \sum \epsilon_i \frac{d}{dt} U_1 = 0.
\]

Since the \( U_1 \) are everywhere linearly independent this is equivalent to the system of \( n \) equations

\[
\frac{d^2}{dt^2} \epsilon_i \frac{d}{dt} U_1 = 0.
\]

We are interested in solutions that vanish at \( t = 0 \). If \( \epsilon_1 > 0 \) then

\[
\epsilon_1(t) = c_1 \sin(\sqrt{\epsilon_1} t), \quad \text{for some constant } c_1.
\]

Then the zeros of \( \epsilon_1(t) \) are at the multiples of \( t = \pi/\sqrt{\epsilon_1} \).

If \( \epsilon_1 = 0 \) then \( \epsilon_1(t) = c_1 t \) and if \( \epsilon_1 < 0 \) then

\[
\epsilon_1(t) = c_1 \sinh(\sqrt{-\epsilon_1} t), \quad \text{for some constant } c_1.
\]

Thus if \( \epsilon_1 \leq 0 \), \( \epsilon_1(t) \) vanishes only at \( t = 0 \). This completes the proof of 20.5.

---

\( K_V \) should not be confused with the Ricci tensor or §19.

In this section we consider a Lie group $G$ with a Riemannian metric which is invariant both under left translations

$$L_{\sigma}: G \rightarrow G, \quad L_{\sigma}(g) = \sigma g$$

and right translation,

$$R_{\tau}(g) = \tau g.$$ If $G$ is commutative such a metric certainly exists. If $G$ is compact then such a metric can be constructed as follows: Let $\langle \quad, \quad \rangle$ be any Riemannian metric on $G$, and let $\mu$ denote the Haar measure on $G$. Then $\mu$ is right and left invariant. Define a new inner product $\langle \langle \quad, \quad \rangle\rangle$ on $G$ by

$$\langle \langle V, W \rangle \rangle = \int_{G} \langle L_{\sigma}^{-1} R_{\tau}(V), L_{\sigma}^{-1} R_{\tau}(W) \rangle \, d\mu(\sigma) \, d\mu(\tau).$$

Then $\langle \langle \quad, \quad \rangle\rangle$ 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 $L_{\tau}$ in any point $\tau \in G$ is given by the formula $L_{\tau}(\sigma) = \tau \sigma^{-1} \tau$.

**Proof:** By hypothesis $L_{\tau}$ and $R_{\tau}$ are isometries. Define a map

$$I_{\tau}: G \rightarrow G$$

by

$$I_{\tau}(\sigma) = \sigma^{-1} \tau.$$

Then $I_{\tau}: T_{e} G \rightarrow T_{e} G$ reverses the tangent space of $e$, so is certainly an isometry on this tangent space. Now the identity

$$I_{\tau} = R_{\sigma^{-1} \tau} L_{\sigma^{-1}}$$

shows that $I_{\tau}: G \rightarrow G^{-1}$ is an isometry for any $\sigma \in G$. Since $I_{\tau}$ reverses the tangent space at $e$, it reverses geodesics through $e$.

Finally, defining $L_{\tau}(\sigma) = \tau \sigma^{-1} \tau$, the identity $I_{\tau} = R_{\sigma^{-1} \tau} L_{\sigma^{-1}}$ shows that each $I_{\tau}$ is an isometry which reverses geodesics through $\tau$.

A 1-parameter subgroup of $G$ is a $C^{\infty}$ homomorphism of $\mathbb{R}$ 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.)

§21. Lie Groups

The geodesics $\gamma$ in $G$ with $\gamma(0) = e$ are precisely the one-parameter subgroups of $G$.

**Proof:** Let $\gamma: \mathbb{R} \rightarrow G$ be a geodesic with $\gamma(0) = e$. By Lemma 20.1 the map $\gamma_{t}: T_{e} G \rightarrow T_{\gamma(t)} G$ takes $\gamma(u)$ into $\gamma(u + 2\pi t)$. Now $\gamma_{t}(T_{e} G) = \gamma(t) \sigma \gamma(t)$ so $\gamma(t) \gamma(u) = \gamma(u + 2\pi t)$. By induction it follows that $\gamma(n \pi t) = \gamma(t)^{n}$ for any integer $n$. If $t' = m' t$ and $t'' = n'' t$ for some $t$ and some integers $m'$ and $n''$, then $\gamma(t' + t'') = \gamma(t') \gamma(t'')$. By continuity $\gamma$ is a homomorphism.

Now let $\gamma: \mathbb{R} \rightarrow G$ be a 1-parameter subgroup. Let $\gamma'$ be the geodesic through $e$ such that the tangent vector of $\gamma'$ at $e$ is the tangent vector of $\gamma$ at $e$. We have just seen that $\gamma'$ is a 1-parameter subgroup. Hence $\gamma' = \gamma$. This completes the proof.

A vector field $X$ on a Lie group $G$ is called left invariant if and only if $(L_{a})_{*}(X_{b}) = X_{ab}$ 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) $\langle\langle X_{\sigma}, Y_{\sigma} \rangle\rangle = \langle\langle X, Y \rangle\rangle$

(b) $R(X, Y)Z = \frac{1}{2} [[X, Y], Z]$

(c) $\langle\langle R(X, Y)Z, W \rangle\rangle = \frac{1}{4} \langle\langle [X, Y], [Z, W] \rangle\rangle$

**Proof:** As in §6 we will use the notation $X \triangleright Y$ for the covariant derivative of $Y$ in the direction $X$. For any left invariant $X$ the identity

$$X \triangleright X = 0$$

is satisfied, since the integral curves of $X$ are left translates of $1$.

Parameter subgroups, and therefore are geodesics.
Therefore
\[(X + Y) \star (X + Y) = (X \star X) + (X \star Y) + (Y \star X) + (Y \star Y)\]
is zero; hence
\[X \star Y + Y \star X = 0.
\]
On the other hand
\[X \star Y - Y \star X = [X, Y]
\]
by §8.5. Adding these two equations we obtain:
2) \[2X \star Y = [X, Y].\]
Now recall the identity
\[\langle X, Z \rangle = \langle Y \star X, Z \rangle + \langle X, Y \star Z \rangle.
\]
(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) \[\langle [Y, X], Z \rangle = \langle X, [Y, Z] \rangle.
\]
By definition, \(R(X, Y)Z\) is equal to
\[-X \star (Y \star Z) + Y \star (X \star Z) + [X, Y] \star Z.
\]
Substituting formula (d), this becomes
\[-\frac{i}{2}[X, [Y, Z]] + \frac{i}{2}[Y, [X, Z]] + \frac{i}{2}[X, Y, Z]
\]
Using the Jacobi identity, this yields the required formula
(b) \[R(X, Y)Z = \frac{i}{2}[X, Y, Z].
\]
The formula (c) follows from (a) and (b).

\section{LIE GROUPS

\subsection{Corollary 21.4}

The sectional curvature \(\langle R(X, Y)X, Y \rangle\) is always \(\geq 0\). Equality holds if and only if \([X, Y] = 0\).

Recall that the center \(c\) of a Lie algebra \(\mathfrak{g}\) is defined to be the set of \(X \in \mathfrak{g}\) such that \([X, Y] = 0\) for all \(Y \in \mathfrak{g}\).

\subsection{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.

\textbf{Proof:} This follows from Meyer's theorem (§19). Let \(X_i\) be any unit vector in \(\mathfrak{g}\) and extend to an orthonormal basis \(X_1, \ldots, X_n\). The Ricci curvature
\[K(X_1, X_1) = \sum_{i=1}^n \langle R(X_1, X_i)X_i, X_i \rangle
\]
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.

\subsection{Corollary 21.6}

A simply connected Lie group \(G\) with left and right invariant metric splits as a Cartesian product \(G' \times \mathbb{R}^k\) where \(G'\) is compact and \(\mathbb{R}^k\) 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 \(G' \times \mathbb{R}^k\) possesses a left and right invariant metric.

\textbf{Proof:} Let \(c\) be the center of the Lie algebra \(\mathfrak{g}\) and let \(\mathfrak{g}' = \{X \in \mathfrak{g} : \langle X, c \rangle = 0\} \text{ for all } c \in c\)
be the orthogonal complement of \(c\). Then \(\mathfrak{g}'\) is a Lie sub-algebra. For if \(X, Y \in \mathfrak{g}'\) and \(C \in c\) 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 \(\mathfrak{g} = \mathfrak{g}' \oplus c\) of Lie algebras. Hence \(G\) splits as a Cartesian product \(G' \times \mathbb{R}^n\), where \(G'\) is compact by 21.5 and \(\mathbb{R}^n\) is simply connected and abelian, hence isomorphic.
to some $\mathbb{R}^k$. (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 $\lambda$-cells for each even value of $\lambda$.

Thus the $X$-th homology groups of $\Omega(G)$ is zero for $\lambda$ odd, and is free abelian of finite rank for $\lambda$ 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 $H_2(S^3)$ is infinite cyclic for all even values of $1$.

**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 $\lambda$ for each geodesic from $p$ to $q$ of index $\lambda$. By §19.4 there are only finitely many $\lambda$-cells for each $\lambda$. Thus it only remains to prove that the index $\lambda$ of a geodesic is always even.

Consider a geodesic $\gamma$ starting at $p$ with velocity vector

$$\gamma = \frac{d}{dt}(t) \in T_p G \cong g$$

According to 620.5, the conjugate points of $p$ along $\gamma$ are determined by the eigenvalues of the linear transformation

$$K_\gamma : T_p G \to T_p G$$

defined by

$$K_\gamma(W) = R(W)\gamma = \frac{1}{4} [V,W] \gamma$$

Defining the adjoint homomorphism

$$\text{Ad } V : g \to g$$

we have

$$K_\gamma = \frac{1}{4} (\text{Ad } V) \circ (\text{Ad } V)$$

The linear transformation $\text{Ad } V$ is skew-symmetric; that is

$$< \text{Ad } V(W) , W' > = - < W, \text{Ad } V(W') >$$

This follows immediately from the identity 21.3a. Therefore we can choose an orthonormal basis $\{0_1, -a_1, 0, a_2, 0, \ldots\}$ so that the matrix of $\text{Ad } V$ takes the form

$$\begin{pmatrix}
0 & a_1 & 0 & a_2 & 0 & \\
-a_1 & 0 & -a_2 & 0 & \\
0 & a_2 & 0 & a_2 & 0 & \\
0 & 0 & -a_2 & 0 & \\
0 & 0 & 0 & -a_2 & 0 \end{pmatrix}$$

It follows that the composite linear transformation $(\text{Ad } V) \circ (\text{Ad } V)$ has matrix

$$\begin{pmatrix}
0 & a_1 & 0 & a_2 & 0 & \\
-a_1 & 0 & -a_2 & 0 & \\
0 & a_2 & 0 & a_2 & 0 & \\
0 & 0 & -a_2 & 0 & \\
0 & 0 & 0 & -a_2 & 0 \end{pmatrix}$$

Therefore the non-zero eigenvalues of $K_\gamma = \frac{1}{4} (\text{Ad } V)^2$ are positive, and occur in pairs.

It follows from 20.5 that the conjugate points of $p$ along $\gamma$ also occur in pairs. In other words every conjugate point has even multiplicity. Together with the Index Theorem, this implies that the index $\lambda$ 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 \( g^d \) of minimal geodesics forms a smooth manifold of dimension \( n \) which can be identified with the equator \( S^d \subset S^{n+1} \). 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 \( d(p, q) = \sqrt{2} \).

**Theorem 22.1.** If the space \( \Omega^d \) 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 \( \pi_i(\Omega^d, \{\alpha_i\}) \) is zero for \( 0 \leq i < \lambda_0 \).

It follows that the inclusion homomorphism
\[ \pi_i(\Omega^d) \to \pi_i(M) \]
is an isomorphism for \( i \leq \lambda_0 - 2 \). But it is well known that the homotopy group \( \pi_1(M) \) is isomorphic to \( \pi_{1+1}(M) \) for all values of \( 1 \). (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 \( 0 \leq i < \lambda_0 - 2 \).

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 \( i < 2n-2 \).

Theorem 22.1 also implies that the homology groups of the loop space \( \Omega^d \) are isomorphic to those of \( \Omega^d \) in dimensions \( \leq \lambda_0 - 2 \). This fact follows from 22.1 together with the relative Huwicz theorem. (See for example H. E. W. p. 106. 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 \( \mathbb{R}^d \); let \( U \) be a neighborhood of \( K \), and let
\[ f: U \to \mathbb{R} \]
be a smooth function such that all critical points of \( f \) in \( K \) have index \( \geq \lambda_0 \).

**Lemma 22.4.** If \( g: U \to \mathbb{R} \) is any smooth function which is "close" to \( f \), in the sense that
\[ \left| \frac{\partial g}{\partial x_i} - \frac{\partial f}{\partial x_i} \right| < \varepsilon, \quad \left| \frac{\partial^2 g}{\partial x_i \partial x_j} - \frac{\partial^2 f}{\partial x_i \partial x_j} \right| < \varepsilon, \quad (1, \ldots, n) \]
uniformly throughout \( K \), for some sufficiently small constant \( \varepsilon \), 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.)
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

\[
e_{i}^{1}, \ldots, e_{n}^{n}: U \to \mathbb{R},
\]
as follows. Let

\[
e_{i}^{1}(x) \leq e_{i}^{2}(x) \leq \ldots \leq e_{n}^{n}(x)
\]
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 \( \lambda \) if and only if the number \( e_{\lambda}(x) \) is negative.

The continuity of the functions \( e_{\lambda}^{i} \) follows from the fact that the A-th eigenvalue of a symmetric matrix depends continuously on the matrix. This in turn follows from the fact that the roots of a polynomial depend continuously on the polynomial. (See §14 of K. Knopp, "Theory of Functions, Part II," published in the United States by Dover, 1947.)

Let \( m_{g}(x) \) denote the larger of the two numbers \( k_g(x) \) and \( -e_{0}^{0}(x) \). Similarly let \( m_{f}(x) \) denote the larger of the corresponding numbers \( k_{f}(x) \) and \( -e_{0}^{0}(x) \). The hypothesis that all critical points of \( f \) in \( K \) have index \( \geq \lambda_0 \) implies that \( -e_{\lambda}^{0}(x) > 0 \) 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

\[
(*) \quad \left| k_{g}(x) - k_{f}(x) \right| < \delta, \quad \left| e_{\lambda}^{0}(x) - e_{\lambda}^{0}(x) \right| < \delta
\]
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 \( nxn \) symmetric matrices. If corresponding entries of the two matrices differ by at most \( \varepsilon \), then corresponding eigenvalues differ by at most \( \varepsilon \). This can be proved using Courant's minimax definition of the \( \lambda \)-th eigenvalue. (See H of Courant, "Über die Abhängigkeit der Schwingungszahlen einer Membran..." Nachrichten, Königlichen Gesellschaft der Wissenschaften zu Göttingen, Math. Phys. Klasse 1919, pp. 259-264.)
IV. APPLICATIONS

Theorem 22.4. Any $g$ which approximates $f$ sufficiently closely, the first and second derivatives also being approximated, will satisfy (2). In fact the compact set $g^{-1}[b,c+\delta]$ can be covered by finitely many compact sets $K_i$, 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 $g^{-1}[b,c+\delta] \subset g^{-1}[b,c+2\delta]$, and all critical points are non-degenerate, with index $\geq \lambda_0$. Hence the manifold $g^{-1}[-m,2\delta]$ has the homotopy type of $g^{-1}[-m,2\delta]$ with cells of dimension $\geq \lambda_0$ attached.

Now consider the map

$$h: \mathbb{R}^n \times [0,1] \to \tilde{\mathbb{R}}, \tilde{\mathbb{R}} \subset g^{-1}[-m,2\delta], \mathbb{R}^n.$$ 

Since $r < \lambda_0$, it follows that $h$ is homotopic within $g^{-1}[-m,2\delta], \mathbb{R}^n$ to a map

$$h': \mathbb{R}^n \times [0,1] \to g^{-1}[-m,2\delta], \mathbb{R}^n.$$ 

But this last pair is contained in $(U, \tilde{\mathbb{R}})$; and $U$ can be deformed into $\mathbb{R}^n$ within $\tilde{\mathbb{R}}$. It follows that $h'$ is homotopic within $(\mathbb{R}^n, \tilde{\mathbb{R}})$ to a map

$$h'': \mathbb{R}^n \times [0,1] \to \mathbb{R}^n, \mathbb{R}^n.$$ 

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

$$\pi_1(\text{Int } \tilde{\mathbb{R}}, \partial \tilde{\mathbb{R}}) = 0$$

for arbitrarily large values of $c$. As in §16 the space $\text{Int } \tilde{\mathbb{R}}$ contains a smooth manifold $\text{Int } \tilde{\mathbb{R}}(t_0, t_1, \ldots, t_k)$ as deformation retract. The space $\tilde{\mathbb{R}}$ of minimal geodesics is contained in this smooth manifold.

The energy function $E: \tilde{\mathbb{R}} \to \mathbb{R}$, when restricted to $\text{Int } \tilde{\mathbb{R}}(t_0, t_1, \ldots, t_k)$, almost satisfies the hypothesis of 22.5. The only difficulty is that $E(\tilde{\mathbb{R}})$ ranges over the interval $d \leq E < c$, instead of the required interval $[0, \delta)$. To correct this, let

$$F: [d, \delta] \to [0, \delta)$$

be any diffeomorphism.

Then

$$F \circ E: \text{Int } \tilde{\mathbb{R}}(t_0, t_1, \ldots, t_k) \to \mathbb{R}$$

satisfies the hypothesis of 22.5. Hence

$$\pi_1(\text{Int } \tilde{\mathbb{R}}(t_0, t_1, \ldots, t_k), \partial \tilde{\mathbb{R}}) \cong \pi_1(\text{Int } \tilde{\mathbb{R}}, \partial \tilde{\mathbb{R}})$$

is zero for $1 < \lambda_0$. 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 $\mathbb{C}^n$ be the space of $n$-tuples of complex numbers, with the usual Hermitian inner product. The unitary group $U(n)$ is defined to be the group of all linear transformations $S: \mathbb{C}^n \to \mathbb{C}^n$ which preserve this inner product. Equivalently, using the matrix representation, $U(n)$ is the group of all $n \times n$ complex matrices $S$ such that $SS^* = I$; 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
\[ \exp A = I + A + \frac{1}{2} A^2 + \frac{1}{3!} A^3 + \ldots \]

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. $U(n)$ is a smooth submanifold of the space of $n \times n$ matrices;
6. the tangent space $T_{I}U(n)$ can be identified with the space of $n \times n$ skew-Hermitian matrices.

Therefore the Lie algebra $\mathfrak{g}$ of $U(n)$ 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 $U(n)$. Computation shows that the bracket product of left invariant vector fields corresponds to the product $[A,B] = AB - BA$ of matrices.

Since $U(n)$ is compact, it possesses a left and right invariant Riemannian metric. Note that the function $\exp: TU(n) \to U(n)$ 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
\[ t \to \exp(tA) \]
defines a 1-parameter subgroup of $U(n)$ (by Assertion (2) above); and hence defines a geodesic.

A specific Riemannian metric on $U(n)$ 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
\[ \text{trace}(AB^*) = \sum_{i=1}^{n} a_{ii}b_{ii} \]
Clearly this inner product is positive definite on $\mathfrak{g}$.

This inner product on $\mathfrak{g}$ determines a unique left invariant Riemannian metric on $U(n)$. To verify that the resulting metric is also right invariant, we must check that it is invariant under the adjoint action of $U(n)$ on $\mathfrak{g}$.

DEFINITION of the adjoint action. Each $S \in U(n)$ determines an inner automorphism
\[ X \rightarrow SXS^{-1} = (L_S R_S^{-1})X \]
of the group $U(n)$. The induced linear mapping
\[ (L_S R_S^{-1}) : TU(n)_I \to TU(n)_I \]
is called $Ad(S)$. Thus $Ad(S)$ is an automorphism of the Lie algebra of $U(n)$. Using Assertion (1) above we obtain the explicit formula
\[ Ad(S)A = SAS^{-1}, \]
for $A \in \mathfrak{g}$, $S \in U(n)$.

The inner product $\langle A, B \rangle$ is invariant under each such automorphism $Ad(S)$. In fact if $A_1 = Ad(S)A$, $B_1 = Ad(S)B$ then the identity
\[ A_1 B_1^* = SAS^{-1}(SBS^{-1})^* = SAS^*S^{-1} \]
implies that
\[ \text{trace } (A_1 B_1) = \text{trace } (3AB^* S^{-1}) = \text{trace } (AB^*) \]
and hence that
\[ \langle A_1, B_1 \rangle = \langle A, B \rangle. \]
It follows that the corresponding left invariant metric on \( U(n) \) is also right invariant.

Given \( A \in \mathfrak{g} \) we know by ordinary matrix theory that there exists \( T \in U(n) \) so that \( T \exp^{-1} \) is in diagonal form
\[ T \exp^{-1} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \]
where the \( \lambda_j \)'s are real. Also, given any \( S \in U(n) \), there is a \( T \in U(n) \) such that
\[ T S T^{-1} = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \]
where again the \( \lambda_j \)'s are real. Thus we see directly that \( \exp : \mathfrak{g} \to U(n) \) is onto.

One may treat the special unitary group \( SU(n) \) in the same way.

\( SU(n) \) is defined as the subgroup of \( U(n) \) 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 \( \mathfrak{g}^1 \), the Lie algebra of \( SU(n) \) is the set of all matrices \( A \) such that \( A + A^* = 0 \) and \( \text{trace } A = 0 \).

In order to apply Morse theory to the topology of \( U(n) \) and \( SU(n) \), we begin by considering the set of all geodesics in \( U(n) \) from \( I \) to \( -I \). In other words, we look for all \( A \in \mathbb{T} U(n) \) such that \( \exp A = -I \). Suppose \( A \) is such a matrix: if it is not already in diagonal form, let \( T \in U(n) \) be such that \( T A T^{-1} \) is in diagonal form. Then
\[ \exp T A T^{-1} = T \exp A T^{-1} = T(-I) T^{-1} = -I \]
so that we may as well assume that \( A \) is already in diagonal form
\[ A = \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \]
In this case,
\[ \exp A = \begin{pmatrix} e^{\lambda_1} & & \\ & \ddots & \\ & & e^{\lambda_n} \end{pmatrix} \]
so that \( \exp A = -I \) if and only if \( A \) has the form
\[ \begin{pmatrix} k_1 e^{i\pi} & & \\ & \ddots & \\ & & k_n e^{i\pi} \end{pmatrix} \]
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 \( |A| = \sqrt{\text{trace } A^2} \), the length of the geodesic determined by \( A \) is \( \pi \sqrt{k_1^2 + \cdots + k_n^2} \). Thus \( A \) determines a minimal geodesic if and only if each \( k_j \) equals \( \pm 1 \), and in that case, the length is \( \pi \sqrt{n} \). Now, regarding such an \( A \) as a linear map of \( \mathbb{C}^n \) to \( \mathbb{C}^n \) observe that \( A \) is completely determined by specifying \( \text{Eigen}(1) \), the vector space consisting of all \( v \in \mathbb{C}^n \) such that \( Av = \text{inv} \); and \( \text{Eigen}(-1) \), the space of all \( v \in \mathbb{C}^n \) such that \( Av = -1v \). Since \( \mathbb{C}^n \) splits as the orthogonal sum \( \text{Eigen}(1) \oplus \text{Eigen}(-1) \), the matrix \( A \) is then completely determined by \( \text{Eigen}(1) \), which is an arbitrary subspace of \( \mathbb{C}^n \). Thus the space of all minimal geodesics in \( U(n) \) from \( I \) to \( -I \) may be identified with the space of all sub-vector-spaces of \( \mathbb{C}^n \).

Unfortunately, this space is rather inconvenient to use since it has components of varying dimensions. This difficulty may be removed by replacing \( U(n) \) by \( SU(n) \) and setting \( n = 2m \). In this case, all the above considerations remain valid. But the additional condition that \( a_1 + \cdots + a_{2m} = 0 \) with \( a_1 = \pm 1 \) restricts \( \text{Eigen}(1) \) to being an arbitrary \( m \) dimensional sub-vector-space of \( \mathbb{C}^{2m} \). This proves the following:
IV. APPLICATIONS

**Lemma 23.1.** The space of minimal geodesics from $I$ to $-I$ in the special unitary group $SU(2m)$ is homeomorphic to the complex Grassmann manifold $G_m(C^{2m})$, consisting of all $m$ dimensional vector subspaces of $C^{2m}$.

We will prove the following result at the end of this section.

**Lemma 23.2.** Every non-minimal geodesic from $I$ to $-I$ in $SU(2m)$ has index $\geq 2m+2$.

Combining these two lemmas with §22 we obtain:

**Theorem 23.3 (Bott).** The inclusion map $G_m(C^{2m}) \rightarrow SU(2m)$ induces isomorphisms of homotopy groups in dimensions $\leq 2m$. Hence

$$\pi_i G_m(C^{2m}) \cong \pi_{i+1} SU(2m)$$

for $i \leq 2m$.

On the other hand using standard methods of homotopy theory one obtains somewhat different isomorphisms.

**Lemma 23.4.** The group $\pi_j G_m(C^{2m})$ is isomorphic to $\pi_{j-1} U(m)$ for $1 \leq j \leq 2m$. Furthermore,

$$\pi_{j-1} U(m) \cong \pi_{j-1} U(m+1) \cong \pi_{j-1} U(m+2) \cong \ldots$$

for $1 \leq j \leq 2m$; and

$$\pi_{j} U(m) \cong \pi_{j} U(m)$$

for $j \neq 1$.

**Proof:** First note that for each $m$ there exists a fibration

$$U(m) \rightarrow U(m+1) \rightarrow S^{2m-1}$$

From the homotopy exact sequence

$$\ldots \rightarrow \pi_1 s^{2m+1} \rightarrow \pi_{1-1} U(m) \rightarrow \pi_{1-1} U(m+1) \rightarrow \pi_{1-1} s^{2m+1} \rightarrow \ldots$$

of this fibration we see that

$$\pi_{1-1} U(m) \cong \pi_{1-1} U(m+1)$$

for $1 \leq m$. (Compare Steenrod, "The Topology of Fibre Bundles," Princeton, 1951, p. 35 and p. 90.) It follows that the inclusion homomorphisms

$$\pi_{1-1} U(m) \rightarrow \pi_{1-1} U(m+1) \rightarrow \pi_{1-1} U(m+2) \rightarrow \ldots$$

are all isomorphisms for $1 \leq m$. These mutually isomorphic groups are called the $(i-1)$-st stable homotopy group of the unitary group. They will be denoted briefly by $\pi_{1-1} U$.

The same exact sequence shows that, for $i = 2m+1$, the homomorphism

$$\pi_{2m} U(m) \rightarrow \pi_{2m} U(m+1) \cong \pi_{2m} U$$

is onto.

The complex Stiefel manifold is defined to be the coset space $U(2m)/U(m)$. From the exact sequence of the fibration

$$U(m) \rightarrow U(2m) \rightarrow U(2m)/U(m)$$

we see that $\pi_i (U(2m)/U(m)) = 0$ for $1 \leq 2m$.

The complex Grassmann manifold $G_m(C^{2m})$ can be identified with the coset space $U(2m)/U(m) \times U(m)$. (Compare Steenrod §7.) From the exact sequence of the fibration

$$U(m) \rightarrow U(2m)/U(m) \rightarrow G_m(C^{2m})$$

we see now that

$$\pi_{1-1} G_m(C^{2m}) \cong \pi_{1-1} U(m)$$

for $1 \leq 2m$.

Finally, from the exact sequence of the fibration

$$SU(m) \rightarrow U(m) \rightarrow S^1$$

we see that $\pi_i SU(m) \cong \pi_i U(m)$ for $i \neq 1$. This completes the proof of Lemma 23.4.

Combining Lemma 23.4 with Theorem 23.3 we see that

$$\pi_{1-1} U = \pi_{1-1} U(m) \cong \pi_{1-1} G_m(C^{2m}) \cong \pi_{1-1} U(m+1) \cong \pi_{1-1} U$$

for $1 \leq 1 \leq 2m$. Thus we obtain:

**Periodicity Theorem.** $\pi_{1-1} U \cong \pi_{1-1} U$ for $1 \geq 1$.

To evaluate these groups it is now sufficient to observe that $U(1)$ is a circle; so that

$$\pi_0 U = \pi_0 U(1) = 0$$

$$\pi_1 U = \pi_1 U(1) \cong \mathbb{Z}$$ (infinite cyclic).

As a check, since $SU(2)$ is a 3-sphere, we have:

$$\pi_2 U = \pi_2 SU(2) = 0$$

$$\pi_3 U = \pi_3 SU(2) \cong \mathbb{Z}.$$

Thus we have proved the following result.
IV. APPLICATIONS

THEOREM 23.5 (Bott). The stable homotopy groups $\pi_j U$ of the unitary groups are periodic with period 2. In fact the groups $\pi_0 U \cong n, U \cong \pi_1 U \cong \cdots$ are zero, and the groups $\pi^1 U \cong \pi^3 U \cong \pi_5 U \cong \cdots$ 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 $SU(n)$, where $n$ is even. Recall that the Lie algebra

$$g' = \mathfrak{sl}(SU(n))$$

consists of all $n \times n$ skew-Hermitian matrices with trace zero. A given matrix $A \in g'$ corresponds to a geodesic from $I$ to $-I$ if and only if the eigenvalues of $A$ have the form $i \pi k_1, \ldots, i \pi k_n$ where $k_1, \ldots, k_n$ are odd integers with sum zero.

We must find the conjugate points to $I$ along the geodesic $t = \exp(tA)$.

According to Theorem 20.5 these will be determined by the positive eigenvalues of the linear transformation

$$K_A : g' \to g'$$

where

$$K_A(W) = R(A,W)A = \frac{1}{2} [[A,W],A].$$

(Compare §21.7.)

We may assume that $A$ is the diagonal matrix

$$
\begin{pmatrix}
1 \pi k_1 \\
\vdots \\
1 \pi k_n
\end{pmatrix}
$$

with $k_1 \geq k_2 \geq \cdots \geq k_n$. If $W = (w_{j\ell})$ then a short computation shows that

$$[A,[A,W]] = \left( \frac{1}{2} (k_j - k_\ell)^2 w_{j\ell} \right),$$

hence

$$[A,W] = \left( \frac{i}{2} (k_j - k_\ell)^2 w_{j\ell} \right).$$

and

$$K_A(W) = \left( \frac{i}{2} (k_j - k_\ell)^2 w_{j\ell} \right).$$

Now we find a basis for $g'$ 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 $(\ell j)$-th place and zeros elsewhere, is in $g'$ and is an eigenvector corresponding to the eigenvalue $\frac{8}{2} (k_j - k_\ell)^2$.

2) Similarly for each $j \neq \ell$ the matrix $E_{j\ell}'$ with $+1$ in the $(j\ell)$-th place and $-1$ in the $(\ell j)$-th place is an eigenvector, also with eigenvalue $\frac{8}{2} (k_j - k_\ell)^2$.

3) Each diagonal matrix in $g'$ is an eigenvector with eigenvalue 0. Thus the non-zero eigenvalues of $K_A$ are the numbers $\frac{i}{2} (k_j - k_\ell)^2$ with $k_j > k_\ell$. Each such eigenvalue is to be counted twice.

Now consider the geodesic $\gamma(t) = \exp tA$. Each eigenvalue $e = \frac{i}{2} (k_j - k_\ell)^2 > 0$ gives rise to a series of conjugate points along $\gamma$ 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

$$t = \left( \frac{k_j - k_\ell}{2} \right), \frac{k_j - k_\ell}{2}, \frac{k_j - k_\ell}{2}, \ldots .$$

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 $j, \ell$ with $k_j > k_\ell$ we obtain two copies of the eigenvalue $\frac{i}{2} (k_j - k_\ell)^2$, and hence a contribution of

$$e \left( \frac{k_j - k_\ell}{2} - 1 \right)$$

to the index. Adding over all $j, \ell$ this gives the formula

$$\lambda = \sum \left( \frac{k_j - k_\ell}{2} \right)$$

for the index of the geodesic $\gamma$.

As an example, if $\gamma$ is a minimal geodesic, then all of the $k_j$
are equal to $\pm 1$. 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 $\geq 3$, and we have

$$\lambda \geq \sum_{j=1}^{m+1} (3 - (-1) - 2) = 2(m+1) .$$

**CASE 2.** $m$ of the $k_i$ are positive and $m$ are negative but not all are $\pm 1$. Then one is $\geq 3$ and one is $\leq -3$ so that

$$\lambda \geq \sum_{j=1}^{m-1} (3 - (-1) - 2) + \sum_{j=1}^{m-1} (1 - (-3) - 2) + (3 - (-3) - 2)$$

$$= 4m \geq 2(m+1) .$$

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 On Clifford modules (to appear), by R. Bott and A. Shapiro, which relates the periodicity theorem with the structure of certain Clifford algebras.

Consider the vector space $R^n$ with the usual inner product. The orthogonal group $O(n)$ consists of all linear maps

$$T : R^n \rightarrow R^n$$

which preserve this inner product. Alternatively $O(n)$ consists of all real $n \times n$ matrices $T$ such that $T^* = T$. This group $O(n)$ can be considered as a smooth subgroup of the unitary group $U(n)$; and therefore inherits a right and left invariant Riemannian metric.

Now suppose that $n$ is even.

**DEFINITION.** A complex structure $J$ on $R^n$ is a linear transformation $J : R^n \rightarrow R^n$, belonging to the orthogonal group, which satisfies the identity $J^2 = -I$. The space consisting of all such complex structures on $R^n$ will be denoted by $\Omega_1(n)$.

We will see presently (Lemma 24.4) that $\Omega_1(n)$ is a smooth submanifold of the orthogonal group $O(n)$.

**REMARK.** Given some fixed $J_1 \in \Omega_1(n)$ let $U(n/2)$ be the subgroup of $O(n)$ consisting of all orthogonal transformations which commute with $J_1$. Then $\Omega_1(n)$ can be identified with the quotient space $O(n)/U(n/2)$.

**LEMMA 24.1.** The space of minimal geodesics from $I$ to $-I$ on $O(n)$ is homeomorphic to the space $\Omega_1(n)$ of complex structures on $R^n$.

**PROOF:** The space $O(n)$ can be identified with the group of $n \times n$ orthogonal matrices. Its tangent space $\mathfrak{o} = T_{O(n)} I$ can be identified with the space of $n \times n$ skew-symmetric matrices. Any geodesic $\gamma$ 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 $T \in O(n)$ so that
$$T A T^{-1} = \begin{pmatrix} 0 & a_1 & 0 & a_2 & 0 & \cdots & 0 & a_m \\ -a_1 & 0 & a_2 & 0 & \cdots & 0 & -a_2 & 0 \\ 0 & -a_2 & 0 & a_3 & 0 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -a_{m-1} & 0 & \cdots & 0 & -a_{m-1} \\ -a_m & 0 & \cdots & 0 & a_m & \cdots & 0 & 0 \\ 0 & a_m & \cdots & 0 & 0 & \cdots & a_m & 0 \\ -a_0 & 0 & \cdots & 0 & 0 & \cdots & 0 & a_0 \end{pmatrix}$$

with $a_1, a_2, \ldots, a_m \geq 0$. A short computation shows that $T(\exp(\pi A))T^{-1}$ is equal to
$$\begin{pmatrix} \cos \pi a_1 & \sin \pi a_1 & 0 & 0 & \cdots & 0 \\ -\sin \pi a_1 & \cos \pi a_1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cos \pi a_2 & \sin \pi a_2 & \cdots & 0 \\ 0 & -\sin \pi a_2 & 0 & \cos \pi a_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\sin \pi a_m & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \cos \pi a_m & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix}$$

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 $\langle A, B \rangle$ is easily seen to be $2(a_1^2 + a_2^2 + \cdots + a_m^2)$.

Therefore the geodesic $\gamma(t) = \exp(\pi t A)$ from $I$ to $-I$ is minimal if and only if $a_1 = a_2 = \cdots = a_m = 1$.

If $\gamma$ is minimal then
$$A^2 = T^{-1} \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & -1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix} T = -I,$$

hence $A$ is a complex structure.

Conversely, let $J$ be any complex structure. Since $J$ is orthogonal we have
$$JJ^* = I$$

where $J^*$ denotes the transpose of $J$. Together with the identity $J^2 = -I$ this implies that $J^* = -J$. Thus $J$ is skew-symmetric. Hence
$$T J T^{-1} = \begin{pmatrix} 0 & a_1 \\ -a_1 & 0 \end{pmatrix}$$

for some $a_1, a_2, \ldots, a_m \geq 0$ and some $T$. Now the identity $J^2 = -I$ implies that $a_1 = a_2 = \cdots = a_m = 1$; and hence that $\exp \pi J = -I$. This completes the proof.

Lemma 24.2. Any non-minimal geodesic from $I$ to $-I$ in $O(2m)$ has index $\geq 2m-2$.

The proof is similar to that of 23.2. Suppose that the geodesic has the form $t \mapsto \exp(\pi t A)$ with
$$A = \begin{pmatrix} 0 & a_1 & 0 & a_2 & 0 & \cdots & 0 & a_m \\ -a_1 & 0 & a_2 & 0 & \cdots & 0 & -a_2 & 0 \\ 0 & -a_2 & 0 & a_3 & 0 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -a_{m-1} & 0 & \cdots & 0 & -a_{m-1} \\ -a_m & 0 & \cdots & 0 & a_m & \cdots & 0 & 0 \\ 0 & a_m & \cdots & 0 & 0 & \cdots & a_m & 0 \\ -a_0 & 0 & \cdots & 0 & 0 & \cdots & 0 & a_0 \end{pmatrix}$$

where $a_1 \geq a_2 \geq \cdots \geq a_m \geq 0$ are odd integers. Computation shows that the non-zero eigenvectors of the linear transformation $K_A = -\frac{1}{4} (\text{Ad} A)^2$ 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
$$\lambda = \sum_{1 < j} (a_i + a_j - 2) + \sum_{a_i > a_j} (a_i - a_j - 2).$$

For a minimal geodesic we have $a_1 = a_2 = \cdots = a_m = 1$ so that
$$\lambda = 0,$$ as expected. For a non-minimal geodesic we have $a_i \geq 3$ so that
$$\lambda \geq \sum_{j=2}^m (3j-2) + 0 = 2m-2.$$

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 $\Omega_i(n) \to \mathfrak{O}(n)$ induces isomorphisms of homotopy groups in dimensions $\leq n-4$. Hence $\pi_i \Omega_i(n) \cong \pi_{i+1} \mathfrak{O}(n)$ for $i \leq n-4$.

Now we will iterate this procedure, studying the space of geodesics from $J$ to $-J$ in $\Omega_i(n)$, 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 $\mathbb{R}^n$ which anti-commute *, in the sense that

$$J_s J_v + J_v J_s = 0$$

for $s \neq v$. 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 $\mathbb{R}^n$ which anti-commute with the fixed structures $J_1, \ldots, J_{k-1}$.

Thus we have

$$\Omega_k(n) \subset \Omega_{k-1}(n) \subset \ldots \subset \Omega_1(n) \subset \mathfrak{O}(n).$$

Clearly each $\Omega_k(n)$ is a compact set. To complete the definition it is natural to define $\Omega_0(n)$ to be $\mathfrak{O}(n)$.

**Lemma 24.4.** Each $\Omega_k(n)$ is a smooth, totally geodesic** submanifold of $\mathfrak{O}(n)$. The space of minimal geodesics from $J_k$ to $-J_k$ in $\Omega_k(n)$ is homeomorphic to $\Omega_{k+1}(n)$, for $0 \leq k < n$.

It follows that each component of $\Omega_k(n)$ is a symmetric space.

For the isometric reflection of $\mathfrak{O}(n)$ in a point of $\Omega_k(n)$ will automatically carry $\Omega_k(n)$ to itself.

---

* These structures make $\mathbb{R}^n$ 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.
Conversely, let \( \gamma \) be any minimal geodesic from \( J_k \) to \(-J_k \) in \( \Omega_k(n) \). Setting \( \gamma(t) = J_k \exp(t A) \), it follows from Assertion 1.2 that \( A \) is a complex structure, and from Assertions 1.2 that \( A \) commutes with \( J_1, \ldots, J_{k-1} \) and anti-commutes with \( J_k \). It follows easily that \( J_k A \) belongs to \( \Omega_{k+1}(n) \). This completes the proof of 24.4.

**Remark.** The point \( J_k A \in \Omega_k(n) \) which corresponds to a given geodesic \( \gamma \) 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 \( \Omega_k(n+1,1) \) as follows. Choose fixed anti-commuting complex structures \( J'_1, \ldots, J'_k \) on \( \mathbb{R}^{n+1} \). Then each \( J \in \Omega_k(n) \) determines a complex structure \( J \oplus J'_k \) on \( \mathbb{R}^n \oplus \mathbb{R}^{n+1} \) which anti-commutes with \( J_\alpha \oplus J'_\alpha \) for \( \alpha = 1, \ldots, k-1 \).

**Definition.** Let \( \Omega_k \) 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 \( \mathcal{O} = \Omega_0 \) is called the infinite orthogonal group.

It is not difficult to see that the inclusions \( \Omega_k(n) \to \Omega_k(n+1) \) give rise, in the limit, to inclusions \( \Omega_k \to \Omega_k \).

**Theorem 24.5.** For each \( k \geq 0 \) this limit map \( \Omega_{k+1} \to \Omega_k \) is a homotopy equivalence. Thus we have isomorphisms
\[
\Omega_k \simeq \Omega_{k+1}, \quad \Omega_k \simeq \Omega_{k+1}, \quad \Omega_k \simeq \Omega_{k+1}.
\]

The proof will be given presently.

Next we will give individual descriptions of the manifolds \( \Omega_k(n) \) for \( k = 0,1,2,\ldots,8 \).

\( \Omega_0(n) \) is the orthogonal group.

\( \Omega_1(n) \) is the set of all complex structures on \( \mathbb{R}^n \).

Given a fixed complex structure \( J_1 \) we may think of \( \mathbb{R}^n \) as being a vector space \( \mathbb{C}^{n/2} \) over the complex numbers.

\( \Omega_2(n) \) can be described as the set of "quaternionic structures" on the complex vector space \( \mathbb{C}^{n/2} \). Given a fixed \( J_2 \in \Omega_2(n) \) we may think of \( \mathbb{C}^{n/2} \) as being a vector space \( \mathbb{H}^{n/4} \) over the quaternions \( \mathbb{H} \). Let \( \text{Sp}(n/4) \) be the group of isometries of this vector space onto itself. Then \( \Omega_2(n) \) can be identified with the quotient space \( \mathbb{U}(n/2)/\text{Sp}(n/4) \).

Before going further it will be convenient to set \( n = 16r \).

**Lemma 24.6 - (3).** The space \( \Omega_2(16r) \) can be identified with the quaternionic Grassmann manifold consisting of all quaternionic subspaces of \( \mathbb{H}^{16r} \).

**Proof:** Any complex structure \( J_3 \in \Omega_3(16r) \) determines a splitting of \( \mathbb{H}^{16r} = \mathbb{R}^{16r} \) into two mutually orthogonal subspaces \( V_1 \) and \( V_2 \) as follows. Note that \( J_1 J_2 J_3 \) is an orthogonal transformation with square \( J_1 J_2 J_3 = J_3 J_2 J_1 = J_1 J_2 J_3 \) equal to \( -1 \). Hence the eigenvalues of \( J_1 J_2 J_3 \) are \( \pm 1 \).

Let \( V_1 \subset \mathbb{R}^{16r} \) be the subspace on which \( J_1 J_2 J_3 \) equals \( +1 \); and let \( V_1 \) be the orthogonal subspace on which it equals \( -1 \). Then clearly \( \mathbb{R}^{16r} = V_1 \oplus V_2 \). Since \( J_1 J_2 J_3 \) 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 \( \mathbb{H}^{16r} = V_1 \oplus V_2 \), into mutually orthogonal quaternionic subspaces, we can define \( J_3 \in \Omega_3(16r) \) by the identities
\[
J_3 V_1 = -J_1 J_2 V_1, \quad J_3 V_2 = J_1 J_2 V_2.
\]

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 \( \mathbb{H}^{16r} \). Henceforth, we will assume that \( J_3 \) has been chosen in this way, so that \( \dim V_1 = \dim V_2 = 2r \).

**Lemma 24.6 - (4).** The space \( \Omega_3(16r) \) can be identified with the set of all quaternionic isometries from \( V_1 \) to \( V_2 \). Thus \( \Omega_3(16r) \) is diffeomorphic to the symplectic group \( \text{Sp}(2r) \).

**Proof:** Given \( J_4 \in \Omega_4(16r) \) note that the product \( J_3 J_4 \) anti-commutes with \( J_1 J_2 J_3 \). Hence \( J_3 J_4 \) maps \( V_1 \) to \( V_2 \) (and \( V_2 \) to \( V_1 \))
IV. APPLICATIONS

Since $J_3 J_4$ commutes with $J$, and $J_2$, we see that

$$J_3 J_4 | V_1 : V_1 \rightarrow V_2$$

is a quaternionic isomorphism. Conversely, given any such isomorphism $T : V_1 \rightarrow V_2$ it is easily seen that $J_2$ is uniquely determined by the identities:

$$\begin{align*}
J_2 | V_1 &= J_3^{-1} T \\
J_2 | V_2 &= -T^{-1} J_3
\end{align*}$$

This proves 24.6 - (4).

**Lemma 24.6 - (5).** The space $\Omega (16r)$ can be identified with the set of all vector spaces $W \subset V$, such that

(1) $W$ is closed under $J$, (i.e., $W$ is a complex vector space) and

(2) $V_1$ splits as the orthogonal sum $W \oplus J_1 W$.

**Proof:** Given $J_2 \in \Omega (16r)$, note that the transformation $J_2 J_4 J_5$ commutes with $J_2 J_3 J_4$ and has square $+1$. Thus $J_2 J_4 J_5$ maps $V_1$ into itself; and determines a splitting of $V_1$ into two mutually orthogonal subspaces. Let $W \subset V$, be the subspace on which $J_2 J_4 J_5$ coincides with $+1$. Since $J_2$ anti-commutes with $J_1 J_4 J_5$, it follows that $J_2 W \subset V$, is precisely the orthogonal subspace, on which $J_2 J_4 J_5$ equals $-1$. Clearly $J_1 W = W$.

Conversely, given the subspace $W$, it is not difficult to show that $J_2$ is uniquely determined.

**Remark.** If $U(2r) \subset Sp(2r)$ denotes the group of quaternion automorphisms of $V$, keeping $W$ fixed, then the quotient space $Sp(2r)/U(2r)$ can be identified with $\Omega (16r)$.

**Lemma 24.6 - (6).** The space $\Omega (16r)$ can be identified with the set of all real subspaces $X \subset W$ such that $W$ splits as the orthogonal sum $X \oplus J_1 X$.

**Proof:** Given $J_2 \in \Omega (16r)$, note that the transformation $J_2 J_4 J_6$ commutes both with $J_2 J_4 J_5$, and with $J_2 J_4 J_5$. Hence $J_2 J_4 J_6$ maps $W$ into its self. Since $(J_2 J_4 J_6)^2 = 1$, it follows that $J_2 J_4 J_6$ determines a splitting of $W$ into two mutually orthogonal subspaces. Let $X \subset W$ be the subspace on which $J_2 J_4 J_6$ equals $+1$. Then $J_1 X$ will be the orthogonal subspace on which it equals $-1$.

Conversely, given $X \subset W$, it is not hard to see that $J_8$ is uniquely determined.

**Remark.** If $0(2r) \subset U(2r)$ denotes the group of complex automorphisms of $W$ keeping $X$ fixed, then the quotient space $U(2r)/O(2r)$ can be identified with $\Omega (16r)$.

**Lemma 24.6 - (7).** The space $\Omega (16r)$ can be identified with the real Grassmann manifold consisting of all real subspaces of $X \subset R^{2r}$.

**Proof:** Given $J_2$, anti-commuting with $J, \ldots, J_6$, note that $J_2 J_4 J_6$ commutes with $J_2 J_4 J_5$, with $J_4 J_2 J_5$, and with $J_2 J_4 J_6$, and has square $+1$. Thus $J_2 J_4 J_6$ determines a splitting of $X$ into two mutually orthogonal subspaces $X$, (where $J_2 J_4 J_6$ equals $+1$) and $X$ (where $J_2 J_4 J_6$ equals $-1$). Conversely, given $X \subset X$, it can be shown that $J_2$ is uniquely determined.

This space $\Omega (16r)$, like $\Omega (16r)$, has components of varying dimension. Again we will restrict attention to the component of largest dimension, by assuming that $\dim X = \dim X_2 = r$.

Thus we obtain:

**Assertion.** The largest component of $\Omega (16r)$ is diffeomorphic to the Grassmann manifold consisting of $r$-dimensional subspaces of $R^{2r}$.

**Lemma 24.6 - (8).** The space $\Omega (16r)$ can be identified with the set of all real isometries from $X$ to $X_2$.

**Proof:** If $J_2 \in \Omega (16r)$ then the orthogonal transformation $J_2 J_8$ commutes with $J_2 J_4 J_5$, $J_2 J_4 J_5$, and $J_2 J_4 J_6$, but anti-commutes with $J_1 J_5 J_7$. Hence $J_2 J_8$ maps $X$ isometrically onto $X_2$. Clearly this isomorphism determines $J_8$ uniquely.

Thus we see that $\Omega (16r)$ is diffeomorphic to the orthogonal
Let us consider this diffeomorphism $\Omega_3(16r) \to \text{O}(r)$, and pass to the limit as $r \to \infty$. It follows that $\Omega_3$ is homeomorphic to the infinite orthogonal group $\text{O}$. Combining this fact with Theorem 24.5, we obtain the following.

**Theorem 24.7 (Bott).** The infinite orthogonal group $\text{O}$ has the same homotopy type as its own 8-th loop space. Hence the homotopy group $\pi_8 \text{O}$ is isomorphic to $\pi_4 \text{Sp}$ for $i \geq 0$.

If $\text{Sp} = \Omega_2$ denotes the infinite symplectic group, then the above argument also shows that $\text{O}$ has the homotopy type of the 8-fold loop space $\Omega_2 \text{Sp}$, and that $\text{Sp}$ has the homotopy type of the 4-fold loop space $\Omega_4 \text{O}$. The actual homotopy groups can be tabulated as follows.

<table>
<thead>
<tr>
<th>$i$ modulo 8</th>
<th>$\pi_1 \text{O}$</th>
<th>$\pi_1 \text{Sp}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\mathbb{Z}_2$</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>$\mathbb{Z}_2$</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>$\mathbb{Z}$</td>
<td>$\mathbb{Z}$</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>$\mathbb{Z}_2$</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>$\mathbb{Z}$</td>
<td>$\mathbb{Z}$</td>
</tr>
</tbody>
</table>

The verification that these groups are correct will be left to the reader. (Note that $\text{Sp}(1)$ is a 3-sphere, and that $\text{SO}(3)$ 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 $\Omega_k \mathbb{S}(r)$. In fact any additional complex structures $J_2, J_6, \ldots, J_4$ on $\mathbb{R}^{16r}$ give rise to anti-commuting complex structures $J_2J_4, J_6J_{10}, \ldots, J_8J_{14}$ on $X_4$, and hence to an element of $\Omega_{k-8}(r)$. However, for our purposes it will be sufficient to stop with $k = 8$.

---

**§24. The Orthogonal Group**

**Definition.** $V$ is a minimal $(J_1, \ldots, J_K)$-space if no proper, non-trivial subspace is closed under the action of $J_1, \ldots, J_K$. Two such minimal vector spaces are isomorphic if there is an isometry between them which commutes with the action of $J_1, \ldots, J_K$.

**Lemma 24.8 (Bott and Shapiro).** For $k \not\equiv 3 \pmod{8}$, any two minimal $(J_1, \ldots, J_K)$-vector spaces are isomorphic.

The proof of 24.8 follows that of 24.6. For $k = 0, 1, 2$ 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 $J_3J_2$ or $-J_3J_2$. 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 $J_7J_6$, $J_6J_7$, or to $-J_7J_6$. 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 $(J_1, \ldots, J_K)$-vector space. From the above discussion we see that:

- $m_0 = 1, m_1 = 2, m_2 = m_3 = 4$.
- $m_4 = m_5 = 8, m_6 = 16$.

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 $J_1, \ldots, J_k$ are anti-commuting complex structures on a vector
space $V$ of dimension $m_k$. Here $r$ can be any positive integer. Then for each unit vector $u \in V$ the $k$ vectors $J_1 u, J_2 u, \ldots, J_k u$ are perpendicular to each other and to $u$. Thus we obtain $k$ linearly independent vector fields on an $(m_{k-1})$-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 $k \neq 2 \pmod{4}$. We must study non-minimal geodesics from $J$ to $-J$ in $\Omega_k(n)$. Recall that the tangent space $\Omega_k(n)$ at $J$ consists of all matrices $JA$ where

1) $A$ is skew
2) $A$ anti-commutes with $J$
3) $A$ commutes with $J, \ldots, J_{k-1}$.

Let $T$ denote the vector space of all such matrices $A$. A given $A \in T$ corresponds to a geodesic $t \mapsto J \exp(tA)$ from $J$ to $-J$ if and only if its eigenvalues are all odd multiples of 1.

Each such $A \in T$ determines a self-adjoint transformation $K_A : T \rightarrow T$. Since $\Omega_k(n)$ is a totally geodesic submanifold of $O(n)$, we can compute $K_A$ by the formula

$$K_A = \frac{1}{2} [A, [A, B]] = \frac{(-A^2 B + 2ABA - BA^2)}{2} \cdot$$

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 $t \mapsto J \exp(tA)$.

Split the vector space $R^n$ 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 $J, \ldots, J_{k-1}$, $J$ and $A$. Then the eigenvalues of $A$ on $M_\alpha$ must be all equal, except for sign. For otherwise $M_\alpha$ would split as a sum of eigenspaces of $A$; and hence would not be minimal. Let $\lambda_0$ be the two eigenvalues of $A|M_\alpha$; where $\lambda_1, \ldots, \lambda_s$ are odd, positive integers.

Now note that $J^4 = \lambda_0 I$ is a complex structure on $M_\alpha$ which anti-commutes with $J_1, \ldots, J_{k-1}$, and $J$. Thus $M_\alpha$ is $(J_1, \ldots, J_{k-1}, J^4)$-minimal. Hence the dimension of $M_\alpha$ is $m_{k+1}$. Since $k + 1 \neq 3 \pmod{4}$ we see that $M_1, M_2, \ldots, M_s$ are mutually isomorphic.

For each pair $\alpha, \beta$ with $\alpha \neq \beta$ we can construct an eigenvector $B : R^n \rightarrow R^n$ of the linear transformation $K_A : T \rightarrow T$ as follows. Let $B|M_\alpha$ be zero for $\alpha \neq \beta$. Let $B|M_\alpha$ be an isometry from $M_\alpha$ to $M_\beta$ which satisfies the conditions

$$B|_\alpha = J_\alpha \overline{B}$$
$$B|_\beta = \overline{B}_\beta.$$
Now consider the geodesic \( t \to J \exp(\pi \mathbf{A}) \). Each eigenvalue \( e^{\theta} \) of \( K_{\mathbf{A}} \) gives rise to conjugate points along this geodesic for \( t = e^{\theta}, 2e^{\theta}, 3e^{\theta}, \ldots \) by 20.5. Thus if \( e^{\theta} \geq h \) 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 \( \geq n/k^+ - 1 \).

It follows that the inclusion map
\[
\Omega_{k+1}(n) \to \Omega_k(n)
\]
induces isomorphisms of homotopy groups in dimensions \( \leq n/k^+ - 1 \). This number tends to infinity with \( n \). Therefore, passing to the direct limit as \( n \to \infty \), it follows that the inclusion map \( i : \Omega_{k+1} \to \Omega_k \) induces isomorphisms of homotopy groups in all dimensions. But it can be shown that both \( \Omega_{k+1} \) and \( \Omega_k \) have the homotopy type of a \( \mathbb{R}^n \)-complex. Therefore, by Whitehead’s theorem, it follows that \( i \) is a homotopy equivalence. This completes the proof of 24.5 providing that \( k \neq 2 \) (mod 4).

**Proof of 24.5**. For \( k \equiv 2 \) (mod 4). The difficulty in this case may be ascribed to the fact that \( \Omega_k(n) \) has an infinite cyclic fundamental group. Thus \( \Omega_k(n) \) 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
\[
f : \Omega_k(n) \to \mathbb{S}^1 \subset \mathbb{C}
\]
as follows. Let \( J_1, \ldots, J_{k-1} \) be the fixed anti-commuting complex structure on \( \mathbb{R}^n \). Make \( \mathbb{R}^n \) into an \((n/2)\)-dimensional complex vector space by defining
\[
vJ = J_1J_2 \cdots J_{k-1}v
\]
for \( v \in \mathbb{R}^n \), where \( i = \sqrt{-1} \in \mathbb{C} \). The condition \( k \equiv 2 \) (mod 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 \( J'J^{-1} \) commutes with \( J \). Thus \( J'J^{-1} \) is a unitary complex linear transformation, and has a well defined complex determinant which will be denoted by \( f(J') \).

Now consider a geodesic \( t \to J \exp(\pi \mathbf{A}) \) from \( J \) to \(-J \) in \( \Omega_k(n) \). Since \( A \) commutes with \( i = J_1J_2 \cdots J_{k-1} \) (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
\[
f(J \exp(\pi \mathbf{A})) = \text{determinant} \ (\exp(\pi \mathbf{A})) = e^{n \text{trace } A}
\]
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 \( \Omega_k(n)J_1, \ldots, J_{k-1} \).

The index \( \lambda \) of this geodesic can be estimated as follows. As before split \( \mathbb{R}^n \) into an orthogonal sum \( M_1 \oplus \cdots \oplus M_j \) where each \( M_h \) is closed under the action of \( J_1, \ldots, J_{k-1} \) and \( A \); and \( M_j \) is minimal. Thus for each \( h \) the complex linear transformation \( A|_{M_h} \) can have only one eigenvalue, say \( \lambda_h \). For otherwise \( M_h \) would split into eigenspaces. Thus \( A|_{M_h} \) coincides with \( \lambda_hJ_1 \cdots J_{k-1}|_{M_h} \). Since \( M_j \) is minimal under the action of \( J_1, \ldots, J_{k-1} \), and \( J \); its complex dimension is \( m_{k/2} \).

Therefore the trace of \( A \) is equal to \( 1(a_1 + \cdots + a_m)m_{k/2} \).

Now for each \( h \neq j \) an eigenvector \( B \) of the linear transformation
\[
B \to K_{\mathbf{A}}B = (\lambda_B^2B + \lambda_BBA - BA^2)/\hbar
\]
Can be constructed much as before. Since \( M_{h} \) and \( M_{j} \) are \( (J_1, \ldots, J_{k-1}, \overline{J}) \)-minimal it follows from 24.8 that there exists an isometry
\[
B|_{M_j} : M_h \to M_j
\]
which commutes with \( J_1, \ldots, J_{k-1} \) and anti-commutes with \( J \). Let \( B|_{M_{j}} \) be the negative adjoint of \( B|_{M_{j}} \); and let \( B|_{M_{j}} \) be zero for \( \theta \neq h, j \). Then an easy computation shows that
\[
K_{\mathbf{A}}B = (\lambda_j - \lambda_{j})^2B/\hbar .
\]
Thus for each \( \alpha_h > \alpha_j \) we obtain an eigenvalue \((\alpha_h - \alpha_j)^2/\hbar \) for \( K_{\mathbf{A}} \).

Since each such eigenvalue makes a contribution of \( (\alpha_h - \alpha_j)^2/\hbar - 1 \) towards the index \( \lambda \), we obtain the inequality
\[
2\lambda \geq \sum_{h \neq j} (\alpha_h - \alpha_j - 2) .
\]
Now let us restrict attention to some fixed component of \( \Omega_k(n) \). That is let us look only at matrices \( A \) such that \( \text{trace } A = 1c m_k/2 \) where \( c \) is some constant integer.

Thus the integers \( a_1, \ldots, a_n \) satisfy

1) \( a_1 \equiv a_2 \equiv \ldots \equiv a_n \equiv 1 \pmod{2} \), (since \( \exp(xA) = -I \)).
2) \( a_1 + \ldots + a_n = c \), and
3) \( \max \left| a_n \right| \geq 3 \) (for a non-minimal geodesic).

Suppose for example that some \( a_n \) 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

\[ 2 \lambda \geq \sum_{a_h > 0} (a_h - a_j - 2) > \sum_{a_h > 0} (a_h - (-3) - 3) = p, \]

hence \( 4 \lambda \geq 2p \geq r + c \); where \( r = n/m_k \) tends to infinity with \( n \). It follows that the component of \( \Omega_k(n) \) 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.

### IV. APPLICATIONS

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 \( \Omega_0 \subset \Omega_1 \subset \Omega_2 \subset \ldots \) of the path space \( \Omega = \Omega(M; p,q) \), and given the information that each \( \Omega_k \) has the homotopy type of a certain CW-complex, we wish to prove that the union \( \Omega \) 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

\[ X = X_0 \times [0,1] \cup X_1 \times [1,2] \cup X_2 \times [2,3] \cup \ldots. \]

This is to be topologized as a subset of \( X \times R \).

**DEFINITION.** We will say that \( X \) is the homotopy direct limit of the sequence \( (X_i) \) if the projection map \( p : X \to X_0 \) defined by \( p(x, \tau) = x \) is a homotopy equivalence.

**EXAMPLE 1.** Suppose that each point of \( X_i \) lies in the interior of some \( X_i \), and that \( X \) is paracompact. Then using a partition of unity one can construct a map

\[ f : X \to R \]

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 \( X \) 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 \( P : X \to X \) 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 \([0] \cup \{1/4,1\}\).

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

\[
\begin{array}{ccccccc}
\emptyset & \subset & \emptyset & \subset & \emptyset & \subset & \ldots \\
\downarrow & & \downarrow & & \downarrow & & \\
K_0 & \subset & K_1 & \subset & K_2 & \subset & \ldots \\
\end{array}
\]

of homotopy equivalences. Since \(\emptyset = \bigcup \emptyset_i\) and \(K = \bigcup K_i\) are homotopy direct limits (compare Examples 1 and 2 above), it follows that the limit mapping \(\emptyset \to K\) is also a homotopy equivalence.

**PROOF** of Theorem A. Define \(\tilde{f}: X_\Sigma \to Y\) by \(\tilde{f}(x,t) = (f(x),t)\). It is clearly sufficient to prove that \(\tilde{f}\) 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 \(f_\Sigma\) is a homotopy equivalence.

**REMARK.** Under these conditions it would be natural to conjecture that \(f_\Sigma\) must actually be homotopic to the identity. However counterexamples can be given.

For each \(n\) let \(h^n_0: X_n \to X_n\) be a one-parameter family of mappings, with \(h^n_0 = f^n\), \(h^n = \text{identity}\). Define the homotopy

\[h^n_t: X_n \to X_n\]

as follows (where it is always to be understood that \(0 \leq t \leq 1\), and \(n = 0,1,2,\ldots\)).

\[
h^n_t(x,n+t) = \begin{cases} 
(h^n_0(x,n+2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\
(h^n_{(3-5t)}(x),n+1) & \text{for } \frac{1}{2} \leq t < \frac{3}{4} \\
(h^n_{(4-5t)}(x),n+1) & \text{for } \frac{3}{4} \leq t \leq 1 
\end{cases}
\]

Taking \(u = 0\) this defines a map \(h_0: X_\Sigma \to X_\Sigma\) which is clearly homotopic to \(f\). The mapping \(h_1: X \to X\) on the other hand has the following properties:

\[h_1(x,n+t) = (x,n+2t) \text{ for } 0 \leq t \leq \frac{1}{2}
\]

\[h_1(x,n+t) \in X_{n+1}(n+1) \text{ for } \frac{1}{2} \leq t \leq 1.
\]

We will show that any such map \(h_1\) is a homotopy equivalence. In fact a homotopy inverse \(g: X \to X_\Sigma\) can be defined by the formula

\[g(x,n+t) = \begin{cases} 
(x,n+2t) & \text{for } 0 \leq t \leq \frac{1}{2} \\
h_1(x,n+\frac{3}{2}-t) & \text{for } \frac{1}{2} \leq t \leq 1
\end{cases}
\]

This is well defined since

\[h_1(x,n+\frac{1}{2}) = h_1(x,n+1) = (x,n+1).
\]

Proof that the composition \(h_1g\) is homotopic to the identity map or \(X_\Sigma\). Note that

\[
h_1g(x,n+t) = \begin{cases} 
(x,n+4t) & \text{for } 0 \leq t \leq \frac{1}{2} \\
h_1(x,n+2t) & \text{for } \frac{1}{2} \leq t \leq 1
\end{cases}
\]

Define a homotopy \(H_1: X_\Sigma \to X_\Sigma\) as follows. For \(0 \leq u \leq \frac{1}{2}\) let

\[
H_1(x,n+t) = \begin{cases} 
(h_1g(x,n+t) & \text{for } 0 \leq t \leq \frac{1}{2} \\
h_1(x,n+1-n) & \text{for } \frac{1}{2} + n \leq t \leq 1
\end{cases}
\]

for \(\frac{1}{2} + n \leq t \leq 1\).
This is well defined since
\[ h_1g(x, n+(1-u)/2) = h_1g(x, n+\frac{2}{3}u) = h_1(x, n+1-u). \]

Now \( F_0 \) is equal to \( h_1 \) and \( H_1 \) is given by
\[
H_2(x, n+t) = \begin{cases} 
(x, n+\frac{1}{3}) & 0 \leq t \leq \frac{1}{3} \\
(x, n+1) & \frac{1}{3} \leq t \leq 1.
\end{cases}
\]

Clearly this is homotopic to the identity.

Thus \( h_1g \) 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
\[
\begin{array}{ccc}
Y_n & \xrightarrow{g_n} & X_n \\
\downarrow{j_n} & & \downarrow{i_n} \\
Y_{n+1} & \xrightarrow{g_{n+1}} & X_{n+1}
\end{array}
\]

(where \( i_n \) and \( j_n \) denote inclusion maps) is homotopy commutative. In fact
\[
i_n g_n \sim g_{n+1} f_{n+1} i_n g_n = g_{n+1} i_n g_n \sim i_{n+1} g_{n+1}.
\]

Choose a specific homotopy \( h_n^0 : Y_n \to X_{n+1} \) with \( h_n^0 = i_n g_n \), \( h_n^1 = g_{n+1} i_n \), and define \( G : Y_n \to X_n \) by the formula
\[
G(y, n+t) = \begin{cases} 
(g_n(y), n+2t) & 0 \leq t \leq \frac{1}{2} \\
(h_{2t-1}(y), n+1) & \frac{1}{2} \leq t \leq 1.
\end{cases}
\]

We will show that the composition \( Gf : X_n \to X \) is a homotopy equivalence.

Let \( X_n^\mathbb{N} \) denote the subset of \( X \) consisting of all pairs \( (x, \tau) \) with \( 0 \leq \tau \leq n \). (Thus \( X_n^\mathbb{N} = X_0 \times \{0, 1\} \cup \ldots \cup X_{n-1}\times\{n-1,n\} \cup X_n \times \{n\}). The composition \( Gf \) carries \( X_n^\mathbb{N} \) into itself by a mapping which is homotopic to the identity. In fact \( X_n^\mathbb{N} \) contains \( X_n \times \{n\} \) as deformation retract; and the mapping \( Gf \) 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 \( (X_n^\mathbb{N}) \), and conclude that \( Gf \) is a homotopy equivalence.

This proves that \( f \) has a left homotopy inverse. A similar argument shows that \( f^{-1} \) is a homotopy equivalence, so that \( f \) has a right homotopy inverse. This proves that \( f \) is a homotopy equivalence (compare page 22) and completes the proof of Theorem A.

**COROLLARY.** Suppose that \( X \) is the homotopy direct limit of \( (X_n) \). If each \( X_n \) has the homotopy type of a CW-complex, then \( X \) itself has the homotopy type of a CW-complex.

The proof is not difficult.