MORSE THEORY

BY

J. Milnor

Based on lecture notes by

M. SPIVAK and R. WELLS

PRINCETON, NEW JERSEY

PRINCETON UNIVERSITY PRESS

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

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

Printed in the United States of America

19 18 17 16 15 14 13 12

PREFACE

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

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

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

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

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

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

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

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

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

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

J.W.M.

Los Angeles, June 1968.

CONTENTS

CONTENTS

PART I

NON-DEGENERATE SMOOTH FUNCTIONS ON A MANIFOLD.

§1. Introduction.

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

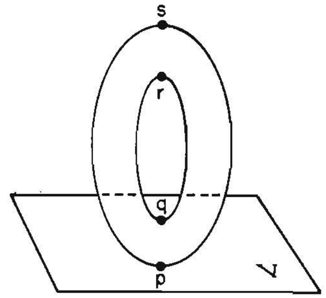


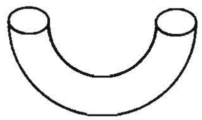
Diagram 1.

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

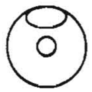
(1) If $a < 0 = f(p)$, then M^a is vacuous.

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

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



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



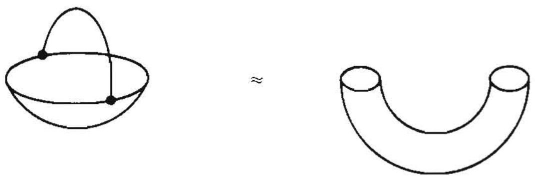
(5) If $f(s) < a$, then M^a is the full torus.

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

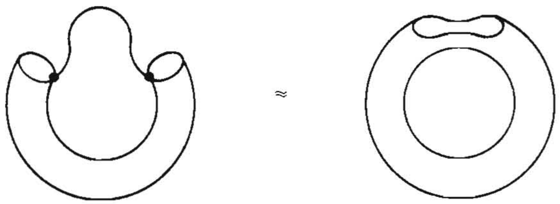
(1) → (2) is the operation of attaching a 0-cell. For as far as homotopy type is concerned, the space M^a, $f(p) < a < f(q)$, cannot be distinguished from a 0-cell:

Here "≈" means "is of the same homotopy type as."

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



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



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

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

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

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

The boundary

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

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

$$Y \cup_g e^k$$

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

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

REFERENCES

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

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

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

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

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

§2. Definitions and Lemmas.

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

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

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

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

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

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

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

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

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

$$\tilde v_p(\tilde w(f)) - \tilde w_p(\tilde v(f)) = [\tilde v,\tilde w]_p(f) = 0$$

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

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

since f has p as a critical point.

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

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

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

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

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

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

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

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

PROOF:

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

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

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

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

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

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

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

then we have

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

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

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

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

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

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

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

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

Therefore, applying 2.1 to the g_j we have

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

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

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

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

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

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

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

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

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

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

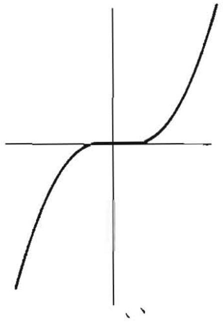
It follows from the inverse function theorem that $v_1,\dots,v_n$ will serve as coordinate functions within some sufficiently small neighborhood U_3 of 0. It is easily verified that f can be expressed as

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

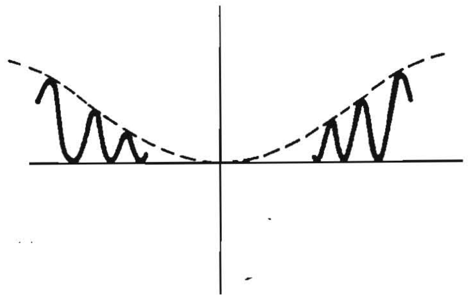
throughout U_3. This completes the induction; and proves Lemma 2.2.

COROLLARY 2.3 Non-degenerate critical points are isolated.

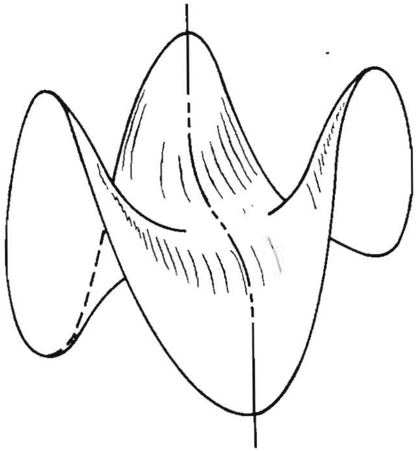
Examples of degenerate critical points (for functions on $\mathbf{R}$ and $\mathbf{R}^2$) are given below, together with pictures of their graphs.



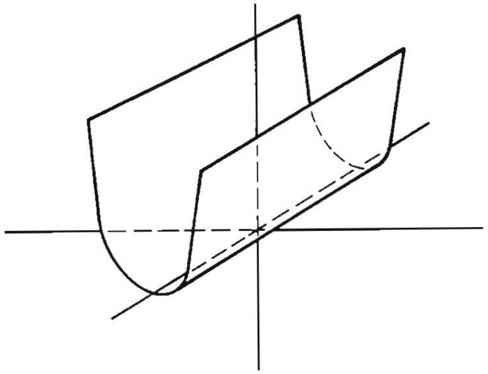
(a) $f(x) = x^3$. The origin is a degenerate critical point.



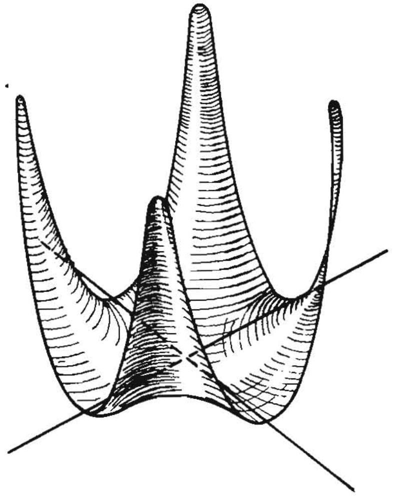
(b) $F(x) = e^{-1/x^2}\sin^2(1/x)$. The origin is a degenerate, and non-isolated, critical point.



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



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



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

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

A <u>1-parameter group of diffeomorphisms</u> of a manifold M is a C^∞ map

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

such that

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

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

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

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

This vector field X is said to generate the group φ.

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

PROOF: Given any smooth curve

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

it is convenient to define the velocity vector

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

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

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

satisfies the differential equation

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

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

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

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

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

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

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

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

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

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

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

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