BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 7, Number 2, September 1982

LECTURES ON MORSE THEORY, OLD AND NEW

BY RAOUL BOTT[1]

Morse Theory is a beautiful and natural extension of the minimum principle for a continuous function on a compact space. In these lectures I would like to discuss it in the context of two problems in analysis which have self-evident geometric interest as well as physical origins.

The first question is simply this. Let M be a compact connected C^∞ manifold endowed with a fixed Riemannian structure. For instance you might think of the two-sphere S^2 with the Riemann structure inherited from an imbedding of S^2 in $\mathbf{R}^3$.

Question. Does such an M always carry a nontrivial closed geodesic?

Recall here first of all that on a compact manifold any two points P and Q can be joined by a geodesic which minimizes the length of all piecewise smooth curves joining P to Q in M. In one way or another this is then an application of the minimum principle, and conceptually you should think of pulling a string confined to M and joining P and Q as tight as possible. When the string has assumed a position in which it cannot be tightened any more, then it describes a geodesic joining P to Q. If it cannot be tightened further even after a "jiggling", then it describes the minimal geodesic in question.

This "pulling tight" principle works also for finding closed geodesics, provided only that we have some constraint to pull against.

Thus if α is a piecewise smooth map of the circle

$$\alpha\colon S^1 \to M$$

which *cannot* be deformed to a point in M, then shortening α in its homotopy class will indeed produce a closed geodesic.

Put differently, let ΛM, denote the space of continuous maps from S^1 to M:

$$\Lambda M = \operatorname{Map}(S^1, M),$$

in the compact open topology.

Also let $\Lambda_* M$ denote the component of the constant maps of S^1 to M. Then a classical theorem going back to Hadamard, Cartan, etc., asserts that

THEOREM. *Every component of ΛM other than $\Lambda_* M$ contains a bona fide closed geodesic.*

Received by the editors November 15, 1981.

1980 *Mathematics Subject Classification*. Primary 58E05.

[1] The following is the text of a series of four lectures which I delivered in Peking during the summer of 1980. However the content is an expanded and somewhat altered version of my lecture at the Poincaré Symposium. Rather than giving a survey I tried in both instances to produce as self-contained an account as possible of two specific problems in the subject.

Let me indicate a proof, once you grant me the following fundamental existence theorem of Riemannian geometry.

LEMMA. *There exists a constant $\varepsilon(M) = \varepsilon > 0$ such that any two points p, q on M with distance $\rho(p, q) < \varepsilon$ are joined by a unique minimizing geodesic segment $s(p, q)$ of length $\rho(p, q)$. Furthermore $s^2(p, q)$ varies smoothly with (p, q) in the region $\rho(p, q) < \varepsilon$ of $M \times M$.*

Armed with this fact, which in turn follows directly from the existence theorems governing elliptic ordinary differential equations, one may argue as follows to establish our theorem.

Let α: $S^1 \to M$, be some point in ΛS, not in the component $\Lambda_* S$. From the continuity of α it follows that we can subdivide the circle S^1 into a finite number of intervals Δ_i, $i = 1, \ldots, n$, such that for $p, q \in \Delta_i$, $\alpha(p)$ and $\alpha(q)$ are within ε of each other. Now let $P_0, P_1, \ldots, P_{n-1}, P_0$, denote the endpoints of the Δ_i, cyclicly arranged on S^1, and let $s(P_0, \ldots, P_{n-1}, P_0)$ be the *geodesic polygon* spanned by the geodesic segments $s(P_i, P_{i+1})$—whose existence follows from our lemma—parametrized proportionally to arc length, and in proportion to the length of Δ_i. Then it should be clear from the picture below that we can deform α in Λ into $s(P_0, \ldots, P_0)$.

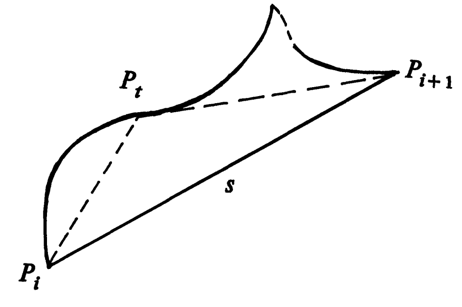


FIGURE 1

Here think of t as a deformation parameter which controls a point P_t on Δ_i moving from P_{i+1} to P_i as t goes from 0 to 1. Now let α_t be the curve which follows α until P_t and then replaces the rest of the curve by $s(P_t, P_{i+1})$.

This is Morse's basic deformation principle and can be used to deform all geodesic problems into finite dimensional ones. In any case at this stage we have seen that:

Each component of ΛM contains a geodesic polygon.

To proceed further choose $0 < \varepsilon < \varepsilon(M)$ and let

$$P_n M \subset M \times M \times \cdots \times M \quad (n \text{ copies})$$

be the subset of n-tuples $(P_1, \ldots, P_n)$ with the property that

$$(1.1) \qquad \rho(P_1, P_2)^2 + \rho(P_2, P_3)^2 + \cdots + \rho(P_n, P_1)^2 \leqslant \varepsilon.$$

Then $P_n M$ is a compact subset of $M^{(n)}$. Further (1.1) implies that each term on the left is $< \varepsilon$, so that every point of $P_n M$ determines a closed n-sided geodesic polygon with vertices at the P_i. If we parametrize the polygons proportionally to arc length, starting at P_1 say, we finally obtain a natural inclusion

$$\iota \colon P_n M \hookrightarrow \Lambda M,$$

which is clearly continuous.

At first sight it might seem that P_n contains only "short" polygons. However observe that by subdividing a polygon, say by introducing new vertices at the midpoints of the edges, the expression on the left of (1.1) is reduced because each term $\rho_i^2 = \rho(P_i, P_{i+1})^2$ is replaced by $(\rho_i/2)^2 + (\rho_i/2)^2 = \rho_i^2/2$. It follows that *any geodesic polygon in ΛM occurs as the image of a point in P_n for n large enough.*

At this stage it is clear that we may confine our search for closed geodesics among the geodesic polygons of P_n in each component of ΛM. For this purpose let

$$E \colon P_n M \to \mathbf{R}$$

be the *energy function*

$$(1.2) \qquad E(P_1,\dots,P_n) = \sum_{\iota=1}^{n} \rho(P_n, P_{i+1})^2; \qquad P_{n+1} \equiv P_1,$$

given by the L.H.S. of (1.1). This energy function is clearly smooth in a vicinity of $P_n \subset M^{(n)}$. Hence E must assume a minimum in each component. Further by increasing n, if necessary, we can arrange it that E takes on this minimum at an interior point, i.e. one with $E < \varepsilon$.

At such a point dE, the differential of E, must therefore vanish. It remains to establish the following assertion: *A critical point of E on $P_n M$ gives rise to a polygon without corners and all of whose edges have equal length. In short, to a closed geodesic.*

This comes about by virtue of the first variation formula for our function ρ^2 in the vicinity of the diagonal in $M \times M$. Indeed in the region $\rho(P, Q)^2 < \varepsilon^2$, one has the following.

LEMMA. (a) *The diagonal $M \subset M \times M$ is a critical submanifold for ρ^2, whose Hessian is nondegenerate in the normal direction to M.*

(b) *At a point (P, Q) of the diagonal in our region, $d\rho^2$ is given by the formula*

$$(1.2) \qquad d\rho^2(Y_P, Y_Q) = \rho\{(X^+, Y_Q) - (X^-, Y_P)\}.$$

Here, X^+, X^- denote the tangents of unit length to $s(P, Q)$ at Q and P respectively, the Y's are tangent vectors at P and Q and $(\ ,\ )$ denotes the inner product.

Summing this expression at the vertices $(P_1,\dots,P_n)$ of a point in $P_n M$ yields

$$(1.3) \qquad dE(Y_1,\dots,Y_n) = \sum_{2}^{n} (Y_i, |S_{i-1}| X_{i-1}^{+} - |S_i| X_i^{-}),$$

where the index $n + 1$ is again to be taken as equal to 1.

At a critical point, therefore, we must have

$$(1.4) \qquad |S_{i-1}|X_{i-1}^{+} = |S_i|X_i^{-}, \qquad i = 2,\ldots,n+1,$$

which precisely expresses the no corner, equal length condition. Q.E.D.

This completely elementary argument therefore establishes the classical Theorem I. An analogous argument could be used to prove the existence of a minimizing geodesic joining two points on M, or the existence of a geodesic joining two submanifolds N_1 and N_2 in M with minimal length.

But consider now the case of a compact simply connected manifold M, for example S^2. Then ΛM has only one component on which the minimum principle only yields the trivial "point paths" of ΛM.

Note by the way if $e\colon \Lambda M \to M$ denotes the evaluation map $\alpha \mapsto \alpha(0)$ then these point paths furnish us with section $\eta\colon M \to \Lambda M$ to e. Technically *e is a fibration in the sense of Serre, with fiber the space of loops ΩM*, that is, the subspace of ΛM consisting of maps α with $\alpha(0)$ some fixed point p of M.

From these two remarks it follows by quite elementary homotopy theory, and Serre's form of the Hurewicz theorem, that the *homotopy groups of ΛM cannot all be trivial.* Indeed, from the homotopy exact sequence of a fibering and the existence of a section to e, it follows that

$$(1.5) \qquad \pi_q(\Lambda M) = \pi_q(M) \oplus \pi_q(\Omega M).$$

Next, from the near tautologous isomorphism $\pi_{q+1}(M) \simeq \pi_q(\Omega M)$, $q \geqslant 1$, it follows that

$$(1.6) \qquad \pi_q(\Lambda M) = \pi_q(M) \oplus \pi_{q+1}(M).$$

Finally the $\pi_q(M)$ cannot all be trivial, by Serre's Hurewicz theorem and Poincaré duality. Q.E.D.

At this stage it suggests itself that one should be able to use the fact that $\pi_q(\Lambda M) \neq 0$ for some q, as a constraint against which one could again minimize and so produce a new extremum. This plan can indeed be carried out and the guiding principle for it was formulated already by G. B. Birkhoff before 1920. It is known as his minimax principle.

To illustrate its application in our present context, let us first simplify matters by once again replacing ΛM by $P_n M$ for n large enough. Indeed the same retraction described earlier, but now done with a compact set of parameters, easily leads to the following [see **[B1]** for details].

LEMMA. *For any fixed q, there exists an n_q such that*

$$(1.7) \qquad \pi_k(P_n M) \simeq \pi_k(\Lambda M) \quad \textit{for all } k \leqslant q \textit{ and } n \geqslant n_q.$$

In short, the P_n approximate ΛM arbitrarily well in homotopy, and therefore in homology as well.

To prove the existence of a classical geodesic in ΛM we now argue as follows. Let $\xi \in \pi_q(M)$ be a nontrivial element of *lowest dimension*. Then according to (1.6), ξ gives rise to a nontrivial element $T\xi$ in $\pi_{q-1}(\Lambda M)$.

Next choose $n > n_q$, so that $P = P_n M$ approximates ΛM to dimension q. Then $T\xi \in \pi_{q-1}(P)$ is also nontrivial.

On P we now again consider our energy function E, whose critical points yield closed geodesics. Hence we will be done once we find a *critical point of E on P other than a point path*, i.e. one with $E > 0$. These point paths of course constitute a submanifold $M \subset P$, on which the energy function assumes an absolute minimum. Assume then—we are out to find a contradiction—that E has no other critical points on P. Then the negative gradient of E, that is, the vector field X on P, defined by the formula

$$-(X, Y) = dE(Y), \tag{1.8}$$

is nonvanishing on $P - M$, and always points downwards. Hence following the flow generated by X will eventually deform P into a tubular neighborhood of M, which in turn can be retracted to M. It follows that under our assumption *all homotopy elements of P come from M*. But this is manifestly not the case for $T\xi$. Indeed, by construction $T\xi \in \pi_q(P) \neq 0$ while $\pi_q(M) \equiv 0$. Q.E.D.

This argument therefore establishes the beautiful theorem of Lyusternik and Fet **[L-F]**:

THEOREM. *Let M be compact and simply connected. Then M carries at least one closed geodesic.*

Let me now explain how this argument is related to the "minimax principle". For that purpose consider the set of maps η: $S_q \to P$ representing $T\xi$, and try to push η as far down, relative to E, as possible. In short consider the real number

$$\kappa = \inf_{\eta} \operatorname{Max}(E, \eta), \qquad [\eta] \in T\xi. \tag{1.9}$$

As we just saw $\kappa > 0$. The minimax principle simply asserts, *that this κ must be a critical value of E*. The proof is again a quite elementary consequence of pushing down in the direction of steepest descent—i.e. along the negative gradient—and I think of it usually as a corollary of what one might call the *first theorem of Morse Theory*. To formulate it and to deduce the minimax principle from it, let us abstract the situation though, so that from now on in this lecture, P will just denote some arbitrary smooth manifold, and E a smooth function on P, whose "half-spaces" $P_a = \{p \in P \mid E(p) < a\}$ *however are assumed to be compact.*

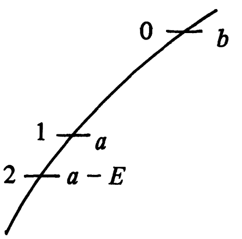
This understood let $a \leqslant b$ be real numbers and consider the inclusion of half-spaces $P_a \subset P_b$.

THEOREM A. *If there is no critical point of E in the region $a \leqslant E \leqslant b$, then*

$$P_a \simeq P_b \tag{1.10}$$

in the sense that they are diffeomorphic.

PROOF. Consider a trajectory of our negative gradient as it leaves the set $E = b$ at time 0. At time $(b - a)$ it is intersecting $E = b$ transversally. Hence by compactness all of them intersect $E = a - \varepsilon$ for some fixed $\varepsilon > 0$. Pictorially each trajectory thus has the three singled out points (see diagram) of intersection with these three level surfaces. Now simply deform the interval $[0, 2]$ into $[1, 2]$ by pushing downwards, but all the time keeping some vicinity of 2 pointwise fixed.



Performing this simultaneously for all of these trajectories, yields the desired diffeomorphism. This argument simultaneously shows that

COROLLARY 1. *Under the conditions of Theorem* I *the inclusion* $P_a \hookrightarrow P_b$ *is a homotopy equivalence.*

COROLLARY 2. *The minimax principle is valid.*

PROOF. Suppose η_n is a sequence of maps with Max $E \mid \eta_n$ tending to κ. Then if κ is not critical, pushing down a fixed ε along the trajectories of X produces a new sequence η_n^1 still representing the same element but with Max $E \mid \eta_n \to \kappa - \varepsilon$. Thus κ is not the inf. Q.E.D.

We have carried through this discussion in terms of the homotopy functor, but notice that any homotopy invariant functor would do just as well in both these corollaries. Thus singular theory, or in the equivariant situation, equivariant singular theory, or K theory, etc. could clearly also be used to predict critical points of a function. On the other hand Theorem A furnishes us with no overall estimate of just how *many* critical points to expect, and in my second lecture I will indicate two quite different steps in this direction, one due to Morse and the other due to Lyusternik and Schnirelmann, both these ideas therefore stemming from the 20's.

Finally a word about the course I steered in this lecture. The polygonal approximation principle is Morse's and otherwise I have followed the account given, say in **[K]**, where the reader will also find a very thorough bibliography. My only contribution is the observation that P_n, defined simply as the half-space $E < \varepsilon$ already approximates ΛM. For the explicit homotopy equivalences the reader is referred to **[B1]**—where they are carried out for the fixed endpoint case. But the argument transparently carries over to our situation. Following Palais and Smale, Klingenberg of course carries out everything in the infinite dimensional context of Hilbert-manifolds. That is, his ΛM is defined as the space of H^1-maps of S^1 to M, and the gradient deformations are then carried out directly in this context.

I know of no aspect of the geodesic question where this approach is essential; however it clearly has some aesthetic advantages, and points the way for situations where finite dimensional approximations are not possible—for instance in the Yang-Mills situation, to be discussed in my third lecture.

Lecture 2. In the last lecture we saw how any change in homotopy type of half-spaces $W_a \subset W_b$ of a smooth function f on a compact manifold W predicts a critical point in the range $a < f < b$, and how pushing a nontrivial homotopy and homology class of P_b—which is not in P_a—down, will lead one to a critical value of the function f. This procedure however lacks any quantitative information, for it is quite possible that two, say, nonhomologous classes "get stuck" at the same critical point.

In Morse Theory this lack is redressed in the following manner: First of all one studies what happens for the "generic function" on M and then refers all other cases to the generic one, by some limiting procedure.

Let me now describe this development in some detail.

Consider then a critical point p of f on W, and let $x_1 \cdots x_n$ be local coordinates on W centered at p. The fact that p is a critical point expresses itself in the vanishing of $df = \Sigma(\partial f/\partial x_i)\, dx_i$ at p. That is

$$\frac{\partial f}{\partial x_i}\Big|_p = 0. \tag{2.1}$$

Consider next the Hessian matrix

$$Hf_p = \frac{\partial^2 f}{\partial x^i \partial x^j}\Big|_p. \tag{2.2}$$

This Hessian of course depends on the local coordinates, but the *rank* of Hf and the number of negative *eigenvalues* of Hf is seen to be invariant under coordinate changes. Morse introduces the terms

$$\text{(2.3)} \qquad \textit{nullity} \text{ of } p \text{ (rel } f) = \dim W - \operatorname{rank} Hf\big|_p,$$

$$\textit{index} \text{ of } p \text{ (rel } f) = \text{number of negative eigenvalues of } Hf\big|_p$$

and calls a function f *nondegenerate* if all its critical points have nullity 0.

These are the generic functions in the sense that *in the vicinity of every function* one may find a generic one. In any case, for a generic f Morse introduces the quantity

$$\mathfrak{M}_t(f) = \sum_p t^{\lambda(p)}, \qquad p \in C(f), \tag{2.4}$$

where the sum is extended over the critical points $C(f)$ of f, and $\lambda(p) =$ index of p relative to f.

This sum turns out to be *finite* because nondegeneracy easily implies the discreteness of the critical points and W was assumed compact. This polynomial, which I will call the *Morse polynomial* (or *series*) of f is then Morse's *quantitative* measure of the critical behavior of f, and he shows that the homology of W sets a definite *lower bound* on it. Precisely let

$$P_t(W) = \sum t^k \dim H_k(W; K) \tag{2.5}$$

be the Poincaré series of W with homology taken relative to some fixed coefficient field K. Then the following inequalities hold.

Morse inequalities. For every nondegenerate f there exists a polynomial $Q_t(f) = q_0 + q_1 t + \cdots$ with *nonnegative coefficients* such that

$$\mathfrak{M}_t(f) - P_t(W) = (1+t)Q_t(f). \tag{2.6}$$

We often write $\mathfrak{M}_t(f) \geqslant P_t(W)$ for (2.6). Clearly this inequality implies that $\mathfrak{M}_t(f)$ majorizes $P_t(W)$ *coefficient by coefficient*. Thus (2.6) predicts at *least* $P_1(M)$ critical points for *any nondegenerate* f on M. However (2.6) is much stronger than this estimate, namely the $(1+t)$ factor on the right implies a feedback relationship between critical points of various indices. The power of this feedback is maybe best illustrated by the following corollary of the Morse inequalities.

MORSE'S LACUNARY PRINCIPLE. *Suppose that no consecutive powers of t occur in* $\mathfrak{M}_t(f)$. *Then* $Q_t(f) \equiv 0$ *so that*

$$\mathfrak{M}_t(f) = P_t(W) \tag{2.7}$$

for every coefficient field K. In particular, W is then free of torsion.

PROOF. The first nonvanishing power of t on the left of (2.6) clearly implies that the next power also occurs on the right and hence by (2.6) must also occur on the left in $\mathfrak{M}_t(f)$. Q.E.D.

The power of this principle is that it sometimes allows one to compute the complete additive homology structure of W, from purely local computations near the critical points of f.

A favorite example of mine is the following. Consider the unit sphere S^{2n+1}

$$\sum_0^n |z_i|^2 = 1, \qquad i = 0,\ldots,n,$$

in $\mathbf{C}^{n+1}$, and on it the function $\varphi(z) = \Sigma_0^n \lambda_i |z_i|^2$ where $\lambda_0 < \lambda_1 < \cdots < \lambda_n$ are a sequence of distinct real numbers. It is clear that φ is invariant under the action

$$e^{i\theta}\colon (z_0,\ldots,z_n) \to (e^{i\theta}z_0,\ldots,e^{i\theta}z_n)$$

of S^1 on S^{2n+1}, and hence descends to CP_n the projective space. Now, by the principle of Lagrange multipliers, if you wish, the extrema of φ correspond to the coordinate axes, and the eigenvalues of the Hessian of φ along the ith-axis are easily seen to be the set

$$\lambda_0 - \lambda_i, \lambda_1 - \lambda_i, \ldots, \lambda_n - \lambda_i$$

with λ_i excluded. Over the reals their multiplicity is 2, and so the index of the ith critical point is $2i$. Thus

$$\mathfrak{M}_t(\varphi) = 1 + t^2 + \cdots + t^{2n}. \tag{2.1}$$

The lacunary principle applies and we conclude that $P_t(\mathbf{C}P_n) = 1 + t^2 + \cdots + t^{2n}$. Q.E.D.

There are two rather different approaches to proving the Morse inequalities, and as both are very instructive, I will say a few words about each of them.

The level surface method 1. Consider a nondegenerate critical point p of our nondegenerate f on W, and assume that it is the only critical point at its level. The "Morse Lemma" now asserts that there is a coordinate system $x_1, \dots, x_n$ about p such that near p

$$f = f(p) - x_1^2 - x_2^2 - \cdots - x_\lambda^2 + x_{\lambda+1}^2 + \cdots + x_n^2$$

with $\lambda = \lambda_p$ the index of p rel f.

Using this explicit description of f near p one proves what I call Theorem B of the Morse theory.

THEOREM B. *Let $a < f(p) < b$ be such that f has no critical points in the range $a < f < b$ other than p.*

Then the diffeomorphism type of M_b differs from that of M_a by the attachment of a thickened λ-cell

$$M_b \simeq M_b \bigcup_{\alpha} e_\lambda \times e_{n-\lambda}, \tag{2.2}$$

and the homotopy type of M_b is therefore that of M_a with λ-cell attached

$$M_b \approx M_a \cup e_\lambda. \tag{2.3}$$

PROOF BY PICTURE. Consider the case of a critical point p of index 1 on a surface so that near p, f can be taken to be $f = -x^2 + y^2$. Then near p the level surfaces of f take the form:

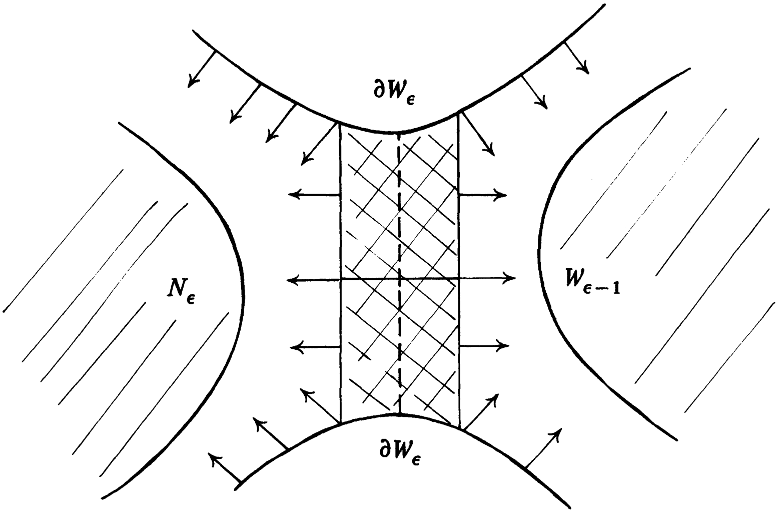


FIGURE 2

Hence if one deletes the cross-hatched region Y from W_ε, then following the gradient lines will deform $W_\varepsilon - Y$ to $W_{-\varepsilon}$. But Y is simply a "thickened 1-cell". The thickening direction is here the y-direction, and the homotopically essential part of Y is already the 1-cell given by its intersection with the X-axis: Thus the homotopy type of W_ε is also described by $(W_\varepsilon - Y) \cup e_1$ as indicated below.

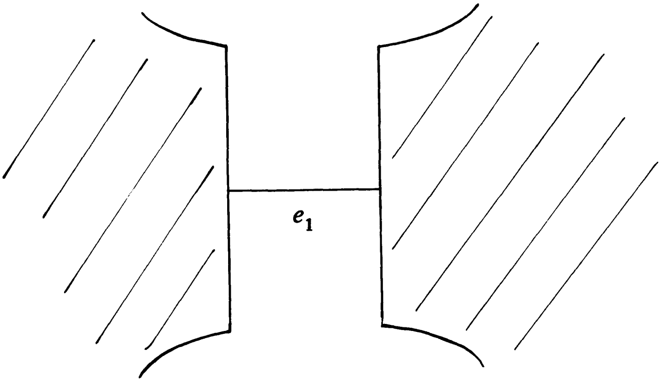


FIGURE 3

So much for a pictorial explanation of Theorem B. Finally a short explanation of how the Morse inequalities follow from Theorems A and B. Consider then the step from $W_{-\varepsilon}$ to W_ε with p the only critical point of f in the range $-\varepsilon < f < \varepsilon$. Assume also that p is nondegenerate of index λ. Let $\mathfrak{M}_t^a(f) = \Sigma_p t^{\lambda_p}$, with $p \in W_a$, $p \in C(f)$, be the Morse polynomial of a half-space W_a, and let P_t be the corresponding Poincaré Polynomial of W_a

$$P_t(W_a) = \sum t^k \dim H_k(W_a).$$

We will actually refine our earlier formulation of the inequalities to the statement that $\mathfrak{M}_t^a(f) \geqslant P_t(W^a)$ for each regular value a, and then proceed from $W_{-\varepsilon}$ to $W_{+\varepsilon}$ by induction. Now the change in $\mathfrak{M}_t$ from $-\varepsilon$ to ε is clearly t^λ.

On the other hand the change from ΔP_t from $P_t(W^{-\varepsilon})$ to $P_t(W^{+\varepsilon})$ can be two fold. Either,

(1) $\Delta P_t = t^\lambda$ or,

(2) $\Delta P_t = -t^{\lambda-1}$.

Once this is granted the inequalities at ε follow from these at $-\varepsilon$. Indeed,

$$\Delta(\mathfrak{M}_t - P_t) = 0 \quad \text{or} \quad t^{\lambda-1}(1+t)$$

depending on the two cases. In either case the Q term of the inequality is augmented by a polynomial with nonnegative coefficients. Q.E.D.

The crucial step is therefore the alternative for ΔP_t above, and this is a standard result in homology theory. It is also a very intuitive one. Consider the boundary ∂e_λ of the attaching cell. It is a $\lambda - 1$ sphere $S_p^{\lambda-1}$ in $W_{-\varepsilon}$. The cycle carried by this sphere either bounds a chain in $W_{-\varepsilon}$ or not. In the first case we cap the chain bounded by $S_p^{\lambda-1}$ with e_λ to create a new nontrivial homology class in W_ε. This corresponds to the alternative: $\Delta P_t = t^\lambda$. In the second case e_λ manifestly has as boundary the nontrivial cycle $S^{\lambda-1}$ in $W_{-\varepsilon}$. Hence in this situation P_t decreases by a $t^{\lambda-1}$. Q.E.D.

Note that to ascertain *which* alternative is valid involves a *global analysis of* $W_{-\varepsilon}$. Note also that if for all critical points the first alternative holds, i.e. $\Delta P_t = t^\lambda$, then $\mathfrak{M}_t(f) = P_t(W)$, that is f has precisely the minimal number of critical points which the topology permits. We often refer to a critical point with $\Delta P_t = t^\lambda$ as *completable*, and to a function for which $\mathfrak{M}_t(f) = P_t(W)$ as a *perfect Morse function* on W.

Note also that a given f can be perfect for one coefficient field and not perfect for another.

Let me conclude this line of proof with the statement of the Morse inequalities in relative form. The proof is the same.

The relative inequalities. Let f be nondegenerate and let $a < b$ be two regular values of f. Then if

$$\mathfrak{M}_t(f)_a^b = \sum t^{\lambda_p}, \qquad p \in C(f) \cap (W_a - W_b),$$

and

$$P_t(W^b, W^a) = \sum \dim H_k(W^b, W^a) t^k$$

the Morse inequalities still hold, that is, $\mathfrak{M}_t(f)_a^b \geqslant P_t(W^b, W^a)$.

The dynamical systems method. Here we pass from f to its gradient $X = \overrightarrow{df}$, relative to some Riemann structure on W, and the action of $\mathbf{R}^1$ on W induced by flowing down, that is along $-X$.

Consider a point $p \in W - C(f)$. Its orbit under this action imbeds $\mathbf{R}^1$ in M, and the closure of the orbit is a segment joining some critical point p to a lower one g. Now, again using the Morse lemma, say, one sees that the trajectories of $\mathbf{R}^1$ which "start" at a fixed critical point p, constitute a cell, W_p, of dimension λ_1, while those which *end* at p constitute a cell W_p^*, of *codimension* λ_p. Furthermore these two cells intersect transversally at p. Indeed their tangent planes at p are precisely the directories of *steepest descent* and *ascent* respectively. In this way, then, one obtains two "stratifications" of W into the "stable" and "unstable" cells;

$$(2.4) \qquad \begin{aligned} W &= \coprod W_p, \quad \dim W_p = \lambda_p, p \in C(f), \\ W &= \coprod W_p^*, \quad \dim W_p^* = \dim W - \lambda_p, \end{aligned}$$

each indexed by the critical points $C(f)$ of f.

Although the closures of these cells can be badly behaved, this decomposition still has sufficient properties to enable one to deduce the Morse inequalities, and I indicate Smale's argument in this direction. By deforming the gradient, X, of f a trifle, if necessary, in the class of vector fields for which

$$Xf \geqslant 0 \quad \text{and} \quad Xf_p > 0 \quad \text{if } p \text{ is not critical,}$$

he shows that these two cell decompositions induced by X will be brought into normal form, in the sense that any two cells W_i and W_j^* will intersect normally at any $p \in W_i \cap W_j^*$. That is, at such a p

$$\dim W_i + \dim W_j^* - \dim W_i \cap W_j^* = n.$$

It follows immediately that if a trajectory starts at p and ends at q then $\dim W_p > \dim W_q$. Indeed, the interior of this trajectory must be in $W_p \cap W_q^*$ and $\dim W_p \cap W_q^* \geqslant 1$. Hence the above formula reads

$$\dim W_p + (n - \dim W_q) \geqslant n + 1 \Rightarrow \dim W_p \geqslant \dim W_q + 1. \quad \text{Q.E.D.}$$

It follows that if K_p is the union of all the cells of $\dim \leqslant p$, then the $\cdots K_p \subset K_{p+1} \subset \cdots$ defines a finite filtration of W by closed sets, and

$$K^p - K^{p-1} = \coprod W_p, \qquad \dim W_p = p,$$

is the union of disjoint open sets.

Since for Cech theory

$$H^*(K^p, K^{p-1}) = \sum H_C^*(K^p - K^{p-1}),$$

where C denotes compact carriers, this implies that,

$$\dim H^q(K^p, K^{p-1}) = \begin{cases} \text{number of } p \text{ cells in 2.3 if } q = p, \\ 0 \quad \text{otherwise.} \end{cases}$$

Now the inequalities follow by quite standard arguments.

This method is less elementary than the one outlined before, but has several advantages. First of all it points the way to extending the Morse inequalities to arbitrary flows, i.e. vector fields X, satisfying certain generic conditions. This led Smale to the so-called Morse-Smale flows and diffeomorphisms (see [SS]).

Secondly a slight modification of our discussion leads us naturally to the Lyusternik-Schnirelmann estimate on the *number* of critical points of a function f on W. Indeed suppose now that p is an arbitrary isolated critical point of f. We still have the set W_p of trajectories of X leaving p, and the corresponding partition (2.3). What we do not know any more is whether W_p is a cell or not. *However W_p will still be contractible* to p, and we can furthermore "thicken" W_p a little so as to preserve this property. It follows that under our assumption on f, W admits a cover $\{\tilde{W}_p\}$ indexed by $C(f)$, by *open contractible* sets. By the very definition of the concept of category, of a space, this implies that

$$\operatorname{Cat}(W) \leqslant \textit{number of critical points of } f.$$

From this in turn we have the cohomological criterion: *Suppose* $\omega_1, \ldots, \omega_m$ *are cohomology classes on W with*

$$\omega_1 \wedge \cdots \wedge \omega_m \neq 0; \qquad \dim \omega_i > 0.$$

Then any function with isolated critical points on W must have at least $(m + 1)$ critical points.

PROOF. If it had fewer, we say only m of them, we could cover W by open sets $\{\tilde{W}_i\}$, $i = 1, \ldots, m$. From the contractability of these and $\dim \omega_i > 0$ it follows that we can choose representations of our ω_i which come from $\tilde{\omega}_i \in H^*(W, \tilde{W}_i)$. But then their product comes from $H^*(W; \ \bigcup_1^m \tilde{W}_i) = 0$. Q.E.D.

Finally let me remark that this principle of Lyusternik-Schnirelmann dually leads to the following refinement of the minimax principle. Suppose again that f has isolated critical points only, and now let z_1 and z_2 be two homology classes. In pushing them down, rel f, they each must get stuck at a critical level say K_1 and K_2, but these might well be equal. The *Lyusternik-Schnirelmann principle* however asserts.

Suppose $z_1 = z_2 \cap \omega$ where ω is a cohomology class of dim > 0 and $\cap$ denotes the Cap product. *Then under our assumptions* $K_1 < K_2$.

One calls the relation $z_1 = z_2 \cap W$, among homology classes *subordination*: z_1 is subordinated to z_2, written $z_1 < z_2$. The L.-S. principle clearly implies that the number of critical points of a smooth function f on W is bounded by 1 plus the cardinality of the longest chain of subordinated classes $z_1 < z_2 < \cdots < z_m$ on W.

This corollary is of course the same as our previous estimate as follows from Poincaré duality.

To sum up, I hope we have learned the following, which in some sense comprises the *Elementary aspects* of critical point theory:

(1) A nondegenerate smooth function on a compact manifold has at least $P_1(W; K) = \Sigma_q \dim H^q(W; K)$ critical points.

(2) A smooth function on W has at least Cat(W) critical points.

(3) If all the critical points of a nondegenerate function are completable, relative to the coefficient field K then f is K-perfect, that is, $\mathfrak{M}_t(f) = P_t(M; K)$, and

(4) if $\mathfrak{M}_t(f)$ is lacunary, then $\mathfrak{M}_t(f) = P_t(M; K)$ for all K.

(5) The gradient flow gives rise to a stratification M, from which one can also deduce the Morse inequalities.

The dates of these various concepts are roughly these: Morse inequalities—1924, Lyusternik-Schnirelmann relations 1929.

The stratification goes back to Thom 1949, Theorem B as stated is due to Smale in the late 50's and in its homotopy version appears in the early 50's in papers of Thom, Pitcher and myself.

Lecture 3. In the last lecture we saw how in the Morse theory a nondegenerate critical point, p, is counted by t^λ, λ being its index. From Theorem B it then also follows that this t^λ is the Poincaré series of the *local relative* homology

$$H^*(W_c \cap U_p, W_c^- \cap U_p)$$

where W_c^- is the open half-space $f < f(p)$ and U_p is an open neighborhood of p. Such a local Poincaré series, can then be used to define a Morse series for quite general critical sets, and in some sense the Morse inequalities will then still be valid.

On the other hand the actual evaluation of the local contribution becomes quite difficult. There is however one extension of nondegeneracy where the local computation carries over practically word for word from our earlier one, and our first aim in this lecture will be to describe this extension.

Accordingly we define a connected submanifold $N \subset W$ to be a *nondegenerate critical manifold of* W if the following conditions are satisfied.

(3.1) *Each point $p \in N$ is a critical point of f.*

(3.2) *The Hessian of f is nondegenerate in the normal direction to N.*

Spelled out this last condition takes this form: Let $p \in N$ and let $(x_1,\ldots,x_k, x_{k+1},\ldots,x_n)$ be a system of local coordinates in W centered at p, such that near p, N is given by the $n-k$ equations N: $x_{k+1} = 0,\ldots,x_n = 0_x$. Then,

$$\det\left(\frac{\partial^2 f}{\partial x^i \partial x^j}\right)\bigg|_p \neq 0 \quad \text{for } i, j = k+1,\ldots,n. \tag{3.3}$$

Alternatively, consider a small tubular ε-neighborhood $W_\varepsilon(N)$ of N, which is fibered over N by the normal discs swept out by geodesics of length $\leqslant \varepsilon$ in the normal direction to N, relative to some Riemann structure on W.

Then (3.3) is equivalent to the assumption: *f restricted to each normal disc is nondegenerate.*

Note that this structure automatically gives us a way of decomposing the normal bundle νN into a *positive* and *negative* part

$$\nu N = \nu^+ N \oplus \nu^- N, \tag{3.4}$$

where $\nu_p^+ N$ and $\nu_p^- N$ are respectively spanned by the *positive and negative Eigen-directions of the Hessian of f.*

The fiber dimension of $\nu^- N$ will be denoted by λ^N and referred to as the index of $N \operatorname{rel} f$. Finally if θ^- denotes orientation bundle of $\nu^- N$, we assert that the proper way to "count" N in the Morse theory is by the polynomial $t^N P_t(N; \theta^-)$.

Precisely, one has the following extension of the Morse inequalities. Suppose f is nondegenerate in the *extended sense that all its critical sets are nondegenerate critical manifolds. Assume also that W is compact. Then if we define the Morse series of f relative to a coefficient field K by*

$$\mathfrak{M}_t(f) = \sum t^{\lambda_N} P_t(N; \theta^- \otimes K), \qquad N \subset C(f), \tag{3.5}$$

the Morse inequalities hold:

$$\mathfrak{M}_t(f) \geqslant P_t(W; K). \tag{3.6}$$

I have time for only two observations to explain this extension. In the context of Theorem B, what happens now, is that as we pass a critical level, we attach *a thickened version of the negative disc-bundle over M to W_a to obtain W_b.*

On the other hand in the context of the flow engendered by the gradient, the cells of the stable and unstable stratifications:

$$W = \bigcup_N W_N, \qquad W^* = \bigcup_N W_N^*, \qquad N \subset C(f),$$

are now replaced by the negative and positive bundles of N:

$$W_N^- = \nu^- N, \qquad W_N^+ = \nu^+ N.$$

In short, everywhere the cells are replaced by the corresponding cell-bundles, and each W_N is counted by its *cohomology with compact support*. Indeed, the Thom-isomorphism

$$H_C^*(\nu^- N) \simeq H^{*-\lambda_N}(N; \theta^-) \tag{3.7}$$

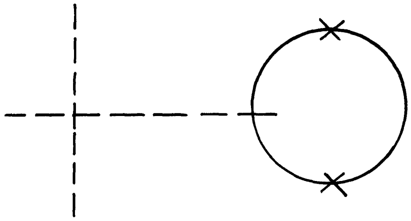
converts this counting procedure into (3.5).

REMARKS. (1) One of the advantages of this extended notion of nondegenerateness is, that the pullback of a function f under a fiber projection $W' \overset{\pi}{\to} W$ *stays nondegenerate*. Note also that under such a pullback a critical manifold N goes over to $\pi^{-1}N$ while its index is preserved:

$$\lambda_N = \lambda_{\pi^{-1}N}. \tag{3.8}$$

On the other hand it is now a trivial statement to say that W admits a perfect nondegenerate Morse function; the constant function is now trivially perfect!

(2) A very simple illustration of this situation is furnished by the function induced by the z-coordinate on the torus obtained by rotating the circle indicated below about the z-axis.



Clearly here the minima and the maxima are nondegenerate circles, with indices 0 and 1. Hence

$$\mathfrak{M}_t(z) = (1 + t) + t(1 + t).$$

Thus z here induces a perfect nondegenerate Morse function.

Actually the example which first motivated me to introduce and exploit this concept, was precisely the closed geodesic problem of my first lecture. For consider the critical point of our energy function E on one of our polygonal space P_n.

Clearly if $(P) = (P_1, \ldots, P_n)$ represents a bona fide closed geodesic, i.e. not a point path, then every *rotation* and *reflection* of the corresponding map

$$\alpha_P \colon S^1 \to M$$

is still in P_n and still critical.

Thus in the closed geodesic problem every bona fide geodesic gives rise to a critical set consisting of *two disjoint circles*. The best one can hope for, then, is that E be nondegenerate in our extended sense. Generically this also turns out to be true, as follows already from general position arguments in Morse's work. Precisely, the following holds.

PROPOSITION. *For a generic Riemann structure on M, and a generic $0 < \varepsilon < \varepsilon(M)$, the energy functional E is nondegenerate on all the spaces $P_n \subset \Lambda M$.*

Actually E is nondegenerate also in highly symmetric situations—such as for instance the spheres in their usual metric. In fact we have the following theorem.

THEOREM. *Let S^n, $n \geqslant 2$, be the standard n-sphere. Then the bona fide closed geodesics are represented by nondegenerate critical manifolds in any P_m in which they occur (ε generic).*

Furthermore these manifolds are of two types:

$$N_0 = S^n \text{ corresponding to the point paths},$$

and

$$N_k = T_1S^n \text{ the unit tangent bundle to } S^n,$$

corresponding to the great circle starting at a unit vector $p \in T_1S^n$ and traversing that circle k times.

Finally, their indices are given by

$$\operatorname{index} N_k = (2k - 1)(n - 1).$$

COROLLARY. *The* mod 2 *Morse series of E on P_m is given by*

$$\mathfrak{M}_t(E; P_m) = (1 + t^n) + \sum_{k=1}^{p} (1 + t^n)(1 + t^{n-1})t^{(2k-1)(n-1)} \tag{3.9}$$

where $p \to \infty$ as $m \to \infty$.

Actually the techniques of **[S-B]** show that all these critical manifolds are also of "completable type" and hence that (3.9) also computes the Poincaré polynomial of P_m. Hence finally in view of Proposition (2.6) we obtain the

COROLLARY. *The* mod 2 *Poincaré series of ΛS^n is given by*

$$P_t(\Lambda S^n) = (1 + t^n) + \frac{t^{n-1}}{1 - t^{2(n-1)}} \cdot (1 + t^n)(1 + t^{n-1}). \tag{3.10}$$

Indeed here we have just let $p \to \infty$.

REMARKS. (1) Actually one can show more, the function E is *perfect* for *all coefficient systems*. For n even ΛS^n therefore inherits the 2-torsion of T_1S^n.

(2) Nowadays the formula (3.10) can also be easily computed from the fibering $\Lambda \xrightarrow{e} S^n$ discussed earlier, however the derivation sketched above moves easily within the Morse theory, and would have been quite accessible to Morse in the 1930's.

(3) Finally let me discuss a consequence of (3.9). Assume then that $M = S^n$ in some Riemannian structure which is generic in the sense of our earlier

proposition. Then the Morse inequalities clearly imply an infinite number of bona fide closed geodesics. Indeed each such contributes $2(1 + t)t^\lambda$ and so we need an infinite number of them to dominate $P_t(\Lambda S^n)$.

Unfortunately one has to admit that this result is not very satisfying, because of the possibility of *iterating* a closed geodesic.

Thus if α: $S^1 \to M$ is bona fide closed geodesic then the composition

$$S^1 \xrightarrow{[n]} S^1 \xrightarrow{\alpha} M$$

where $[n]$ denotes the n-fold covering $z \to z^n$, $S^1 = \{|z| = 1\}$, is a new one, called the nth iterate of α and denoted by α^n. Geometrically the iterates of course seem redundant, however they must be counted in the Morse calculus. In short if one defines α to be a *prime* closed geodesic if it is not the iterate of any other closed geodesic, then the natural and as it seems, very difficult question is to estimate the number of these prime geodesics on M. For some reason the large literature on this subject teems with mistakes. Morse in his attack introduced the circular connectivities, to be discussed later, and made a mistake which I myself followed later on.

The later results of Alber, for instance as explained in Klingenberg's book **[K]**, are false. The ambitious main result of that book to the effect that there are infinitely many closed prime geodesics on any compact manifold is now also conceded to have gaps in its proof, etc., etc.

Let me conclude this lecture therefore by discussing the most modest but hopefully correct implication of the Morse inequalities to this question.

A first step in this direction is clearly to find some estimate of the behavior of the index $\lambda(\alpha^n)$ under iteration. In 1956 I investigated this question in **[B1]** and came up with the following really quite elementary estimate. To formulate it we must first recall the "Index Theorem of Morse" in the present context.

Suppose then that α is a bona fide closed geodesic, and think of α as a periodic map

$$\alpha\colon \mathbf{R} \to M, \qquad \alpha(x + 1) = \alpha.$$

The normal bundle to α is then a vector bundle ν over $\mathbf{R}$, and the Jacobi equations of α are naturally to be considered as a second order differential equation L on the space of smooth sections of ν. Explicitly L takes the form

$$LY \equiv -\nabla_X^2 Y + R(X, Y)X, \tag{3.11}$$

where X is the tangent field along α, Y a normal field, that is $Y \in \Gamma(\nu_\alpha)$ and R is the curvature tensor.

In any case (3.11) is an elliptic ordinary differential equation, which relative to a parallel frame along α takes the form

$$L(\eta) \equiv -\frac{d^2\eta_i}{dt^2} + k_{ij}(t)\eta_j, \qquad i, j = 1, \ldots, (n-1),$$

with $k_{ij}(t)$ a periodic function of t.

For instance on the sphere the matrix $\|k_{ij}\|$ reduces to $k \cdot$ (identity where $1/k$ is the constant curvature of S^n).

Consider now the eigenvalue problem

$$LY = \lambda Y \tag{3.12}$$

subject to the *periodic boundary condition*

$$Y(t+1) = Y(t). \tag{3.13}$$

This is then a well-posed Sturm-Liouville type problem, which in view of the positivity of the operator d^2/dt^2, has only a finite number of algebraic eigenvalues.

Morse defines the index and nullity of α by the formula

$$\begin{aligned} &\text{index } \alpha = \text{number of negative eigenvalues of (3.9) subject to (3.10)} \\ &\text{nullity } \alpha = \text{multiplicity of 0 as an eigenvalue in (3.9).} \end{aligned} \tag{3.14}$$

This definition is pertinent for our purposes, as it is not difficult to show that (see for instance [**B1**])

PROPOSITION. *If $p \in P_nM$ is any polygonal representative of the closed geodesic α, then the index and nullity of p* rel *E are given by* (3.14).

REMARK. The definitions (3.11) are the natural candidates for—and turn out to be—the nullity and index of the energy function in the Hilbert-manifold approach to ΛM. Hence our proposition just expresses the compatability of the finite dimensional approximations to the Hilbert-manifold approach.

Using this equivalence, one now proceeds as follows. First of all, let

$$J_\alpha = \text{Space of solutions of } LY = 0. \tag{3.15}$$

This is then the $2(n-1)$-dimensional vector space of *Jacobi-fields* along α.

The periodicity of the equation $LY = 0$ now implies that the transformation $t \to t+1$ induces a linear map $T_\alpha\colon J_\alpha \to J_\alpha$ called the *linear Poincaré map* of α. This transformation plays a fundamental role in all questions concerning α, and also enters our question concerning the index of the iterates of α.

Indeed as is shown in [**B3**], (3.15) determines a *nonnegative integer valued function* Λ on the circle $|z| = 1$, which *jumps only at the points* $\{\xi_i\}$ *of* $|z| = 1$ *in the spectrum of T_α—and the jumps of Λ are there bounded by the multiplicity of ξ_i—such that*

$$\lambda(\alpha^n) = \sum \Lambda(\omega), \tag{3.16}$$

where the ω ranges over the nth roots of $+1$ *or* -1 *depending on whether α is orientable or not.*

Here of course α is called orientable if the parallel transport of a frame along α preserves the orientation.

Immediate consequences of (3.16) are:

If the spectrum of T_α is off the unit circle then

$$\lambda(\alpha^n) = n\lambda(\alpha). \tag{3.17}$$

If the spectrum intersects the unit circle at the points $e^{2\pi i \rho_k}$, with ρ_k irrational, then

$$\lambda(\alpha^n) = a_{-1} + a_0 n + \sum a_i [n\rho_k] \tag{3.18}$$

where the a_i are integers and [ ] denotes the greatest integer less than.

In any case we have

$$\lambda^*(\alpha) = \frac{\lim(\alpha^n)}{n} = \frac{1}{2\pi}\int_0^{2\pi} \Lambda(\omega)\, d\omega = a_0 + \sum a_k \rho_k \tag{3.19}$$

with the $\{\rho_k\}$ as above.

A less immediate corollary, due to Ziller, is that

$$na - b \leqslant \lambda(\alpha^n) \leqslant na + b \tag{3.20}$$

where a and b are numbers determined by the intervals into which the spectrum of T_α partitions the unit circle.

These matters are discussed in great detail in Klingenberg's book, where he also brings independent proofs and relates these questions with the Maslov cycle, etc. In that context T_α of course arises as the interesting part of the differential of the map

$$G_\alpha \colon T_1 M \to T_1 M,$$

given by applying the geodesic flow with time equal to the length of α. Then every point p on α is clearly a fixed point of G_α whose differential has normal and tangential components given by

$$dG_{\alpha,p} = T_\alpha \oplus 1. \tag{3.21}$$

In this connection let me mention the classification of α as hyperbolic, elliptic, and parabolic, according to whether the spectrum of T_α, is (i) off the unit circle, (ii) on the unit circle but not at ±1, with dG_α equivalent to a rotation, or (iii) concentrated at ±1.

Returning to our main concern let us try and estimate the number of prime geodesics in the nondegenerate case. The following now follows trivially.

THEOREM. *Let M be a compact simply connected manifold and assume that the Betti numbers* dim $H_k(\Lambda M;\, K)$ *tend to* ∞ *with k, for some coefficient field K. Then in any generic Riemann structure, M will have an infinite number of prime geodesics.*

PROOF. Generic implies that all bona fide closed geodesics correspond to two nondegenerate critical circles in any P_n approximating ΛM. It follows that the combination of any α and all its iterates α^n, is given by an expression of the form

$$2(1+t)\sum_{n=1}^{\infty} t^{\lambda(\alpha^n)},$$

and in view of (3.21) the coefficients of this series remain bounded.

Hence if M had only a finite number of prime closed geodesics the Morse series of E on P_n, n large, would also, contradicting the Morse inequalities. Q.E.D.

Note that if one drops the nondegeneracy condition, then the Morse inequalities by themselves will not yield an infinite number of prime geodesics. One needs a bound on the number of nondegenerate critical points into which the iterates of a degenerate one bifurcates. Such a bound was obtained by Gromoll and Meyer, in 1964 **[G-M]** and led them to the beautiful result that Theorem 5 is valid, *for all Riemann structures on M.* I will not have time, unfortunately, to comment on their theorem here if I am to speak about the *equivariant* theory in some detail next time.

Let me therefore stick to the nondegenerate situation and see what, if anything, the Morse inequalities predict for the spheres.

In view of our preceeding computations, we already know that the Morse series

$$\mathfrak{M}_t = \lim_{n\to\infty} \mathfrak{M}_t(E \text{ on } P_n \text{ wth } E > 0)$$

will take the form $\mathfrak{M}_t = 2(1+t)\overline{\mathfrak{M}}_t$ in *any nondegenerate* situation. For instance, with $M = S^{n+1}$, $n \geqslant 2$, the Morse inequalities therefore take the form

$$(3.22)\qquad 2(1+t)\overline{\mathfrak{M}}_t - \frac{t^n}{1-t^{2n}}(1+t^n)(1+t^{n+1}) \geqslant (1+t)Q(t).$$

Unfortunately, already the iterates of a single prime geodesic, could satisfy these inequalities so that (3.22) does not advance this cause at all.

In my next lecture I would like to discuss an equivariant version of the Morse theory and explain its connection, both to this problem and the Yang-Mills theory.

Lecture 4. The equivariant case. The most telling criticism of our results so far is the following.

If one deforms the n-sphere into an ellipsoid

$$(4.1)\qquad \sum_{i=1}^{n+1} a_i^2 x_i^2 = 1$$

with $a_1 < a_2 < \cdots < a_{n+1}$, the first critical manifold, T_1S^n, decomposes into the $n(n+1)/2$ geodesics given by the intersection of the coordinate planes with (4.1). On the other hand T_1S^n contributes

$$t^{n-1}P_t(T_1S^n) = t^{n-1}(1+t^{n-1})(1+t^n)$$

to the Morse series of ΛS^n. Hence under small perturbations T_1S^n should contribute no more than $P_1(T_1S^n) = 4$ critical points.

The correct diagnosis of this ailment is that our energy function has a built-in *symmetry* which has to be taken into account before the proper correspondence between geometry and topology is realized.

The symmetry in question is of course due to the fact, that the energy-integral

$$(4.2)\qquad E = \int_0^1 |\dot{\alpha}|^2\, dt$$

is *invariant under rotations and reflections of* S^1.

Thus if we consider a model for ΛM, on which E is well defined—for instance the Hilbert-manifold of H^1-maps of S^1 to M, or even the simpler model of piecewise smooth maps of S^1 to M, parametrized proportionately to arc length, then E is *not* an *arbitrary* function on ΛM—it is a priori invariant under the natural action of $O(2)$ on ΛM, induced by the action of $O(2)$ on S^1.

Correspondingly in the context of our polygonal approximations $P_n(M)$, E is seen to be a priori invariant under the finite group generated by cyclic permutations and reversal, of the vertices of our polygons.

In both cases one is therefore led to the general question of *how the Morse theory is to be altered to take into account a priori symmetries of a function f under the action of a compact Lie group G on a manifold W.*

There are two cases to be considered.

Case 1. *The action of G on W is free.*

In this case the quotient space W/G is itself a manifold, and the function f naturally descends to a smooth function

$$f/G\colon M/G \to \mathbf{R}. \tag{4.3}$$

It is clear therefore, that in this situation the appropriate theory is simply the old or usual Morse theory of f/G.

Case 2. *The action is not free.*

In this case W/G fails to be a manifold, so that to apply the Morse theory to f/G, one would first of all have to extend it intelligently to nonmanifolds.

Actually this can be done to a certain extent; and this procedure has been the main tool in the past especially for the closed geodesic problem.

I would like to champion a quite different approach here, which is a natural extension of Case 1 from the topologists point of view even though it might seem bizarre from an analysts point of view.

We have already seen how Theorems A and B yield the Morse inequalities via the standard properties of the homology functor. In the present context one should therefore be able to obtain the appropriate extensions of these inequalities via these same theorems, by simply replacing the functor H_*, by the *equivariant homology functor* H_*^G. This program works very nicely and yields the following result.

In accordance with the principle $H_* \mapsto H_*^G$, define the *equivariant Poincaré series* by

$$P_t(W) = \sum t^k \dim H_k^G(W), \tag{4.4}$$

and similarly the *equivariant Morse series*, for a *nondegenerate f*, *by*

$$\mathfrak{M}_t^G(f) = \sum_N t^{\lambda_N} P_t^G(N; \theta^-), \qquad N \in C(f). \tag{4.5}$$

With this understood, we have the following consequences of Theorems A and B.

THEOREM V. *The equivariant Morse inequalities*: *For any nondegenerate G-invariant f on the compact manifold W, the Morse inequalities hold also in the equivariant sense*

$$\mathfrak{M}_t^G(f) - P_t^G(W) = (1 + t)Q_t^G(f). \tag{4.6}$$

To apply this principle, one of course has to learn to compute with the equivariant homology, and I would like to say a few introductory words for the nonspecialists on this score. In homotopy theory it was realized a long time ago, that the quotient construction $W \to W/G$ works properly *only* when the action is free. Now, starting with this premise what is to be done about nonfree actions? Well first of all it is a triviality that if *U is any space on which G acts freely, then the diagonal action of G on $W \times U$ is free*. This suggests that as far as homotopy is concerned, one should find a space U on which (1) G acts freely, and (2) whose homotopy is trivial, (i.e., U is contractible).

Such spaces turned out to *exist*, *be essentially unique*, and play an absolutely essential role in all of modern topology.

In any case granting the existence of such a U for G, the *homotopy quotient W_G of any action* is defined by

$$W_G = U \times W/G, \tag{4.7}$$

where here G of course acts *diagonally* on the product.

Elementary properties of this construction are: *W_G projects naturally on W/G and U/G; and*

(4.8) *If G acts freely on W, then the projection*

$$W_G \to W/G$$

is a homotopy equivalence.

(4.9) *The projection*

$$W_G \to U/G$$

is always a fibering with fiber W.

Note by the way, that in this calculus U/G plays the role of the *homotopy quotient of the trivial action of G on a point*. This space is again of fundamental importance in topology, is usually denoted by BG, and is referred to as the *classifying space of G*. It is a topological space which somehow reflects both the algebraic and the topological properties of G.

In any case, all this granted, the *equivariant version*, F^G, *of any functor F on spaces is now simply defined by*

$$F^G(W) \equiv F(W_G). \tag{4.10}$$

In particular note that

$$H^G_*(\text{point}) = H_*(BG), \tag{4.11}$$

so that *in the equivariant Morse series, a nondegenerate critical point, contributes the expression*

$$P^G_t(\text{point}) = P_t(BG). \tag{4.12}$$

More generally one proves that: *If N is a nondegenerate critical manifold consisting of a single orbit $N = G/H$ then*

$$P^G_t(G/H) = P^H_t(H) = P_t(BH). \tag{4.13}$$

In short these classifying spaces and their *ordinary homology*, play an essential role in the equivariant theory.

Below I list a few examples, which are of course absolutely standard in modern topology.

Examples

	G	U	$U/G = BG$	$P_t(BG)$
	$\mathbf{Z}$	$\mathbf{R}$	$\mathbf{R}/\mathbf{Z} = S^1$	$(1+t)$
	$\mathbf{Z}^n$	$\mathbf{R}^n$	$\mathbf{R}^n/\mathbf{Z}^n = \underset{(n)}{S^1 \times \cdots \times S^1}$	$(1+t)^n$
(4.14)	$\mathbf{Z}_2$	$S(\mathfrak{H})$	$\mathbf{R}P_\infty$	$1/(1-t)$
	$U(1) = S^1$	$S(\mathfrak{H})$	$\mathbf{C}P_\infty$	$1/(1-t^2)$
	$U(2)$	2-frames in H	$\mathbf{G}_2(H)$	$\dfrac{1}{(1-t^2)(1-t^4)}$
	$U(n)$	n frames in H	$\mathbf{G}_n(H)$	$\dfrac{1}{(1-t^2)\cdots(1-t^{2n})}$

Here $\mathbf{Z}$ denotes the integers, $\mathbf{Z}^n$ the direct product of $\mathbf{Z}$ with itself n times, $\mathbf{Z}_2$ the group $\{\pm 1\}$, and $U(n)$ of course the unitary group. The first two U's come to mind immediately, on the other hand the rest may strike nonspecialists as surprising. In all of these $\mathfrak{H}$ denotes a complex infinite-dimensional Hilbert space, and $S(\mathfrak{H})$ its unit sphere. The space of n-frames on $\mathfrak{H}$ is then the space of n-tuples $\{x_1,\ldots,x_n\}$ of elements in $S(\mathfrak{H})$ which are mutually orthogonal. $U(n)$ clearly acts on these, i.e.

$$\{x_i\} \to \left\{\sum_j U_{ij}x_j\right\}$$

and the quotient gives precisely the Grassmannian $\mathbf{G}_n(\mathfrak{H})$ of all n-dimensional subspaces of $\mathfrak{H}$. For $n = 1$, this is simply the projective space.

All these examples then rely on the beautiful fact that the unit sphere in an infinite-dimensional Hilbert Space is contractible! In the final column, P_t is computed with any field K, except for the $\mathbf{Z}_2$ case. There P_t is given for the field $\mathbf{Z}_2$.

With these basics out of the way we can test this new calculus in some simple examples.

Consider then the action of the circle S^1 on the two sphere S^2, given by rotation about the z-axis in $\mathbf{R}^3$,

$$S^2\colon x^2 + y^2 + z^2 = 1.$$

Also, let f be the height function z, on S^2. Note that in this case S^1 does not act freely at the points $z = \pm 1$, i.e., precisely at the critical points f. Note also that the s of interval $-1 \leqslant z \leqslant 1$ now parametrizes the quotient S^2/S^1, so that

$$P_t(S^2/S^1) = 1. \tag{4.15}$$

Thus the quotient has a quite *inappropriate homology structure*. On the other hand

$$P_t^G(S^2) = \frac{1+t^2}{1-t^2}, \tag{4.16}$$

as follows easily from (4.9).

Correspondingly note that

$$\mathfrak{M}_t^G(z) = \frac{1}{1-t^2} + \frac{t^2}{1-t^2}, \qquad G = S^1,$$

corresponding to the fixed points $z = -1$ and $z = +1$ respectively. Thus, our function z is perfect, both, equivariantly and in the normal sense.

If we pass to the function z^2 on S^2,

$$\mathfrak{M}_t^G(z^2) = 1 + \frac{2t^2}{1-t^2}, \qquad G = S^1,$$

with one corresponding to the minimum circle orbit at $z = 0$. Because S^1 acts *freely here*, this orbit counts in terms of its index and the ordinary Poincaré polynomial of $N/G = pt$.

This function is, strangely enough, still perfect equivariantly, however it is not in the normal sense; indeed

$$\mathfrak{M}_t(z^2) = (1+t) + 2t^2.$$

Let us next apply this principle to the closed geodesic question. For instance what is the contribution to $\mathfrak{M}_t^G(E)$ of the first critical manifold T_1S^n. Clearly $O(2)$ acts freely in this instance, so that, by (4.8),

$$P_t^G(T_1S^n) = P_t(T_1S^n/O(2)).$$

Now $(T_1S^n)/O(2)$ is simply the Grassmannian $G_2(n+1)$ of 2-planes in $\mathbf{R}^{n+1}$. In particular,

$$P_1\{G_2(n+1)\} = \frac{n(n+1)}{2}.$$

Thus the equivariant theory predicts the right number of nondegenerate closed geodesics into which T_1S^n splits under any small deformation. Notice on the other hand that the higher critical sets, i.e. the kth iterate of the geodesics in T^1S_n, contribute by

$$t^{(2k-1)(n-1)} \cdot P_t^G(T_1S^n), \qquad G = O(2), \tag{4.17}$$

where however $O(2)$ *now acts* with $\mathbf{Z}/k\mathbf{Z}$ in the kernel! Over the rationals Q, this has *no effect* but mod k, say, (4.17) therefore introduces all manner of torsion into ΛS^n, and in one way or another it was this torsion which was improperly accounted for in Morse's attempts on this question long ago.

Indeed Morse's *circular connectivities* of S^n were given by the series

$$\frac{t^{n-1}P_t(G_2(n+1))}{1-t^{2(n-1)}}, \tag{4.18}$$

with $P_t(G_2, n)$ taken over the field $\mathbf{Z}_2$. On the other hand, the rational equivariant Morse series, in our present sense, is precisely given by the above expression, with $P_t(G_2, n)$ *computed over the rational* $\mathbf{Q}$.

The equivariant approach to this problem has been worked out in detail by my student Nancy Hingston, and it seems to us that it yields all correct known results in this direction more directly than any other method.

But let me now finally turn to the case where this equivariant theory really works per excellence and where Michael Atiyah and I were first led to apply it. This is the case of the Yang-Mills theory in dimension 2.

Although this theory is not directly pertinent to physics, it is of great interest in the theory of "stable bundles" in algebraic geometry. In particular it recaptures and refines results on the topology of the space of moduli of these bundles due to Harder [H]. Furthermore Harder's results originated in deep theorems in number theory, so that our contribution can at least be thought of as one more testimonial for the unity of mathematics. Let me therefore quickly sketch the barest outline of this application.

Recall first, that the Yang-Mills functional $A \to S(A)$ is defined on the space $\mathfrak{A}(P)$ of connections on some principal bundle P over a compact manifold M. More precisely:

$$S(A) = \int_M \|F_A\|^2 \, dv \tag{4.19}$$

where F_A is the curvature of A and the norm is taken relative to a fixed Riemann structure on M, and an Ad-invariant positive quadratic form on the Lie algebra of the structure group G of P. We only consider the case G compact, so that such a form always exists.

Our theory is especially appropriate here, as $\mathfrak{a}(P)$ is contractible(!) and S has a larger group of symmetries. Indeed S is invariant under the group

$$\mathcal{G}(P) = \text{Aut } P \text{ over the identity on } M,$$

of Gauge transformations, which acts naturally on $\mathfrak{a}(P)$. From our point of view, the appropriate topological invariant is therefore

$$P_t^{\mathcal{G}}(\mathfrak{A}) \equiv P_t(B\mathcal{G}). \tag{4.20}$$

Note that this equality follows from the contractability of $\mathfrak{a}(P)$ and (4.10).

The Poincaré series on the right is computable when M is a compact Riemann surface of genus g, and the G is the unitary group $U(n)$.

THEOREM.

$$\mathcal{P}_t(B\mathcal{G}; K) = \frac{\{(1+t)(1+t^3)\cdots(1+t^{2n-1})\}^{2g}}{[(1-t^2)(1-t^4)\cdots(1-t^{2n-2})]^2(1-t^{2n})} \tag{4.21}$$

when $\mathcal{G}$ is the group of Gauge transformations of any principal bundle P over M, with structure group $U(n)$, and K is any coefficient field.

We can now finally state and apply the main assertion of Atiyah's and mine.

THEOREM. *In the situation envisaged above, the Yang-Mills functional is perfect in the equivariant sense.*

In short,

$$\mathfrak{M}_t(S) = P_t^G(\mathfrak{A}) = P_t(B\mathcal{G}). \tag{4.22}$$

I have time here for only a hint at the proof. The perfection of this functional is in the present case closely related to a principle we call "self-completion", which I will explain here briefly in the context of the previous sections. Let then G act on the manifold W as before and let f be G-invariant, with N a nondegenerate critical manifold which we assume to consist of the G-orbit of $p \in W$,

$$(4.23) \qquad N = G/H, \qquad H \text{ the stabilizer of } p.$$

Then H acts on the normal space $\nu_p(N)$, and also on the negative normal space $\nu_p^-(N)$, relative to some H-invariant metric on W. We call this the *negative isotropy representation*, and denote it by λ_N. This representation is the proper analogue of the integer λ_N—the index of N—in the standard theory. Consider now the vector bundle over BH associated to the universal bundle U, by λ_H, and let $e_N \in H^*(BH)$ be its Euler class if it is orientable. (Note that if H is connected, this will automatically be the case indeed then BH is simply connected.) With this understood, we call e_N *K-injective if the multiplication by* e_N

$$(4.24) \qquad Ue_N\colon H^*(BH;K) \to H^*(BH;K)$$

is injective.

THEOREM. *Suppose $C(f)$ consists only of orbits $N = G/H$ as above, with H connected and e_N κ-injective for every $N \subset C(f)$. Then the nondegenerate G-invariant function f on the compact manifold W is K-perfect*:

$$(4.25) \qquad \mathfrak{M}_t^G(f) = P_t^G(W;K).$$

For example consider the S^1 action on S^2 of our earlier example and the function z on S^2. At the minimum ν_N^- is the *zero dimensional bundle and there our condition is* always to be considered as verified. At the maximum, λ_p is the *standard representation* of S^1 on $\mathbf{R}^2$, whose Euler class e generates $H^2(BS^1) = \mathbf{Z}[e]$. Thus our principle applies and immediately implies that z was *equivariantly perfect.*

In some sense this simple phenomenon occurs over and over again for the Yang-Mills functional, for unitary bundles. Partly this is due to the fact that *the stability groups of a connection A for P, are always of the form*

$$H_A = U(n_1) \times \cdots \times U(n_k), \qquad \sum n_i = n.$$

Hence all $H^*(BH)$'s are polynomial rings over which the injectivity criterion is easily checked. Note by the way how much simpler these cohomology groups are compared to what one encounters in the closed geodesic problem. Nevertheless this principle is also applicable there and helps in the computation of $H_*^{O(2)}(\Lambda S^n)$.

Finally a word on the application of (4.22). As we are primarily concerned with $U(n)$-bundles we will discuss them in terms of the vector bundles of rank n they define.

In view of the work of Narasimhan and Seshadri, the following assertions are then easily verified.

(4.26) Let L be a line bundle with first Chern-class n. Then the only extremum of S is the minimum of S and corresponds one-to-one to the Jacobian torus of M: $J = (S^1)^{2g}$.

(4.27) Let E be a vector bundle of rank 2, and first Chern-class 1. Then the minimum of S corresponds to a smooth variety isomorphic to the *space of moduli* $\mathbf{M}_2$ *of stable bundles* in the sense of algebraic geometry.

(4.28) All other extrema correspond to products of extrema of the 1-dimensional case. Thus they correspond to varieties $J \times J$ one for each decomposition

$$E = L_1 \oplus L_2, \quad c_1(L_1) < c_1(L_2), \quad c_1(L_1) + c_1(L_2) = 1.$$

One next computes the index of each of these varieties, to be given by

$$\operatorname{index}(J \times J) = 2 \dim H^1(M; L_2^* \otimes L_1) = 2g + 4(c_1(L_2) - 1).$$

(Here the cohomology is taken in the holomorphic sense.) So that finally one computes

$$\text{(4.29)} \qquad \mathfrak{M}_t^g(S) = \frac{P_t(\mathbf{M}_2)}{1 - t^2} + \frac{t^{2g}(1+t)^{4g}}{(1-t^2)^2(1-t^4)}.$$

Note that here $1/(1 - t^2)$ should be thought of as $P_t(BU_1)$, with U_1 the centralizer of points in $\mathbf{M}_2$, while $1/(1 - t^2)^2$ corresponds to the centralizer of the $J \times J$ varieties. Finally $1/(1 - t^4)$ occurs in the second factor instead of $1 + t^4 + t^8 + \cdots$ and thus accounts for the *sum over all* critical points of the (4.27) type.

Now, (4.21), (4.22) and (4.29) yield a formula for $P_t(\mathbf{M})$:

$$\text{(4.30)} \qquad P_t(\mathbf{M}_2) = \frac{(1+t)^{2g}(1+t^3)^{2g}}{(1-t^2)(1-t^4)} - \frac{t^{2g}(1+t)^4}{(1-t^2)(1-t^4)}.$$

Similarly for higher n, because (4.22) leads to a relation which recursively determines $P_t(\mathbf{M}_n)$.

The remarkable fact is, as mentioned before, that this recursive procedure for $P_t(\mathbf{M}_n)$—at least over $\mathbf{Q}$—was already known when M. Atiyah and I discovered it in the context of Morse theory. The formula (4.30) is implicit in Newsteads work and comes from a direct computation [N]. Precisely the formula (4.30) then occurs in Harder's paper [H], but deduced from theorems of C. L. Siegel and the Weil conjecture, which at the time were still conjectural. The general case was then taken up by Narasimhan and Harder—again from this number theoretic point of view.

Actually (4.22) refines all these results in the sense that it implies that all the varieties $\mathbf{M}_n$ are *free of torsion*, and such statements are out of range of the Harder methods.

This has of necessity been a very brief glimpse into the Morse theory in this context, and I have not had time to go into details, or to even come to grips with the now essential infinite dimensionality of the problem.

A paper with M. Atiyah in preparation will hopefully be finished soon on the subject, where the $P_t(\mathbf{M}_n)$ are also computed by a direct equivariant stratification of a space equivalent to $\mathfrak{a}(P)$. In a sense one can use the basic structure theorems of Narasimhan and Seshadri to describe the negative bundle stratification of this space, without even mentioning the Morse theory.

Still, in the final analysis, these two points of view fit together beautifully, and are then seen to be analogous to the Morse theory and Algebraic Geometry approach to the cell structures on the compact homogeneous spaces G/P, of the complex semisimple Lie groups.

In any case, the interested reader can find some of the details in a preliminary version of our paper, **[A-B]** available as a preprint.

Bibliography

[A-B] M. Atiyah and R. Bott, *On the Yang-Mills equations over Riemann surfaces*, Inst. Hautes Études Sci. Publ. Math., Preprint.

[B1] R. Bott, *Nondegenerate critical manifold*, Ann. of Math. (2) **60** (1954), 248–261.

[B2] ______, *The periodicity theorem for the classical groups*, Ann. of Math. (2) **70** 2 (1959), 179–203.

[B3] ______, *On the iteration of closed geodesics and the Sturm intersection theory*, Comm. Pure Appl. Math. **9** (1956), 171–206.

[G-M] D. Gromoll and W. Meyer, *On differentiable functions with isolated critical points*, Topology **8** (1969), 361–369.

[H] G. Harder, *Eine Bemerkung Zu einer Arbeit von P. E. Newstead*, J. für Math. **242** (1970), 16–25.

[H-N] G. Harder and M. S. Narasimhan, *On the cohomology groups of moduli spaces of vector bundles over curves*, Math. Ann. **212** (1975), 215–248.

[K] W. Klingenberg, *Lectures on closed geodesics*, Die Grundlehren der Mat. Wissenschaften, vol. 230, Springer-Verlag, Berlin and New York, 1978.

[L-S] L. Lyusternik and L. Schnirelmann, *Topological methods in the calculus of variations*, Gosndarstv. Izdat. Tehn-Teor. Lit., Moscow, 1930.

[N] P. E. Newstead, *Stable bundles of rank* 2 *and odd degree over a curve of genus* 2, Topology **7** (1968), 205–215.

[S] C. S. Seshadri, *Space of unitary vector bundles on a compact Riemann surface*, Ann. of Math. (2) **85** (1967), 303–336.

[S-B] H. Samelson and R. Bott, *Applications of the theory of Morse to symmetric spaces*, Amer. J. Math. **80** (1968), 965–1029.

[S-S] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc. **73** (1967), 747–817.

[T] R. Thom, *Sur une partition en cellules associèe àune fonction sur une varité*, C. R. Acad. Sci. Paris Sér. A-B **228** (1949), 973–975.

Department of Mathematics, Harvard University, Cambridge, Massachusetts 02138