1

TOPICS IN MORSE THEORY: LECTURE NOTES

Ralph L. Cohen
Kevin Iga
Paul Norbury

August 9, 2006

[1]The first author was supported by an NSF grant during the preparation of this work

Preface

These notes are an expanded version of lecture notes for a graduate course given at Stanford University during the Autumn of 1990. The course was geared to students who had completed a one year course in Algebraic Topology and had some familiarity with basic Differential Geometry.

There were two basic goals in the course and in these notes. The first was to give an introduction to Morse theory from a topological point of view. Besides classical Morse theory on a compact manifold, topics discussed included equivariant Morse functions, and more generally nondegenerate functions having critical submanifolds, as well as Morse functions on infinite dimensional Hilbert manifolds that satisfy the Palais–Smale condition (C). The general theme of these discussions was an attempt to understand, in as precise terms as possible, how the topology of the manifold is determined by the critical points of a Morse function and the gradient flow lines between them. Toward this end we discussed recent work of myself, J. Jones, and G. Segal which defines the concept of the classifying space of a Morse function, which is a simplicial space whose n - simplices are parameterized by n - tuples of gradient flow lines; the main theorem of which being that this space is homeomorphic to the underlying manifold. In order to have the machinery to discuss this work in detail I spent time in the course and in the notes discussing several relevant topics from Algebraic Topology, including framed cobordism, simplicial spaces, and spectral sequences.

The second goal of the course was to discuss several examples of relatively recent work in Gauge theory where Morse theoretic ideas and techniques have been applied. In these cases critical points tend to form moduli spaces, and we again attempted to emphasize the question of how the topology of the ambient manifold is determined by the topology of the critical sets and the spaces of gradient flows. We discussed several algebraic topological techniques for studying these questions including adaptations of the classifying space construction mentioned above.

These notes are in a rather rough form. Many details have been omitted, but references to the literature where they may be found are given at the end of every chapter.

I would like to take this opportunity to thank the students in the course for their interest and enthusiasm. In particular I would like to thank M. Betz, P. Norbury, and M. Sanders for going through various rough drafts of these

notes and in some cases finding and correcting errors and making suggestions. I am also grateful to S. Kerckhoff, T. Mrowka, and H. Samelson for helpful discussions.

R.L. Cohen
April, 1991

Introduction

Let M be a C^∞, compact, Riemannian manifold (without boundary), and

$$f : M \longrightarrow \mathbb{R}$$

a C^∞ map. A point $p \in M$ is a *critical point* of f if the differential $df_p : T_pM \longrightarrow \mathbb{R}$ is zero. (Here T_pM denotes the tangent space of M at p.)

Morse theory arises from the recognition that the number of critical points of f is constrained by the topology of M. For instance, since M is compact, then f attains its maximum value, and hence, f has at least one critical point. Morse theory typically deals with more non-trivial examples, where the differentiability of M and f play a larger role.

At first, one might guess this is primarily an application of topology to optimization theory. After all, optimization is concerned with finding critical points, and Morse theory would say that even if we understand only the topology of M, we might be able to make important predictions about how many critical points we might have.

But the main interest in Morse theory has been the reverse: using a discussion of critical points of f to uncover information about the topology of M. A more accurate statement is that we can study general classes of manifolds using the perspective obtained by considering the critical points of various functions f. This is crucial in the program to classify manifolds, though it has also increased our understanding of a wide range of subjects from symplectic geometry to non-abelian gauge theory.

Typically, we do not simply count the number of critical points, but consider more detailed information, such as the second derivative test, to count which critical points are maxima, which are minima, and so on.

The second derivative test at each critical point $p \in M$ involves a symmetric bilinear form, the Hessian $\mathrm{Hess}_p(f)$, on the tangent space $\mathrm{Hess}_p(f)$ is most properly viewed as a symmetric, bilinear form on the tangent space

$$\mathrm{Hess}_p(f) : T_pM \times T_pM \longrightarrow \mathbb{R}.$$

We will define $\mathrm{Hess}_p(f)$ in this context later. However in terms of local coordinates $\{x_1, \ldots, x_n\}$ of a neighborhood of $p \in M$, $\mathrm{Hess}_p(f)$ is the familiar matrix of second order partial derivatives

$$\mathrm{Hess}_p(f) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(p) \right).$$

A critical point $p \in M$ is said to be *nondegenerate* if the Hessian $\text{Hess}_p(f)$ is nonsingular. The *index* of a nondegenerate critical point p, $\lambda(p)$, is defined to be the dimension of the negative eigenspace of $\text{Hess}_p(f)$. Thus a local minimum has index 0, and a local maximum has index n. In an undergraduate multivariable calculus class, students typically focus on dimension $n = 2$, where local minima have index 0, saddles have index 1, and local maxima have index 2. We will use examples from dimension $n = 2$ to build geometric intuition, but we will be interested in applying Morse theory to arbitrary dimension n.

Now, instead of only counting the number of critical points of f, we can be more delicate and ask how many critical points there are of each index. We define the integers $c_0, \ldots, c_n$ so that c_i is the number of critical points of index i. As an example of the sort of relation Morse theory predicts between the c_i and the topology of M, we give the following theorem, Morse's theorem (which will be proved in chapter 5):

Theorem 1 (Morse's Theorem) *Let $f : M \longrightarrow \mathbb{R}$ be a C^∞ function so that all of its critical points are nondegenerate. Then the Euler characteristic $\chi(M)$ can be computed by the following formula:*

$$\chi(M) = \sum (-1)^i c_i(f)$$

where $c_i(f)$ is the number of critical points of f having index i.

A function $f : M \longrightarrow \mathbb{R}$ whose critical points are all nondegenerate is called a *Morse function*. The kinds of theorems we would like to prove in Morse theory will typically only apply to Morse functions. As we will see in chapter 4, however, "most" smooth functions are Morse. Thus in the hypothesis of the previous theorem, we could have said that f is a C^∞ Morse function.

Recall that the Euler characteristic of M is

$$\chi(M) = \sum (-1)^i b_i(M)$$

where $b_i(M)$ is the ith Betti number of M (the ith Betti bumber of M is the rank of the ith homology group of M). Thus the above theorem is astonishing in that on the right side of the equation we have an expression that involves f, while the left side of the equation does not depend on f at all, and in fact only depends on topological information about M.

There are other theorems, mentioned in chapter 5, that relate numbers of critical points of each index with the Betti numbers of the manifold; and we will see how studying other questions related to f give rise to more delicate topological information about M than just the Betti numbers.

There are a number of ways these sorts of theorems can be proved, corresponding to the several ways of approaching Morse theory.

0.1 The Classical approach to Morse Theory

Historically, Morse's original idea was to consider $f^{-1}((-\infty, a])$, for various real numbers a, and compare these manifolds as manifolds with boundary. This is

already explained clearly in Milnor's book *Morse Theory* [**?**]. This approach proves that compact manifolds have the homotopy type of a CW complex, and the CW complex structure allows us to compute the Euler characteristic.

0.2 Smale's approach

The second approach due to Smale, will be introduced in chapter 6. In this approach, we consider *gradient flow lines*. A *gradient flow line* (or *integral curve*) is a curve

$$\gamma : \mathbb{R} \longrightarrow M$$

that satisfies the following differential equation (the *flow equation*):

$$\frac{d\gamma}{dt} + \nabla_\gamma(f) = 0.$$

Here $\nabla(f)$ is the gradient vector field determined by f. Given the Riemannian metric $\langle \cdot, \cdot \rangle$, $\nabla(f)$ is determined by the property that

$$\langle \nabla_p(f), v \rangle = df_p(v)$$

where v is any tangent vector $v \in T_pM$. Hence if viewed as trajectories, gradient flow lines are the paths of steepest descent. Note that this approach depends explicitly on a Riemannian metric, unlike the previous approach.

Through these flows, we can actually decompose M into cells, and in such a way reconstruct the CW complex itself. In this way, the decomposition arises naturally from the Morse function f. Central to this approach are the *stable manifold* and *unstable manifold* of a critical point.

Definition 0.1 *Let M be a manifold, $f : M \longrightarrow \mathbb{R}$ a Morse function, and g a metric on M. Let p be a critical point of f. Then the* stable manifold *of p, $W^s(p)$, is the set of points in M that lie on gradient flow lines $\gamma(t)$ (defined using f and g) so that $\lim_{t\to+\infty} \gamma(t) = p$. Similarly, the* unstable manifold *of p, $W^u(p)$, is the set of points in M that lie on gradient flow lines $\gamma(t)$ so that $\lim_{t\to-\infty} \gamma(t) = p$.*

If M is compact, then every point in M lies on the stable manifold for some critical point, and on the unstable manifold for some critical point. This partitions M in two ways: as a collection of stable manifolds and as a collection of unstable manifolds.

Furthermore, as we will see, the stable and unstable manifold of a critical point are both balls: the unstable manifold of a critical point of index λ is a ball of dimension λ, and the stable manifold is a ball of dimension $n - \lambda$ where n is the dimension of the manifold.

This allows us to view the manifold as a collection of cells (unstable manifolds), one for each critical point, where the dimension of the cell is the index of the critical point. In this way, we can view the manifold as a CW complex.

In particular, we can derive the Strong Morse Inequalities. This can also reveal more topological information about the manifold than just homology, since we have at our disposal information about the manifold as a CW complex.

This approach is quite elegant and we will have much to say about it.

0.3 The Moduli Space approach

Related to this is an approach that fixes two critical points $p, q \in M$, and considers the set of gradient flow lines γ that "go from p to q" in the sense that

$$\lim_{t \longrightarrow -\infty} \gamma(t) = p,$$
$$\lim_{t \longrightarrow \infty} \gamma(t) = q.$$

Under the proper topology, this set of gradient flow lines from a to b forms a manifold called a *moduli space* (generally not compact), and it is possible to construct much of the topology of M from the moduli space. This approach has been particularly fruitful in studying a generalization of Morse theory on certain infinite-dimensional manifolds.

It also motivates a generalization due to M. Betz and R. Cohen that studies maps of graphs into the manifold, and this can compute the cup product and other homotopy invariants. We will discuss this in chapter ???.

0.4 Witten's supersymmetric approach

There is also an approach due to Witten that relates Morse theory to supersymmetry. Although considered by many mathematicians to be the most obscure of the approaches, it is an approach that physicists have found extremely fruitful in suggesting conjectures for the infinite-dimensional setting. A rigorous proof of the Morse inequalities using this method can be found in [**?**]. It is not clear if more topological information than cohomology can be obtained in this method.

0.5 Overview of the book

The first part of this book explains the Morse–Smale approach, and shows how it encodes a great deal of topological information about the manifold. For instance, we find the attaching maps for the CW complex, and show how cup products and Steenrod squares arise from generalizing gradient flow lines to graphs mapped in to the manifold.

The second part of this book deals with topological applications of Morse theory, particularly when generalized to certain infinite dimensional settings. Again, the idea is to use information about the critical points (or gradient flow lines) of a Morse function to obtain deeper information about the global topology of the manifold.

Contents

Chapter 1

Preliminaries

Chapter 2

Framed cobordisms

One of the most important topological applications of transversality theory was due to R. Thom and J. Pontrjagin in the mid 1950's who used it to systematically describe the classification of closed manifolds up to cobordism in terms of homotopy theory. In our discussion of Morse theory *framed cobordism* will be useful and so we will end this chapter by developing this concept and describing the Thom–Pontrjagin theorem in this context. For details about this we refer the reader to [**?**].

Let M^n be a smooth, closed n - manifold embedded in $\mathbb{R}^{n+k}$ with normal bundle $\nu^k(M)$. A *framing* of M (or more precisely a framing of $\nu^k(M)$) is a trivialization

$$\phi : \nu^k(M) \xrightarrow{\cong} M \times \mathbb{R}^k.$$

Two framed manifolds (M^n, ϕ) and (N^n, ψ) in $\mathbb{R}^{n+k}$ are said to be *framed cobordant* if there is a framed $n+1$ - manifold (W, Φ) (the "cobordism") embedded in $\mathbb{R}^{n+k+1}$ with

$$\partial W = M^n \sqcup N^n$$

and the framing Φ of $\nu(W)$ restricts to ϕ and ψ on the boundary components.

Cobordism is an equivalence relation, so let η_n^k be the set of cobordism classes of framed n - manifolds in $\mathbb{R}^{n+k}$. Notice there is a natural map

$$j_k : \eta_n^k \longrightarrow \eta_n^{k+1}$$

given by the inclusion of $\mathbb{R}^{n+k} \subset \mathbb{R}^{n+k+1}$ as the first $n+k$ coordinates. We let

$$\eta_n = \varinjlim_k \eta_n^k$$

where the direct limit is taken with respect to the maps j_k. η_n is the set of cobordism classes of *stably framed* n - manifolds. Notice that η_n is an abelian group under disjoint union and the direct sum

$$\eta_* = \oplus_n \eta_n$$

is a graded ring, where the multiplication is induced by cartesian product of manifolds. The Thom–Pontryagin theorem uses transversality theory to relate this ring to the homotopy groups of spheres.

To make this precise let $\pi_m(S^k)$ be the m^{th} homotopy group of the sphere S^k. This is defined to be the homotopy classes of maps from $S^m = \mathbb{R}^m \cup \infty$ to $S^k = \mathbb{R}^k \cup \infty$ that preserve ∞. There is a natural stabilization map

$$\sigma : \pi_m(S^k) \longrightarrow \pi_{m+1}(S^{k+1})$$

defined by *suspending* a class $[\alpha] \in \pi_m(S^k)$. That is, if $\alpha : S^m \longrightarrow S^k$, we define $\sigma(\alpha)$ to be the composition

$$\sigma(\alpha) : S^{m+1} = (\mathbb{R}^m \times \mathbb{R}) \cup \infty \xrightarrow{\alpha \times 1} (\mathbb{R}^k \times \mathbb{R}) \cup \infty = S^{k+1}.$$

By taking the limit over these maps we define the n^{th} *stable stem*, π_n^s to be

$$\pi_n^s = \varinjlim_k \pi_{n+k}(S^k).$$

When $n = 0$, $\pi_0^s \cong \mathbb{Z}$, determined by the degree of a map, and it is well known that for $n \geq 1$, π_n^s is a finite abelian group. The direct sum

$$\pi_*^s = \oplus_{n=0}^{\infty} \pi_n^s$$

is well known to be a commutative, graded ring. The multiplication is given by composition of maps.

The Thom–Pontrjagin theorem asserts that every stably framed manifold can be represented by an element in the stable stems.

Theorem 2.1 (Thom–Pontrjagin) *There is a natural isomorphism of graded rings*

$$\tau : \pi_*^s \longrightarrow \eta_*.$$

Proof: [Outline of Proof] Let $[\alpha] \in \pi_n^s$ be represented by the map

$$\alpha : S^{n+k} \longrightarrow S^k.$$

By the transversality theorem (Theorem 3.7) we may assume that $\alpha \pitchfork \{0\}$ where $0 \in \mathbb{R}^k \subset S^k$. (Notice that in this case transversality is simply the statement that $0 \in S^k$ is a regular value.) Let $M = \alpha^{-1}(\{0\})$. By Theorem 3.5, $codim_{S^{n+k}}(M) = codim_{S^k}(\{0\}) = k$, and so M is an n - dimensional submanifold of the sphere. Now the normal bundle of $\{0\} \in S^k$ has a unique framing and therefore by Corollary 3.6 there is an induced framing ϕ on $\nu(M)$. We define

$$\tau[\alpha] = (M^b, \phi) \in \eta_n.$$

The fact that τ is a well defined homomorphism of abelian groups for each n is a straightforward exercise. It uses the fact that any homotopy between

representatives of the same class in π_n^s can be assumed to be transverse to $\{0\} \in S^k$. The preimage of $\{0\}$ under such a homotopy will be a cobordism between the framed manifolds defined by applying τ to the two representatives. The fact that τ preserves the product structure and is therefore a homomorphism of rings is standard. Details can be found in Stong's book [**?**].

We define an inverse

$$\theta : \eta_n \longrightarrow \pi_n^s$$

to τ as follows. Let (M, ϕ) be a framed n - manifold in $\mathbb{R}^{n+k}$. By viewing the normal bundle $\nu(M)$ as a tubular neighborhood of M, the framing ϕ induces a homeomorphism

$$\phi_* : \nu(M)/\partial\nu(M) \xrightarrow{\cong} M \times D^k/(M \times \partial D^k) = M \times S^k/(M \times \{\infty\}).$$

Now define a map

$$\rho : S^{n+k} \longrightarrow \nu(M)/\partial\nu(M)$$

to be the identity on all points lying in the tubular neighborhood $\nu(M)$, and to be the constant map (at the basepoint $= \partial\nu(M)$) on all points lying outside the tubular neighborhood. We define $\theta(M, \phi)$ to be the composition

$$\theta(M,\phi) : S^{n+k} \xrightarrow{\rho} \nu(M)/\partial\nu(M) \xrightarrow{\phi_*} M \times S^k/(M \times \{\infty\}) \longrightarrow S^k$$

where the last map is the obvious projection. Again, one needs to check that θ is a well defined group homomorphism and is inverse to τ. We leave this as an exercise for the reader; it is carried out in detail in [**?**]. □

2.1 Manifolds with corners

Definition 2.2 *A paracompact Hausdorff topological space M is a* manifold with corners *of dimension n if there exists an open covering $\{U_\alpha\}$ of M so that for each U_α there is an open set $V_\alpha \subset \mathbb{R}^n_+$ and a homeomorphism $\phi_\alpha : U_\alpha \longrightarrow V_\alpha$, and so that for every α and β with $U_\alpha \cap U_\beta \neq \emptyset$, $\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta^{-1}(U_\alpha \cap U_\beta) \longrightarrow V_\alpha$ is smooth.*

Lemma 2.3 *If $U \subset \mathbb{R}^n_+$ is an open set, and if $f : U \longrightarrow \mathbb{R}^n_+$ is smooth, and $p \in U$ is written in coordinates $(p_1, \ldots, p_n)$, and k of these coordinates are zero, then $f(p)$ has k coordinates as zero.*

Proposition 2.4 *If M is a manifold with corners of dimension n, then there is a well-defined function $c : M \longrightarrow \mathbb{N}$ so that if $p \in M$, then $c(p)$ is the number of zeros in $\phi_\alpha(p)$, where $p \in U_\alpha$ and ϕ_α is as in the definition of manifolds with corners.*

Proof: Let U_α and U_β both contain p. Since $\phi_\alpha \circ \phi_\beta^{-1}$ is smooth, by the previous lemma, the number of coordinates that are zero is the same in $\phi_\alpha(p)$ and $\phi_\beta(p)$. □

remark: c is lower semicontinuous (upper?) so $c^{-1}(0)$ open
remark: cobordism: needs faces
In [**?**], Jänich defines manifolds with faces:

Definition 2.5 *If M is a manifold with corners, consider $c^{-1}(1)$. It is a union of connected components. If F is a connected component of $c^{-1}(1)$, the closure of F in M is called a* connected face *of M. If $\{F_1, \ldots, F_j\}$ is a collection of disjoint connected faces of M, then their union $\cup F_i$ is called a* face *of M.*

Definition 2.6 *$(M, \partial_0 M, \ldots \partial_{k-1} M)$ is a $\langle k \rangle$-faced n-manifold with faces if M is an n-dimensional manifold with corners, $\partial_i M$ is a face of M for each i, $0 \leq i \leq k-1$, $\cup_{i=0}^{n-1} \partial_i M = M - c^{-1}(0)$, and if $\imath \neq j$, then for any $x = in \partial_i M \cap \partial_j M$, $c(x) \geq 2$.*

Note: $\emptyset$ is considered a face.
To prove:

Theorem 2.7 *If a and b are critical points of M, then $\overline{\mathcal{M}}_{a,b}$ is a manifold with corners.*

Proof: By **??**, $\mathcal{M}_{a,b}$ is a manifold, so there is an open cover of $\mathcal{M}_{a,b}$ of charts mapped to $\mathbb{R}^n$. These open sets continue to be open in $\overline{\mathcal{M}}_{a,b}$ (since $\mathcal{M}_{a,b}$ is an open subset of $\overline{\mathcal{M}}_{a,b}$) and cover the parts of $\overline{\mathcal{M}}_{a,b}$ corresponding to unbroken flow lines.

Suppose $p \in \overline{\mathcal{M}}_{a,b}$ corresponds to the broken flow line

$$(\gamma_1, \ldots, \gamma_k)$$

where γ_i is a flow from a critical point a_{i-1} to a different critical point a_i. Note that we consider $a = a_0$ and $b = a_k$. Then by the Gluing theorem **??**, there is a smooth map

$$\# : \mathcal{M}_{a_0,a_1} \times \mathcal{M}_{a_1,a_2} \times \cdots \times \mathcal{M}_{a_{k-1},a_k} \times [0,1)^{k-1} \longrightarrow \overline{\mathcal{M}}_{a,b}$$

which is a diffeomorphism onto its image, so that $\#(\gamma_1, \ldots, \gamma_k, 0, \ldots, 0) = p$. For each $1 \leq i \leq k$, since $\mathcal{M}_{a_{i-1},a_i}$ is a manifold, there is an open coordinate chart U_i around γ_i homeomorphic to a subset V of $\mathbb{R}^{q_i}$ for some q_i.

Let $a = a_0, a_1, \ldots, a_{k-1}, a_k = b$ be critical points so that for each i, $\mathcal{M}_{a_{i-1},a_i}$ is non-empty. Then by the Gluing theorem **??**, there is a smooth map

$$\# : \mathcal{M}_{a_0,a_1} \times \mathcal{M}_{a_1,a_2} \times \cdots \times \mathcal{M}_{a_{k-1},a_k} \times [0,1)^{k-1} \longrightarrow \overline{\mathcal{M}}_{a,b}$$

which is a diffeomorphism onto its image. For each $1 \leq i \leq k$, $\mathcal{M}_{a_{i-1},a_i}$ is a manifold, and $[0,1)$ is a manifold with corners, so the domain of the gluing map above is a manifold with corners. In particular, by composing coordinate chart maps with $\#^{-1}$, we see that the image of $\#$ is a manifold with corners.

...
atlas covering $\mathcal{M}_{a_{i-1},a_i}$,

□

Note: need $\mathcal{M}_{a,b}$ is open in $\overline{\mathcal{M}}_{a,b}$. Need dense too? gluing map is a smooth map

cpt manfd w corners -¿ finitely many faces? product of mfd w corners = mfd w corners differentiation on mfds w corners

Part I

Classical Morse theory

Chapter 3

Digression: Transversality

Crucial to Morse theory is the concept of transversality. Readers who want a more in-depth discussion on transversality can refer to Hirsch's book *Differential Topology* [**?**].

We will need transversality at several points. We will use it to prove Morse functions exist, by considering transversality on the space of functions. Later, we will need various submanifolds of M to intersect in a nice, expected way, and here we consider transversality on M.

3.0.1 Transversality with linear functions

We begin with the notion of transversality in the realm of linear algebra, and in the next section we will apply this to general differentiable functions using the tangent space approximation.

Let V and W be finite dimensional real vector spaces and $S \subset W$ a subspace. Let

$$\Phi : V \longrightarrow W$$

be a linear transformation.

Definition 3.1 *Φ is transverse to $S \subset W$ (we write: $\Phi \pitchfork S$) if*

$$\Phi(V) + S = W.$$

Lemma 3.2 *If $\Phi \pitchfork S$, and if $\alpha_1, \ldots, \alpha_q$ are linearly indepedent forms on W so that*

$$S = \{\vec{w} \in W \mid \alpha_1(\vec{w}) = 0, \ldots, \alpha_q(\vec{w}) = 0\}$$

then

$$\beta_1 = \alpha_1 \circ \Phi, \ldots, \beta_q = \alpha_q \circ \Phi$$

are linearly independent forms on V.

Proof: Suppose $c_1, \ldots, c_q$ are real numbers, not all zero. Then since the α_i's are linearly independent, there is some vector $\vec{w} \in W$ so that

$$\sum_{i=1}^{q} c_i \alpha_i(\vec{w}) \neq 0.$$

Since $\Phi(V) + S = W$, we can write $\vec{w}$ as $\Phi(\vec{v}) + \vec{s}$ for some $\vec{v} \in V$ and $\vec{s} \in S$. Then

$$\sum_{i=1}^{q} c_i \alpha_i(\vec{w}) = \sum_{i=1}^{q} c_i \alpha_i(\Phi(\vec{v})) + c_i \alpha_i(\vec{s}) = \sum_{i=1}^{q} c_i \beta_i(\vec{v}) + 0$$

so we see that if not all the c_i are zero, then $\sum c_i \beta_i$ is not zero. Thus the $\beta_1, \ldots, \beta_q$ are independent. □

Proposition 3.3 *If $\Phi \pitchfork S$, then $\Phi^{-1}(S) \subset V$ is a subspace of codimension equal to $\operatorname{codim}_W(S)$.*

Proof: Let α_i, β_i be as in the previous proposition. First notice that

$$\Phi^{-1}(S) = \{\beta_1 = 0, \ldots, \beta_q = 0\}.$$

This proves that $\Phi^{-1}(S)$ is a linear subspace of V. By the above proposition, the β_i are independent. This proves that the codimension of $\Phi^{-1}(S)$ in V is q. □

3.0.2 Transversality with smooth functions

Now suppose M and N are smooth manifolds $S \subset N$ a submanifold. Let

$$f : M \longrightarrow N$$

be a smooth map.

Definition 3.4 *f is said to be transverse to S ($f \pitchfork S$) if and only if for every $x \in M$, either $f(x) \notin S$, or if $f(x) \in S$ then $Df_x \pitchfork T_{f(x)}S$ (i.e. $Df_x(T_xM) + T_{f(x)}S = T_{f(x)}N$.)*

With smooth maps as well as vector spaces, a basic property of transversality is that it preserves codimension of submanifolds.

Theorem 3.5 *If $f \pitchfork S$, then $f^{-1}(S)$ is a submanifold of M such that*

$$\operatorname{codim}_M f^{-1}(S) = \operatorname{codim}_N S.$$

Proof: Let $x \in f^{-1}(S)$. Let V be a neighborhood of $f(x) \in N$ that supports functions

$$\phi_1, \ldots, \phi_p : V \longrightarrow \mathbb{R}$$

satisfying

- $S \cap V = \{\phi_1 = 0, \ldots, \phi_p = 0\}$
- $d\phi_1, \ldots, d\phi_p$ are independent linear forms at each point in V.

Now consider the open set $U = f^{-1}(V) \subset M$. This is a neighborhood of $x \in M$ that supports the functions

$$\psi_1 = \phi_1 \circ f, \ldots, \psi_p = \phi_p \circ f$$

whose set of common zeros is $f^{-1}(S) \cap U$. Moreover since $f \pitchfork S$, we have by Proposition 3.3 that $d\psi_1, \ldots, d\psi_p$ are linearly independent forms at each point in U. This defines the local manifold structure of $f^{-1}(S)$. $\square$

One consequence of this proof is the following:

Corollary 3.6 *If $f : M \longrightarrow N$ is transverse to the submanifold $S \subset N$, then it induces a bundle isomorphism*

$$\nu(f^{-1}(S)) \cong f^*(\nu(S)).$$

where $\nu(S) \subset TN$ denotes the normal bundle to S in N; and similarly for $\nu(f^{-1}(S)) \subset TM$.

Furthermore, if the normal bundle $\nu(S)$ comes equipped with some structure (e.g. an orientation, a complex structure, a framing) then the normal bundle $\nu(f^{-1}(S))$ has an induced structure of the same kind.

One of the main features of transversality is that it is a generic condition. That is, "nearly all" maps are transverse to a given submanifold. To make this precise we first adopt some notation. Let $C^r(M, N)$ denote the space of C^r-differentiable maps with the C^r-compact-open topology. In this topology a basic neighborhood around $f \in C^r(M, N)$ is given as follows. Let (ϕ, U), and (ψ, V) be charts on M and N respectively, $K \subset f^{-1}(V) \cap U$ be a compact subset, and $\epsilon > 0$. Consider the set

$$\Gamma_{K,V,\epsilon} = \{g \in C^r(M, N) \mid g(K) \in V \text{ and } \|D^k(\psi f \phi^{-1})(x) - D^k(\psi g \phi^{-1})(x)\| < \epsilon \text{ for all } 0 < k \leq r\}$$

for all $x \in K$. That is, $g : M \longrightarrow N$ is in this set $\Gamma_{K,V,\epsilon}$ if in terms of local coordinates f and g have their values, and the values of their first k derivatives within ϵ of each other at every point of K. These $\Gamma_{K,V,\epsilon}$ sets are the basic open sets in the topology of $C^r(M, N)$. For more details on the topologies of function spaces we refer the reader to [**?**][ch. 2][1]

Suppose M and N are manifolds, with M compact. Let $f : M \longrightarrow N$ be smooth, and $S \subset N$ be a submanifold. We define the function space

$$\pitchfork^r(M, N; S) = \{f \in C^r(M, N) \mid f \pitchfork S\}.$$

[1]The reference cited here, Hirsch, describes this as the "weak C^r topology", and defines a "strong C^r topology" with more open sets. For most of our purposes, M will be compact and the two topologies coincide. But even when M is not compact, we will continue to use the weak topology.

Theorem 3.7 (Transversality theorem) *Let M, N, and S be as above. Then $\pitchfork^r(M,N;S) \subset C^r(M,N)$ is an open and dense subspace.*

Remark 3.1 *If M is not compact, and K is a compact subset of M, we can define $\pitchfork^r_K(M,N;S)$ to be the subset of $C^r(M,N)$ of functions f so that at every $x \in K$, $Df(T_x(M)) + T_{f(x)}(S) = T_{f(x)}N$. Then the correct statement is that $\pitchfork^r_K(M,N;S)$ is open and dense in $C^r(M,N)$. This more general statement is proved in Hirsch [**?**], and the proof is more or less the same, but it is more confusing. We don't need this more general version so we omit it here.*

The first step in the proof is a result of R. Thom proved in 1956.

Let M, N, S, be as above, and P another smooth manifold. (It will be viewed as a parameter space.)

Lemma 3.8 *Let $F : M \times P \longrightarrow N$ be transverse to S. Then for a dense subset $A \subset P$, $p \in A$ implies that*

$$F_p : M = M \times \{p\} \longrightarrow N$$

is transverse to S.

Proof: Let $\Sigma = F^{-1}(S) \subset M \times P$. Then by Theorem 3.5 Σ is a submanifold of $M \times P$ of codimension equal to the codimension of S in N. Let

$$\pi : M \times P \longrightarrow P$$

be the projection. We consider the restriction $\pi|_\Sigma$:

$$\pi|_\Sigma : \Sigma \longrightarrow P.$$

Sard's theorem states that the set of regular values of $\pi|_\Sigma$ is dense in P.

We will now prove that if $p \in P$ is a regular value of $\pi|_\Sigma$, then $F_p \pitchfork S$:

Suppose $p \in P$ is a regular value of $\pi|_\Sigma$. Then $D_p\pi|_\Sigma$ is surjective. Similarly, since $F : M \times P \longrightarrow N$ is transverse to S, we have that

$$D_{(m,p)}F(T_{(m,p)}M \oplus T_{(m,p)}P) + T_{F(m,p)}S = T_{F(m,p)}N$$

for all m so that $F(m,p) \in S$.

Let $\vec{s} \in T_{F(m,p)}N$. Since $F \pitchfork S$, we can write

$$\vec{s} = D_{(m,p)}F(\vec{v}) + \vec{s}_1$$

where $\vec{v} \in T_{(m,p)}M$ and $\vec{s}_1 \in T_{F(m,p)}S$. Since $p \in P$ is a regular value of $\pi|_\Sigma$, there is a vector $\vec{w} \in T_{(m,p)\Sigma}$ so that $D\pi(\vec{w}) = D\pi(\vec{v})$. Then $\vec{v} - \vec{w} \in T_{(m,p)}M \times \{p\}$ and

$$DF_p(\vec{v} - \vec{w}) + \vec{s}_1 = \vec{s}.$$

Thus, $F_p \pitchfork S$.

Incidentally, the converse is true: $F_p \pitchfork S$ implies p is a regular value for $\pi|_\Sigma$, and the proof is similarly straightforward. If $\vec{v} \in T_pP$, and an $m \in M$ exists so

that $(m,p) \in \Sigma$. By the transversality of F_p, there is a vector $\vec{x} \in T_{(m,p)}M \times \{p\}$ so that

$$DF_p(\vec{x}) - DF(\vec{v}) \in TS$$

Then $\vec{v} - \vec{x}$ projects under $D\pi$ to $\vec{v}$, and DF maps it to TS. $\square$

As a consequence of this result we get the following local version of Theorem 3.7.

Lemma 3.9 *Let M be a compact manifold, and let V be an open subset of $\mathbb{R}^n$. Let $\mathbb{R}^k \subset \mathbb{R}^n$ be a linear subspace. Then*

$$\pitchfork^r(M, V; \mathbb{R}^k \cap V)$$

is open and dense in $C^r(M,V)$.

Proof: Step 1: open

For every $x \in M$, consider a coordinate neighborhood U. Inside this coordinate neighborhood is an open coordinate ball around x, $B(x) \subset U$, so that the closure $\overline{B}$ is contained in U. The collection of $B(x)$ is an open cover of M, and since M is compact, there are a finite number of points $\{x_1, \ldots, x_n\}$ so that the $B(x_i)$ still cover M. For each i, let D_i be the closure of B_i. Then each D_i is compact.

Consider the maps

$$C^r(M,V) \xrightarrow{r} C^r(D_i,V) \xrightarrow{i} C^r(D_i,\mathbb{R}^n)$$

where r is the map that restricts functions from M to V to functions from D_i to V, and i is the map that composes a function from D_i to V with the inclusion of V into $\mathbb{R}^n$.

We now prove that r is continuous.

Let $s : M \longrightarrow V$ be an element of $C^r(M,V)$. Consider a basic neighborhood $\Gamma_{K,U,\epsilon}$ in $C^r(D_i,V)$ around $r(s) = s|_{D_i}$. The preimage of $\Gamma_{K,U,\epsilon}$ under r is the set of C^r functions $w : M \longrightarrow V$ so that $w(K) \subset U$ and $|Dw - Ds| < \epsilon$. This is the neighborhood $\Gamma_{K,U,\epsilon}$ around s. Therefore r is continuous. The proof that i is continuous is very similar and is left as an exercise.

Exercise 3.1 *Prove that i is continuous.*

Now let $Q(D_i, \mathbb{R}^n; \mathbb{R}^k)$ be the subset of $C^r(D_i, \mathbb{R}^n)$ consisting of those functions $f : D_i \longrightarrow \mathbb{R}^n$ that are transverse to $\mathbb{R}^k$. This means that at each point the partial derivatives matrix, together with extra columns for $e_1, \ldots, e_k$ (the basis for $\mathbb{R}^k$), is rank n. The set of such matrices is open, so $Q(D_i, \mathbb{R}^n; \mathbb{R}^k)$ is an open subset of $C^r(D_i, \mathbb{R}^n)$.

Now $\pitchfork_{D_i}(M, V; \mathbb{R}^{\cap} V)$ is the preimage of $Q(D_i, \mathbb{R}^n; \mathbb{R}^k)$ under $i \circ r$, so it is open. Finally, $\pitchfork(M, V; \mathbb{R}^k \cap V)$ is the intersection of these sets, so it is open.

Step 2: dense

To complete the proof of the lemma it suffices to prove that $\pitchfork^r(M, \mathbb{R}^n; \mathbb{R}^k)$ is dense in $C^r(M, \mathbb{R}^n)$. To do this let $f : M \longrightarrow \mathbb{R}^n$ and consider the map

$$F : M \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

defined by

$$F(m, w) = f(m) + w.$$

F is clearly a submersion and hence is transverse to every subspace. In particular it is transverse to $\mathbb{R}^k \subset \mathbb{R}^n$. By the previous lemma, the set of $w \in \mathbb{R}^n$ such that

$$F_w : M \longrightarrow \mathbb{R}^n \quad m \longrightarrow f(m) + w$$

is transverse to $\mathbb{R}^k$ is dense in $\mathbb{R}^n$. Thus there is a sequence of w_i converging to 0 so that F_{w_i} is transverse to $\mathbb{R}^k$.

To prove that F_{w_i} converges to $f = F_0$ in the C^r topology, we note that a basic open neighborhood of f in C^r is defined by a compact set $K \subset M$ and an open set $V \subset \mathbb{R}^n$, such that $f(K) \subset V$, and an $\epsilon > 0$. Now $f(K)$ is compact and $\mathbb{R}^n - V$ is closed, so the minimum distance $d(K, \mathbb{R}^n - V)$ is greater than zero. Let N be a number so that $|w_i| < d(K, \mathbb{R}^n - V)$ for all $i \geq N$. Then for all $i \geq N$, $F_w(K) = f(K) + w \subset V$. The derivatives of f and F_{w_i} are identical, so F_{w_i} is in the basic open neighborhood of C^r defined by K, V, and ϵ, whenever $i \geq N$.

Therefore, F_{w_i} converges to f in the C^r topology. $\square$

Proof: [Proof of Theorem 3.7] We now proceed with the proof of the transversality Theorem 3.7. We prove it in the case $\partial N = \emptyset$ and $S \subset N$ is a closed submanifold. For a proof in the general case we refer the reader to Hirsch's book [**?**]. Under these assumptions if we take locally finite, countable atlases $\mathcal{U}$ of M and $\mathcal{V}$ of N, such that the closure of any element in $\mathcal{U}$ is diffeomorphic to a closed ball in $\mathbb{R}^n$, then for any $U \in \mathcal{U}$ and $V \in \mathcal{V}$ the previous lemma implies that

$$\pitchfork^r(cl(U), V; V \cap S) \subset C^r(cl(U), V)$$

is open and dense. Since M is compact we can assume it is covered by finitely many of the open sets in $\mathcal{U}$. The openness assertion of the theorem follows immediately.

To prove the denseness assertion, let $f \in C^r(M, N)$, and for $U \in \mathcal{U}$ and $V \in \mathcal{V}$ let $T_{U,V}$ be the closure of $U \cap f^{-1}(V)$. Notice that $\pitchfork^r(T_{U,V}, V; S \cap V)$ is dense in $C^r(T_{U,V}, V)$. Thus we can approximate the restriction $f|_{T_{U,V}}$ arbitrarily closely (in the C^r sense) by maps transverse to $S \cap V$. Let $\lambda : T_{U,V} \longrightarrow [0, 1]$ be a smooth map with support in a compact subset of $U \cap f^{-1}(V)$ such that $\lambda = 1$ near U. Then if $g_j \in \pitchfork^r(T_{U,V}, V; V \cap S)$ is a sequence of transversal maps converging to f on $T_{U,V}$, then the maps

$$h_j(x) = \begin{cases} f(x) + \lambda(x)(g_j(x) - f(x)), & \text{for } x \in T_{U,V} \\ f(x), & \text{for } x \in M - T_{U,V} \end{cases}$$

converge to f on all of M. Notice that since $h_j = g_j$ on U, it follows that each $h_j \in \pitchfork^r_{K \cap U}(M, N; S)$. This shows that each $\pitchfork^r_U(M, N; S)$ is dense in $C^r(M, N)$.

We already know they are all open. Thus by the Baire category theorem their common intersection (over all $U \in \mathcal{U}$) is dense. This implies that $\pitchfork^r(M, N; S)$ is dense. $\square$

Let $E \longrightarrow M$ be a smooth vector bundle and let $C^r(M, E)$ denote the space of C^r sections of E. Let $\sigma \in C^\infty(M, E)$ be a particular, fixed section and let $\pitchfork^r(M, E; \sigma) \subset C^r(M, E)$ denote the subspace of sections transverse to $\sigma(M) \subset E$. We leave it to the reader to adapt the above arguments to prove the following bundle version of the transversality theorem.

Theorem 3.10 *Let $E \longrightarrow M$ and $\sigma \in C^\infty(M, E)$ be as above. Then*

$$\pitchfork^r(M, E; \sigma) \subset C^r(M, E)$$

is open and dense.

This is the crucial theorem that allows us to prove that, in some sense, "most" smooth functions are Morse, as advertised in the introduction. This is left as an exercise in the next chapter.

Before we leave transversality theory we discuss one nice application of the transversality theorem.

Let M be a smooth manifold of dimension n, and $f : M \longrightarrow \mathbb{R}^q$ a smooth map. Consider its differential

$$Df : M \longrightarrow \mathrm{Lin}(TM, \mathbb{R}^q)$$

where $\mathrm{Lin}(TM, \mathbb{R}^q) \longrightarrow M$ is the bundle of linear transformations. Its fiber at a point $p \in M$ is the group of linear transformations $\mathrm{Lin}(T_pM, \mathbb{R}^q)$, which via a choice of basis can be thought of as the group of $n \times q$ real matrices. Notice that $\mathrm{Lin}(TM, \mathbb{R}^q)$ is a manifold of dimension $n(q + 1)$.

Let $\Sigma_r \subset \mathrm{Lin}(TM, \mathbb{R}^q)$ be the subbundle consisting of linear transformations of rank r.

Exercise 3.2 *Show that Σ_r is a submanifold of $\mathrm{Lin}(TM, \mathbb{R}^q)$ of codimension $(n - r)(q - r)$.*

Thus if $q > 2n$, the codimensions of all of the Σ_r's are greater than n. Hence in this setting $D(f) \pitchfork \Sigma_r$ if and only if $Df(M) \cap \Sigma_r = \emptyset$. Thus $D(f) \pitchfork \Sigma_r$ for every $r \geq 1$ if and only if f is an immersion (i.e. $D(f)$ has maximal rank at every point in M). The following is then an easy consequence of the transversality theorem.

Theorem 3.11 *If M is a compact manifold of dimension n and $q > 2n$, then the space of immersions of M into $\mathbb{R}^q$ is open and dense in $C^r(M, \mathbb{R}^q)$.*

Chapter 4

Critical points and Gradient flow lines

4.1 The index of critical points

Let M be a manifold, and $f : M \longrightarrow \mathbb{R}$ a C^2 function. As explained in the introduction, a point $p \in M$ is called a *critical point* of f if $df_p = 0$.

Let $(U, \phi : U \longrightarrow \mathbb{R}^n)$ be a coordinate chart around p, so that $\phi(p) = 0$. Write ϕ as $(x_1, \ldots, x_n)$. Write tangent vectors v and w in T_pM as $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_n)$, respectively (specifically, $d\phi_p(v) = (v_1, \ldots, v_n)$ and similarly for w). Then the Hessian of f at p, using the coordinate chart (U, ϕ) is given by the formula

$$\mathrm{Hess}_p(f)(v, w) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} v_i w_j$$

Since f is C^2, this is defined and symmetric in v and w. This is also bilinear in v and w.

Proposition 4.1 *When p is a critical point for $f : M \longrightarrow \mathbb{R}$, the Hessian at p is independent of the coordinate chart.*

Proof: Now suppose we had a different coordinate chart around p, $(V, \psi : V \longrightarrow \mathbb{R}^n)$, with $\psi(p) = 0$. Write ψ as $(y_1, \ldots, y_n)$. Then $Q = \psi \circ \phi^{-1} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a diffeomorphism, and $dQ(e_i) = \sum_{j=1}^{n} \frac{\partial x_i}{\partial y_j} e_j$ (where $e_1, \ldots, e_n$ is the standard basis in $\mathbb{R}^n$), then the Hessian defined for this new coordinate

chart is

$$\begin{aligned}
\mathrm{Hess}_p(f)(v,w) &= \sum_{i,j=1}^{n} \frac{\partial}{\partial y_i}\left(\frac{\partial}{\partial y_j}(f)\right) v_i w_j \\
&= \sum_{i,j=1}^{n} \frac{\partial x_k}{\partial y_i}\frac{\partial}{\partial x_k}\left(\frac{\partial x_m}{\partial y_j}\frac{\partial}{\partial x_m}(f)\right) v_i w_j \\
&= \sum_{i,j=1}^{n} \frac{\partial x_k}{\partial y_i}\frac{\partial}{\partial x_k}\left(\frac{\partial x_m}{\partial y_j}\right)\frac{\partial}{\partial x_m}(f) v_i w_j + \sum_{i,j=1}^{n} \frac{\partial x_k}{\partial y_i}\frac{\partial x_m}{\partial y_j}\frac{\partial^2}{\partial x_k \partial x_m}(f) v_i w_j \\
&= \sum_{i,j=1}^{n} \frac{\partial x_k}{\partial y_i}\frac{\partial}{\partial x_k}\left(\frac{\partial x_m}{\partial y_j}\right)\frac{\partial}{\partial x_m}(f) v_i w_j + \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_k \partial x_m}(f) dQ(v)_k dQ(w)_m
\end{aligned}$$

Now note that the first term is zero when p is a critical point, so as a bilinear form on T_pM, the Hessian is well-defined. □

Remark 4.1 *If p is not a critical point of f, then the Hessian at p is not well-defined, in that using the above notation, it would depend on the coordinate chart. There are ways to extend the Hessian to all of M: by patching together coordinate charts and using partitions of unity; by choosing a metric on M, then using the Levi–Civita connection corresponding to this metric to take the covariant derivative of df at p, and so on. But these approaches all require some extra data. In this book we will only be concerned with the Hessian at critical points.*

4.2 Morse functions

Definition 4.2 *If $p \in M$ is a critical point for a C^2 function $f : M \longrightarrow \mathbb{R}$, then we call p* nondegenerate *if $\mathrm{Hess}_p(f)$ is nondegenerate as a bilinear form. If all critical points of M are nondegenerate, we say that f is* Morse.

Remark 4.2 *Recall that a bilinear form $B(v,w) : V \times V \longrightarrow \mathbb{R}$ is nondegenerate if for every non-zero $v \in V$, there exists a w so that $B(v,w) \neq 0$. Equivalently, if V is finite dimensional, we can choose any basis for V and write B as a matrix M using this basis, as $B(v,w) = v^T M w$. Then B is nondegenerate if and only if $\det(M) \neq 0$. Also, since B is symmetric, we can choose a basis in which the matrix M is diagonal, and then the criterion that B is non-degenerate is equivalent to the statement that M has no zero eigenvalues. These facts can be found in any linear algebra text.*

Let X be a vector field on a manifold M with a zero at $p \in M$. Recall that p is an *elementary zero* of X if and only if the Jacobian of X at p is invertible. In terms of local coordinates around p, if we write X as

$$X = (X_1(x_1, \ldots, x_n), \ldots X_n(x_1, \ldots, x_n))$$

then the Jacobian of X is given by

$$\left(\frac{\partial X_i}{\partial x_j}(p)\right).$$

Lemma 4.3 *The vector field X has only elementary zeros if and only if $X \in \pitchfork^r(M, TM; \zeta)$, where ζ is the zero section.*

Proof: X has only elementary zeros if and only if at every point p where $X(p) = 0$, the Jacobian is invertible. In other words, at every point where $X(p) = \zeta(p)$, $dX|_p$ never maps anything in T_pX to the tangent space of the ζ section in $T_{(p,0)}(TX)$. Therefore this is equivalent to the statement that X is transverse to ζ. □

Thus Theorem 3.10 implies the classical result that "most" vector fields have only elementary zeros.

Exercise 4.1 *Let $f : M \longrightarrow \mathbb{R}$ be a C^2 function, and let $p \in M$ be a critical poitn of f. Prove that the Jacobian of the gradient vector field $\nabla(f)$ at p is the Hessian at p.*

Exercise 4.2 *Prove that f is Morse if and only if $\nabla(f)$ has only elementary zeros.*

Corollary 4.4 *Let M be a compact n-manifold. Let $r \geq 2$. The set of C^r Morse functions from M to $\mathbb{R}$ is open and dense in $C^r(M, \mathbb{R})$.*

Proof: Use the result of the previous exercise and Theorem 3.10. □

4.3 The gradient flow equation

Let M be a manifold, g a Riemannian metric on M, and $f : M \longrightarrow \mathbb{R}$ be a Morse function. As explained in the introductory chapter, a *(gradient) flow line* is a curve

$$\gamma : (a, b) \longrightarrow M$$

that satisfies the differential equation

$$\frac{d\gamma}{dt} + \nabla_\gamma(f) = 0. \tag{4.1}$$

If we imagine a particle that travels along γ, with t describing time, the particle travels in the path of steepest descent, with velocity given by the gradient. If we imagine f to be "height", the particle could be a "sticky" ball that travels down along the surface but has too much friction with the surface to build up any speed.[1]

[1]These equations do not correspond exactly to such a physical system, but it is a good visual aid.

Note also that the gradient flow equation depends on the Riemannian metric g, since $\nabla_\gamma(f)$ depends on the metric in the following way: $g(v, \nabla(f)) = df(v)$. The typical gradient seen in undergraduate calculus classes occurs on $\mathbb{R}^n$ with the standard flat metric.

Exercise 4.3 *Verify that if $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a differentiable function on $\mathbb{R}^n$, and if we use the flat metric on $\mathbb{R}^n$, then*

$$\nabla(f) = \frac{\partial f}{\partial x_1} e_1 + \cdots + \frac{\partial f}{\partial x_n} e_n.$$

Exercise 4.4 *Let f be as in the previous exercise, but suppose the metric is given by an arbitrary symmetric matrix g (that is, $g(e_i, e_j) = g_{ij}$). Find the formula for $\nabla(f)$ in terms of f and g.*

Remark 4.3 *Now note that the property of $p \in M$ being a critical point of f does not depend on the metric. As a bilinear form, the Hessian does not depend on the metric either, and therefore so is the property of p being a non-degenerate critical point, and the index of the critical point.*

Example 4.1 *If a is a critical point of f, then the constant curve $\gamma(t) = a$ satisfies the flow equations, so γ is a flow line. Note that by the uniqueness of solutions of ODEs, if any flow line contains a critical point, it must be the constant one.*

Example 4.2 *Let $M = \mathbb{R}^2$ with the flat metric, and let $f(x, y) = x^2 + y^2$. Then we can solve the gradient flow equations:*

$$\dot{x} = -2x$$
$$\dot{y} = -2y$$

and therefore the gradient flow lines are $(x, y) = (ae^{-2t}, be^{-2t})$ for some fixed a and b. For any such line, y/x is a constant, so each lies in a line. In fact, it is the open radial ray from the origin. See figure 4.1.

Example 4.3 *Let $M = \mathbb{R}^2$ with the flat metric, and let $f(x, y) = x^2 - y^2$. Then it turns out that the gradient flow lines are $(x, y) = (ae^{-2t}, be^{2t})$ for some fixed a and b. For any such line, xy is a constant, so the gradient flow lines are hyperbolas of the form $xy = c$. See figure 4.2.*

Example 4.4 *Let $M = S^2 \subset \mathbb{R}^3$ with the standard round metric, and let $f(x, y, z) = z$ (the so-called "height function" defined by the embedding of S^2 into $\mathbb{R}^3$). Then there are two critical points: one minimum at $(0, 0, -1)$, and one maximum at $(0, 0, 1)$. The flow lines are "lines of longitude". See figure 4.2.*

Figure 4.1: Flow lines for $f(x, y) = x^2 + y^2$

Figure 4.2: Flow lines for $f(x, y) = x^2 - y^2$

Figure 4.3: Flow lines for the height function on S^2

Figure 4.4: Flow lines for the height function on the torus

Example 4.5 *Let T^2 be the torus in $\mathbb{R}^3$, embedded as follows:*

$$(\theta, \phi) = (b\cos(\phi), (a + b\sin(\phi))\cos(\theta), (a + b\sin(\phi))\sin(\theta)$$

where $0 < b < a$. The picture looks like a donut standing on its edge, as in figure 4.4. Again, take for f the "height function" z. Then there are four critical points: $(\theta, \phi) = (\pm\pi/2, \pm\pi/2)$, as you can check. The index for $(\pi/2, \pi/2)$ is 2, the index for $(\pi/2, -\pi/2)$ and $(-\pi/2, \pi/2)$ is 1, and the index for $(-\pi/2, -\pi/2)$ is 0.

There are two natural choices for a metric on T^2: either the metric induced from the embedding from $\mathbb{R}^3$, or the flat metric defined by $ds^2 = d\theta^2 + d\phi^2$. Although pictorially it may help to ponder the resulting gradient flow lines from the metric induced by $\mathbb{R}^3$ (these are the actual flows of steepest descent on a physical donut), it is easier to calculate the flow lines when the flat metric is used. The flow lines can be described explicitly, or else you can verify that there are flows with $\theta = \pm\pi/2$ for which θ is constant, and flows with $\phi = \pm\pi/2$ for which ϕ is constant. These flows give rise to two flows from the index 2 critical point to one of the index 1 critical points, two flows from one index 1 critical point to the other, and two flows from the lower index 1 critical point to the index 0 critical point. The other flows are in a one-parameter family of flows which go from the index 2 critical point to the index 0 critical point.

Exercise 4.5 *Work out the details of the above examples. Find the closed form solutions to the gradient flow equations and find which critical points they connect to.*

4.4 Basic properties of gradient flow lines

Lemma 4.5 *The function $f : M \longrightarrow \mathbb{R}$ is nonincreasing along flow lines. f is strictly decreasing along any flow line which does not contain a critical point.*

Proof: Let $\gamma : (a,b) \longrightarrow M$ be a flow line. Consider the composition $f \circ \gamma : (a,b) \longrightarrow \mathbb{R}$. Its derivative is given by

$$\begin{aligned}\frac{d}{dt} f(\gamma(t)) &= \langle \nabla_{\gamma(t)}(f), \frac{d\gamma(t)}{dt} \rangle \\ &= \langle \nabla_{\gamma(t)}(f), -\nabla_{\gamma(t)}(f) \rangle \\ &= -\left|\nabla_{\gamma(t)}(f)\right|^2 \leq 0.\end{aligned}$$

The only way this can be zero is if $\gamma(t)$ is on a critical point of f. In particular, if $\gamma(t)$ does not contain in its image a critical point of f, then $f(\gamma(t))$ is strictly decreasing.

□

Remark 4.4 *In the above proof, we showed*

$$\frac{d}{dt} f(\gamma(t)) = -\left|\nabla_{\gamma(t)}(f)\right|^2.$$

We can also show

$$\begin{aligned}\frac{d}{dt} f(\gamma(t)) &= \langle \nabla_{\gamma(t)}(f), \frac{d\gamma(t)}{dt} \rangle \\ &= \langle -\frac{d\gamma(t)}{dt}, \frac{d\gamma(t)}{dt} \rangle \\ &= -\left|\frac{d\gamma(t)}{dt}\right|^2 \leq 0.\end{aligned}$$

and this would also prove that $f(\gamma(t))$ is nonincreasing.

Remark 4.5 *Now if $\gamma(t)$* does *contain a critical point p, then by example 4.1 the flow must be a constant flow, and $f(\gamma(t))$ is constant on this flow.*

Thus there are two kinds of flow lines: constant flows that stay at a critical point, and flows that descend for all t, and do not contain a critical point.

Theorem 4.6 *Suppose that M is a closed manifold. Then given any $x \in M$ there is a unique flow line defined on entire real line*

$$\gamma_x : \mathbb{R} \longrightarrow M$$

that satisfies the initial condition

$$\gamma_x(0) = x.$$

Furthermore the limits

$$\lim_{t \to -\infty} \gamma_x(t) \quad \textit{and} \quad \lim_{t \to +\infty} \gamma_x(t)$$

converge to critical points of f. These are referred to as the starting *and* ending points *of the flow γ_x.*

The flow map

$$T : M \times \mathbb{R} \longrightarrow M$$

defined by $T(x,t) = \gamma_x(t)$ is smooth.

Proof: Let $x \in M$. By the existence and uniqueness of solutions to ordinary differential equations, there is an $\epsilon > 0$ and a unique path

$$\gamma_x : (-\epsilon, \epsilon) \longrightarrow M$$

satisfying the flow equation

$$\frac{d\gamma_x(t)}{dt} + \nabla_{\gamma_x(t)}(f) = 0$$

for all $|t| < \epsilon$, and the initial condition $\gamma_x(0) = x$. By the compactness of M we can choose a uniform ϵ for all $x \in M$. Notice therefore that for $|t| < \epsilon$ we can define a self map of M,

$$\gamma_t : M \longrightarrow M$$

by the formula $\gamma_t(x) = \gamma_x(t)$. Notice that $\gamma_0 = id$, the identity map. By uniqueness it is clear that

$$\gamma_{t+s} = \gamma_t \circ \gamma_s$$

providing that $|t|$, $|s|$, $|t+s| < \epsilon$. Among other things this implies that each γ_t is a *diffeomorphism* of M because $\gamma_t^{-1} = \gamma_{-t}$.

Now suppose that $|t| \geq \epsilon$. Write $t = k(\epsilon/2) + r$ where $k \in \mathbb{Z}$ and $|r| < \epsilon/2$. If $k \geq 0$ we define

$$\gamma_t = \gamma_{\frac{\epsilon}{2}} \circ \gamma_{\frac{\epsilon}{2}} \circ \ldots \gamma_{\frac{\epsilon}{2}} \circ \gamma_r$$

where the map $\gamma_{\frac{\epsilon}{2}}$ is repeated k times. If $k < 0$ then replace $\gamma_{\frac{\epsilon}{2}}$ by $\gamma_{\frac{-\epsilon}{2}}$. Thus for every $t \in \mathbb{R}$ we have a map $\gamma_t : M \longrightarrow M$ satisfying $\gamma_t \circ \gamma_s = \gamma_{t+s}$, and hence each γ_t is a diffeomorphism.

The curves

$$\gamma_x : \mathbb{R} \longrightarrow M$$

defined by $\gamma_x(t) = \gamma_t(x)$ clearly satisfy the flow equations and the initial condition $\gamma_x(0) = x$. This means that the gradient flow equations can be solved for all $t \in \mathbb{R}$, and in particular, we will from now on require that gradient flow lines be defined as functions $\gamma : \mathbb{R} \longrightarrow M$ instead of being defined only on an open interval.

Now let γ be a flow line. Consider the composition $f \circ \gamma : \mathbb{R} \longrightarrow \mathbb{R}$. By the Fundamental Theorem of Calculus, if $a < b$, then

$$(f \circ \gamma)(b) - (f \circ \gamma)(a) = \int_a^b \frac{d}{dt}(f \circ \gamma)(t)\, dt.$$

Since M is compact $f \circ \gamma$ has bounded image, so the left side is bounded. By Lemma 4.5, $\frac{d}{dt}(f \circ \gamma) < 0$. Therefore

$$\lim_{t \to \pm\infty} \frac{d}{dt}(f \circ \gamma)(t) = 0.$$

By the proof of Lemma 4.5 we know that

$$0 = \lim_{t \to \pm\infty} \frac{d}{dt} f(\gamma(t)) = \lim_{t \to \pm\infty} -\left|\nabla_{\gamma(t)}(f)\right|^2.$$

Let U be any union of small disjoint open balls around the critical points. By the compactness of M, $M - U$ is compact, so $|\nabla_x(f)|^2$ has a minimum value on $M - U$. Since $M - U$ has no critical points, this minimum value is strictly positive. But since the above limit is zero, we know that for sufficiently large $|t|$, $\gamma(t) \in U$. Since the balls are disjoint and $\gamma(t)$ is continuous, there is a critical point p so that for any open ball around p, $\gamma(t)$ is in that ball for sufficiently large t. Therefore $\lim_{t\to\infty} \gamma(t)$ exists and is equal to p; similarly, $\lim_{t\to-\infty} \gamma(t)$ exists and is equal to a critical point.

The differentiability of the flow map $T(x,t) = \gamma_x(t)$ with respect to t follows because $\gamma_x(t)$ satisfies the differential equation. The differentiability of T with respect to x follows from Peano's theorem (the differentiable dependence of solutions to ODEs with respect to initial conditions). This is proved in Hartman's book on ODEs [**?**] in chapter V, Theorem 3.1. □

4.5 Height-parameterized Gradient Flow Lines

Let $\gamma(t)$ be a non-constant gradient flow line from p to q. Then by Lemma 4.5, we know that $h(t) = f(\gamma(t))$ is strictly decreasing, and in particular, is a diffeomorphism from $\mathbb{R}$ to the open interval $(f(q), f(p))$. We can therefore consider the smooth curve $\eta(t) = \gamma(h^{-1}(t))$ from $(f(q), f(p))$ to M. Then it is easy to check that $f(\eta(t)) = t$. So γ and η have the same image, but the parameter in η represents height (that is, the value of f).

Exercise 4.6 *Prove that $f(\eta(t)) = t$ as claimed above.*

We can also extend η to a continuous map from the closed interval $[f(q), f(p)]$ to M by defining $\eta(f(q)) = q$ and $\eta(f(p)) = p$.

Exercise 4.7 *Prove that the extension of η to the closed interval $[f(q), f(p)]$ is continuous.*

Definition 4.7 *If $\gamma(t)$ is a non-constant gradient flow line for f, and $h(t) = f(\gamma(t))$, then*

$$\eta(t) = \gamma(h^{-1}(t)) : [f(q), f(p)] \longrightarrow \mathbb{R}$$

is the height-reparameterization *of γ, and such a curve is a* height-parameterized gradient flow *of f.*

Remark 4.6 *This reparameterization of γ is a direction-reversing one, since h is strictly decreasing. This is to be expected since $f(\gamma(t))$ is decreasing but $f(\eta(t) = t$ is increasing.*

We now differentiate η.

Exercise 4.8 *Prove*

$$\frac{d}{dt}\eta(t) = \frac{\nabla_{\eta(t)}(f)}{\left|\nabla_{\eta(t)}(f)\right|^2}$$

Therefore, $\eta(t)$ is the solution to another differential equation which may be described as follows:

Lemma 4.8 *Away fron the critical points of f, we may consider the vector field*

$$X(x) = \frac{\nabla_x(f)}{|\nabla_x(f)|^2}.$$

Then we define curves $\zeta : (s_1, s_2) \longrightarrow M$ that satisfy

$$\frac{d}{dt}\zeta(t) = X(\zeta(t))$$

Then ζ is a height-reparameterized flow line.

Proof: We insist that (s_1, s_2) be maximal. We then can show that $\frac{d}{dt}f(\zeta(t)) = 1$ as usual (do this now if you wish). Pick a number $s \in (s_1, s_2)$, and consider the gradient flow line $\gamma(t)$ so that $\gamma(0) = \zeta(s)$. We do the height-reparameterization to γ to get a height-reparameterized curve η. Now η satisfies the same differential equation as ζ, and $\eta(f(\zeta(s))) = \zeta(s)$, so we translate the domain as follows: $\eta_0(t) = \eta(t + f(\zeta(s)) - s)$ satisfies the same differential equation as ζ and $\eta_0(s) = \zeta(s)$ so by the uniqueness of solutions to ODEs, $\eta_0 = \zeta$.

Therefore solutions to $\frac{d}{dt}\zeta(t) = X(\zeta(t))$ are precisely those that are height-parameterized flows. □

Therefore $X(x)$ and $\nabla(f(x))$ have the same integral curves, although with different parameterizations.

Chapter 5

The Classical Approach to Morse Theory

In the introduction (section 0.2), we mentioned that a Morse function on a closed manifold produces a CW complex decomposition of the manifold, with a cell of dimension λ for each critical point of index λ of f. In this chapter, we prove this statement up to homotopy. That is, we construct a homotopy equivalence of the manifold to a CW complex of the kind just described. This is the original approach to Morse theory by Morse [**?**] and we follow the approach of Milnor [**?**] in this chapter.

Throughout this chapter, we will assume M is a closed manifold and $f : M \longrightarrow \mathbb{R}$ is a smooth Morse function. We will also consider the following manifolds (with boundary):

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

where a is any real number. If a is less than the minimum value of f, then M^a is the empty set. If a is larger than the maximum value of f, then M^a is M. The values of a in between will provide, up to homotopy, the necessary cell decomposition.

There are a number of technical details, but the intuition is simple: Let M be a surface embedded in $\mathbb{R}^3$, and f be the vertical coordinate z. We initially let a be less than the minimum value of f so that $M^a = \emptyset$, and gradually increase a (see Figure 5.1). This is analogous to gradually filling the surface with water, so that M^a is the part of the surface that is under water. Now if a increases from a_1 to a_2 without passing through critical values, then M^{a_1} and M^{a_2} are diffeomorphic (see Figure 5.2). But if, by increasing from a_1 to a_2, we pass through one critical point, then at that point the water may do something more interesting. Up to homotopy, this turns out to be an attaching of a cell of dimension λ, where λ is the index of the critical point (see Figure 5.3).

So as we pass critical points one by one, the manifold is created by successively attaching cells (up to homotopy type). This demonstrates that the

Figure 5.1: M^a for different values of a

Figure 5.2: M^{a_1} and M^{a_2} are diffeomorphic if there are no critical values between a_1 and a_2.

Figure 5.3: When there is one critical value between a_1 and a_2, M^{a_2} is homotopy equivalent to M^{a_1} with a cell attached.

manifold is homotopy equivalent to a CW complex of the type described above.

In this chapter we prove the details of the above intuition. First we prove that nothing happens if there is no critical point between two levels, using the results of gradient flow lines from chapter 4. Then we show that if there is one critical point between the two levels, the homotopy type changes by adding a cell. We prove this via the Morse Lemma (Theorem 5.3), which studies the behavior of f near a critical point. We conclude by producing the homotopy equivalence between the manifold and the CW complex, and giving some interesting applications to topology.

Exercise 5.1 *Let M be a manifold and let $f : M \longrightarrow \mathbb{R}$ be a Morse function. Prove that $f^{-1}(\{a\})$, the boundary of M_a, is a manifold if a is a regular value of f.*

5.1 The Regular Interval Theorem

We first show that if we increase M^a from M^{a_1} to M^{a_2}, and there are no critical values between a_1 and a_2, then M^{a_1} and M^{a_2} are diffeomorphic.

The main point is the following theorem:

Theorem 5.1 (Regular interval theorem) *Let $f : M \longrightarrow [a, b]$ be a smooth map on a compact Riemannian manifold with boundary. Suppose that f has no critical points and that $f(\partial M) = \{a, b\}$. Then there is a diffeomorphism*

$$F : f^{-1}(a) \times [a, b] \longrightarrow M$$

making the following diagram commute:

$$\begin{array}{ccc} f^{-1}(a) \times [a,b] & \xrightarrow{F} & M \\ {\scriptstyle proj.}\downarrow & & \downarrow{\scriptstyle f} \\ [a,b] & \xrightarrow[=]{} & [a,b]. \end{array}$$

In particular all the level surfaces are diffeomorphic.

Proof: Since f has no critical points we may consider the vector field

$$X(x) = \frac{\nabla_x(f)}{|\nabla_x(f)|^2}.$$

defined in Lemma 4.8. Let $\eta_x(t)$ be a curve through x satisfying

$$\frac{d}{dt}\eta_x(t) = X(\eta_x(t))$$

and $f(\eta_x(t)) = t$.

Let I be a maximal interval on which η_x is defined. We wish to show that $I = [a, b]$. First, since M is compact, $f(\eta_x(I)) = I$ is bounded.

Let $d = \sup(I)$. Then by the compactness of M, there is a point $x \in M$ that is a limit point of $\eta_x(d - 1/n)$. Since $\eta_x'(t) = X(\eta_x(t))$ is bounded, this limit point is unique, and $\lim_{t\to d^-} \eta_x(t) = x$. We can extend η_x to d by making $\eta_x(d) = x$.

Now $\lim_{t\to d} \eta_x'(t) = \lim_{t\to d} X(\eta_x(t)) \to X(\eta_x(d))$, and let v be this limit. We will now show that $\eta_x'(d) = v$. In particular, we will show that for every $\epsilon > 0$, there exists a $\delta > 0$ so that for all h with $0 < h < \delta$,

$$\left| \frac{\eta_x(d) - \eta_x(d-h)}{h} - v \right| < \epsilon.$$

Note that a coordinate chart is chosen near $\eta_x(d)$ to allow the subtraction here.

So let $\epsilon > 0$ be given. By the definition of v, there exists a δ_1 so that for all w with $0 < h < \delta_1$,

$$|\eta_x'(d-h) - v| < \epsilon$$

By the fundamental theorem of calculus,

$$\begin{aligned}
\eta_x(d-h) - \eta_x(d) &= \int_{d-h}^{d} \eta_x'(t)\,dt \\
\eta_x(d-h) - \eta_x(d) + vh &= \int_{d-h}^{d} (\eta_x'(t) - v)\,dt \\
|\eta_x(d-h) - \eta_x(d) + vh| &\le \int_{d-h}^{d} |\eta_x'(t) - v|\,dt \\
&\le \int_{d-h}^{d} \epsilon\,dt \\
&\le \epsilon h \\
\left| \frac{\eta_x(d-h) - \eta_x(d)}{h} + v \right| &\le \epsilon \\
\left| \frac{\eta_x(d-h) - \eta_x(d)}{-h} - v \right| &\le \epsilon
\end{aligned}$$

Therefore $\eta_x'(d) = v$, and since $v = X(\eta_x(d))$, the flow equation is satisfied by η_x at d.

By maximality of I, $d \in I$. Similarly with $c = \inf(I)$, we see that $c \in I$. Therefore I is closed.

If $\eta_x(s) \notin \partial M$, then by the existence of solutions of ODEs, there is an interval $(s - \epsilon, s + \epsilon)$ around s on which η_x satisfies the differential equation $\eta_x'(t) = X(\eta_x(t))$. Therefore $\eta_x(c)$ and $\eta_x(d)$ are in ∂M. Thus $c = f(\eta_x(c))$ and $d = f(\eta_x(d))$ may be either a or b. Since the derivative of $f \circ \eta_x$ is one, we see that $c = a$ and $d = b$. Therefore $I = [a, b]$.

Since $x \in M$ was arbitrary, and $a \le f(x) \le b$, we see that $f(M) = [a, b]$. Furthermore, if $x \notin \partial M$, then by the existence of solutions to ODEs, as above,

we have η_x defined in a small neighborhood of $t = f(x)$, so that $a < f(x) < b$. Therefore $f^{-1}(a)$ and $f^{-1}(b)$ are unions of boundary components.

Define a map

$$F : f^{-1}(a) \times [a, b] \longrightarrow M$$

by the formula

$$F(x, t) = \eta_x(t).$$

The differentiability of F follows from the same argument as in Theorem 4.6 to prove the differentiability of T, but with η_x instead of γ_x.

Define

$$G : M \longrightarrow f^{-1}(a) \times [a, b]$$

as

$$G(x) = (\eta_x(a), f(x)).$$

The differentiability of G follows in the same way as the differentiability of F. We claim that F and G are inverses. To prove this, note that the integral curves through x and $\eta_x(t)$ are the same, that $f(\eta_x(t)) = t$ and by uniqueness of solutions to ODEs, we have $F(G(x)) = x$ and $G(F(x, t)) = (x, t)$. This proves that F is a diffeomorphism. □

Corollary 5.2 *Let M be a compact manifold, and $f : M \longrightarrow \mathbb{R}$ a smooth Morse function. Let $a < b$ and suppose that $f^{-1}[a, b] \subset M$ contains no critical points. Then M^a is diffeomorphic to M^b. Furthermore, M^a is a deformation retract of M^b.*

Proof: First we prove that M^a is a deformation retract of M^b. By the regular interval theorem (Theorem 5.1), there is a natural diffeomorphism F from $f^{-1}([a, b])$ to $f^{-1}(a) \times [a, b]$. Since $f^{-1}(a) \times \{a\}$ is a deformation retract of $f^{-1}(a) \times [a, b]$, we see that $f^{-1}(a)$ is a deformation retract of $f^{-1}([a, b])$. We can now paste this deformation retraction with the identity on M_a to obtain the deformation retracton from M_b to M_a.

To prove that M^a is diffeomorphic to M^b we apply the same principle, but we need to be more careful to preserve smoothness during the patching process.

Since the set of critical points of f is a closed subset of the compact set M (and hence is compact), the set of critical values of f is compact. Therefore there are real numbers c and d with $c < d < a$ so that there are no critical values in $[c, b]$.

By Theorem 5.1 there is a natural diffeomorphism F from $f^{-1}([c, b])$ to $f^{-1}(c) \times [c, b]$, that maps $f^{-1}([c, a])$ diffeomorphically onto $f^{-1}(c) \times [c, a]$. There is also a diffeomorphism $H : f^{-1}(c) \times [c, b] \longrightarrow f^{-1}(c) \times [c, a]$, and we can insist that it be the identity on $f^{-1}(c) \times [c, d]$ (finding this function is an easy exercise in one-variable analysis, and in case you are interested, is listed as an exercise below). Thus

$$F^{-1} \circ H \circ F : f^{-1}([c, b]) \longrightarrow f^{-1}([c, a])$$

is a diffeomorphism that is the identity on $f^{-1}([c, d])$, and thus we can patch it together with the identity on M_d to create a diffeomorphism from M_b to M_a. □

This corollary says that the topology of the submanifolds M^a does not change with $a \in \mathbb{R}$ so long as a does not pass through a critical value.

Exercise 5.2 *Fill in the detail of the proof of Corollary 5.2 that finds a diffeomorphism $H : f^{-1}(c) \times [c, b] \longrightarrow f^{-1}(c) \times [c, a]$ that is the identity on $f^{-1}(c) \times [c, d]$.*

5.2 Passing through a critical value

We now examine what happens to the topology of these submanifolds when one does pass through a critical value. For this, we will need to understand the function f in the neighborhood of a critical point. This is what the Morse lemma provides us:

Theorem 5.3 (Morse Lemma) *Let p be a nondegenerate critical point of index λ of a smooth function $f : M \longrightarrow \mathbb{R}$, where M is an n-dimensional manifold. Then there is a local coordinate system $(x_1, \ldots, x_n)$ in a neighborhood U of p with $x_i(p) = 0$ with respect to which*

$$f(x_1, \ldots, x_n) = f(p) - \sum_{i=1}^{\lambda} x_i^2 + \sum_{j=\lambda+1}^{n} x_j^2.$$

The proof given here is essentially that in Milnor's famous book on Morse theory [**?**].

Proof: Since this is a local theorem we might as well assume that $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ with a critical point at the origin, $p = 0$. We may also assume without loss of generality that $f(0) = 0$. Given any coordinate system for $\mathbb{R}^n$ we can therefore write

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

for $(x_1, \ldots, x_n)$ in a neighborhood of the origin. In this expression we have

$$g_j(x_1, \ldots, x_n) = \int_0^1 \frac{\partial f}{\partial x_j}(tx_1, \ldots, tx_n)dt.$$

Now since 0 is a critical point of f, each $g_j(0) = 0$, and hence we may write it in the form

$$g_j(x_1, \ldots, x_n) = \sum_{i=0}^{n} x_i h_{i,j}(x_1, \ldots, x_n).$$

Let $\phi_{i,j} = (h_{i,j} + h_{j,i})/2$. Hence we can combine these equations and write

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

where $(\phi_{i,j})$ is a symmetric matrix of functions. By doing a straightforward calculation one sees furthermore that the matrix

$$(\phi_{i,j}(0)) = \left(\frac{1}{2}\frac{\partial^2 f}{\partial x_i \partial x_j}(0)\right)$$

and hence by the nondegeneracy assumption is nonsingular. From linear algebra we know that symmetric matrices can be diagonalized. The Morse lemma will be proved by going through the diagonalization process with the representation of f as $\sum x_i x_j \phi_{i,j}$.

Assume inductively that there is a neighborhood U_k of the origin and coordinates $\{u_1, \ldots, u_n\}$ with respect to which

$$f = \pm(u_1)^2 \pm \cdots \pm (u_k)^2 + \sum_{i,j \geq k+1} u_i u_j \psi_{i,j}(u_1, \ldots, u_n)$$

where $(\psi_{i,j})$ is a symmetric, $n-k \times n-k$ matrix of functions. By a linear change in the last $n-k$ coordinates if necessary, we may assume that $\psi_{k+1,k+1}(0) \neq 0$.

Let

$$\sigma(u_1, \ldots, u_n) = \sqrt{|\psi_{k+1,k+1}(u_1, \ldots, u_n)|}$$

in perhaps a smaller neighborhood $V \subset U_k$ of the origin. Now define new coordinates

$$v_i = u_i \quad \text{for } i \neq k+1$$

and

$$v_{k+1}(u_1, \ldots u_n) = \sigma(u_1, \ldots, u_n)\left[u_{k+1} + \sum_{i=k+2}^{n} u_i \frac{\psi_{i,k+1}(u_1, \ldots, u_n)}{\psi_{k+1,k+1}(u_1, \ldots, u_n)}\right].$$

The v_i's give a coordinate system in a sufficiently small neighborhood U_{k+1} of the origin. Furthermore a direct calculation verifies that with respect to this coordinate system

$$f = \sum_{i=1}^{k+1} \pm(v_i)^2 + \sum_{i,j=k+2}^{n} v_i v_j \theta_{i,j}(v_1, \ldots, v_n)$$

where $(\theta_{i,j})$ is a symmetric matrix of functions. This completes the inductive step. The only remaining point in the theorem is to observe that the number of negative signs occuring in the expression for f as a sum and difference of squares is equal to the number of negative eigenvalues (counted with multiplicity) of $Hess_0(f)$ which does not depend on the particular coordinate system used. □

Remark 5.1 *The Morse Lemma describes the behavior of the function f near a critical point, but it does not describe the behavior of the gradient flow lines. The reason for this is that the gradient depends on the Riemannian metric, and if we use the coordinate system given by the Morse Lemma, we do not know how this metric behaves. See section 6.2 in chapter refch:cw.*

Corollary 5.4 *If M is a manifold and $f : M \longrightarrow \mathbb{R}$ is Morse, then the set of critical points of f is a discrete subset of M.*

Proof: Suppose there were a sequence of critical points x_n converging to some point $a \in M$. Since df is a continuous one-form on M, we know that a is a critical point of f. Then apply the Morse Lemma above to a, which gives a formula for f in a neighborhood of a. But there are no critical points in this neighborhood as can be seen directly by calculating df in these coordinates. This is a contradiction. □

Exercise 5.3 *Prove the converse of Exercise 5.1; that is, if M is a compact manifold and $f : M \longrightarrow \mathbb{R}$ is a Morse function, and if a is not a regular value of f, then $f^{-1}(\{a\})$ is not a manifold.*

Definition 5.5 *Let $f : M \longrightarrow [a, b]$ be a Morse function on a compact manifold. We say that f is* admissible *if $\partial M = f^{-1}(a) \cup f^{-1}(b)$, where a and b are regular values. This implies that each of $f^{-1}(a)$ and $f^{-1}(b)$ are unions of connected components of ∂M.*

Theorem 5.6 *Let $f : M \longrightarrow \mathbb{R}$ be an admissible Morse function on a compact manifold. Suppose f has a unique critical point z of index λ. Say $f(z) = c$. Then there exists a λ - dimensional cell D^λ in the interior of M with $D^\lambda \cap f^{-1}(a) = \partial D^\lambda$, and there is a deformation retraction of M onto $f^{-1}(a) \cup D^\lambda$.*

Proof: [Proof, following [**?**], with a few errors corrected] By replacing f by $f(x) - c$ we can assume that $f(z) = 0$. Notice that by the regular interval theorem Theorem 5.1 it is sufficient to prove the theorem for the restriction of f to the inverse image of any closed subinterval of $[a, b]$ around $c = 0$.

Let (ϕ, U) be an chart around z with respect to which the Morse lemma is satisfied. Write $\mathbb{R}^n = \mathbb{R}^\lambda \times \mathbb{R}^{n-\lambda}$. ϕ maps U diffeomorphically onto an open set $V \subset \mathbb{R}^\lambda \times \mathbb{R}^{n-\lambda}$, and

$$f \circ \phi^{-1}(x, y) = -|x|^2 + |y|^2.$$

Notice that $\phi(z) = (0, 0)$. Put $g(x, y) = -|x|^2 + |y|^2$.

We will use gradient flows, which depend on the metric on M. We choose a metric for M by pulling back the flat metric on $\mathbb{R}^n$ by ϕ, and extending the metric arbitrarily to the rest of M. In this way, ϕ will be a local isometry, and

$$D\phi(u)(\nabla_u(f)) = \nabla_v(g),$$

for any $u \in U$ such that $\phi(u) = v \in V$.

Let $0 < \delta < 1$ be such that V contains $\Lambda = B^\lambda(\delta) \times B^{n-\lambda}(\delta)$ where

$$B^i(\delta) = \{x \in \mathbb{R}^i \mid \sum_{j=1}^{n} x_j^2 \leq \delta$$

is the closed coordinate ball around the origin of radius δ.

Let $\epsilon > 0$ be small enough that $\sqrt{4\epsilon} < \delta$, and let

$$c^\lambda = B^\lambda\left(\sqrt{\epsilon}\right) \times \{0\} \subset V$$

and we define

$$D^\lambda = \phi^{-1}(c^\lambda) \subset M.$$

A deformation of $f^{-1}[-\epsilon, \epsilon]$ to $f^{-1}(\epsilon) \cup D^\lambda$ is made by patching together two deformations. First consider the set

$$\Lambda_1 = B^\lambda\left(\sqrt{\epsilon}\right) \times B^{n-\lambda}\left(\sqrt{2\epsilon}\right).$$

Consider the following figure for the case $\lambda = 1$, $n = 2$.

Note that inside Λ_1, $f(x, y) = -|x|^2 + |y|^2 > -\epsilon + |y|^2 > -\epsilon$. Furthermore, since $x \in B^\lambda\left(\sqrt{\epsilon}\right)$, we have that $(x, 0) \in c^\lambda$.

In $\Lambda_1 \cap g^{-1}[\epsilon, \epsilon]$ a deformation is obtained by moving (x, y) at constant speed along the interval joining (x, y) to the point $(x, 0) \in g^{-1}(-\epsilon) \cup B^\lambda$, by $(x, (1-t)y)$. This deformation then induces a deformation of $\phi^{-1}(\Lambda_1)$.

Outside the set

$$\Lambda_2 = B^\lambda(\sqrt{2\epsilon}) \times B^{n-\lambda}(\sqrt{3\epsilon})$$

the deformation moves each point along the vector field $-\nabla(g)$ so that it reaches $g^{-1}(-\epsilon)$ in unit time. (The speed of each point is chosen to equal the length of its path under the deformation.) See the following figure for a pictorial description of this deformation.

This deformation is transported to $U - \phi^{-1}(\Lambda_2)$ by ϕ, and is then extended over $M - \phi^{-1}(\Lambda_2)$ by following the gradient flow lines of f.

Now if such a flow enters V, we now show it may not enter Λ_2: Suppose we have a flow that enters V from the outside at time t. Then since the closure of Λ_2 is in V, there is a time arbitrarily close to t where the point is (x, y) which is not in Λ_2. Then at this time either $|x|^2 > 2\epsilon$ or $|y|^2 > 3\epsilon$. But if $|y|^2 > 3\epsilon$ then because for$g^{-1}([-\epsilon, \epsilon])$, we have $\epsilon > -|x|^2 + |y|^2 > -|x|^2 + 3\epsilon$ so that $|x|^2 > 2\epsilon$. Therefore, either way, $|x|^2 > 2\epsilon$. But for x non-zero, $|x|$ increases along flow lines. Therefore (x, y) will not be in Λ_2 for any later time until it leaves V (and by repeating the argument for future visits to V, it never enters Λ_2).

In $f^{-1}([-\epsilon, \epsilon]) - \phi^{-1}(\Lambda_2)$, then, the downward gradient flow is defined, and since we assume there are no other critical points than z, the methods of the proof of Theorem 5.1 show that the flows defined there flow downward to $f^{-1}(-\epsilon)$.

On $f^{-1}([-\epsilon, \epsilon]) - \phi^{-1}(\Lambda_2)$, then, we can define the deformation to flow along the gradient flow with constant speed, with speed equal to the length of the flow line from the point to its destination on $f^{-1}(-\epsilon)$. In this way, after unit time, everything in $f^{-1}([-\epsilon, \epsilon]) - \phi^{-1}(\Lambda_2)$ is deformed into $f^{-1}(-\epsilon)$.

To extend the deformation to points of $\Lambda_2 - \Lambda_1$ it suffices to find a vector field on Λ which agrees with X in Λ_1 and with $-\nabla(g)$ in $\Lambda - \Lambda_2$. Such a vector field is

$$Y(x, y) = 2(\mu(x, y)x, -y)$$

where the map $\mu : \mathbb{R}^\lambda \times \mathbb{R}^{n-\lambda} \longrightarrow [0, 1]$ vanishes in Λ_1 and equals 1 outside Λ_2. The fact that each integral curve of Y which starts at a point of

$$(\Lambda_2 - \Lambda_1) \cap g^{-1}[-\epsilon, \epsilon]$$

must reach $g^{-1}(-\epsilon)$ because $|x|$ is nondecreasing along integral curves.

The global deformation of $f^{-1}[-\epsilon, \epsilon]$ into $f^{-1}(-\epsilon) \cup D^\lambda$ is obtained by moving each point of Λ at constant speed along the flow line of Y until it reaches $g^{-1}(-\epsilon) \cup B^\lambda$ in unit time and transporting this motion to M via ϕ; while each point of $M - \phi^{-1}(\Lambda)$ moves at constant speed along the flow line of $\nabla(f)$ until it reaches $f^{-1}(-\epsilon)$ in unit time. Points on $f^{-1}(-\epsilon) \cup D^\lambda$ stay fixed. □

5.3 Homotopy equivalence to a CW complex

Theorem 5.7 *Let M be a closed manifold, and $f : M \longrightarrow \mathbb{R}$ a Morse function on M. Then M has the homotopy type of a CW complex, with one cell of dimension λ for each critical point of index λ.*

Proof: Without loss of generality, the critical points of f all have different values under f (if $f(p) = f(q)$ and p and q are critical points, then let $B_1 \subset B_2$ be balls around q small enough that in $B_2 - B_1$, we have $|\nabla f|$ bounded away from zero by some ϵ, and add a small bump function to f supported in B_2 and

constant in B_1 whose gradient is bounded above by ϵ, and which does not raise the value of $f(q)$ high enough to reach another critical value of f).

Now let $a_0 < \cdots < a_k$ be a sequence of real numbers so that a_0 is less than the minimum value of f, a_k is greater than the maximum value of f, and between a_i and a_{i+1} there is exactly one critical point. By Theorem 5.6 we have a homotopy equivalence h_i between $M^{a_{i+1}}$ and $M^{a_i} \cup D^{\lambda_i}$ (where the union is an attaching map as in a CW complex). By composing the h_i's, we obtain a homotopy equivalence from $M = M^{a_k}$ to a union of disks attached by CW attaching maps.

□

Corollary 5.8 *Given $f : M \longrightarrow \mathbb{R}$ as above there is a chain complex referred to as the* Morse–Smale complex

$$\ldots \longrightarrow C_\lambda \xrightarrow{\partial_\lambda} C_{\lambda-1} \longrightarrow \ldots \xrightarrow{\partial_1} C_0 \tag{5.1}$$

whose homology is $H_*(M;\mathbb{Z})$, where C_λ is the free abelian group generated by the critical points of f of index λ.

Proof: This is the cellular chain complex coming from the CW complex in Theorem 5.7. □

We can now prove some of the results promised in the introduction, that relate the topology of M to the numbers of critical points of f:

Corollary 5.9 (Morse's Theorem) *Let $f : M \longrightarrow \mathbb{R}$ be a C^∞ function so that all of its critical points are nondegenerate. Then the Euler characteristic $\chi(M)$ can be computed by the following formula:*

$$\chi(M) = \sum (-1)^i c_i(f)$$

where $c_i(f)$ is the number of critical points of f having index i.

Proof: The Euler characteristic $\chi(M)$ can be computed as the alternating sum of the ranks of the chain groups of any CW decomposition of M. □

Corollary 5.10 (Weak Morse Inequalities) *Let c_p be the number of critical points of index p and let β_p be the rank of the homology group $H_p(M)$. Then*

$$\beta_p \leq c_p.$$

Proof: The chain group $C_p \otimes \mathbb{R}$ generated by the c_p cells of dimension p is a vector space of dimension c_p. The group of cycles is of dimension at most c_p. After quotienting by the boundaries, we see that $H_p(M;\mathbb{R})$ is a vector space of dimension at most c_p. □

Corollary 5.11 (Strong Morse Inequalities) *Let M, f, $c_i(f)$, and $b_i(M)$ be as above. Then for all natural numbers i,*

$$\sum_{k=0}^{i}(-1)^{i-k}c_i(f) \geq \sum_{k=0}^{i}(-1)^{i-k}b_i(M).$$

Proof: The proof is similar except we take a closer look at the boundaries. Tensoring the chains with $\mathbb{R}$, so that we write $V_k = C_k \otimes \mathbb{R}$, we get the following chain complex of vector spaces:

$$\ldots \longrightarrow V_i \xrightarrow{\partial_i} V_{i-1} \longrightarrow \ldots \xrightarrow{\partial_1} V_0$$

We write V_k as $Im(\partial_{k+1}) \oplus H_k(M;\mathbb{R}) \oplus (V_k/\ker(\partial_k))$ and note that $Im(\partial_{k+1})$ is of the same dimension as $V_{k+1}/\ker(\partial_{k+1})$. Thus if we define d_k to be the dimension of $V_k/\ker(\partial_k)$, we have

$$c_k = d_{k+1} + b_k + d_k$$

and applying the alternating sum above we get

$$\sum_{k=0}^{i}(-1)^{i-k}c_i(f) = d_{i+1} + \sum_{k=0}^{i}(-1)^{i-k}b_i(M)$$

(where here we need that $d_0 = 0$). This proves the strong Morse inequalities. □

To see that the strong Morse inequalities prove the weak Morse inequalities, write down the strong Morse inequality for i and for $i+1$, and subtract the two inequalities. To see that the strong Morse inequalities imply Morse's theorem, apply the strong Morse inequality for i and for $i+1$ for i larger than the dimension of the manifold M, noting that $c_j = 0$ and $b_j = 0$ for all $j > \dim(M)$.

It is instructive to work out the following:

Exercise 5.4 *Show that the strong Morse inequalities is "strictly stronger" than the weak Morse inequalities together with Morse's theorem. What I mean is: given the $n+1$-tuple of natural numbers $(b_0, \ldots, b_n)$, we can find another $n+1$-tuple of natural numbers $(c_0, \ldots, c_n)$ so that these numbers satisfy the weak Morse inequality and the Morse theorem but not the strong Morse inequalties.*

A typical application of these result is to use homology calculations to deduce critical point data. For example we have the following.

Application 5.12 *Every Morse function on the complex projective space*

$$f : \mathbb{CP}^n \longrightarrow \mathbb{R}$$

has at least one critical point in every even dimension $\leq 2n$.

The following is a historically important application of Morse theory, due to Reeb, that follows from the techniques we have mentioned so far.

Application 5.13 *Let M^n be a closed manifold admitting a Morse function*

$$f : M \longrightarrow \mathbb{R}$$

with only two critical points. Then M is homeomorphic to the sphere S^n.

Remark 5.2 *This theorem does* not *imply that M is diffeomorphic to S^n. In [?] Milnor found an example of a manifold that is homeomorphic, but not diffeomorphic to S^7. Indeed he proved that there are 28 distinct differentiable structures on S^7! Milnor actually used this fact to prove that the manifolds he constructed were homeomorphic to S^7.*

Proof: [Proof of Theorem 5.13] Let S and N be the critical points. By the compactness of M we may assume that S is a minimum and N is a maximum. (Think of them as the eventual south and north poles of the sphere.) Let $f(S) = t_0$ and $f(N) = t_1$. By the Morse lemma there are coordinates $(x_1, \ldots, x_n)$ in a neighborhood U_+ of N with respect to which f has the form

$$-x_1^2 + \cdots + -x_n^2 + t_1.$$

Therefore there is a $b < t_1$ so that if we let $D_+ = f^{-1}[b, t_1]$ then there is a diffeomorphism

$$D_+ \cong D^n$$

with $\partial D_+ = f^{-1}(b) \cong S^{n-1}$. Repeating this process with the minimum point P we obtain a point $a > t_0$ and a diffeomorphism of the space $D_- = f^{-1}[t_1, a]$,

$$D_- \cong D^n$$

with $\partial D_- = f^{-1}(a) \cong S^{n-1}$. By Theorem 5.1 we have that

$$f^{-1}[a, b] \cong f^{-1}(a) \times [a, b] \cong S^{n-1} \times [a, b].$$

Hence we have a decomposition of the manifold

$$\begin{aligned} M = f^{-1}[t_0, t_1] &= f^{-1}[t_0, a] \cup f^{-1}[a, b] \cup f^{-1}[b, t_1] \\ &\cong D^n \cup S^{n-1} \times [a, b] \cup D^n \end{aligned}$$

where the attaching maps are along homeomorphisms of S^{n-1}. We leave it as an exercise to now construct a homeomorphism from this manifold to S^n.

□

Exercise 5.5 *Finish the proof of Theorem 5.13 by showing that the resulting space*

$$D^n \cup S^{n-1} \times [a, b] \cup D^n$$

is homeomorphic to S^n. Hint: Start by embedding one D^n into S^n, then embed $S^{n-1} \times [a, b]$ into S^n to match the first embedding, then to put the last D^n in, you must think of D^n as the cone on S^{n-1}. This last part is why the proof does not prove that this is diffeomorphic to S^n.

In general, there are many applications of this work to the problem of classifying manifolds of dimensions 5 and higher, leading to the h-cobordism theorem and the s-cobordism theorem, and surgery theory. There are many books that describe these developments of the 1960s and 1970s, the old classics being Milnor's book on the h-cobordism theorem, which hides the Morse theoretic motivation [**?**], Wall's book on surgery theory [**?**], and Browder's book [**?**] that covers the same topics but is more readable because he restricts to the simply-connected case. More friendly introductions to the subject, recommended for students, are [**?**], [**?**] and [**?**].

Part II

Spaces of gradient flows

Chapter 6

Morse theory using unstable manifolds

The previous chapter described how any closed manifold M is homotopy equivalent to a CW complex given by a Morse function f. It would be nice to show more explicitly where the cells of the CW complex come from. In particular, it would be nice if we could describe the manifold as a CW complex directly, instead of referring to homotopy equivalences.

This is actually explained in chapter 10, but in this chapter we set the stage by decomposing M into open disks. Each of these open disks comes from a critical point of the function f.

6.1 Stable and unstable manifolds of a critical point

As before, for any point $x \in M$, let $\gamma_x(t)$ be the flow line through x, i.e. it satisfies the differential equation

$$\frac{d}{dt}\gamma = -\nabla_\gamma(f)$$

with the initial condition $\gamma(0) = x$. We know by Theorem 4.6 that $\gamma_x(t)$ tends to critical points of f as $t \to \pm\infty$. So for any critical point a of f we define the *stable manifold* $W^s(a)$ and the *unstable manifold* $W^u(a)$ as follows:

Definition 6.1 *Let M be a manifold, and f a smooth function on M. Let a be a critical point for f. We define the two subsets of M:*

$$W^s(a) = \{x \in M : \lim_{t\to+\infty} \gamma_x(t) = a\}$$
$$W^u(a) = \{x \in M : \lim_{t\to-\infty} \gamma_x(t) = a\}.$$

and call $W^s(a)$ the stable manifold *of a and $W^u(a)$ the* unstable manifold *of a.*

Figure 6.1: The stable and unstable manifolds of a critical point.

In other words, $W^s(a)$ is the set of points on M that flow down to a, and $W^u(a)$ is the set of points on M that would flow "up" to a if the gradient flow were reversed. The use of the term "manifold" is justified by the stable manifold theorem:

Theorem 6.2 (Stable Manifold Theorem) *Let M be an n-dimensional manifold, and $f : M \longrightarrow \mathbb{R}$ a Morse function. Let a be a critical point of f of index λ. Then $W^u(a)$ and $W^s(a)$ are smooth submanifolds diffeomorphic to the open disks D^λ and $D^{n-\lambda}$, respectively.*

This will be proved in Section 6.3 below for a large class of metrics (though it is in general true for all metrics).

Proposition 6.3 *If M is a compact manifold with Riemannian metric g, and $f : M \longrightarrow \mathbb{R}$ is a Morse function, then*

$$M = \bigcup_a W^u(a)$$

is a partition of M into disjoint sets, where the union is taken over all critical points a of f.

Proof: The fact that the union of the $W^u(a)$ is M comes from the fact that every point of M lies on a flow line γ, and we can always find $\lim_{t\to-\infty}\gamma(t)$.

The fact that the $W^u(a)$ and $W^u(b)$ are disjoint when $a \neq b$ is due to the fact that γ is unique. □

Exercise 6.1 *Find the unstable manifolds for each critical point in Example 4.4.*

Exercise 6.2 *Consider the contour drawing of a Morse function f shown in Figure 6.2. Imagine that the domain is S^2, and the contour drawing illustrates a coordinate patch of S^2 that contains all critical points of f except for one*

Figure 6.2: A contour drawing of a Morse function

minimum. Copy the contour drawing, and on this drawing, sketch the unstable manifolds for each critical point.

Assume that there are no other critical points except the minimum and those that are visually apparent by the contour drawing. Assume also that the metric in this coordinate patch is the flat Euclidean metric corresponding to the coordinates, so that the contour drawing may be viewed as on a region in $\mathbb{R}^2$.

Exercise 6.3 *Find the unstable manifolds for each critical point in Example 4.5.*

From these exercises you can see that this decomposition of M makes M look like a CW complex, with one cell of dimension λ for each critical point of index λ. The torus example is problematic because an edge gets attached to the middle of another edge, but consider the following fix:

Exercise 6.4 *Consider the torus in* $\mathbb{R}^3$ *as before, but with a slight perturbation. That is, tilt the torus by pulling it down slightly toward its hole, so that it is not standing on its outside edge, as in Figure 6.4. Find the unstable manifolds for each critical point here.*

The point is that with this example, we have a decomposition of M into cells, with a cell of dimension λ for each critical point of index λ. These are essentially the cells D^λ in Theorem 5.6.

The disks appearing in this result and those appearing in Theorem 5.6 are related in the following way. Suppose that $[t_0, t_1] \subset \mathbb{R}$ has the property that $f^{-1}([t_0, t_1]) \subset M$ has precisely one critical point a of index λ with $f(a) = c \in (t_0, t_1)$. Then by Theorem 5.6 there is a disk $D^\lambda \subset M^{t_1}$ and a homotopy equivalence

$$M^{t_1} \simeq M^{t_0} \cup D^\lambda.$$

Now note that $W^u(a) \cap f^{-1}([t_0, t_1])$ is, under a Euclidean metric defined by the Morse coordinate chart, equal to the D^λ mentioned in the proof of Theorem 5.6.

[See Chap 3.7; there is more of a statement there about diffeomorphisms that bring $W^u(a) \cap f^{-1}([t_0, t_1])$ to D^λ. Or: in case of nice metric Also, draw a picture.]

This strengthening of Theorem 5.7 makes it intuitively clear why the Morse equality (Theorem 5.9) and the Weak Morse inequalities (Corollary 5.10) hold. In addition, the Strong Morse inequalities (Corollary 5.11) also follow quickly.

There are several problems: first, we need to prove the Stable manifold theorem. Next, we need to prove that this decomposition into open cells is actually a CW complex. Now, a CW complex is described as a collection of *closed* disks, where the boundaries of these closed disks are identified with points that lie in other disks, via functions called *attaching maps*. So we need to turn the open disks $W^u(a)$ into closed disks, and describe how they are attached.

As we saw in the first torus exercise above, it does not actually always work. There is a condition (called the *Morse–Smale* condition) under which this program works. We will describe this condition in section 7. Under this condition, we will see in Chapter 10 how to view M as a CW complex.

At this stage, we should view this as roughly a CW complex decomposition, but with open disks instead of closed disks. These disks and other related spaces will play a major role in the next several chapters.

We will then study the attaching maps for the CW complex decomposition in some detail, using framed cobordism. This allows us to find more topological information about M than is given in the Morse inequalities. For instance, it allows us to compute the homology of M explicitly.

6.2 Nice metrics

In chapter 5, we proved the Morse Lemma (Theorem 5.3), which says that locally, around any nondegenerate critical point, we can choose a coordinate chart so that

$$f(x_1, \ldots, x_n) = f(p) - \sum_{i=1}^{\lambda} x_i^2 + \sum_{j=\lambda+1}^{n} x_j^2. \tag{6.1}$$

In other words, we have a local explicit formula for f around a critical point, no matter what f is, as long as the critical point is non-degenerate.

What does the gradient vector field look like around such a critical point? Based on the above equation (6.1), you might expect the gradient to be this:

$$\nabla(f) = (-2x_1, \ldots, -2x_\lambda, 2x_{\lambda+1}, \ldots, 2x_n) \tag{6.2}$$

But because the metric is not described, it is possible (even likely) that the gradient vector field is *not* this at all. Recall that the gradient is obtained by $g(v, \nabla(f)) = df(v)$ (see the discussion at the beginning of chapter 4, especially Exercises 4.3 and 4.4).

Since we are dealing with gradient vector fields, and their corresponding flow lines, it would make sense for us to want to choose coordinates to standardize the gradient vector field so that equation (6.2) is true, rather than equation (6.1). This is especially the case, since if equation (6.2) is true, then the gradient flow equation

$$\frac{d}{dt}\gamma(t) = -\nabla_{\gamma(t)}(f)$$

would take the form (if we write $\gamma(t) = (x_1(t), \ldots, x_n(t))$):

$$\begin{aligned}
\dot{x}_1 &= 2x_1 \\
&\vdots \\
\dot{x}_\lambda &= 2x_\lambda \\
\dot{x}_{\lambda+1} &= -2x_{\lambda+1} \\
&\vdots \\
\dot{x}_n &= -2x_n
\end{aligned}$$

which is easily solved since each equation only deals with one variable.

If the metric is anything else, we might still hope to diagonalize this system of differential equations, choosing coordinates $(x_1, \ldots, x_n)$ so that $\dot{\gamma}(t) = -\nabla_{\gamma(t)}(f)$ looks like

$$\dot{x}_i = c_i x_i \tag{6.3}$$

for some non-zero real constants $c_1, \ldots, c_n$. Then the c_i would be negatives of the eigenvalues of the Hessian of f at the critical point, and the corresponding eigenvectors would be the standard basis vectors $\partial/\partial x_i$ in this coordinate chart.

Unfortunately, it is in general impossible to choose coordinates so that (6.3) holds, as the following exercises show:

Exercise 6.5 *Solve the system of differential equations (6.3).*

Exercise 6.6 *Solve the system of differential equations*

$$\dot{x} = 2x \tag{6.4}$$
$$\dot{y} = -y \tag{6.5}$$
$$\dot{z} = z + xy \tag{6.6}$$
$$\tag{6.7}$$

and show that there is no change of coordinates that transform it into the form (6.3).

Exercise 6.7 *Let $f(x, y, z) = x^2 - y^2 + z^2$. Find a metric $g(x, y, z)$ on a neighborhood of $(0, 0, 0) \in \mathbb{R}^3$ so that the gradient flow equations near the origin are as in equation (6.7). Hence prove that it is in general impossible to choose coordinates so that the gradient flow equations look like equation (6.3) in a neighborhood of the critical point. Note that the metric must be symmetric and positive definite in the neighborhood.*

Note that in this exercise, what goes wrong is a kind of "resonance" phenomenon that occurs in ordinary differential equations when two eigenvalues are the same. By analogy, we would expect this kind of problem to be rare, and we might hope that for most situations, we can choose coordinates to put the gradient flow equations in the standard form of equation (6.3), but to address this will take us rather far afield (see [**?**]).

Instead, we choose to follow Hutchings [**?**] to modify the metric to the standard metric so that equation (6.1) gives rise to the gradient in equation (6.2), which in turn gives rise to the gradient flow equations in equation (6.3).

This motivates the following definition, due to Hutchings [**?**]:

Definition 6.4 *Let M be a manifold and f be a Morse function. A metric is said to be* nice *if there exist coordinate neighborhoods around each critical point of f so that for each such neighborhood there are non-zero real numbers $c_1, \ldots, c_n$ so that the gradient flow equations are*

$$\dot{x}_i = c_i x_i,$$

as in (6.3).

Proposition 6.5 *Let M be a compact manifold and f a Morse function. There exists a nice metric on (M, f). In fact, these are dense in the L^2 space of metrics.*

Proof: Let g_0 be any smooth metric on M. Consider the set of critical points of f. Apply the Morse lemma (Lemma 5.3), to find nonoverlapping coordinate neighborhoods of each critical point of f in M, each with coordinates $x_1, \ldots, x_n$ so that the Morse function in each neighborhood is

$$f(x_1, \ldots, x_n) = f(p) - \sum_{i=1}^{\lambda} x_i^2 + \sum_{j=\lambda+1}^{n} x_j^2.$$

For each critical point a of f, let U_a be the coordinate neighborhood given by the Morse lemma, let B_1 be a coordinate ball around a that is completely inside U_a, and let B_2 be another coordinate ball around a of smaller radius than B_1. (By *coordinate ball* I mean the set whose coordinates $(x_1, \ldots, x_n)$ satisfy $x_1^2 + \cdots + x_n^2 < r$ for some r.)

Let $\phi : U_a \longrightarrow \mathbb{R}$ be a smooth function so that ϕ is 1 on B_2 and 0 outside B_1. Let g_E be the standard Euclidean metric with respect to the$x_1, \ldots, x_n$ coordinates. Define g to be

$$g = g_0(x)(1 - \phi(x)) + g_E(x)\phi(x).$$

Since the set of symmetric positive definite bilinear forms is a convex set, this convex linear combination of the two metrics will be a metric on U_a. Extend g by setting it equal to g_0 on the rest of M. Then g is a metric for which a is nice.

Now proceed inductively through the other critical points of M. This creates a metric g so that there is a coordinate neighborhood metric ball B around each critical point where both f and the metric are in a standard form. Then the gradient flow equation

$$\frac{d\gamma}{dt} = \nabla_\gamma(f)$$

looks like equation (6.3).

By taking B_2 smaller and smaller, we see that the difference between g and g_0 is supported on an arbitrarily small set, and by the boundedness of the metric on M, we know that this difference is arbitrarily small in L^2. □

6.3 The proof of the stable/unstable manifold theorem

We now prove the Stable manifold theorem for nice metrics:

Theorem 2 (Stable Manifold Theorem) *Let M be an n-dimensional manifold, with nice metric g, and $f : M \longrightarrow \mathbb{R}$ a Morse function. Let a be a critical point of f of index λ. Then $W^u(a)$ and $W^s(a)$ are smooth submanifolds diffeomorphic to the open disks D^λ and $D^{n-\lambda}$, respectively.*

Remark 6.1 *This theorem is still true if the metric g is not nice, but to prove this would take too long and we don't need it in this generality. Curious readers can see [?] for the proof.*

Proof: [Proof of the Stable Manifold Theorem] If g is a nice metric, then there is a coordinate neighborhood B around each critical point where the gradient flow equations are

$$\frac{d\gamma_i}{dt} = c_i\gamma_i(t)$$

where $\gamma_i(t)$ is the i-th coordinate of γ. Note that the c_i are the negatives of eigenvalues of the Hessian, corresponding to the directions given by the standard basis in the coordinate chart. Reorder the coordinates so that the first λ eigenvalues are the negative ones (so that the first λ values of c_i are positive).

Then explicitly,

$$\gamma_i(t) = \begin{cases} \gamma_i(0)e^{|c_i|t}, & i \leq \lambda \\ \gamma_i(0)e^{-|c_i|t}, & i > \lambda \end{cases} \tag{6.8}$$

inside B.

We prove the theorem for $W^s(a)$. The proof for $W^u(a)$ is exactly analogous, and besides, it follows from the $W^s(a)$ case, applied to the function $-f$. We will first prove that $W^s(a)$ is smooth in a small neighborhood of a.

Let W_0 be the subset of B consisting of those points where $x_1 = x_2 = \cdots = x_\lambda = 0$. Then from the explicit solution (6.8), we see that $W_0 \subset W^s(a)$.

Now W_0 is an open disk of dimension $n - \lambda$ centered on a, and hence is a manifold, and is furthermore a submanifold of M.

Recall from Theorem 4.6 that the flow map defined as

$$T : M \times \mathbb{R} \longrightarrow M$$

$$T(x, t) = \gamma_x(t)$$

is smooth. Apply this flow backward in time by some time t: define $W_t = T(W_0, -t)$. This will be diffeomorphic to W_0 and a subset of $W^s(a)$. As t goes to infinity, we span a larger and larger subset of $W^s(a)$.

Let $x \in W^s(a)$, and γ the corresponding gradient flow line with $\gamma(0) = x$. Since $\lim_{t\to\infty} \gamma(t) = a$, we know that for some $t_0 > 0$, $\gamma(t) \in B$ for all $t \geq t_0$. I will now show that $\gamma(t_0) \in W_0$.

Suppose $\gamma(t_0) \notin W_0$. The translated flow $\eta(t) = \gamma(t + t_0)$ is a gradient flow line, with the property that $\eta(0) \notin W_0$, and $\eta(t) \in B$ for all $t > 0$. Then for some coordinate $i > \lambda$, $\eta_i(0) \neq 0$. By the explicit solution (6.8), $\eta_i(t)$ will grow indefinitely, so that eventually η (and hence γ) leaves the coordinate ball B. This is a contradiction. Therefore, $\gamma(t_0) \in W_0$.

Since every element of $W^s(a)$, when flowed forward, eventually is in W_0, we know that $\cup_t W_t = W^s(a)$.

Let $\psi : [0, 1) \longrightarrow \mathbb{R}$ be a smooth monotonic function with $\psi(0) = 0$ and $\lim_{t\to 1} \psi(t) = +\infty$. Using $|x|$ as $\sqrt{x_1^2 + \cdots + x_n^2}$, and r_0 as the radius of the coordinate ball B, we see that $T(x, \psi(|x|/r_0))$ maps W_0 diffeomorphically onto $W^s(a)$. Recall that W_0 is a submanifold of M which is a disk of dimension $n - \lambda$. Therefore, $W^s(a)$ is a submanifold of M and diffeomorphic to $D^{n-\lambda}$.

□

Exercise 6.8 *Prove the Stable Manifold Theorem (Theorem 2) for the unstable manifold $W^u(a)$, without applying the theorem to stable manifolds of $-f$. Instead, carefully go through the proof for $W^s(a)$ and write out the corresponding proof that would work for $W^u(a)$.*

Proposition 6.6 *The tangent space of $W^s(a)$ at a is the positive eigenspace of the Hessian of f at a. Similarly, the tangent space of $W^u(a)$ at a is the negative eigenspace of the Hessian of f at a.*

Proof: Again, we are assuming the metric is nice, but this is unnecessary.

Now $W^s(a)$ is a smooth submanifold of M, so its tangent space at a is well-defined. Define W_0 as in the previous proof, as

$$\{(x_1, \ldots, x_n) \mid x_1 = \cdots = x_\lambda = 0\}.$$

The tangent space to W_0 is therefore the span of $\partial/\partial x_i$ for $i = \lambda + 1$ to n. This is the positive eigenspace of the Hessian.

On the other hand $W_0 \subset W^s(a)$, and since they are of the same dimension, W_0 is an open neighborhood of a in $W^s(a)$. Therefore W_0 and $W^s(a)$ have the same tangent space at a.

The proof for $W^u(a)$ can be done similarly, or if you wish, you may use the result for $W^s(a)$ on $-f$. □

6.4 Structure of f restricted to stable/unstable manifolds

Let a be a critical point of f. Let us consider the function f restricted to $W^u(a)$. Since $W^u(a)$ is defined to be the set of points which in some sense lie "below" a on gradient flow lines, we expect a to be a maximum of f on $W^u(a)$, and level sets to be spheres around a.

Theorem 6.7 *Let (M, g) be a Riemannian manifold and $f : M \longrightarrow \mathbb{R}$ a Morse function. Let a be a critical point of f. Let $h : W^u(a) \longrightarrow \mathbb{R}$ be the restriction of f to $W^u(a)$. Then a is the unique critical point of h, and it is the absolute maximum. If $\epsilon > 0$ is small enough, and $f(a) - \epsilon < c < f(a)$, then $h^{-1}(c)$ is diffeomorphic to a $\lambda - 1$ dimensional sphere in $W^u(a)$ around a.*

Similarly, let $j : W^s(a) \longrightarrow \mathbb{R}$ be the restriction of f to $W^s(a)$. Then a is the unique critical point of j, and it is the absolute minimum. If $\epsilon > 0$ is small enough, and $f(a) < c < f(a) + \epsilon$, then $j^{-1}(c)$ is diffeomorphic to a $n - \lambda - 1$ dimensional sphere in $W^s(a)$ around a.

Proof: We will prove this for $W^u(a)$, and the result for $W^s(a)$ is the same using $-f$ instead of f.

Let $x \in W^u(a)$, and $x \neq a$. Let $\gamma(t)$ be the unique gradient flow line with $\gamma(0) = x$. Since $x \in W^u(a)$, we have that $\lim_{t \to -\infty} \gamma(t) = a$.

According to Lemma 4.5, $f(\gamma(t))$ is strictly decreasing. By the continuity of f, $\lim_{t\to-\infty} f(\gamma(t)) = f(a)$. So $f(a) > f(x)$. Therefore, a is the absolute maximum of h.

Now, $\gamma(t) \in W^u(a)$ for all t, so $\gamma'(0) \in T_x W^u(a)$. Since $f(\gamma(t))$ is strictly decreasing, $\gamma'(0) \neq 0$ (if it were, $\frac{d}{dt} f(\gamma(t)) = \nabla(f) \cdot \gamma'(0)$ would be zero). By the gradient flow equation $\gamma'(t) = -\nabla_{\gamma(t)}(f)$, the $-\nabla_x(f) \neq 0$. Therefore, x is not a critical point of h. Since x was arbitrary, except for not equalling a, there are no critical points of h except for a.

Now we consider the Hessian of h at a. Find a coordinate chart of M around a so that $W^u(a)$ is given by the equations $x_{\lambda+1} = \cdots = x_n = 0$. By the invariance of the Hessian under coordinate change (Proposition 4.1), the Hessian of f can be computed in such a coordinate chart. Since $T_a W^u(a)$ is the negative eigenspace of the Hessian of f (Proposition 6.6) we conclude that the matrix

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{ij}$$

is negative definite. Since $W^u(a)$ is given by setting $x_{\lambda+1}, \ldots, x_n$ to be constant (in fact, zero), we see that for $i, j \leq \lambda$, this matrix is the same as

$$\left(\frac{\partial^2 h}{\partial x_i \partial x_j}\right)_{ij}.$$

Therefore the Hessian of h at a is negative definite. In particular, a is a non-degenerate critical point of h, and h is Morse.

We now consider the preimages $h^{-1}(c)$.

For this, we use the Morse Lemma (Theorem 5.3) applied to h on the manifold $W^u(a)$. The Morse Lemma states that there exist a coordinate neighborhood U around a with coordinates $x_1, \ldots, x_\lambda$ on $W^u(a)$ so that

$$h(x_1, \ldots, x_\lambda) = f(a) - x_1^2 - \cdots - x_\lambda^2.$$

Let $\epsilon > 0$ be given so that the ball

$$B = \{(x_1, \ldots, x_\lambda) | x_1^2 + \cdots + x_\lambda^2 < \epsilon\}$$

is contained in U. Within this ball it is clear that the preimages $h^{-1}(c)$ (when $f(a) - \epsilon < c < f(a)$) are coordinate spheres around a. We will now verify that there are no other parts to $h^{-1}(c)$ which are outside B.

Suppose $x \in W^u(a)$, and $x \notin B$. As earlier in the proof, let $\gamma(t)$ be the gradient flow with $\gamma(0) = x$. As before, $\lim_{t\to-\infty} \gamma(t) = a$. But B is an open set around U. Therefore, for some $t < 0$, $\gamma(t) \in B$. Since $x = \gamma(0)$ is not in B, the generalized Jordan curve theorem says that there exists some $T < 0$ for which $\gamma(T)$ is on the boundary of B. Since $f(\gamma(t))$ is strictly decreasing,

$$f(x) = f(\gamma(0)) < f(\gamma(T)) = f(a) - \epsilon.$$

So $f(x) < f(a) - \epsilon$. Therefore, if $f(a) - \epsilon < c < f(a)$, then $h^{-1}(c)$ is a susbet of B, and is therefore the coordinate spheres we found earlier. $\square$

Chapter 7

Morse–Smale functions: intersecting stable and unstable manifolds

7.1 The Morse–Smale condition

Consider Exercise 6.3. One of the edges did not attach to vertices, but to the midpoint of another edge. In Exercise 6.4, a perturbation of this situation, this problem is fixed, and both edges end at the bottom vertex. This indicates that it is not enough that f be Morse for the unstable manifold picture to work well. We need a further transversality condition, which we define now.

Definition 7.1 *Suppose $f : M \longrightarrow \mathbb{R}$ is a Morse function that satisfies the extra condition that for any two critical points a and b the unstable and stable manifolds $W^u(a)$ and $W^s(b)$ intersect transversally. This is the* Morse–Smale condition*, and if f satisfies this condition, we call f a* Morse–Smale function*.*

Smale [**?**] showed that Morse–Smale functions exist. More specifically, given a metric g and function $f : M \longrightarrow \mathbb{R}$, there exists another metric g' and another function $f' : M \longrightarrow \mathbb{R}$ so that f' is Morse–Smale with respect to g'. His proof also demonstrates that f and f' and g and g' can be made arbitrarily close to each other. Hence the set of configurations of functions and metrics so that the functions are Morse–Smale with respect to that metric is dense.

Actually, more is true: if f is Morse, then for an open, dense set of metrics g, f is Morse–Smale. This can be proved using the same techniques that are used in the proofs in Smale's paper. We will sketch out a proof at the end of this chapter that the set of such metrics is dense. In the meantime we will first study some properties of Morse–Smale functions.

Exercise 7.1 *Show that the example in Exercise 6.4 is Morse–Smale, and the example in Exercise 4.5 is not.*

Exercise 7.2 *Suppose f is Morse (not necessarily Morse–Smale) and suppose b is a critical point of f. Do $W^u(b)$ and $W^s(b)$ always intersect transversally?*

7.2 Intersections of stable and unstable manifolds: $W(a,b)$

The main purpose of the Morse–Smale condition is that it allows us to see how stable and unstable manifolds of different critical points intersect. For every pair of critical points a and b, let

$$W(a,b) = W^u(a) \cap W^s(b).$$

$W(a,b)$ is the space of all points in M that lie on flow lines starting from a and ending at b.

Proposition 7.2 *Let (M,g) be a Riemannian manifold of dimension n, let $f : M \longrightarrow \mathbb{R}$ be Morse–Smale, and a and b be two critical points of f. Then $W(a,b)$ is a smooth manifold of dimension*

$$a) - b).$$

Proof: If f is Morse–Smale, then $W^u(a)$ and $W^s(b)$ intersect transversally. So by Theorem 3.5, the intersection $W^u(a) \cap W^s(b) = W(a,b)$ is a manifold of dimension $\dim(W^u(a)) + \dim(W^s(b)) - n = a) + (n - b)) - n = a) - b)$. □

Corollary 7.3 *Let $f : M \longrightarrow \mathbb{R}$ be a Morse–Smale function, and let a and b be two distinct critical points of f. If $a) \leq b)$, then $W(a,b) = \emptyset$.*

Proof: If $a) < b)$, then the previous proposition shows that $W(a,b)$ is a manifold of negative dimension, so it must be empty.

If $a) = b)$, then similarly $W(a,b)$ must be a manifold of dimension 0, but since the gradient flow acts freely on elements of $W(a,b)$, the dimension of $W(a,b)$ must be at least one. Therefore it must be empty. □

Definition 7.4 *We refer to the number*

$$a) - b)$$

as the relative index *of a and b.*

Exercise 7.3 *Find $W(a,b)$ for each pair of critical points a and b for Exercise 6.4.*

Exercise 7.4 *Suppose a and b are critical points of f and $a \neq b$. Are a and b in $W(a,b)$? If there are other critical points of f, is it possible that these are in $W(a,b)$? Now consider the case $a = b$. What is $W(a,b)$?*

7.3 Morse–Smale moduli spaces: $W(a,b)^t$

The fundamental object of study will not usually be $W(a,b)$, but a particular "horizontal" slice. If we view the gradient flow as an action of $\mathbb{R}$ on $W(a,b)$, then we can study the orbit space (called the Moduli space) $W(a,b)/\mathbb{R}$. For good intuition and for practical considerations it is useful to instead pick out a representative of each $\mathbb{R}$ orbit in $W(a,b)$. One way to do this is to select a real number t between a and b and pick out the representative in $f^{-1}(t)$. This is the approach used in our first definition of the moduli space (there will be other equivalent definitions soon).

Definition 7.5 *Pick a value $t \in \mathbb{R}$ between $f(a)$ and $f(b)$, and let $W(a,b)^t$ to be the set $W(a,b) \cap f^{-1}(t)$.*

Proposition 7.6 *If a and b are distinct critical points of f, then $W(a,b)^t$ is a smooth submanifold of M.*

Proof: First, we see that $f|_{W(a,b)} : W(a,b) \longrightarrow \mathbb{R}$ is transverse to the point $\{t\} \subset \mathbb{R}$. This is because for any point $x \in W(a,b)$ so that $f(x) = t$, $\nabla_x(f)$ is not zero, and so neither is $df_x(\nabla_x(f)) = \|\nabla_x(f)\|^2$. Therefore $f|_{W(a,b)} \pitchfork \{t\}$.

Therefore, we may apply Theorem 3.5, and get that $\left(f|_{W(a,b)}\right)^{-1}(\{t\}) = W(a,b)^t$ is a submanifold of $W(a,b)$ of codimension one. □

Proposition 7.7 *Let a and b be distinct critical points of f. The function*

$$\phi : W(a,b)^t \times \mathbb{R} \longrightarrow W(a,b)$$

defined by

$$\phi(p,s) = T_s(p)$$

is a diffeomorphism.

Proof: We begin by proving ϕ is onto. Let $x \in W(a,b)$. Let γ be the flow line that has $\gamma(0) = x$. Since $\lim_{t\to\infty} f(\gamma(t)) = f(b)$ and $\lim_{t\to-\infty} f(\gamma(t)) = f(a)$, by continuity we have that for some s, $f(\gamma(-s)) = t$. Then $\gamma(-s) = p$ and $T_s(p) = x$.

Now to show ϕ is one-to-one, suppose $x = \phi(p_1,s_1) = \phi(p_2,s_2)$. Then $T_{-s_1}(x) = p_1$ and $T_{-s_2}(x) = p_2$, meaning that the unique flow line γ with $\gamma(0) = x$ also has $\gamma(s_1) = p_1$ and $\gamma(s_2) = p_2$. Since $f(p_1) = t = f(p_2)$, and $\frac{d}{ds} f(\gamma(s)) < 0$, it must be that $s_1 = s_2$ and therefore $p_1 = p_2$.

Therefore ϕ^{-1} is defined as a set map. To show that ϕ^{-1} is continuous, it is necessary to show that if U is an open neighborhood of $(p,s) \in W(a,b)^t \times \mathbb{R}$, then there exists an open neighborhood of $\phi(p,s)$ in $W(a,b)$ that is a subset of $\phi(U)$. It suffices to show this for open neighborhoods U of the form $B_p(\epsilon) \times (s-\epsilon, s+\epsilon)$. Since T_{-s} is a diffeomorphism of M that maps neighborhoods of $\phi(p,s)$ to neighborhoods of $\phi(p,0)$, it suffices to prove this for $s = 0$.

So what we need to show is if $\epsilon > 0$ is sufficiently small, and $p \in W(a,b)^t$, then there exists a δ so that whenever $d(p,y) < \delta$, then writing $y = \phi(q,r)$ gives us $|r| < \epsilon$ and $d(p,q) < \epsilon$.

Since p is not a critical point, there is a δ_1 so that $B_p(2\delta_1)$ does not contain critical points. In this ball, $m = \inf |\nabla f|^2$ is strictly greater than zero and $\sup |\nabla f|$ is finite. If $\sup |\nabla f| > 1$, then let $M = \sup |\nabla f|$, but otherwise let $M = 1$. By continuity of f there is a δ_2 so that $|f(p) - f(B_p(\delta_2))| < m\epsilon/2M$. Choose δ to be smaller than $\min(\delta_1, \delta_2, \epsilon/2)$.

Now in the proof of Lemma 4.5, we saw that

$$\frac{d}{dt} f(\gamma(t)) = -\left|\nabla(f)\right|^2.$$

Integrating and using the fundamental theorem of calculus, we get

$$|f(\gamma(-r)) - f(\gamma(0))| \geq |r| \inf |\nabla f|^2$$

which leaves us with

$$|r|m = |r| \inf |\nabla f|^2 \leq |f(p) - f(y)| < m\epsilon/2M$$

so that $|r| < \epsilon/2M < \epsilon$.

Now,

$$\begin{aligned} d(q, y) &\leq \int |\gamma'(t)|\, dt \\ &= \int |\nabla(f)|\, dt \\ &\leq Mr < \epsilon/2. \end{aligned}$$

So by the Triangle inequality, $d(p, q) \leq d(p, y) + d(q, y) < \delta + \epsilon/2 < \epsilon$. Therefore ϕ^{-1} is continuous.

To prove ϕ^{-1} is smooth, we estimate $d\phi$ and show it is non-degenerate. Let $(p, s) \in W(a, b)^t \times \mathbb{R}$ and let $v_1, \ldots, v_k$ be a basis for the tangent space of $W(a, b)^t$ at p, and let $\partial/\partial t$ be the tangent vector to $\mathbb{R}$. Now if $d\phi$ is degenerate at (p, s), then $d\phi(v_1), \ldots, d\phi(v_k), d\phi(\partial/\partial t)$ would be linearly dependent. Now since $\phi|_{W(a,b)^t \times \{s\}}$ is just the flow map T_s, and this flow map is a diffeomorphism, we know that $d\phi(v_1), \ldots, d\phi(v_k)$ are linearly independent. Therefore any linear dependence would involve $d\phi(\partial/\partial t)$, so that

$$d\phi(\partial/\partial t) = \sum c_k d\phi(v_k)$$

for some real numbers c_k.

Now since $\phi(p, s) = T_s(p)$, $d\phi(\partial/\partial t)$ at (p, s) is $\frac{\partial}{\partial s} T_s(p) = \gamma'(s)$, where γ is the flow with $\gamma(0) = p$. Then if we compose with T_{-s},

$$\begin{aligned} dT_{-s} d\phi(\partial/\partial t) &= \sum c_k dT_{-s} d\phi(v_k) \\ dT_{-s}\gamma'(s) &= \sum c_k v_k \\ \gamma'(0) &= \sum c_k v_k. \end{aligned}$$

But we know $\gamma'(0)$ is transverse to $TW(a,b)^t$, which is a level set of f. Therefore, we have a contradiction, and $d\phi$ is non-degenerate. Therefore ϕ^{-1} is smooth. □

If we use the notation $+a$ to denote the function $+a : \mathbb{R} \longrightarrow \mathbb{R}$ with $+a(x) = x + a$, then the following diagram commutes:

$$\begin{array}{ccc} W(a,b)^t \times \mathbb{R} & \xrightarrow{\phi} & W(a,b) \\ \downarrow{\scriptstyle (1,+s)} & & \downarrow{\scriptstyle T_s} \\ W(a,b)^t \times \mathbb{R} & \xrightarrow{\phi} & W(a,b) \end{array}$$

7.4 Existence and denseness of Morse–Smale metrics

We now sketch a proof that the set of metrics for which a Morse function is Morse–Smale is dense.

Theorem 7.8 *Let M be a manifold. Let $f : M \longrightarrow \mathbb{R}$ be a Morse function. For a dense set of metrics g, f is Morse–Smale.*

Proof: (Sketch of proof) We suppose a Riemannian metric g is given, and show that there exists a Riemannian metric g' arbitrarily close to g so that f is Morse–Smale with respect to g'. For the purposes of this proof ∇_g refers to the gradient using the metric g.

We start by finding a vector field X close to $\nabla_g f$ that agrees with $\nabla_g f$ near the critical points of f but so that the ascending and descending manifolds are transverse (step 1). We then show that for some metric g' close to g, $X = \nabla_{g'}(f)$ (step 2).

Step 1: finding the vector field X

The details of this step are found in Smale's proof of Theorem A in the work just cited above ([**?**]).

Let the critical values of f be $c_1 < \cdots < c_k$. Choose $\epsilon > 0$ arbitrary, but small enough so that for each i, $c_{i+1} > c_i + 4\epsilon$, and in fact, small enough so that for each critical point p, Theorem 6.7 gives us that $W^s(p) \cap f^{-1}((-\infty, c])$ is a ball for all $f(p) < c < f(p) + 4\epsilon$.

We first let $X = \nabla g$. Then we proceed by induction on i, starting at c_1 and ending at c_k, at each stage altering X in $f^{-1}(c_i + \epsilon, c_i + 3\epsilon)$.

At stage i in the induction, we consider each critical point p so that $f(p) = c_i$. In a neighborhood of p, we consider

$$Q = f^{-1}(c_i + 2\epsilon) \cap W^s(p).$$

Since $-\nabla(f)$ is transverse to level sets of f, the gradient flow can be integrated in a small neighborhood of Q so that there is a coordinate z with $-m \leq z \leq m$ so that $\partial/\partial z$ is $-\nabla(f)$ and $z = 0$ coinciding with Q. Here m is chosen so that this keeps us in $f^{-1}(c_i + \epsilon, c_i + 3\epsilon)$. By the coordinate structure of f near p, a

tubular neighborhood U of Q is a trivial λ-disk bundle. So if P is a λ dimensional disk of radius 1, then there is a diffeomorphism sending $[-m,m] \times P \times Q$ onto this tubular neighborhood of Q, so that the first coordinate is the coordinate z, and $0 \times 0 \times Q$ is mapped to Q by the identity function. From now on, we will identify U with $[-m,m] \times P \times Q$ in our notation.

Consider all critical points q with $f(q) > c_i$. Let

$$S = \cup_{q, f(q)>c_i, \nabla_q(f)=0} (0 \times P \times Q) \cap W^s(q)$$

and let $g : S \longrightarrow P$ be the restriction of $\pi_P : [-m,m] \times P \times Q \longrightarrow 0 \times P \times 0$ to S. By Sard's theorem there exist $v \in P$ arbitrarily close to zero so that $2v$ is a regular value of g.

Now construct $\beta : [-m,m] \longrightarrow \mathbb{R}$ so that $\beta(z) \geq 0$, $\beta(z) = 0$ in a neighborhood of $\partial[-m,m]$, and $\int_0^{\pm m} \beta(z)\, dz = \pm|v|$. If v was chosen small enough, $\beta(z)$ and $|\beta'(z)|$ can be kept smaller than ϵ.

Let $P_0 \subset P$ be a λ-dimensional disk of radius $1/3$.

We also construct a smooth $\gamma : P \longrightarrow \mathbb{R}$ so that $0 \leq \gamma \leq 1$, $\gamma = 0$ in a neighborhood of P, $\gamma = 1$ on P_0, and $|\partial\gamma/\partial x_i| \leq 2$.

Let X' be the vector field on M that equals X outside U, and on $[-m,m] \times P \times Q$ let X' be given by

$$X' = -\frac{\partial}{\partial z} - \beta(z)\gamma(x)\frac{v}{|v|}.$$

We use the bounds on β and γ to ensure that $df(X') > 0$.

To see that the new stable and unstable manifolds $W'^s(p)$ and $W'^u(q)$ intersect transversally, we examine any point of intersection, and flow by X' until it is in $f^{-1}(c_i + 2\epsilon)$. It will then be at a point $\{0\} \times P \times Q \subset [-m,m] \times P \times Q$. The flow X' for time $\pm m$ carries $(0,x,y) \in [-m,m] \times P \times Q$ to $(\pm m, x \pm v, y)$, as can be seen by explicitly integrating out X'.

If q is any critical point with $f(q) > c_i$, then consider the new stable manifold $W'^s(q)$ of q under X'. It agrees with the old stable manifold $W^s(q)$ on $(m,0,y)$, and after flowing by $-m$ we get to $(0,-v,y)$.

Also, the new unstable manifold $W'^u(p)$ agrees with the old unstable manifold $W^u(p)$ for $z = -m$, and flowing by X' for time m from here shows that $W'^u(p) \cap (0 \times P \times Q)$ is

$$\{(0, x+v, y) | (0,x,y) \in W^u(p)\}.$$

So their intersection is the set

$$\{(0,-v,y) | (0,2v,y) \in W^u(p)\}$$

and since $2v$ is a regular value of g, this intersection is transverse.

We do this for all the critical points with critical value c_i, and these do not interfere with each other as long as ϵ is small enough that the neighborhoods U do not intersect.

We then proceed with larger and larger i, until we have constructed a new X'.

Step 2: finding the metric g'

Note that X is unchanged (it still equals $\nabla_g f$) near critical points of f. So near critical points of f we define g' to equal g. Outside these neighborhoods we define, at each point $x \in M$, a linear transformation A_x on T_xM that is the identity on the kernel of df, and sends X to

$$\frac{\sqrt{df(X)}}{\|df\|_g} \nabla_g(f).$$

Since $df(X) > 0$, this is invertible, and if X is close to $\nabla_g(f)$, then A_x is close to the identity. Let $g'(v, w) = g(Av, Aw)$. Then g' is close to g.

Now if we write an arbitrary vector $w \in T_x(M)$ as $w = w_0 + aX$ where $df(w_0) = 0$, then it is a matter of computation to verify that $g'(X, w) = df(w)$. By definition of gradient, this means $X = \nabla_{g'}(f)$. □

So the set of metrics under which f is Morse–Smale is dense. To show that it is open is harder.

[I can't find it in the literature. Is it proved?]

Corollary 7.9 *Given a Morse function $f : M \longrightarrow \mathbb{R}$, there exists a metric g so that f is Morse–Smale.*

Chapter 8

Spaces of flow lines

As before, $f : M \longrightarrow \mathbb{R}$ is a smooth Morse–Smale function, and a and b are critical points.

The space $W(a, b)$ is a subset of M, namely, the set of points that lie on flows that go from a to b, but by identifying the point in $W(a, b)$ with the flow itself, we will see in this section that $W(a, b)$ can be identified as the space of gradient flow lines starting from a and ending at b. That is, let $P_{a,b}$ be the set of C^1 paths from a to b:

$$P_{a,b} = \{\gamma \in C^1(\mathbb{R}, M) \mid \lim_{t\to-\infty} \gamma(t) = a, \lim_{t\to+\infty} \gamma(t) = b\}$$

and let $\mathcal{F}_{a,b}$ be the subset of $P_{a,b}$ consisting of gradient flows:

$$\mathcal{F}_{a,b} = \{\gamma \in P_{a,b} \mid \gamma'(t) = -\nabla_{\gamma(t)}(f).$$

We will show $\mathcal{F}_{a,b}$ and $W(a, b)$ are homeomorphic, by identifying the point $p \in W(a, b)$ with the gradient flow $\gamma(t)$ so that $\gamma(0) = p$. Once we make this identification, we will use $W(a, b)$ or $\mathcal{F}_{a,b}$ interchangably, or rather, $W(a, b)$ will be viewed as a subset of M and as a subset of $P(a, b)$ as the situation demands.

One reason for thinking of $W(a, b)$ as a function space is that this is the most effective way to generalize the results of Morse theory to many infinite-dimensional situations which have been of interest since the 1980s, like Floer homology. Another reason is that sometimes working with $\mathcal{F}_{a,b}$ is the easiest way to prove theorems about the topology of the space of flows, especially when we prove theorems relating to compactifications and gluing.

8.1 The space of gradient flows $\mathcal{F}_{a,b}$

As above, we let $P_{a,b}$ be the set of C^1 paths from a to b, and let $\mathcal{F}_{a,b} \subset P_{a,b}$ be the set of paths that are gradient flows. More specifically, consider the map L that sends an element of $P_{a,b}$ to the vector field

$$\frac{d\gamma}{dt}(t) + \nabla_{\gamma(t)}(f)$$

which is a vector field over the image of γ. More accurately, it is a section of the pullback bundle $\gamma^* TM$. Overall, L is a section of a fiber bundle whose base is $P_{a,b}$ and where the fiber over $\gamma \in P_{a,b}$ is C^0 sections of $\gamma^* TM$ which converge to zero at $\pm\infty$.

Then $\mathcal{F}_{a,b}$ is $L^{-1}(0)$.

Now the Morse–Smale condition is equivalent to the condition that L be transverse to zero, that is, that dL is surjective at $\mathcal{F}_{a,b}$. This fact is not often proved, though it is widely believed.

[Find somewhere that does it so we can reference it]

Theorem 8.1 *Let $f : M \longrightarrow \mathbb{R}$ be Morse, and let g be a nice metric. Let a and b be critical points of f. Then $W^u(a)$ is transverse to $W^s(b)$ if and only if dL is surjective at $\mathcal{F}_{a,b}$.*

Proof: Let $\gamma \in \mathcal{F}_{a,b}$ be given, and let $p = \gamma(t_0)$ for some real t_0. We will show that dL is surjective at γ if and only if $W^u(a)$ and $W^s(b)$ intersect transversally at p. This will prove the theorem since every point in $W(a,b) = W^u(a) \cap W^s(b)$ is $\gamma(t_0)$ for some $\gamma \in \mathcal{F}_{a,b}$ and some $t_0 \in \mathbb{R}$.[1]

Since $L(\gamma)$ can be written $\frac{d\gamma}{dt} + \nabla_{\gamma(t)}(f)$, then using an arbitrary coordinate patch around the image of the curve γ,

$$L(\gamma)_i = \frac{d\gamma_i}{dt} + \sum_j g^{ij}(\gamma(t)) \frac{\partial f}{\partial x_j}$$

where g^{ij} is the matrix corresponding to the inverse of the metric matrix g_{ij}. Taking derivatives we see that

$$dL(\xi)_i = \frac{d\xi_i}{dt} + \sum_{j,k} \frac{\partial g^{ij}}{\partial x_k} \xi_k \frac{\partial f}{\partial x_j} + g^{ij} \frac{\partial^2 f}{\partial x_j \, \partial x_k} \xi_k.$$

The point is that in these coordinates,

$$dL(\xi) = \frac{d\xi}{dt} + A(t)\xi$$

where $A(t)$ is a linear transformation on $T_{\gamma(t)}M$. Note that $\lim_{t\to-\infty} A(t)$ is the Hessian at a, and $\lim_{t\to+\infty} A(t)$ is the Hessian at b, but for finite values of t, $A(t)$ is not in general symmetric or even diagonalizable. Note also that if we change basis using a family of matrices $C(t)$,

$$C^{-1} dL(C\xi) = \frac{d\xi}{dt} + C^{-1}\frac{dC}{dt}\xi + C^{-1}AC$$

[1] Of course we could have fixed t_0 to be anything we like. The reader will be best served thinking of p as far from a and b, and t to be of moderate size, such as zero, since this is where the intuition for all the steps of the proof are clearest, even though it applies to all of $W(a,b)$.

so A becomes $C^{-1}C' + C^{-1}AC$.

Since g is nice, we can choose a coordinate neighborhood U around a with coordinates $x_1, \ldots, x_n$ so that in this coordinate the gradient flow equations are

$$\frac{dx_i}{dt} + \lambda_i x_i = 0.$$

Here λ_i are the (non-zero) eigenvalues of the Hessian of f at a. In particular, the curve $\gamma(t)$ for t small enough that $\gamma(t)$ is in the coordinate chart satisfies

$$L(\gamma)_i = \frac{d\gamma_i}{dt} + \lambda_i \gamma_i(t) = 0.$$

In this neighborhood, define the basis $e_i = \frac{\partial}{\partial x_i}$ of TM. Then in this basis, if we write $\xi \in T_{\gamma(t)} P_{a,b}$ as $\xi = \sum_i \xi_i e_i$, then

$$dL(\xi)_i = \frac{d\xi_i}{dt} + \lambda_i \xi_i$$

so that

$$dL(e_i) = \lambda_i e_i.$$

Now extend e_i to $e_i(t)$ a basis on $T_{\gamma(t)}M$ for all t by insisting that

$$dL(e_i(t)) = \lambda_i e_i(t)$$

for all t. This is well-defined because $\gamma(t)$, being strictly decreasing for f, is one-to-one. For some T, $e_i(t) = e_i$ (the basis in the neighborhood U of a) for all $t < -T$.

We now wish to show that the $\{e_i\}$ are linearly independent at $T_{\gamma(t)}$. Suppose some non-zero linear combination $\phi = \sum c_{i,0} e_i(t_1)$ were zero at some value of t_1. Then this extends to $\phi(t)$ satisfying $dL(\phi(t)) = 0$ for all t as follows. Since dL is linear, if we write ϕ in the $e_i(t)$ basis: $\phi = \sum c_i(t) e_i(t)$, then $dL(\phi) = \frac{d\phi}{dt} + A(t)\phi(t) = 0$ implies

$$\sum \frac{dc_i}{dt} e_i(t) + c_i \lambda_i e_i(t)$$

so that

$$\frac{dc_i}{dt} = \lambda_i c_i$$

which has the solution

$$c_i(t) = c_{i,0} e^{\lambda_i (t - t_1)}$$

for all i. On the other hand, at t_1, $\phi(t_1) = 0$. By the uniqueness of solutions to first-order ordinary differential equations, $\phi(t) = 0$ for all t. Therefore

$$\phi(t) = \sum_i c_i(t) e_i(t) = 0$$

is zero when $t < -T$ (so that we are in the neighborhood of a where we know e_i is a basis). Then since e_i is a basis, $c_i(t) = 0$ at such a t. By the formula

for $c_i(t)$ above, $c_{i,0} = 0$. Therefore, $\phi(t_1) = 0$. Therefore for all t, $\{e_i\}$ are linearly independent. Since the dimension of the tangent space is n, we have that $\{e_i(t)\}$ is a basis for $T_{\gamma(t)}M$.

Now in the coordinate neighborhood U of a, $W^u(a)$ consists of the subspace

$$\text{Span}\{e_i|\lambda_i < 0\}.$$

In fact, if we take, for every e_i with $\lambda_i < 0$ and real number $h > 0$, curves

$$\gamma_{h,i}(t) = \gamma(t) + he^{-\lambda_i t}e_i(t)$$

in U, then

$$L(\gamma_{h,i}) = L(\gamma) + h\,L(e^{-\lambda_i t}e_i(t)) = 0 + 0 = 0$$

and $\gamma_{h,i}(t) \to 0$ as $t \to -\infty$. We can extend $\gamma_{h,i}(t)$ to all t to satisfy the gradient flow equations $L(\gamma_{h,i}) = 0$. Then $\gamma_{h,i}(t)$ is a gradient flow from the critical point a, so $\gamma_{h,i}(t) \in W^u(a)$ for all t. Now consider $\frac{\partial}{\partial h}\gamma_{h,i}(t)$. Since $L(\gamma_{h,i}(t)) = 0$, we have

$$dL(\frac{\partial}{\partial h}\gamma_{h,i}) = 0.$$

If t is small enough that $\gamma_{h,i}(t) \in U$, then we can directly compute $\frac{\partial}{\partial h}\gamma_{h,i}(t) = e^{-\lambda_i t}e_i$. On the other hand, for $t = t_0$, $\frac{\partial}{\partial h}\gamma_{h,i}(t_0) \in T_pW^u(a)$. Therefore for each i with $\lambda_i < 0$, $e_i(t_0) \in T_pW^u(a)$.

To summarize, we have, for each t, a basis $\{e_i(t)\}$ of $T_{\gamma(t)}M$ so that $Le_i(t) = \lambda_i e_i(t)$, and for $t < -T$, $e_i(t)$ is a constant basis. We also have that if $\lambda_i < 0$, then $e_i(t_0) \in T_pW^u(a)$.

Similarly, we can examine the neighborhood V around b with coordinates $y_1, \ldots, y_n$ so that the gradient flow equations $L(y) = 0$ satisfy

$$\frac{dy_i}{dt} + \mu_i y_i = 0$$

where μ_i are the non-zero eigenvalues of the Hessian of f at b. We can similarly produce basis vectors $f_i(t)$ at $T_{\gamma(t)}M$ so that $Lf_i(t) = \mu_i f_i(t)$ for all t, and when $t > T$ for some large T, $\gamma(t) \in V$ and $f_i(t) = \frac{\partial}{\partial y_i}$. Also, if $\mu_i > 0$, then $f_i(0) \in T_pW^s(b)$.

Therefore, at p,

$$T_pW^u(a) = \text{Span}\{e_i(t_0)|\lambda_i < 0\}$$

and

$$T_pW^s(b) = \text{Span}\{f_i(t_0)|\mu_i > 0\}.$$

Suppose $W^u(a)$ and $W^s(b)$ do not intersect transversally at p. Then there exists a non-zero $v \in T_pM$ that is perpendicular to $T_pW^u(a) + T_pW^s(b)$. Conversely, if there exists a non-zero vector v perpendicular to $T_pW^u(a) + T_pW^s(b)$, then $W^u(a)$ and $W^s(b)$ do not intersect transversally. So the existence of such a v is equivalent to $W^u(a)$ and $W^s(b)$ interesecting transversally.

Similarly, if on $T_\gamma P_{a,b}$ we define the inner product

$$\langle \xi, \eta \rangle = \int_{-\infty}^{\infty} g(\xi(t), \eta(t))\, dt$$

where g is the Riemannian metric, then dL at γ would not be surjective if and only if there is a η so that

$$\langle \eta, dL(\xi) \rangle = 0$$

for all $\xi \in T_\gamma P_{a,b}$.[2] [Be more precise about the codomain of dL.]

Associated to the basis $e_i(t)$, construct the basis $\tilde{e}_i(t) \in T_{\gamma(t)}M$ so that $g(e_i(t), \tilde{e}_j(t)) = \delta_{i,j}$[3]. It is straightforward to prove that $\{\tilde{e}_i(t)\}$ is linearly independent, so it forms a basis.

Exercise 8.1 *Prove $\{\tilde{e}_i(t)$ is a basis for $T_{\gamma(t)}M$.*

Exercise 8.2 *In the Euclidean plane, draw the vectors $e_1 = (1, 0)$ and $e_2 = (2, 3)$. Find $\tilde{e}_1$ and $\tilde{e}_2$ (for the standard Euclidean metric).*

Similarly, let $\{\tilde{f}_i(t)\}$ be the basis for $T_{\gamma(t)}M$ so that $g(f_i(t), \tilde{f}_j(t)) = \delta_{i,j}$. Astute readers might recall that given a basis for a finite-dimensional vector space V we can form a dual basis for V^*, and that an inner product also allows us to relate vectors in V with those in V^*. The composition of these two is the process described in taking e_i to $\tilde{e}_i$ or f_i to $\tilde{f}_i$.

Now suppose η is perpendicular to $dL\xi$ for all $\xi \in T_\gamma P_{a,b}$. We can write ξ using the $e_i(t)$ basis or the $f_i(t)$ basis as follows:

$$\xi(t) = \sum_i \xi_i(t) e_i(t) = \sum_i \zeta_i(t) f_i(t)$$

and we can write η using the $\tilde{e}_i(t)$ basis or the $\tilde{f}_i(t)$ basis:

$$\eta(t) = \sum_i \eta_i(t) \tilde{e}_i(t) = \sum_i \theta_i(t) \tilde{f}_i(t).$$

[2] There is the possible objection that $\langle \cdot, \cdot \rangle$ is not defined on all of $T_\gamma P_{a,b}$ since the integral might not converge. For that matter we were not careful in defining the codomain of dL. We will define the topology on the codomain of dL so that L^2 is dense there...

[3] $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

Furthermore, we can write $dL\xi$ as

$$\begin{aligned}
dL\xi(t) &= \left(\frac{d}{dt} + A(t)\right)\sum_i \xi_i(t)e_i(t) = \left(\frac{d}{dt} + A(t)\right)\sum_i \zeta_i(t)f_i(t) \\
&= \sum_i \xi_i'(t)e_i(t) + \xi_i(t)e_i'(t) + \xi(t)A(t)e_i(t) = \sum_i \zeta_i'(t)f_i(t) + \zeta_i(t)f_i'(t) + \zeta(t)A(t)f_i(t) \\
&= \sum_i \xi_i'(t)e_i(t) + \xi_i(t)dL(e_i(t)) = \sum_i \zeta_i'(t)f_i(t) + \zeta_i(t)dL(f_i(t)) \\
&= \sum_i \xi_i'(t)e_i(t) + \xi_i(t)\lambda_i e_i(t) = \sum_i \zeta_i'(t)f_i(t) + \zeta_i(t)\mu_i f_i(t) \\
&= \sum_i \left(\xi_i'(t) + \xi_i(t)\lambda_i\right)e_i(t) = \sum_i \left(\zeta_i'(t) + \zeta_i(t)\mu_i\right)f_i(t)
\end{aligned}$$

We then write the relation $\langle \eta, L\xi\rangle = 0$ as

$$\begin{aligned}
\langle \eta, L\xi\rangle &= \int_{-\infty}^{\infty} g(\eta, L\xi)\,dt \\
&= \int_{-\infty}^{t_0} g(\eta, L\xi)\,dt + \int_{t_0}^{\infty} g(\eta, L\xi)\,dt \\
&= \int_{-\infty}^{t_0} \sum_{i,j} \eta_i(\xi_j' + \xi_j\lambda_j)g(\tilde{e}_i, e_j)\,dt + \int_{t_0}^{\infty} \sum_{i,j} \theta_i(\zeta_j' + \zeta_j\mu_j)g(\tilde{f}_i, f_j)\,dt \\
&= \int_{-\infty}^{t_0} \sum_{i,j} \eta_i(\xi_j' + \xi_j\lambda_j)\delta_{i,j}\,dt + \int_{t_0}^{\infty} \sum_{i,j} \theta_i(\zeta_j' + \zeta_j\mu_j)\delta_{i,j}\,dt \\
&= \int_{-\infty}^{t_0} \sum_{i} \eta_i\xi_i' + \eta_i\xi_i\lambda_i\,dt + \int_{t_0}^{\infty} \sum_{i} \theta_i\zeta_i' + \theta_i\zeta_i\mu_i\,dt \\
&= \int_{-\infty}^{t_0} \sum_{i} -\eta_i'\xi_i + \eta_i\xi_i\lambda_i\,dt + \eta_i(t)\xi_i(t)|_{-\infty}^{t_0} + \int_{t_0}^{\infty} \sum_{i} -\theta_i'\zeta_i + \theta_i\zeta_i\mu_i\,dt + \theta_i(t)\zeta_i(t)|_{t_0}^{\infty} \\
&= \int_{-\infty}^{t_0} \sum_{i} \xi_i(-\eta_i' + \eta_i\lambda_i)\,dt + g(\eta(t), \xi(t))|_{-\infty}^{t_0} + \int_{t_0}^{\infty} \sum_{i} \zeta_i(-\theta_i' + \theta_i\mu_i)\,dt + g(\eta(t), \xi(t))|_{t_0}^{\infty} \\
&= \int_{-\infty}^{t_0} \sum_{i} \xi_i(-\eta_i' + \eta_i\lambda_i)\,dt + \int_{t_0}^{\infty} \sum_{i} \zeta_i(-\theta_i' + \theta_i\mu_i)\,dt + g(\eta(t), \xi(t))|_{-\infty}^{\infty}
\end{aligned}$$

If this is to be zero for all ξ, then by taking ξ_i to be zero except in small intervals and for particular values of i, we can see that

$$\begin{aligned}
-\eta_i' + \eta_i\lambda_i &= 0, && \text{for } t < t_0 \\
-\theta_i' + \theta_i\mu_i &= 0, && \text{for } t > t_0.
\end{aligned}$$

This has the solution

$$\eta_i = \eta_{i,0} e^{\lambda_i t}, t < t_0$$
$$\theta_i = \theta_{i,0} e^{\mu_i t}, t > t_0$$

and if $g(\eta(t), \xi(t))$ is to also be zero as $t \to \pm\infty$, we must insist that $\eta_{i,0} = 0$ for $\lambda_i < 0$ and $\theta_{i,0} = 0$ for $\mu_i > 0$.

So $\eta(t_0)$ will be in the span of $\tilde{e}_i$ with $\lambda_i > 0$ and in the span of $\tilde{f}_i$ with $\mu_i < 0$. That is, it is perpendicular to e_i for $\lambda_i < 0$ and perpendicular to f_i for $\mu_i > 0$. In other words, $\eta(t_0)$ is perpendicular to $T_p W^u(a)$ and $T_p W^s(b)$.

Conversely, if v is perpendicular to $T_p W^u(a)$ and $T_p W^s(b)$, this defines a solution $\eta(t)$ to $dL(\eta) = 0$ with $\eta(t_0) = v$, and as we saw above, this means η is perpendicular to the image of dL. □

Proposition 8.2 *Let $f : M \longrightarrow \mathbb{R}$ be Morse. The set of metrics g so that f is Morse–Smale is a dense G_δ set.*

This is proved in Mattias Schwarz's book *Morse Homology*[**?**].

[Need non-compact version of transversality theorem]

Proof: We will apply the Transversality lemma (Lemma 3.8) with $M = P_{a,b}$, P the set $\mathrm{Met}(M)$ of C^1 metrics on our original manifold (taken with the C^1 topology), N the set of vector fields in $\gamma^* TM$, and S the zero section in N. Let F be dL.

Since γ is an embedding of an interval, its tubular neighborhood B is diffeomorphic to a ball, and thus we can choose coordinates $x_1, \ldots, x_n$ so that a is $(1, 0, \ldots, 0)$, b is $(0, 0, \ldots, 0)$, and the image of γ is $x_2 = \cdots = x_n = 0$, with $0 < x_1 < 1$.

We first note that in coordinates the function F for a given metric g are:

$$F^i_g(x) = \frac{dx^i}{dt} + g^{ij} \frac{\partial}{\partial x^j} f(x(t)) = 0. \tag{8.1}$$

If we fix a metric g_0, and use it to turn g into a linear transformation g^*, we get:

$$F_g(x) = \frac{dx}{dt} + g^* \nabla_{g_0} f(x(t)) = 0. \tag{8.2}$$

The set of metrics g is thus in bijective correspondence with the subset of sections $\Gamma(Ad(TX))$ consisting of (g_0-) positive-definite, symmetric linear transformations. This is an open set of the Banach space of sections consisting of symmetric linear transformations, so that this Banach space is a good model for the tangent space to P.

We first show that $F : P_{a,b} \times \mathrm{Met} \to N$ is a submersion on $\mathcal{F}_{a,b}$. We first fix $\gamma(t) \in \mathcal{F}_{a,b}$, and fix a target $\eta \in TN$, and show that it is in the image of

$dF_g(\gamma) : TP_{a,b} \times T\,\mathrm{Met} \to TE$.

$$dF_g(x)(\xi, \alpha) = \frac{d\xi}{dt} + \alpha^* \nabla_{g_0} f(x(t)) + g^* H_f(x)\xi + dg^*(\xi)\nabla_{g_0} f(x(t)) = \eta. \quad (8.3)$$

Here H_f is a matrix of second derivatives of f, which converges to the Hessian of f as t approches $\pm\infty$. Also note that $dg^*(\xi)\nabla_{g_0} f(x(t))$ approaches zero as t goes to $\pm\infty$. If we define

$$H(t)(\xi) = g^* H_f(x)\xi + dg^*(\xi)\nabla_{g_0} f(x(t))$$

then the equation is

$$\frac{d\xi}{dt} + \alpha^* \nabla_{g_0} f(x(t)) + H(t)(\xi) = \eta$$

where $H(t)$ converges to the Hessian of f as t goes to $\pm\infty$.

We first identify a ball around each critical point, inside which the matrix H_f is nondegenerate: by the Morse condition, the Hessian is nondegenerate at the critical points, and by the continuity of the second derivatives of f and of the determinant function, there is a neighborhood of the critical points for which H_f is nondegnerate. We furthermore require that the balls for two different critical points are disjoint. Fix a radius r smaller than the necessary radii of all such balls.

Since $\gamma(t)$ converges to a and b, for $t \to -\infty$ and $t \to +\infty$, respectively, there is a T for which $\gamma(t) \in B(a, r)$ for $t < -T$, and $\gamma(t) \in B(b, r)$ for $t > T$. Then it follows from the above that along $\gamma(t)$, the matrix H_f is nondegenerate for $|t| > T$.

For $|t| > T$, we split the equation into the following system:

$$\begin{aligned} \frac{d\xi}{dt} + H(t)\xi &= \eta\alpha^* \nabla_{g_0} f(\gamma(t)) \\ &= 0. \end{aligned} \quad (8.4)$$

We will approach the first equation in (8.4). But first we prove it in the special case where H_f is constant.

Lemma 8.3 *Suppose H_0 is a constant $n \times n$ real matrix, symmetric and nondegenerate. Pick any real number $T > 0$. Suppose $\eta(t)$ is a continuous function to $\mathbb{R}^n$ on $-\infty < t \leq -T$, such that $\lim_{t\to-\infty} \eta(t) = 0$. Then there exists a solution to the differential equation*

$$\frac{dx}{dt} + H_0 x = \eta \quad (8.5)$$

for $-\infty < t \leq -T$ so that $\lim_{t\to-\infty} x(t) = 0$. Furthermore

$$\|x\|_{C^1} \leq K\|\eta\|_{C^0}$$

for some K.

Proof: We solve for x as follows:

$$\begin{aligned}
\frac{dx}{dt} + H_0 x &= \eta(t) \\
\frac{d}{dt}\left(e^{H_0 t}x(t)\right) &= e^{H_0 t}\eta(t) \\
e^{H_0 t}x(t) &= \int_{t_0}^{t} e^{H_0 s}\eta(s)\,ds + C \\
x(t) &= e^{-H_0 t}\int_{t_0}^{t} e^{H_0 s}\eta(s)\,ds + e^{-H_0 t}C \\
x(t) &= \int_{t_0}^{t} e^{(s-t)H_0}\eta(s)\,ds + e^{-H_0 t}C
\end{aligned}$$

First, we note that this formula implies the long-term existence of solutions. Also, as usual, the parameter t_0 can be changed, with a corresponding change in C without affecting the solution.

Now we only need $\lim_{t\to-\infty} x(t) = 0$.

To do this, we first define, for every real number $s \le -T$, a space $C^k([-\infty, s])$ to be the space of C^k vector fields along $[-\infty, s]$ which converge to zero near $-\infty$. We will write C_0^k for $C^k([-\infty, -T])$ where we are assuming our ordinary domain is $(-\infty, -T]$, but the zero subscript reminds us that functions in this space must converge to zero at $-\infty$. As before, these spaces have supremum norms

$$\|f(t)\|_{C^k([-\infty,s])} = \sum_{i=0}^{k} \sup_{x\in(-\infty,s]} \left|D^i f(x)\right|.$$

The norm C_0^k when $s = -T$ is just C^k, and will be written as such. Note that the space C_0^k is a closed subspace of $C^k((-\infty, -T])$ and is therefore complete under the C^k norm.

We also split $\mathbb{R}^n$ into the positive and negative eigenspaces for H_0: $C^1([-\infty, s]) = C^1_+([-\infty, s]) \oplus C^1_-([-\infty, s])$, $\eta = \eta_+ + \eta_-$, and so on. We also denote by λ the minimum of the absolute values of the eigenvalues of H_0, and Λ is the maximum of the absolute values of the eigenvalues of H_0.

For $t < T$, we choose t_0 and C to find the following solutions:

$$\begin{aligned}
x_+ &= \int_{-\infty}^{t} e^{(s-t)H_0}\eta_+(s)\,ds \\
x_- &= \int_{-T}^{t} e^{(s-t)H_0}\eta_-(s)\,ds
\end{aligned}$$

so that x satisfies the differential equation (8.5). We now need to prove $x_+(t)$ and $x_-(t)$ converge to zero as $t \to -\infty$. The first equation involves a $-\infty$, so we first need to show the integral exists. First note that since in the integral, $s < t$, and since H_0 on this space has positive eigenvalues, that the range of eigenvalues for $(s-t)H_0$ is from $(s-t)\lambda$ to $(s-t)\Lambda$. Since $(s-t)\lambda > (s-t)\Lambda$, we have $|e^{(s-t)H_0}\eta_+| \le e^{(s-t)\lambda}|\eta_+|$.

Now let $t_0 < t < -T$.

$$\begin{aligned}\left|\int_{t_0}^{t} e^{(s-t)H_0}\eta_+(s)\,ds\right| &\le \int_{t_0}^{t}\left|e^{(s-t)H_0}\eta_+(s)\right|\,ds\\ &\le \int_{t_0}^{t} e^{(s-t)\lambda}\,|\eta_+(s)|\,ds\\ &\le |\eta_+|_{C^0}\frac{1}{\lambda}\left(e^{(t_0-t)\lambda}-1\right)\\ &\le |\eta_+|_{C^0}\frac{1}{\lambda}\end{aligned}$$

so that $x_-(t)$ is bounded as $t_0 \to -\infty$. Furthermore we can use the Cauchy criterion to prove the existence of the limit as follows. Let t_1 and t_2 be numbers both less than t_0, and without loss of generality $t_1 < t_2$. Then

$$\begin{aligned}\left|\int_{t_1}^{t} e^{(s-t)H_0}\eta_+(s)\,ds - \int_{t_2}^{t} e^{(s-t)H_0}\eta_+(s)\,ds\right| &= \left|\int_{t_1}^{t_2} e^{(s-t)H_0}\eta_+(s)\,ds\right|\\ &\le \int_{t_1}^{t_2}\left|e^{(s-t)\lambda}\right|\,|\eta_+(s)|\,ds\\ &\le \|\eta_+\|_{C^0}\int_{t_1}^{t_2} e^{(s-t)\lambda}\,ds\\ &= \|\eta_+\|_{C^0}\frac{1}{\lambda}\left(e^{\lambda(t_2-t)}-e^{\lambda(t_1-t)}\right)\\ &\le \|\eta_+\|_{C^0}\frac{1}{\lambda}\left(e^{\lambda(t_0-t)}-0\right)\\ &= \|\eta_+\|_{C^0}\frac{1}{\lambda}e^{\lambda t_0}e^{-\lambda t}\end{aligned}$$

which as t_0 goes to $-\infty$, goes to zero. Thus, by the Cauchy criterion, the limit

$$x_+ = \int_{-\infty}^{t} e^{(s-t)H_0}\eta_+(s)\,ds$$

exists. Now we need to show that as $t \to -\infty$, $x_+(t)$ goes to zero. Again we estimate

$$\begin{aligned}|x_+(t)| &\le \int_{-\infty}^{t}\left|e^{(s-t)\lambda}\right|\,|\eta_+(s)|\,ds\\ &\le \|\eta_+\|_{C^0([-\infty,t])}\int_{-\infty}^{t} e^{(s-t)\lambda}\,ds\\ &= \|\eta_+\|_{C^0([-\infty,t])}\frac{1}{\lambda}.\end{aligned}$$

Now as t goes to $-\infty$, $\|\eta_+\|_{C^0([-\infty,t])}$ goes to zero, so $x_+(t)$ also goes to zero. Also as a corollary,

$$\|\eta_+\|_{C^0} \le \|\eta_+\|_{C^0}\frac{1}{\lambda}. \tag{8.6}$$

Similarly, we compute for the negative eigenspace, for $t < -T < 0$. Again, $|e^{(s-t)H_0}\eta_-| \leq e^{-(s-t)\lambda}|\eta_-|$ since $s - t > 0$ and the eigenvalues for $(s-t)H_0$ range from $-(s-t)\Lambda$ to $-(s-t)\lambda$. Again we compute:

$$\begin{aligned}
|x_-(t)| &= \left| \int_{-T}^{t} e^{(s-t)H_0}\eta_-(s)\, ds \right| \\
&\leq \int_{t}^{-T} \left| e^{(s-t)H_0}\eta_-(s) \right| ds \\
&\leq \int_{t}^{-T} e^{-(s-t)\lambda} \left|\eta_-(s)\right| ds \\
&= \int_{t}^{t/2} e^{-(s-t)\lambda} \left|\eta_-(s)\right| ds \\
&+ \int_{t/2}^{-T} e^{-(s-t)\lambda} \left|\eta_-(s)\right| ds \\
&\leq \|\eta_-\|_{C^0([-\infty,t])} \int_{t}^{t/2} e^{-(s-t)\lambda}\, ds \\
&+ \|\eta_-\|_{C^0} \int_{t/2}^{-T} e^{-(s-t)\lambda}\, ds \\
&= \|\eta_-\|_{C^0([-\infty,t])} \frac{1}{\lambda}\left(1 - e^{\lambda t/2}\right) \\
&+ \|\eta_-\|_{C^0} \frac{1}{\lambda}\left(e^{\lambda t/2} - e^{-\lambda T}e^{\lambda t}\right) \\
&\leq \|\eta_-\|_{C^0([-\infty,t])} \frac{1}{\lambda} + \|\eta_-\|_{C^0} \frac{1}{\lambda}\left(e^{\lambda t/2} - e^{-\lambda T}e^{\lambda t}\right)
\end{aligned}$$

so that as t goes to $-\infty$, $\|\eta_-\|_{C^0([-\infty,t])}$ goes to zero and $e^{\lambda t/2}$ and $e^{\lambda t}$ go to zero, so that the above computation shows that $x_-(t)$ goes to zero.

Now using these computations we find

$$\begin{aligned}
\|x_-\|_{C^0} &\leq \|\eta_-\|_{C^0} \frac{1}{\lambda} + \|\eta_-\|_{C^0} \frac{1}{\lambda}\left(e^{\lambda t/2} - e^{-\lambda T}e^{\lambda t}\right) \\
&\leq \|\eta_-\|_{C^0} \frac{1}{\lambda} + \|\eta_-\|_{C^0} \frac{1}{\lambda}(1-0) \\
&= \|\eta_-\|_{C^0} \frac{2}{\lambda}.
\end{aligned}$$

Putting this estimate together with equation (8.6) we get

$$\|x\|_{C^0} \leq \|x_+\|_{C^0} + \|x_-\|_{C^0} \leq \|\eta\|_{C^0} \frac{3}{\lambda}.$$

We can also get a C^1 bound on x using the differential equation itself.

$$\begin{aligned}\frac{d}{dt}x(t) &= \eta(t) - H_0 x(t)\\ \|x'(t)\|_{C^0} &\le \|\eta\|_{C^0} + \|H_0\|\|x\|_{C^0}\\ &\le \|\eta\|_{C^0} + \Lambda\|x\|_{C^0}\\ &\le \|\eta\|_{C^0} + \frac{3\Lambda}{\lambda}\|\eta\|_{C^0}\\ &\le \left(1 + \frac{3\Lambda}{\lambda}\right)\|\eta\|_{C^0}\end{aligned}$$

Therefore

$$\|x\|_{C^1} = \|x\|_{C^0} + \|x'\|_{C^0} \le \left(1 + \frac{2+3\Lambda}{\lambda}\right)\|\eta\|_{C^0}.$$

□

Corollary 8.4 *For each T, $\frac{d}{dt} + H_0 : C^1([-\infty, -T]) \to C^0([-\infty, -T])$ has a bounded right inverse, whose operator bound is bounded above with a bound independent of T.*

Proof: Define $G : C^0 \to C^1$ by the equations already given in the proof of the previous lemma:

$$\begin{aligned}G_+(\eta) &= \int_{-\infty}^{t} e^{(s-t)H_0}\eta_+(s)\,ds\\ G_-(\eta) &= \int_{T}^{t} e^{(s-t)H_0}\eta_-(s)\,ds\end{aligned}$$

and $G(\eta) = G_+(\eta) \oplus G_-(\eta)$.

The previous proof showed that G is a right inverse, and that it is bounded with operator norm

$$1 + \frac{2+3\Lambda}{\lambda}$$

which is independent of T. □

Clearly, we can apply these results to $[T, \infty]$ given η defined for $t \ge T$, also.

We will now tackle the original (non-constant) version of the differential equation (8.4) on $|t| > T$.

Lemma 8.5 *The equation*

$$\frac{d\xi}{dt} + H(t)\xi = \eta \tag{8.7}$$

has a solution on $|t| > T$ so that as $t \to \pm\infty$, $\xi(t) \to 0$.

Proof: Since the domain $|t| > T$ is disconnected, we can prove this for $t > T$, and the proof for $t < T$ will be analogous.

We also know that $H(t)$ converges to H_0 when $t \to -\infty$ since H is the Hessian and $t \to -\infty$ means we are moving toward a particular point on the manifold: the critical point we start from.

We write $H(t) = H_0 + R(t)$ and compute:

$$\begin{aligned}\left(\frac{d}{dt} + H\right) G\eta &= \left(\frac{d}{dt} + H_0 + R\right) G\eta \\ &= \eta + RG\eta \\ &= (1 + RG)\eta\end{aligned}$$

Now $R(t) = H(t) - H_0$ is an operator from C_0^1 to C_0^0 and, in this operator norm, converges to zero as T goes to $-\infty$. In particular, there is a T so that

$$\|R\| < \frac{1}{2}\left(1 + \frac{2 + 3\Lambda}{\lambda}\right)^{-1}$$

in operator norm. For this T take the G described above so that $\|RG\| < 1/2$ in operator norm. For such T (or smaller), (1+RG) is invertible, a fact that can be proved by the contraction principle as follows: if $y \in C_0^0$ and we wish to solve

$$y = (1 + RG)x$$

for $x \in C_0^0$, we can rewrite this as

$$x = y - RGx$$

and define $A : C_0^0 \longrightarrow C_0^0$ by

$$A(x) = y - RGx.$$

Then $\|A(x_1) - A(x_2)\| = \|RG(x_2 - x_1)\| \leq \frac{1}{2}\|x_2 - x_1\|$, so A is a contraction mapping. Thus there is a unique fixed point x of A which is the unique solution to $x = y - RGx$. Furthermore, $\|x\| \leq \|y\| - \|RGx\| \leq \|y\| - \frac{1}{2}\|x\|$ so $\|x\| \leq 2\|y\|$, so the inverse of $1 + RG$ is well-defined and bounded.

If we let $\xi = G(1 + RG)^{-1}\eta$, then

$$\left(\frac{d}{dt} + H(t)\right)\xi = \eta.$$

We furthermore have the following bound on ξ:

$$\begin{aligned}\|\xi\|_{C^1} &\leq \|G\| \left\|(1 + RG)^{-1}\right\| \|\eta\|_{C^1} \\ &\leq 2\left(1 + \frac{2 + 3\Lambda}{\lambda}\right)\|\eta\|_{C^1}\end{aligned}$$

□

We therefore have ξ for $|t| > T$ which satisfies the equation (8.7) and converges to zero as $t \to \pm\infty$.

We then define $\alpha = 0$ for $t < -T$, and note that (α, ξ) solves (8.3) for $t < -T$.

We apply the same procedure for $t \to +\infty$ at the critical point there, in that neighborhood. In this way, we have (α, ξ) defined for $|t| > T$ which satisfies (8.3) for $|t| > T$. We can extend ξ arbitrarily (using, say, a cutoff function) over all t.

We extend α in a small neighborhood of the critical points a and b, whose intersection with the image of γ is $\gamma((-\infty, -T])$ and $\gamma([T, \infty))$, to also be zero.

The idea next is to extend α to satisfy (8.3) for $|t| \leq T$. Intuitively, since $\nabla f(x(t))$ is never zero, we simply choose any g_0-symmetric endomorphism that sends $\nabla f(x(t))$ to

$$\eta - \frac{d\xi}{dt} - H(t)\xi.$$

To show we can do this globally, we first need a general lemma:

Lemma 8.6 *Let B be a manifold, and $E \to B$ be a fiber bundle where fibers are affine spaces, and transition functions are affine maps. Then E has a section.*

Proof: We fix a trivialization $\{U_\alpha, E_\alpha, \phi_{\alpha\beta}\}$. Over each neighborhood U_α we take a constant section s_α. We then take a partition of unity Φ_α over $\{U_\alpha\}$ and the section we take is

$$s = \sum_\alpha \Phi_\alpha s_\alpha.$$

Since $\sum_\alpha \Phi_\alpha(x) = 1$, this sum remains in the affine space. So s is a section of E. $\square$

Corollary 8.7 *If (B, g_0) is a Riemannian manifold (not necessarily compact) and v a nonvanishing vector field, and w any vector field, there exists a section A of $\mathrm{Hom}(TB, TB)$ which is g_0-symmetric so that $Av = w$.*

Proof: Consider the vector bundle of g_0-symmetric bilinear forms in $\mathrm{Hom}(TB, TB)$. At each point, the space of such that assigns w to v forms an affine space, nonempty and of the correct dimension if v is nonzero. As a sub-fiber bundle, this is an affine bundle as above. Applying the lemma yields the result. $\square$

Now extend $\eta - \frac{d\xi}{dt} - H(t)\xi$ to a vector field in a small tubular neighborhood U around the image of γ. This neighborhood can be chosen to be the neighborhood around the critical points a and b where we have already chosen α to be zero, union a small neighborhood around the image $\gamma([-T, T])$ (small enough to be far from other critical points). Since $\nabla f(x)$ is nonzero in this second neighborhood piece, then if we use this neighborhood for B above, and $v = \nabla f(x)$ and w to be the extension of

$$\eta - \frac{d\xi}{dt} - H(t)$$

we just defined to our new neighborhood U, we have, by the above corollary, a g_0-symmetric α so that (8.3) holds for all t. We extend α to all of X using

a cutoff function to zero outside the neighborhood U. In this way, we have a $(\xi, \alpha) \in TP \times TB$ so that $dF_{(x,g)}(\xi, \alpha) = \eta$.

So we have that F is transverse to the zero section, and we apply the Transversality lemma to show that for a G_δ dense set of metrics g, F_g is transverse to the zero section. □

Proposition 8.8 *Suppose the Riemannian metric g is fixed. The set of functions f that are Morse–Smale is G_δ and dense.*

Remark 8.1 *Modifying f is somewhat problematic since doing so may in fact move the set of critical points, and in particular, move a or b. We will show we can insist on modifying f without changing its critical set at all and still achieve this proposition.*

Proof: Similarly with the previous proposition, we will use the Transversality lemma (Lemma 3.8). This time we replace P with the set of Morse functions from M to $\mathbb{R}$.

Again, we choose a coordinate chart B around the image of γ so that a is at $(1, 0, \ldots, 0)$, b is at $(0, 0, \ldots, 0)$, and the image of γ is $x_2 = \cdots = x_n = 0$ with $0 < x_1 < 1$. In these coordinates,

$$dF(\xi, h) = \frac{d\xi}{dt} + H(t)\xi + \nabla(h).$$

If we have a given η so that

$$\frac{d\xi}{dt} + H(t)\xi + \nabla(h) = \eta, \tag{8.8}$$

we again choose $\xi \in C_0^1$ as in Lemma 8.5 so that

$$\frac{d\xi}{dt} + \xi H(t) = 0$$

for $|t| > T$, and extend ξ in a C^1 fashion to all t via a cutoff function.

We now need to find h.

First, define the vector field $\psi = \frac{d\xi}{dt} + H(t)\xi$ along the image of γ. Note that $\psi = 0$ for $|t| > T$. Now extend ψ to $x_2 = \cdots = x_n = 0$ by defining it to be zero when $x_1 < 0$ or $x_1 > 1$. Then extend ψ to B by $\psi(x_1, x_2, \ldots, x_n) = \eta(x_1)$. Note that $\psi = 0$ in a neighborhood of the critical points a and b. Also note that ψ is in C^1.

Then define $h(x_1, \ldots, x_n) = \int_C \psi \dot{d}r$ using a curve C that goes from $(0, \ldots, 0)$ to $(x_1, 0, \ldots, 0)$, then from there linearly to $(x_1, \ldots, x_n)$. Then $\nabla(h) = \psi$, so the equation (8.8) is satisfied.

Now extend h to the rest of M using a cutoff function to bring it to zero outside B.

Examine ∇h in a neighborhood of the critical points. In a neighborhood around each critical point, $\psi = 0$. So h is constant in those neighborhoods. In particular, $f + \epsilon h$ has a and b as nondegenerate critical points. Furthermore, h is in C^2 since ψ is in C^1. □

Proposition 8.9 *If f is Morse–Smale, then $\mathcal{F}_{a,b}$ is a smooth manifold of dimension*

$$a) - b).$$

Note that we will soon prove that $\mathcal{F}_{a,b}$ is homeomorphic to $W(a,b)$, so this will soon be unnecessary, but we do this anyway so that we can see how these ideas can be proved using analytic and ODE techniques.

Proof: The fact that $\mathcal{F}_{a,b}$ is smooth follows from the implicit function theorem and the surjectivity of dL at $\mathcal{F}_{a,b}$. The dimension can be computed by the tangent space $T_\gamma \mathcal{F}_{a,b}$. In the proof of Theorem 8.1, we saw that we can view $T_\gamma P_{a,b}$ as vector fields in $\gamma^* TM$ which vanish at $\pm\infty$, and these are in $T_\gamma \mathcal{F}_{a,b}$ if and only if they are in the kernel of dL, that is, if they satisfy the equation $dL(\xi) = 0$. Such as solution ξ has a value at p, and is determined by this value.

Let T be large enough that $\gamma((-\infty, T])$ is in the coordinate neighborhood U around a, and so that $\gamma([T, \infty))$ is in the coordinate neighborhood V around b. For the solution to converge to zero at $-\infty$, it must be that $\xi(-T)$ in the span of e_i for i with $\lambda_i < 0$. Therefore $\xi(t_0) \in W^u(a)$. Similarly, for the solution to converge to zero at $+\infty$, it must be that $\xi(T)$ is in the span of f_i for i with $\mu_i > 0$. Therefore $\xi(t_0) \in W^s(b)$.

Therefore the kernel of dL at γ is represented by the set of initial values $\xi(t_0)$ which are in $W^u(a) \cap W^s(b)$. If this intersection is transverse, its dimension is $\dim W^u(a) + \dim W^s(b) - n = a) + (n - b)) - n = a) - b)$. □

[[Does the following theorem work for diffeomorphism? Comment on $d\phi$.]]

8.2 Equivalence of $W(a,b)$ and $\mathcal{F}_{a,b}$

Proposition 8.10 *Define $\phi : \mathcal{F}_{a,b} \longrightarrow M$ by $\phi(\gamma) = \gamma(0)$. Then ϕ maps $\mathcal{F}_{a,b}$ homeomorphically onto $W(a,b)$.*

Proof: First, ϕ is continuous by the definition of the compact-open topology. It maps $\mathcal{F}_{a,b}$ onto $W(a,b)$ by the definition of $W(a,b)$. We have $\phi^{-1}(p) = \gamma_p(t)$, the unique flow line that has $\gamma_p(0) = p$ (see Theorem 4.6). By Theorem 4.6, ϕ^{-1} is continuous. □

[[Give example]]

8.3 The moduli space of gradient flows $\mathcal{M}(a,b)$

Now since the gradient flow equations do not depend on time, we note that if $\gamma(t) \in W(a,b)$, then the time translate $\gamma_s(t) = \gamma(t+s)$ is also in $W^(a,b)$, for all $s \in \mathbb{R}$. Furthermore, this time translation can be described in terms of the flow map T defined in the statement of Theorem 4.6 as follows: $T(\gamma(t), s) = \gamma_s(t)$.

In other words, $\mathbb{R}$ acts smoothly on $W(a, b)$. Assuming $a \neq b$, $\mathbb{R}$ acts without fixed points (as can be seen by considering $f(T(\gamma(t), s))$ and the fact that $f(\gamma(t))$ is a decreasing function).

This action can be viewed in terms of $W(a, b) \cong W(a, b)^t \times \mathbb{R}$ as $(p, r) \to (p, r + s)$, so there is nothing unusual about this action (orbits are closed, the quotient is a Hausdorff manifold, and so on). We can therefore perform the quotient $W(a, b)/\mathbb{R} \cong W(a, b)^t$. Equivalently, we can consider $\mathcal{F}_{a,b}/\mathbb{R}$, which identifies gradient flow lines without distinguishing time translates of the same flow. We define

$$\mathcal{M}(a, b) = \mathcal{F}_{a,b}/\mathbb{R} \cong W(a, b)/\mathbb{R} \cong W(a, b)^t \tag{8.9}$$

and define this to be the *moduli space of flow lines* from a to b, or the space of *unparameterized flow lines*.

By the diffeomorphism $\mathcal{M}(a, b) \cong W(a, b)^t$, we can view $\mathcal{M}(a, b)$ as a subspace of M if we so choose.

Exercise 8.3 *Describe explicitly the diffeomorphism $\mathcal{M}(a, b) \longrightarrow W(a, b)^t$ and its inverse, without referring to the intermediate space $\mathcal{F}_{a,b}/\mathbb{R}$ or $W(a, b)/\mathbb{R}$. You need not prove that it, or its inverse, is smooth or even continuous.*

Corollary 8.11 *If t_1 and t_2 are values strictly between $f(b)$ and $f(a)$, then $W(a, b)^{t_1}$ and $W(a, b)^{t_2}$ are diffeomorphic. That is, the diffeomorphism type of $W(a, b)^t$ does not depend on t.*

Example 8.1 *Consider the tilted torus from Exercise 6.4:*

Again, we have tilted the torus toward its hole, and we let f be the height function.

There are four critical points; a has index 2, b and c have index 1, and d has index 0. As the figure depicts, the moduli spaces $\mathcal{M}(a, b)$, $\mathcal{M}(a, c)$, $\mathcal{M}(b, d)$, and $\mathcal{M}(c, d)$ are all spaces consisting of two distinct points each. We will denote these flows by α_i, β_i, γ_i, and δ_i respectively. All points on the torus not lying on any of these flows is on a flow in $\mathcal{M}(a, d)$. This moduli space is one dimensional, and indeed is the disjoint union of four open intervals. If the torus is viewed in

the usual way as a square with opposite sides identified, then these flows can be depicted as follows.

Example 8.2 *Consider the sphere example from Exercise 4.4. There are two critical points: a maximum at $N = (0, 0, 1)$ and a minimum at $S = (0, 0, -1)$. Each flow line from N to S is a longitude line. The moduli space $\mathcal{M}(N, S)$ is one dimensional and parameterized by longitude. It is homeomorphic to the circle S^1. Hence we see that moduli spaces need not be cells.*

8.4 Height and arclength parameterizations

We now consider another way of thinking of $\mathcal{M}(a, b)$ than simply flow lines modulo translation. We can, instead, fix the translation altogether by parameterizing by height:

Proposition 8.12 *The set of height-parameterized flows from a to b, $H_{a,b}$, is diffeomorphic to $\mathcal{M}(a, b)$.*

Proof: We view $\mathcal{M}(a, b)$ as $W(a, b)^t$ for some fixed value t between $f(a)$ and $f(b)$. Analogously to the gradient flow T, we define the height-parameterized gradient flow $Z_h(x)$ as follows: let $x \in W(a, b)^t$ and h a value between $f(a)$ and $f(b)$. We define $Z_h(x)$ to be the solution to the differential equation

$$\frac{d}{dh} Z_h = \frac{\nabla_{Z_h}(f)}{|\nabla_{Z_h}(f)|^2}$$

subject to the initial condition $Z_t = x$, whenever $x \in W(a, b)^t$. By section 4.5, we know that for fixed x, $Z_h(x)$ is a height-parameterized flow, and is defined for $h \in (f(b), f(a))$. By smooth dependence on initial conditions, we know that the function that assigns $x \mapsto Z_h(x) \in H_{a,b}$ is smooth.

We now consider the map $\xi : H_{a,b} \to W(a, b)^t$ defined by $\xi(\eta) = \eta(t)$. This is the inverse of the previous function, and is also smooth; therefore, ξ is a diffeomorphism. $\square$

It will also be convenient to talk about arclength-parameterized flows. These are flows that are reparameterized so that the parameter is arclength. They satisfy the differential equation

$$\frac{d}{ds}\eta(s) = -\frac{\nabla_{\eta(s)}(f)}{|\nabla_{\eta(s)}(f)}. \tag{8.10}$$

As with height-parameterized flows, they are only defined on a bounded interval. This is because each gradient flow is of finite length. Unlike height-parameterized flows, though, the interval may vary among flows, even flows between the same critical points. But as we will now prove, there is a uniform upper bound on this length. This will be important later in this section when we discuss compactness issues.

Proposition 8.13 *Let M be a compact manifold, with Riemannian metric g and Morse function $f : M \longrightarrow \mathbb{R}$. The set of lengths of gradient flow lines is bounded.*

Proof: We will prove this only when g is "nice", though this is true in general.

For each critical point p_i let B_i be a ball around p_i such that the flow lines are given by

$$\frac{dx_i}{dt} = c_i x_i$$

in some coordinate chart where B_i is the coordinate ball around p_i. Let B be the union of these B_i. Outside B, by the compactness of M and continuity of $|\nabla f|$, there is a minimum value m for $|\nabla f|$, with $m > 0$.

Now let $\gamma(t)$ be a gradient flow line, and let a and b be the limits as t goes to $-\infty$ and ∞, respectively. For all critical points p_i let $S_i \subset \mathbb{R}$ be set of real numbers where $\gamma(t) \in B_i$, and let $S_0 = \mathbb{R} - \cup S_i$. Then the length of γ on S_0 is

$$\begin{aligned}
\int_{S_0} |\gamma'(t)|\, dt &\leq \int_{S_0} \frac{|\nabla f|}{m} |\gamma'(t)|\, dt \\
&\leq \frac{1}{m} \int_{S_0} \langle \nabla f, \gamma'(t) \rangle\, dt \\
&= \frac{1}{m} \int_{S_0} \frac{d}{dt} f(\gamma(t))\, dt \\
&\leq \frac{1}{m} \int_{\mathbb{R}} \frac{d}{dt} f(\gamma(t))\, dt \\
&= \frac{1}{m}(f(b) - f(a)) \\
&\leq \frac{1}{m}(\sup_{x \in M} f(x) - \inf_{x \in M} f(x))
\end{aligned}$$

Now we compute the length of γ in each of the S_i. First, note that γ cannot go through a given ball B_i more than once, since the outgoing flows cross the

boundary of B_i at points c where $f(c) < f(p_i)$, and enter at points d where $f(d) > f(p_i)$. We see, then, that each S_i is connected.

On B_i consider the Euclidean metric E given by the coordinate chart under which the flow in B_i is "nice". Let $g' = kE$ be a constant scaling of this metric such that $g'(v, w) > g(v, w)$ for all tangent vectors v and w at any point $x \in B_i$.

The length of γ in S_i is

$$\begin{aligned}\int_{S_i} |\gamma'(t)|\,dt < \int_{S_i} |\gamma'(t)|_{g'}\,dt &= \int_{S_i} \sqrt{\sum_i x_i'(t)^2}\,dt \\ &\leq \int_{S_i} \sum_i |x_i'(t)|\,dt \\ &\leq \sum_i \int_{S_i} |x_i'(t)|\,dt.\end{aligned}$$

By examining the explicit solution

$$x_i = a_i e^{-c_i t}$$

to the flow equations, we see that inside B_i, each x_i is monotonic. Therefore for each i, $x_i'(t)$ is either always $|x_i'(t)|$ or always $-|x_i'(t)|$. We can therefore switch the absolute value sign and the integral sign. If the g' radius of B_i is R_i, then

$$\begin{aligned}&\leq \sum_i \left| \int_{S_i} x_i'(t)\,dt \right| \\ &= \sum_i |x_i(\sup(S_i)) - x_i(\inf(S_i))| \\ &\leq \sum_i 2R_i = 2nR_i\end{aligned}$$

If c is the number of critical points of f, then the total length of the flow is less than

$$\frac{1}{m}(\sup_{x\in M} f(x) - \inf_{x\in M} f(x)) + 2n\sum_{p_i} R_i.$$

This provides us with a uniform bound on the length of a gradient flow. □

The fact that the actual interval varies depending on the flow is irksome, so when we talk about reparameterizing a gradient flow to be arclength-parameterized, we define $\eta(s) = a$ for all s to the left of the interval on which it would be defined, and $\eta(s) = b$ for all s to the right of that interval. Then $\eta(s)$ is constant at a, then flows with constant speed 1, then is constant at b.

Analogously to $H_{a,b}$, we can define the space of arclength-parameterized flows $A_{a,b,t}$ to be the space of continuous curves $\eta(s)$ so that

- There is an interval $[c, d] \subset \mathbb{R}$,
- For $s \leq c$, $\eta(s) = a$,

- For $s \geq d$, $\eta(s) = b$,
- For $s \in (c, d)$, $\eta(s)$ satisfies (8.10), and
- $\eta(0) \in W(a, b)^t$.

Exercise 8.4 *Analogously to the proof of Proposition 8.12, prove that $A_{a,b,t}$ is diffeomorphic to $W(a, b)^t$.*

8.5 Compactness issues concerning moduli spaces

From Example 8.1, we see that some moduli spaces, like $\mathcal{M}(a, b)$, are compact, and others, like $\mathcal{M}(a, d)$, are not. Consider the situation with $\mathcal{M}(a, d)$, which is a union of four open intervals. Take the interval corresponding to the northeast corner of the figure. As we move along the space of flows to the top, we seem to approach a union of two flow lines: one from a to b and one from b to d. If we move instead to the bottom, we appear to approach another union of two flow lines: from a to c then from c to d.

These unions might be called "piecewise flows" or "broken flows" in analogy to phrases like "piecewise continuous" or "piecewise linear". Of course, the analogy is not complete since the gradient flow from a to b has as its domain the entire real line, and if you wanted to tack on the flow from b to d you would have no more domain to use. But if we take arclength parameterization or height-parameterization this problem disappears.

Along these lines, consider the following (erroneous) view of the convergence we are observing: let $\eta_1, \eta_2, ...$ be a sequence in $\mathcal{M}(a, d)$ that goes upward in the example we are considering, and for each η_i, let γ_i be a gradient flow in $\mathcal{F}_{a,b}$ representing $\eta_i \in \mathcal{F}_{a,b}/\mathbb{R}$. It is not true that γ_i converges to a piecewise flow, or indeed, that it converges at all. We can choose γ_i so that it converges to the flow from a to b; or we can choose it so that it converges to the flow from b to d; or we can choose it so that it converges to the constant flow at b; or the constant flow at a; or the constant flow at d. We can choose the lift so that it does not converges at all.

Exercise 8.5 *Convince yourself that all these choices can be made.*

The sense that a sequence of flows in the moduli space $\mathcal{M}(a, d)$ may converge to piecewise flows must involve the space $\mathcal{M}(a, b)$ directly, without choosing lifts back to $\mathcal{F}_{a,b}$. This can be done through height-parameterization or arclength-parameterization, and in that context piecewise flows, and the convergence to piecewise flows, are easy to describe.

In this section, we will prove that the only way the moduli spaces can fail to be compact is because of the broken flow lines. Later, we will form a compactification of the moduli space that includes these broken flow lines, and show that this compactification is compact, and that the neighborhood of these broken flow lines in the compactification is well-behaved, in the sense that the compactification is a manifold with corners.

To state these results, we first need a definition.

Definition 8.14 *Suppose that a and b are critical points of f. Then we say that $a > b$ if $W(a,b)$ is nonempty; equivalently if there is a flow line $\gamma(t)$ with the property that*

$$\lim_{t\to\infty} \gamma(t) = a, \qquad \lim_{t\to-\infty} \gamma(t) = b.$$

Define $l(a,b)$ to be the largest integer l for which there is a chain of critical points $a = a_0 > a_1 > \cdots > a_l > a_{l+1} = b$ and we say that b is a successor *of a if $l(a,b) = 1$, i.e. there is no critical point c with $a > c > b$.*

Lemma 8.15 *Let $\{\alpha_n\}$ be a sequence in the moduli space $\mathcal{M}(a,b)$ and suppose that α_n does not have a convergent subsequence in $\mathcal{M}(a,b)$ as $n \to \infty$. Then there is a subsequence of α_n and a finite sequence of critical points*

$$a = a_0 > a_1 > \cdots > a_l > a_{l+1} = b$$

with $l \geq 1$ and flow lines γ_i joining a_{i-1} to a_i, where $1 \leq i \leq l$, with the following property. Given $\epsilon > 0$ there is an integer $N > 0$ such that

$$d(\gamma_i, \alpha_n) = \inf_t d(\gamma_i(0), \alpha_n(t)) < \epsilon$$

for all $n \geq N$ and in the subsequence.

Proof: Let t be a fixed value between $f(b)$ and $f(a)$. We view $\mathcal{M}(a,b)$ as the space of arclength-parameterized flows $A_{a,b,t}$.

We first show that α_n is equicontinuous. If $\epsilon > 0$, we take $\delta = \epsilon$, and if $|s-s_0| < \delta$, then by the Mean Value theorem, $|\alpha_i(s)-\alpha_i(s_0)| \leq \max|\dot{\alpha}_i||s-s_0| < \delta = \epsilon$.

Since the manifold is compact, the sequence is uniformly bounded. By the Arzela-Ascoli theorem, the sequence has a convergent subsequence, uniformly on compact sets. By Proposition 8.13, the domain can be taken to be such a compact set. Without loss of generality, we will assume that our sequence is such a convergent subsequence. The only question is whether the limit is a piecewise flow.

If $\alpha_i(s_0)$ converges to a critical point, we will say nothing of regularity. If it does not converge to a critical point, we can find an interval $[a,b]$ around s_0 and a subsequence so that $\alpha_i(s_0)$ is bounded away from a critical point.

Away from critical points, $V(\alpha) = -\frac{\nabla_\alpha(f)}{|\nabla_\alpha(f)|}$ is a continuous vector field. As such, if $\alpha_i(s)$ converges uniformly on $[a,b]$ to $\alpha(s)$, then we consider the convergence of $V(\alpha_i(s))$ on $[a,b]$.

Certainly, $V(\alpha)$ is uniformly continuous on the compact subset of M containing $\alpha_i([a,b])$ but excluding small balls around the critical points of f. If $\epsilon > 0$, we choose the δ involved in uniform continuity of $V(\alpha)$ in the domain just described. Then by the uniform convergence of α_i, we can choose an N so that $|\alpha_i(s) - \alpha(s)| < \delta$ for all $i > N$.

So

$$|V(\alpha_i(s)) - V(\alpha(s))| < \epsilon.$$

In other words, $\dot{\alpha}_i(s) = -\frac{\nabla f}{|\nabla f|} \circ \alpha_i(s)$ converges uniformly to $-\frac{\nabla f}{|\nabla f|} \circ \alpha(s)$.

The fact that $\alpha_i(s)$ converges uniformly to $\alpha(s)$ and $\dot{\alpha}_i(s)$ converges uniformly implies that $\lim_{i\to\infty} \dot{\alpha}_i(s) = \dot{\alpha}(s)$. Since $\dot{\alpha}_i(s)$ converged uniformly to $-\frac{\nabla f}{|\nabla f|} \circ \alpha(s)$, we have that

$$\dot{\alpha}(s) = -\frac{\nabla f}{|\nabla f|} \circ \alpha(s),$$

in other words, that $\alpha(s)$ is on a flow line except where $\alpha(s)$ is a critical point. □

Let M be a compact Riemannian manifold without boundary, and $f : M \to \mathbb{R}$ a C^2 Morse function. Let $M(a, b)$ be the moduli space of Morse flows between critical points a and b (with parametrization). Note that by associating each with their value at 0, we can associate $M(a, b)$ with the set of points of M which lie on flows from a to b. Let $\bar{M}(a, b)$ be the union

$$\bar{M}(a, b) = \bigcup_{c_1, c_2, \ldots, c_k} M(a, c_1) \times \cdots \times M(c_k, b)$$

topologized as a subset of M.

Theorem 8.16 *The space $\bar{M}(a, b)$ is compact.*

Proof:

[... in chap3.8?]

[[From chap 5:]]

One way of viewing this retraction is in terms of the *flow energy* of a curve

$$\alpha : (s_0, s_1) \longrightarrow M$$

which is defined to be

$$E_{s_1}^{s_0}(\alpha) = \frac{1}{2} \int_{s_0}^{s_1} \left(|\frac{d\alpha}{dt}(s)|^2 + |\nabla_{\alpha(s)}(f)|^2 \right) ds.$$

Lemma 8.17

$$E_{s_1}^{s_0}(\alpha) = f(\alpha(s_0)) - f(\alpha(s_1)) + \frac{1}{2} \int_{s_0}^{s_1} \left(|\frac{d\alpha}{dt}(s) + \nabla_{\alpha(s)}(f)|^2 \right) ds.$$

Proof: This lemma follows from integrating the equality

$$\begin{aligned}\left|\frac{d\alpha}{dt}(s)+\nabla_{\alpha(s)}(f)\right|^2 &= \left\langle \frac{d\alpha}{dt}(s)+\nabla_{\alpha(s)}(f),\ \frac{d\alpha}{dt}(s)+\nabla_{\alpha(s)}(f)\right\rangle \\ &= \left|\frac{d\alpha}{dt}(s)\right|^2+\left|\nabla_{\alpha(s)}(f)\right|^2+2\left\langle \frac{d\alpha}{dt}(s),\ \nabla_{\alpha(s)}(f)\right\rangle \\ &= \left|\frac{d\alpha}{dt}(s)\right|^2+\left|\nabla_{\alpha(s)}(f)\right|^2+2df\left(\frac{d\alpha}{dt}(s)\right).\end{aligned}$$

□

Now let a and b be critical points of f and let $P_{a,b}$ be the space of L^2 - convergent paths $\alpha : \mathbb{R} \longrightarrow M$ such that

$$\lim_{t\to-\infty}\alpha(t)=a, \quad \lim_{t\to+\infty}\alpha(t)=b.$$

L^2 convergence means that $\int_{-\infty}^{\infty}|\frac{d\alpha}{dt}(s)|^2ds<\infty$. This lemma shows that the absolute minima of $E_{-\infty}^{\infty}: P_{a,b}\longrightarrow\mathbb{R}$ are precisely those parameterized paths in $P_{a,b}$ which satisfy the flow equations. This is the space we referred to as $\mathcal{F}_{a,b}$.

[

Chap 3.7

Attaching map Franks' theorem ? Maybe later? Case where critical points are successors? Case where relative dimension is 1

Chap 3.8

Gluing? $\mathcal{F}_{a,b}$ tubular neighborhood in $P_{a,b}$ Gluing Behavior of ends Manifold with corners

Chap 3.9

Hutchings compactification of $W^u(a)$ examples (including hexagon) Compactness Behavior of ends CW complex explicitly

]

8.6 The Morse–Smale chain complex

We have seen that a smooth manifold can be described as a CW complex, with a cell of dimension λ for each critical point of index λ. We also know that the homology of a CW complex can be computed from the cells of the CW complex by defining a λ-chain to be a formal sum of cells of dimension λ (taken with orientation), and by defining boundary homomorphisms to describe the geometric boundary of the cells in question.

We can therefore compute the homology of the manifold using a Morse–Smale function $f : M^n \longrightarrow \mathbb{R}$ directly, by defining the following complex:

For each integer i with $0 \le i \le n$ (where n is the dimension of the manifold), let C_i be the free abelian group generated by critical points of f of index i. Let

$\partial_i : C_i \longrightarrow C_{i-1}$ be defined on c a critical point of index i as

$$\partial_i c = \sum_b n_{c,b} b$$

where the sum is taken over all critical points b of index $i-1$, and $n_{c,b} \in \mathbb{Z}$ is the number of flow lines from c to b, taken with sign. To define ∂_i on more general chains (sums of critical points) we extend it linearly.

[Explain why $n_{c,b}$ is finite (0-dimensional, compact).
Explain where the sign comes from (orientation)]

Proposition 8.18 $\partial^2 = 0$

Proof: [Need gluing theorem] Let c be a critical point of f of index i. We

will compute $\partial\partial c$:

$$\begin{aligned}
\partial\partial c &= \partial \sum_b n_{c,b} b \\
&= \sum_b n_{c,b} \partial b \\
&= \sum_b n_{c,b} \sum_a n_{b,a} a \\
&= \sum_a \left(\sum_b n_{c,b} n_{b,a} \right) a
\end{aligned}$$

where b ranges over critical points of index $i-1$ and a ranges OBover critical points of index $i-2$. The sum

$$\sum_b n_{c,b} n_{b,a}$$

can be taken over all critical points b of index $i-1$ for which there exist flows from c to b and flows from b to a.

By the compactness theorem, $\mathcal{M}(c,a)$ is a compact manifold with corners. Since the relative index is two, there can only be corners of codimension one. Therefore, $\mathcal{M}(c,a)$ is a manifold with boundary. The boundary is

$$\bigcup_b \mathcal{M}(c,b) \times \mathcal{M}(b,a)$$

which is a set of points whose number is

$$\sum_b n_{c,b} n_{b,a}$$

so this sum, taken with sign, is zero. [Explain more with sign] Therefore, $\partial^2(c) = 0$. For more general chains, we simply apply this linearly. □

Definition 8.19 *The complex*

$$0 \xrightarrow{0} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{0} 0$$

is called the Morse–Smale *complex.*

Chapter 9

Compactification of the Moduli space of flows

9.1 The Compactification and a Simplicial Decomposition of a Torus

Lemma 8.15 describes what happens when one approaches the end of a moduli space of flows $\mathcal{M}(a,b)$. Namely, one approaches a sequence of flows (which we will refer to as a *composite flow* $\{\gamma_1,\ldots,\gamma_m\}$ where $\gamma_i \in \mathcal{M}(a_{i-1},a_i)$ where $\{a=a_0,a_1,\ldots,a_{m-1},a_m=b\}$ is a sequence of critical points. This suggests the following compactification theorem. Recall from chapter 6 the partial ordering on critical points defined by setting $a_1 > a_2$ if there exists a flow from a_1 to a_2, that is if $\mathcal{M}(a_1,a_2)$ is nonempty. A sequence $\mathbf{a}=\{a_0,\ldots,a_n\}$ is ordered if $a_i > a_{i+1}$ for all i. Given such a sequence we let $s(\mathbf{a})=a_0$ and $e(\mathbf{a})=a_n$. We define the *length* of this sequence, $l(\mathbf{a})$, to be $n-1$. Finally we define

$$\mathcal{M}(\mathbf{a}) = \mathcal{M}(a_0,a_1)\times\cdots\times\mathcal{M}(a_{n-1},a_n).$$

Theorem 9.1 *A compactification of the moduli space of flows $\mathcal{M}(a,b)$ is given by the following space:*

$$\overline{\mathcal{M}}(a,b) = \mathcal{M}(a,b)\cup\bigcup_{\mathbf{a}}\mathcal{M}(\mathbf{a})$$

where the union is taken over all ordered sequences of critical points $\mathbf{a}$ with $s(\mathbf{a})=a$ and $e(\mathbf{a})=b$.

Theorem 9.2 *Let $f: M\longrightarrow\mathbb{R}$ be a Morse function on a compact manifold as above, and a, a_1, and b be critical points for f with $a>a_1>b$. Then there exists an $\epsilon>0$ and maps*

$$\mu:(0,\epsilon]\times\mathcal{M}(a,a_1)\times\mathcal{M}(a_1,b)\longrightarrow\mathcal{M}(a,b)$$

which we write as

$$(t, \gamma_1, \gamma_2) \longrightarrow \gamma_1 \circ_t \gamma_2$$

that satisfies the following properties:

1. *μ is a diffeomorphism onto its image.*

2. *μ satisfies the following associativity law:*

$$(\gamma_1 \circ_s \gamma_2) \circ_t \gamma_3 = \gamma_1 \circ_s (\gamma_2 \circ_t \gamma_3)$$

   for all s, $t \leq \epsilon$.

3. *This associativity property defines maps*

$$\mu : (0, \epsilon]^l \times \mathcal{M}(\mathbf{a}) \longrightarrow \mathcal{M}(a, b)$$

   where $(\mathbf{a})$ is any ordered sequence of critical points of length l with $s(\mathbf{a}) = a$ and $e(\mathbf{a}) = b$. These maps are also diffeomorphisms onto their images.

4. *Define $\mathcal{K}(a, b) \subset \mathcal{M}(a, b)$ to be*

$$\mathcal{K}(a, b) = \mathcal{M}(a, b) - \bigcup \mu \left((0, \epsilon)^l \times \mathcal{M}(\mathbf{a}) \right)$$

   where the union is taken over all ordered sequences $(\mathbf{a})$ of length ≥ 1 having $s(\mathbf{a}) = a$ and $e(\mathbf{a}) = b$. Then $\mathcal{K}(a, b)$ is compact.

5. *Define the compactification $\overline{\mathcal{M}}(a, b)$ to be the union along μ*

$$\overline{\mathcal{M}}(a, b) = \mathcal{M}(a, b) \cup_{\mathbf{a}} \bigcup_{\mu} [0, \epsilon]^l \times \mathcal{M}(\mathbf{a}).$$

   Then $\overline{\mathcal{M}}(a, b)$ is homeomorphic to $\mathcal{K}(a, b)$.

This theorem says that the ends of the moduli space $\mathcal{M}(a, b)$ consist of spaces of half open cubes parameterized by composable sequences of flow lines. The compact space $\mathcal{K}(a, b)$ is formed by removing the associated *open* cubes. The compactification $\overline{\mathcal{M}}(a, b)$ is formed by formally closing the cubes. It should therefore not be surprising that they are homeomorphic. If $\gamma_1 \in \mathcal{M}(a, a_1)$ and $\gamma_2 \in \mathcal{M}(a_1, b)$, then the parameter $t \in (0, \epsilon]$ in the flow $\gamma_1 \circ_t \gamma_2 \in \mathcal{M}(a, b)$ can be viewed as a measure of how close this flow comes to the critical point a_1. Thus the fact that the pairing μ is a diffeomorphism onto its image allows us to view the space $\mathcal{K}(a, b)$ as the space of flows that stay at least ϵ away from all critical points other than a and b (in this undefined measure). On the other hand, the homeomorphic space $\overline{\mathcal{M}}(a, b)$ can be viewed as formally adjoining composite (or "broken") flows to $\mathcal{M}(a, b)$. From now on we will rescale the metric if necessary so as to be able to assume $\epsilon = 1$.

9.2 Gluing Flow Lines and the Compactification Theorem

In this section we will discuss the gluing of flow lines

$$\mu : (0, \epsilon] \times \mathcal{M}(a, a_1) \times \mathcal{M}(a_1, b) \longrightarrow \mathcal{M}(a, b)$$

and will outline the proof of the compactification Theorem 9.2. Details are to be found in [**?**].

To describe the gluing procedure we first need to pick an $\epsilon > 0$ with the property that if x and y are two points in M with $d(x, y) = \delta < \epsilon$, where $d(x, y)$ means the geodesic distance, then there is a unique geodesic $g_{x,y} : [-\delta/2, \delta/2] \longrightarrow M$, parameterized by arclength, joining x to y. Now let a, a_1, and b be critical points of f and let γ_1 be a flow line joining a to a_1, and γ_2 a flow line from a_1 to b. For $t \in (0, \epsilon/3]$ we define the "patched curve"

$$\gamma_1 *_t \gamma_2 : \mathbb{R} \longrightarrow M$$

as follows. Let x be the last point on the curve γ_1 whose distance from a_1 is t (there may be more than one point on γ_1 whose distance from a_1 is t). Let y be the first point on the curve γ_2 whose distance from a_1 is t. Now let δ be the geodesic distance between x and y and parameterize γ_1 so that it satisfies the initial condition $\gamma_1(-\delta/2) = x$ and parameterize γ_2 so that it satisfies $\gamma_2(\delta/2) = y$. Define $\gamma_1 *_t \gamma_2$ to be the following curve:

$$\gamma_1 *_t \gamma_2(s) = \begin{cases} \gamma_1(s) & \text{if } s \leq -\delta/2 \\ g_{x,y}(s) & \text{if } s \in [-\delta/2, \delta/2] \\ \gamma_2(s) & \text{if } s \geq \delta/2 \end{cases} \tag{9.1}$$

The following is a picture of the patched curve $\gamma_1 *_t \gamma_2$.

We observe the following associativity relation.

Lemma 9.3 *Let a, a_1, a_2, and a_3 be critical points of f and let γ_1, γ_2, and γ_3 be flow lines in $\mathcal{M}(a, a_1)$, $\mathcal{M}(a_1, a_2)$, and $\mathcal{M}(a_2, b)$ respectively. Then for any t_1, $t_2 \in (0, \epsilon/3]$ there exists a constant k such that*

$$((\gamma_1 *_{t_1} \gamma_2) *_{t_2} \gamma_3)(s) = (\gamma_1 *_{t_1} (\gamma_2 *_{t_2} \gamma_3))(s + k).$$

[Note: Since $*_t$ does not give us flows, and the inputs to $*_t$ are assumed to be flows, is doing it twice defined?]

Exercise 9.1 *Prove Lemma 9.3.*

We observe that the patched flows $\gamma_1 *_t \gamma_2$ are not actually flow lines in that they do not satisfy the flow equations for all time. However we will prove that they are "approximate " flows in the sense that they lie very close to the space of flow lines $\mathcal{F}_{a,b}$ inside the space $P_{a,b}$ of all curves from a to b. We will show that a sufficiently small neighborhood of the space of flows $\mathcal{F}_{a,b}$ actually retracts onto $\mathcal{F}_{a,b}$ (in much the same way as a tubular neighborhood of an embedded submanifold retracts onto the submanifold). The glued flow $\gamma_1 \circ_t \gamma_2$ will be the image under this retraction of $\gamma_1 *_t \gamma_2$.

[–removed flow energy]

One way of viewing the idea behind the gluing of flows is to first prove that the patched flow $\gamma_1 *_t \gamma_2$ lies in an arbitrarily small neighborhood of $\mathcal{F}_{a,b} \cong W(a,b)$ in $P_{a,b}$ so long as $t < \epsilon$ and ϵ is sufficiently small. One then uses calculus of variations to show that for any curve $\alpha \in P_{a,b}$ that is sufficiently close to $\mathcal{F}_{a,b}$ (i.e. sufficiently close to a global minimum of the energy functional E) there is a unique flow line ϕ_α of the energy functional E satisfying

1. $\phi_\alpha(0) = \alpha$
2. $\lim_{s\to\infty} \phi_\alpha(s) = \alpha_\infty$ exists and is an absolute minimum of E. That is, $\alpha_\infty \in \mathcal{F}_{a,b}$.

The glued flow line would be defined by

$$\gamma_1 \circ_t \gamma_2 = (\gamma_1 *_t \gamma_2)_\infty.$$

Rather than go through the calculus of variations arguments to show that this gluing procedure is well defined and satisfies Theorem 9.2, we will instead use a different approach involving linearizing the flow equations. This technique has broad generalizations outside the context of Morse functions on compact manifolds. Indeed it is motivated by work of Taubes [**?**] in the gluing of "instantons" in gauge theory. Gluing of instantons will be discussed later in these notes.

To set up this approach, let $\gamma \in P_{a,b}$ and consider the tangent space $T_\gamma P_{a,b}$. It is easy to see that

$$T_\gamma P_{a,b} = \{\alpha : \mathbb{R} \longrightarrow TM \text{ such that } p \circ \alpha = \gamma, \text{ and } \lim_{t\to\pm\infty} \alpha(t) = 0.\}$$

Here $p : TM \longrightarrow M$ is the projection of the tangent bundle of the underlying compact manifold M. We refer the reader to Milnor's book [**?**] for a detailed description of the tangent bundle $TP_{a,b}$.

Now consider the vector field on the path space

$$\mathcal{S} : P_{a,b} \longrightarrow T(P_{a,b})$$

defined by

$$\mathcal{S}(\gamma) = \frac{d\gamma}{dt} + \nabla_\gamma f.$$

Notice that the space of flows $\mathcal{F}_{a,b}$ is precisely the set of zeros of the vector field $\mathcal{S}$. Using the underlying metric on M, one has a connection on $TP_{a,b}$ (the associated Levi–Civita connection) that allows us to differentiate the section $\mathcal{S}$ to obtain a family of operators

$$D_\gamma \mathcal{S} : T_\gamma P_{a,b} \longrightarrow T_\gamma P_{a,b} \quad \gamma \in P_{a,b}.$$

The precise definition of the operator $D\mathcal{S}$ is as the image of the section $\mathcal{S} \in C^\infty(TP_{a,b})$ under the covariant derivative D induced by the connection,

$$C^\infty(TP_{a,b}) = \Omega^0(P_{a,b}; TP_{a,b}) \xrightarrow{D} \Omega^1(P_{a,b}; TP_{a,b}).$$

Here we may interpret the one forms

$$\begin{aligned}\Omega^1(P_{a,b}; TP_{a,b}) &= C^\infty(T^*(P_{a,b}) \otimes TP_{a,b})) \\ &= C^\infty(Hom(TP_{a,b}, TP_{a,b}))\end{aligned}$$

where an element $\phi \in C^\infty(Hom(TP_{a,b}, TP_{a,b}))$ assigns to a curve $\gamma \in P_{a,b}$ a linear operator $\phi(\gamma) : T_\gamma P_{a,b} \longrightarrow T_\gamma P_{a,b}$.

The following is little more than a check of the definitions.

Lemma 9.4 *The vector field $\mathcal{S}$ on $P_{a,b}$ is transverse to the zero section if and only if for every flow line $\gamma \in \mathcal{F}_{a,b}$ the operator*

$$D_\gamma \mathcal{S} : T_\gamma P_{a,b} \longrightarrow T_\gamma P_{a,b}$$

is surjective.

The operator $D\mathcal{S}$ can be described explicitly as follows. Let

$$Hess(f) : TM \longrightarrow TM$$

be the covariant derivative (with respect to the Levi–Civita connection) of the gradient vector field

$$\nabla(f) : M \longrightarrow TM.$$

Given a critical point $a \in M$ and a basis for T_aM, $Hess_a(f) : T_aM \longrightarrow T_aM$ is represented by the matrix of 2^{nd} order partial derivatives of f. From this point of view of the Hessian it is easy to see that the operator $D\mathcal{S}$ has the form

$$\begin{aligned}D_\gamma \mathcal{S} : T_\gamma P_{a,b} &\longrightarrow T_\gamma P_{a,b} \\ \alpha &\longrightarrow \frac{d\alpha}{dt} + Hess_{\gamma(t)}(f) \circ \alpha.\end{aligned}$$

Now according to Theorem 8.1, the Morse–Smale condition is precisely that this operator is surjective. So if f is Morse–Smale, we are assured that the operator $D\mathcal{S}$ is surjective for every pair of critical points a and b. Now since $\mathcal{F}_{a,b}$ is the space of zeros of the vector field $\mathcal{S}$, the surjectivity of the covariant derivative implies that for any $\gamma \in \mathcal{F}_{a,b}$,

$$T_\gamma \mathcal{F}_{a,b} = Ker(D_\gamma \mathcal{S}) \subset T_\gamma P_{a,b}.$$

One parameter families of operators of the form $\frac{d}{dt} + A_t$, where A_t is a self adjoint linear operator have been studied in great detail by Atiyah, Patodi, and Singer in [**?**]. When these operators are Fredholm (i.e. have finite dimensional kernel and cokernel) then the index (the difference in the dimension of the kernel and cokernel) is given by the *spectral flow* of A_t. That is the number of eigenvalues of A_t that "flow" from being negative to positive over time t. That is,

$$(\frac{d}{dt} + A_t) = \lim_{t \to -\infty} A_t) - \lim_{t \to +\infty} (A_t)$$

where (A_t) is the dimension of the negative eigenspace of A_t. In our case this result confirms what we already knew for simpler reasons; that

$$\begin{aligned} \dim(W_{a,b}) &= \dim(\mathcal{F}_{a,b}) = \dim(T_\gamma \mathcal{F}_{a,b}) \\ &= \frac{d}{dt} + Hess_{\gamma(t)}) = ind(a) - ind(b). \end{aligned}$$

Even though in our case we knew the dimension of $W(a,b)$ for easier reasons (transversality), being able to relate the dimension of the zeros of a vector field with the index of its covariant derivative is an important general tool.

The argument then continues and shows, using basically only the implicit function theorem, that there exists a small neighborhood, $P^\eta_{a,b}$ of the space of flows $\mathcal{F}_{a,b}$ in the full path space $P_{a,b}$ and a retraction map

$$\rho : P^\eta_{a,b} \longrightarrow \mathcal{F}_{a,b}$$

so that on the level of derivatives,

$$D\rho_\gamma : T_\gamma P_{a,b} \longrightarrow T_\gamma \mathcal{F}_{a,b} = Ker(D_\gamma \mathcal{S})$$

is the orthogonal projection onto the kernel of the operator $D_\gamma \mathcal{S}$.

One then shows that for ϵ sufficiently small, the image of the patching map

$$\begin{aligned} (0, \epsilon] \times \mathcal{F}_{a,c} \times \mathcal{F}_{c,b} &\longrightarrow P_{a,b} \\ (t, \gamma_1, \gamma_2) &\longrightarrow \gamma_1 *_t \gamma_2 \end{aligned}$$

takes values in $P^\eta_{a,b}$. If one composes with the projection map $\rho : P^\eta_{a,b} \longrightarrow \mathcal{F}_{a,b}$ and divides out by the translation action of the real numbers $\mathbb{R}$, one gets an induced map

$$\begin{aligned} \mu : (0, \epsilon] \times \mathcal{M}(a,c) \times \mathcal{M}(c,b) &\longrightarrow \mathcal{M}(a,b) \\ (t, \gamma_1, \gamma_2) &\longrightarrow \gamma_1 \circ_t \gamma_2 \end{aligned}$$

This is the map used to prove Theorem 9.2 in [**?**]. The associativity law that μ is required to satisfy (part (2) of Theorem 9.2) follows from Lemma 9.3 and the naturality of the projection map ρ. The compactness properties follow easily from Lemma 8.15 and the fact that μ is a local diffeomorphism. This is technically the most difficult property to verify. It is done by first observing that since $(0,\epsilon] \times \mathcal{M}(a,c) \times \mathcal{M}(c,b)$ and $\mathcal{M}(a,b)$ have the same dimension it is sufficient to prove that μ is locally one to one. By the inverse function theorem, to do this it is sufficient to check that its differential is one to one at every point (t,γ_1,γ_2). But since the patching map

$$\begin{aligned}(0,\epsilon] \times \mathcal{F}_{a,c} \times \mathcal{F}_{c,b} &\longrightarrow P^{\eta}_{a,b} \\ (t,\gamma_1,\gamma_2) &\longrightarrow \gamma_1 *_t \gamma_2\end{aligned}$$

is clearly one to one, and since the derivative of the projection map $\rho : P^{\eta}_{a,b} \longrightarrow \mathcal{F}_{a,b}$ is the orthogonal projection onto $Ker(D\mathcal{S})$, it is sufficient to show that for ϵ chosen sufficiently small, the image of the derivative of the patching map is never orthogonal to $Ker(D\mathcal{S})$. This is done by a relatively straightforward calculation of the Hessian of f along a patched curve. Details of this argument will appear in [**?**].

Chapter 10

An explicit CW structure on a manifold

In Chapter 5 we saw that a Morse function f on a compact manifold M can be used to show that M is homotopy equivalent to a CW complex. This is obtained by examining $M^t = f^{-1}((-\infty, t])$ for various values of t. For t sufficiently small, M^t is empty, and as we pass critical points, we attach cells (up to homotopy equivalence), and as we move t between critical points, we get a diffeomorphism.

In Chapter 6 we saw how the unstable manifold of each critical point is an open disk, and if f is Morse–Smale, this partitions M into what appears to be a CW complex. We even saw in Chapter 11 how to extract information about the attaching maps.

It is tempting, therefore, to wonder if this partitioning of M into cells is really a CW complex decomposition. In this chapter we will see that it is, using a technique due to Hutchings[**?**].

10.1 The Hutchings closed cell

The main issue to resolve is that in a CW complex, the cells are closed disks, while the unstable manifold $W^u(a)$ of a critical point $a \in M$ is an open disk. To fix this situation, consider the following compactification of $W^u(a)$.

Definition 10.1 *If $a \in M$ is a critical point of $f : M \longrightarrow \mathbb{R}$, let $\overline{W}^u(a)$ be the following union:*

$$\bigcup_{c_1,\ldots,c_r} \mathcal{M}(a, c_1) \times \cdots \times \mathcal{M}(c_{r-2}, c_{r-1}) \times W(c_{r-1}, c_r)$$

with the identifications given by the compactness theorem.

[I know that should have been more explicit. Later.]

[Show it is compact.
Show it is a disk.
Talk about attaching maps again.
Prove it is a CW complex]

Chapter 11

The Attaching maps

We have shown how a manifold M is homotopy equivalent to a CW complex (Theorem 5.7), using a Morse function f. We have also seen how, using a metric, we can obtain the cells of this CW complex as the unstable manifolds $W^u(a)$ for each critical point. But a CW complex is determined not only by its cells, but by its attaching maps. In this section we explain how to find these attaching maps.

Let a be a critical point of M, and let t be a real number $t < f(a)$ with $f(a) - t$ sufficiently small.

Proposition 11.1 *There is a diffeomorphism of M that sends M^t onto itself and so that the attached disk D^λ in Theorem 5.6 gets sent to $W^u(a) \cap f^{-1}[t, f(a)]$. In other words, up to a diffeomorphism of M,*

$$D^\lambda = W^u(a) \cap f^{-1}[t_0, t_1]. \tag{11.1}$$

Proof: [Proof goes here] □

This has the following implication for the CW structure of M. Let $M(f)$ be the CW complex associated to the Morse function $f : M \longrightarrow \mathbb{R}$ as in Theorem 5.7. Let $M(f)^{(q)}$ denote the q-skeleton. It consists of one cell of dimension p for every critical point of f having index $p \leq q$.

[don't use (q) skeleton. Instead, r-skeleton.]

Recall that the data associated with a CW complex is as follows: a collection of closed λ-dimensional cells $\overline{D}_i^\lambda$, and for each λ-dimensional cell, an attaching map $\phi_i^\lambda : S_i^{\lambda-1} \longrightarrow \cup_{k<\lambda} \overline{D}_i^\lambda$ from its boundary to the $\lambda - 1$-skeleton. In this way, we view a CW complex as growing inductively, by first taking a bunch of zero-cells, then attaching the one-cells, and so on.

The way we have used a Morse function to show the manifold has the homotopy type of a CW complex (Theorem 5.7) is similar: we examine $M^t =$

$f^{-1}((-\infty, t])$ for various increasing values of t. we start with a zero-cell when t is the minimum value of f. As we increase t, if we do not pass through a critical value, the homotopy type of M^t is unchanged; but if we increase it through a critical value, we attach a cell whose dimension is the index of the critical point. Thus, inductively, we attach cells to whatever we have already built.

From the perspective of using unstable manifolds for cells, we might hope to replace the notion of increasing t with the notion of adding unstable manifolds directly, in the order of increasing values for f. Now, it may be disconcerting that the order that we are attaching cells is by value for f instead of dimension (as would be the case with CW complexes). But this is irrelevant, since if $\mathrm{ind}(a) \geq \mathrm{ind}(b)$, and (f, g) is Morse–Smale, then there are no flows from b to a, and in particular, the unstable manifold from b does not approach cells of higher dimension.

[There might be a need to cite compactness theorems from chap 3.8]

11.1 Consecutive critical points: Franks' theorem

Suppose a and b are consecutive critical points, in the sense that there are flows from a to b but no other intermediate flows. That is, there does not exist a critical point c so that there are flows from a to c and flows from c to b.

Let the index of a be p and let $r < p$ be the index of b. Let $D^p_a \subset M(f)^{(p)}$ be the associated attaching disk to the critical point a. It is attached by a map

$$\phi_a : S^{p-1}_a \longrightarrow M(f)^{(p-1)}.$$

The following is an immediate corollary of (11.1):

Corollary 11.2 *The attaching map $\phi_a : S^{p-1}_a \longrightarrow M(f)^{(p-1)}$ factors through the r-skeleton:*

$$\phi_a : S^{p-1}_a \longrightarrow M(f)^{(r)}.$$

Notice that the quotient of the skeleta is a wedge of r-dimensional spheres indexed by the critical points of index r:

$$M(f)^{(r)}/M(f)^{(r-1)} \simeq \bigvee_{\beta \in \mathrm{Crit}_r} S^r_\beta$$

where Crit_r denotes the set of critical points of index r. Given such a critical point $b \in \mathrm{Crit}_r$ we consider the relative attaching map

$$\phi_{a,b} : S^{p-1} \xrightarrow{\phi_a} M(f)^{(r)} \xrightarrow{p} M(f)^{(r)}/M(f)^{(r-1)} \simeq \bigvee_{\beta \in \mathrm{Crit}_r} S^r_\beta \xrightarrow{\pi_b} S^r_b \quad (11.2)$$

where p and π_b are the obvious projection maps. We will consider the maps $\phi_{a,b}$ as elements in the homotopy group $\pi_{p-1}(S^r)$. As in chapter 2 we can stabilize

to get an element to get an element in the stable homotopy groups of spheres which, by abuse of notation we call by the same name

$$\phi_{a,b} \in \pi^s_{p-r-1}.$$

By Theorem 2.1 (The Thom–Pontrjagin theorem) we can identify this group with the framed cobordism group η_{p-r-1}. In [**?**] J. Franks showed how $\phi_{a,b}$ is represented by $\mathcal{M}(a,b)$, the moduli space of flow lines between the two critical points a and b, as a framed moduli space.[1]

Notice that Lemma 8.15 says that $\mathcal{M}(a,b)$ is a compact manifold if and only if b is a successor of a. Notice that in the context of studying the relative attaching maps (as in equation (11.2)) we are assuming that this is indeed the case. So $\mathcal{M}(a,b)$ is a closed, $p-r-1$ dimensional manifold. $\mathcal{M}(a,b)$ can be given a framing as follows. Let t be any regular value between $f(a)$ and $f(b)$. By (11.1) the intersection

$$W^u(a) \cap f^{-1}(t)$$

is a sphere of dimension $p-1$ which we denote by $S^u(a)$. Similarly, we define

$$S^s(b) = W^s(b) \cap f^{-1}(t)$$

which is a sphere of dimension $n-p-1$.

By construction we have

$$\begin{aligned}\mathcal{M}(a,b) \cong W(a,b)^t = W(a,b) \cap f^{-1}(t) \\ = W^s(b) \cap W^u(a) \cap f^{-1}(t) = W^s(b) \cap S^u(a).\end{aligned}$$

Thus we have a natural embedding $\mathcal{M}(a,b) \hookrightarrow S^u(a)$ which has codimension r. The normal bundle of this embedding is the pull-back of the normal bundle of the embedding

$$W^s(b) \hookrightarrow M$$

which comes equipped with a unique framing because $W^s(b)$, being diffeomorphic to the disk D^{n-r} is contractible. This induces a framing α on the normal bundle of $\mathcal{M}(a,b) \hookrightarrow S^u(a)$. Hence the pair $(\mathcal{M}(a,b),\alpha)$ determine an element in the framed cobordism group

$$(\mathcal{M}(a,b),\alpha) \in \eta_{p-r-1}$$

Theorem 11.3 (Frank's theorem [?]) *Under the Thom–Pontrjagin construction, the relative attaching map*

$$\phi_{a,b} \in \pi^s_{p-r-1}$$

[1]Note that this is not a complete description of the attaching map: only the "component" that corresponds to the piece that wraps around the unstable manifold for the one critical point b. Even if we took all these data for each critical point, we still do not have the attaching map becaues $\pi_i(X \vee Y)$ is not necessarily isomorphic to $\pi_i(X) \times \pi_i(Y)$. Furthermore, we are ignoring the data that describes how the attaching map maps to the s-skeleton where $s < r$. But we can have some fun at the r stage, and so we will.

corresponds to the framed moduli space of flows

$$(\mathcal{M}(a,b),\alpha) \in \eta_{p-r-1}.$$

Proof: By (11.1) the source sphere in the attaching map

$$\phi_{a,b} : S^{p-1} \longrightarrow S^r,$$

being the boundary of the disk represented by the critical point a, is $S^u(a)$. Furthermore, the composition given in (11.2) defining $\phi_{a,b}$ is homotopic to the composition (which by abuse of notation we also call $\phi_{a,b}$,

$$\phi_{a,b} : S^u(a) \xrightarrow{\phi_a} M(f)^{(r)} \xrightarrow{proj} M(f)^{(r)}/(M(f)^{(r)} - W^s(b)) = S^r.$$

The transversality condition (i.e. the Morse–Smale condition) implies that the base point $* \in S^r$ is a regular value of $\phi_{a,b}$. Furthermore the inverse image

$$\phi_{a,b}^{-1}(*) = S^u(a) \cap S^s(b) = \mathcal{M}(a,b).$$

The induced framing on $\mathcal{M}(a,b)$ from the framing of $* \in S^r$ is clearly the framing α described above. Hence $(\mathcal{M}(a,b),\alpha)$ is the image of $\phi_{a,b}$ under the Thom–Pontrjagin construction. □

11.2 Relative index one

Now suppose a and b are critical points of relative index one. Say $(a) = p$ and $(b) = p-1$. Then the space of flows, $(\mathcal{M}(a,b),\alpha)$ is a zero dimensional, framed, compact manifold; that is a finite set of points (flow lines) with signs attached to them induced by the framing. This is what we normally call an "orientation". And indeed, an orientation is precisely what we need on a CW complex to define homology.

Let $n(a,b) \in \mathbb{Z}$ denote the signed number of flow lines:

$$n(a,b) = \sum_{\gamma \in \mathcal{M}(a,b)} \alpha(\gamma)$$

where $\alpha(\gamma) = \pm 1$ is the sign associated to the flow line γ by the framing α.

$$n(a,b) \in \mathbb{Z} = \eta_0 = \pi_0^s$$

is therefore the integer given by the degree of the relative attaching map

$$\phi_{a,b} : S^{p-1} \longrightarrow S^{p-1}.$$

This allows us to compute the boundary homomorphism in the Morse–Smale complex (5.1). Recall that in this complex, C_p is the free abelian group generated by Crit_p. Let $[a] \in C_p$ denote the generator corresponding to the critical point a.

Corollary 11.4 *The coefficient of* $[b] \in C_{p-1}$ *of the boundary* $\partial_p[a]$, $\langle \partial_p[a], [b] \rangle$, *is given by the formula*

$$\langle \partial_p[a], [b] \rangle = n(a, b) \in \mathbb{Z}.$$

We note that this formula for the Morse–Smale complex is implicit in the work of Smale (see [**?**]). As described above it should be viewed as a special case of the attaching map formulae of Franks [**?**]. However the formulae in corollary 11.4 seems to have first been explicitly given in the literature by Witten [**?**].

Part III

Category of gradient flows

Chapter 12

The Flow Category of a Morse Function

Let $f : M \longrightarrow \mathbb{R}$ be a Morse function on a compact manifold. In this chapter we will construct a topological category $\mathcal{C}_f$ whose objects are the critical points of f and where the space of morphisms between two critical points a and b is a compactification, $\overline{\mathcal{M}}(a,b)$ of the moduli space of flows $\mathcal{M}(a,b)$. We will then prove that the classifying space $B\mathcal{C}_f$ is homeomorphic to the manifold M. This will give M the structure of an explicit simplicial space (i.e. the nerve of $\mathcal{C}_f$) and will complete our goal of recovering the topology of the manifold directly and explicitly in terms of the space of flows of the gradient vector field.

[Removed: Compactification argument, now in Chap 3.8]

We will state this compactification theorem more precisely later, and in particular describe the topology of $\overline{\mathcal{M}}(a,b)$. Then the theorem will be proved in the next section. In this section we will describe how it will be used to define the category induced by a Morse function and state how its classifying space is related to the underlying manifold. We will then consider the examples of two dimensional tori. We will begin by stating the main theorem to be discussed in this chapter. The details of the proof of this theorem will appear in [**?**].

Theorem 12.1 *Let $f : M \longrightarrow \mathbb{R}$ be a Morse function on a compact manifold satisfying the Morse–Smale transversality conditions. Consider the topological category $\mathcal{C}_f$ whose objects are the critical points of f, and whose spaces of morphisms are the compactified moduli spaces,*

$$Mor(a,b) = \overline{\mathcal{M}}(a,b).$$

Composition in this category is given by the inclusion

$$\overline{\mathcal{M}}(a,b) \times \overline{\mathcal{M}}(b,c) \hookrightarrow \overline{\mathcal{M}}(a,c)$$

described in the compactification Theorem 9.1. Then there is a natural homeomorphism of the classifying space of this category with the underlying manifold

$$B\mathcal{C}_f \xrightarrow{\cong} M.$$

Notice that this theorem defines an explicit simplicial space description (as the nerve of the category $\mathcal{N}(\mathcal{C}_f)$) of the manifold M in terms of the moduli spaces of flow lines of the gradient vector field of the Morse function f. Indeed the k-simplices of this decomposition are parameterized by the space of k-tuples of "composable" flow lines. We will discuss implications of this theorem in following chapters. The main goal of this section is to illustrate this simplicial decomposition for a particular example of a Morse–Smale flow on a torus. In order to do this we need a more precise version of the compactification theorem. This theorem will give a detailed description of the ends of moduli spaces.

12.1 Torus example

The homeomorphism between $\mathcal{K}(a,b)$ and $\overline{\mathcal{M}}(a,b)$ allows us to define the category $\mathcal{C}_f$ in two equivalent (isomorphic) ways. The first way, as described in Theorem 12.1, is to let the space of morphisms between critical points a and b be $\overline{\mathcal{M}}(a,b)$, where the composition law is given by

$$\begin{aligned}\overline{\mathcal{M}}(a,b) \times \overline{\mathcal{M}}(b,c) &\longrightarrow \overline{\mathcal{M}}(a,c)\\ (\gamma_1, \gamma_2) &\longrightarrow \gamma_1 \circ_0 \gamma_2.\end{aligned}$$

The other equivalent definition is to let the space of morphisms between critical points a and b be $\mathcal{K}(a,b)$, where the composition law is given by

$$\begin{aligned}\mathcal{K}(a,b) \times \mathcal{K}(b,c) &\longrightarrow \mathcal{K}(a,c)\\ (\gamma_1, \gamma_2) &\longrightarrow \gamma_1 \circ_\epsilon \gamma_2.\end{aligned}$$

For the purposes of the following diagrams we will use the second way of defining $\mathcal{C}_f$.

Consider the following figure depicting a Morse–Smale flow on the torus.

Here we view the two-torus as embedded in ordinary three-space, standing

on one of its ends with the hole facing the reader, but tilted slightly toward the reader. We let f be the height function.

There are four critical points; a has index 2, b and c have index 1, and d has index 0. As the figure depicts, the moduli spaces $\mathcal{M}(a,b)$, $\mathcal{M}(a,c)$, $\mathcal{M}(b,d)$, and $\mathcal{M}(c,d)$ are all spaces consisting of two distinct points each. We will denote these flows by α_i, β_i, γ_i, and δ_i respectively. All points on the torus not lying on any of these flows is on a flow in $\mathcal{M}(a,d)$. This moduli space is one dimensional, and indeed is the disjoint union of four open intervals. Thus $\mathcal{K}(a,d)$ is the disjoint union of four closed intervals. If the torus is viewed in the usual way as a square with opposite sides identified, then these flows can be depicted as follows.

Now consider the simplicial description in the classifying space $B\mathcal{C}_f$. The vertices correspond to the objects of the category $\mathcal{C}_f$, that is the critical points. Thus there are four vertices. There is one one simplex (interval) for each morphism (flow line), glued to the vertices corresponding to the starting and endpoints of the flows. Notice that the points in $\mathcal{K}(a,d)$ index a one parameter family of one simplices attached to the vertices labelled by a and d. Finally observe that there is a two-simplex for every pair of composable flows. There are eight such pairs (coming from the four points in each of the product moduli spaces $\mathcal{M}(a,b) \times \mathcal{M}(b,d)$ and $\mathcal{M}(a,c) \times \mathcal{M}(c,d)$.) A two-simplex labelled by a pair of flows, say (α, β) will have its three faces identified with the one simplices labelled by α, β, and $\alpha \circ_1 \beta$ respectively. Notice that all higher dimensional simplices in the nerve $\mathcal{N}(\mathcal{C}_f)$ are degenerate and so do not contribute to the geometric realization. The following figure depicts the resulting simplicial structure of the classifying space and illustrates Theorem 12.1 that this space

is homeomorphic to the underlying manifold.

12.2 Proof of Theorem 12.1

In this section we give a proof of the main theorem of this chapter, Theorem 12.1. That is, we identify the manifold M with the classifying space of the category $\mathcal{C}_f$. Recall that this theorem proposes to describe M as a simplicial space, where the simplices are parameterized by cartesian products of the moduli spaces, $\mathcal{M}(a,b)$, or equivalently, $\mathcal{K}(a,b)$. We begin by describing a filtration of the spaces $\mathcal{M}(a,b)$.

By scaling the metric of M if necessary, we can assume that the constant ϵ in the statement of Theorem 9.2 is 1 and so we will now drop this from the notation. We can think of the set $\mathcal{K}(a,b)$ of Theorem 9.2 as the space of flow lines from a to b which keep distance of at least 1 from any critical points c with $a > c > b$. More generally we can filter the space $\overline{\mathcal{M}}(a,b)$ by saying that a curve in $\overline{\mathcal{M}}(a,b)$ has filtration k if it gets within distance less than 1 of at most k intermediate critical points. Precisely, we define

$$\mathcal{K}^{(k)}(a,b) = \bigcup_{l \le k} \bigcup_{a > a_1 > \cdots > a_l > b} \mu\left([0,1]^l \times \mathcal{K}(a,a_1) \times \ldots \mathcal{K}(a_l,b)\right).$$

so that $\mathcal{K}^{(0)}(a,b) = \mathcal{K}(a,b)$ and

$$\mathcal{K}^{(k-1)}(a,b) \subseteq \mathcal{K}^{(k)}(a,b).$$

Thus γ is in $\mathcal{K}^{(k)}(a,b)$ if and only if γ can be decomposed as

$$\gamma = \gamma_0 \circ_{s_1} \cdots \circ_{s_l} \gamma_l$$

where $\gamma_i \in \mathcal{K}(a_{i-1}, a_i)$, $0 \le s_i \le 1$ for all i, and $l \le k$.

Lemma 12.2

$$\mathcal{K}^{(k)}(a,b) \setminus \mathcal{K}^{(k-1)}(a,b) \cong \bigsqcup_{a > a_1 > \cdots > a_k > b} [0,1)^k \times \mathcal{K}(a,a_1) \times \ldots \mathcal{K}(a_k,b)$$

Proof: This is obvious, I hope. □

Lemma 12.3

$$\bigcup \mathcal{K}^{(k)}(a,b) = \mathcal{M}(a,b)$$

Proof: This too is obvious, I hope. □

Let $\mathbf{a} = (a_0, \ldots, a_{l+1})$ denote an arbitrary decreasing sequence of critical points with length $l(\mathbf{a}) = l$, starting point $s(\mathbf{a}) = a_0 = a$ and ending point $e(\mathbf{a}) = a_{l+1} = b$. We define

$$\mathcal{K}(\mathbf{a}) = \mathcal{K}(a_0, a_1) \times \cdots \times \mathcal{K}(a_l, a_{l+1}).$$

Using these lemmas we see that the map

$$\bigsqcup_l \bigsqcup_{l(\mathbf{a})=l} [0,1]^l \times \mathcal{K}(\mathbf{a}) \longrightarrow \overline{\mathcal{M}}(a,b)$$

defined by

$$(s_1, \ldots, s_l; \gamma_0, \ldots, \gamma_l) \longrightarrow \gamma_0 \circ_{s_1} \cdots \circ_{s_l} \gamma_l$$

is onto and therefore $\overline{\mathcal{M}}(a,b)$ can be recovered by imposing an equivalence relation on the above disjoint union. It is straightforward to extract this equivalence relation; it is generated by

$$(s_1, \ldots, s_{i-1}, 1, s_{i+1}, \ldots, s_l; \gamma_1, \ldots, \gamma_l) \simeq$$
$$(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_l; \gamma_1, \ldots, \gamma_{i-1} \circ_1 \gamma_i, \ldots, \gamma_l).$$

Thus the relations only involve the faces of the cubes which do not contain the point $(0, \ldots, 0)$.

From this argument we draw the following conclusion.

Theorem 12.4

$$\overline{\mathcal{M}}(a,b) = \bigsqcup_l \bigsqcup_{l(\mathbf{a})=l} [0,1]^l \times \mathcal{K}(\mathbf{a}) / \simeq$$

The next step is to go from this description of the spaces $\overline{\mathcal{M}}(a,b)$ to one of the manifold M. Let $\gamma \in \mathcal{M}(a,b)$ be a flow line. Then we may compactify γ by adding the points a and b to form the curve $\overline{\gamma}$. This curve is closed in the sense that it contains all of its limit points. The function f is decreasing along flow lines and so it gives a natural diffeomorphism

$$f : \overline{\gamma} \longrightarrow [f(b), f(a)].$$

Now suppose that $\gamma = \gamma_0 \circ_0 \cdots \circ_0 \gamma_l$ is a point of $\overline{\mathcal{M}}(a,b)$ which is not in $\mathcal{M}(a,b)$. Thus $\gamma_0, \ldots, \gamma_l$ is a sequence of flow lines joining critical points $a > a_1 > \cdots > a_l > b$. In this case define

$$\overline{\gamma} = \overline{\gamma}_0 \cup \cdots \cup \overline{\gamma}_l,$$

so $\overline{\gamma}$ is a curve in M joining a to b; in an obvious sense $\overline{\gamma}$ is a *piecewise flow line* joining a to b. The function f defines a diffeomorphism

$$f : \overline{\gamma}_i \longrightarrow [f(a_i), f(a_{i-1})]$$

and these diffeomorphisms piece together to define a homeomorphism

$$f : \overline{\gamma} \longrightarrow [f(b), f(a)].$$

This shows that each element in $\overline{\mathcal{M}}(a,b)$ can be identified with a well defined curve $\gamma : [f(b), f(a)] \longrightarrow M$ parameterized by the inverse of the above homeomorphism. However, with this parameterization, none of these curves satisfy the flow equations. In any case we get a map

$$\phi : [f(b), f(a)] \times \overline{\mathcal{M}}(a,b) \longrightarrow M$$

whose image is the closure of the space $W(a,b) \subset M$ since we have added to $W(a,b)$ all points on the curves $\overline{\gamma} : [f(b), f(a)] \longrightarrow M$ where $\gamma \in \overline{\mathcal{M}}(a,b)$. Therefore the map

$$\bigsqcup_{\mathbf{a}} [f(a_{l+1}), f(a_0)] \times [0,1]^l \times \mathcal{K}(\mathbf{a}) \longrightarrow M$$

defined by

$$(t; s_1, \ldots, s_l; \gamma_0, \ldots, \gamma_l) \longrightarrow (\gamma_0 \circ_{s_1} \cdots \circ_{s_l} \gamma_l)(t)$$

is onto. The disjoint union is taken over all decreasing sequences of critical points $\mathbf{a} = (a_0, \ldots, a_{l+1})$. Once more it is not difficult to extract the appropriate equivalence relation on the disjoint union. Given a sequence $\mathbf{a}$ of critical points as above with $l(\mathbf{a}) = l$, we define

$$J_{\mathbf{a}} = [f(a_{l+1}), f(a_0)], \quad I^{l(\mathbf{a})} = [0,1]^l.$$

Now we define

Definition 12.5

$$\mathcal{R}_f = \bigsqcup_{\mathbf{a}} J_{\mathbf{a}} \times I^{l(\mathbf{a})} \times \mathcal{K}(\mathbf{a}) / \sim$$

where the relations $\sim$ are given by

$$(t; s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots, s_l; \gamma_0, \ldots, \gamma_l) \sim \begin{cases} (t; s_1, \ldots, s_{i-1}; \gamma_0, \ldots, \gamma_{i-1}), & \text{if } t \in [f(a_i), f(a_0)] \\ (t; s_{i+1}, \ldots, s_l; \gamma_{i+1}, \ldots, \gamma_l), & \text{if } t \in [f(a_{l+1}), f(a_i)] \end{cases} \tag{12.1}$$

and

$$(t; s_1, \ldots, s_{i-1}, 1, s_{i+1}, \ldots, s_l; \gamma_0, \ldots, \gamma_l) \sim (t; s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_l; \gamma_0, \ldots, \gamma_{i-1} \circ_1 \gamma_i, \ldots, \gamma_l) \tag{12.2}$$

The map ϕ respects the equivalence relation $\sim$ so gives a well defined map

$$\mathcal{R}_f \longrightarrow M.$$

Theorem 12.6 *The map*

$$\phi : \mathcal{R}_f \longrightarrow M$$

is a homeomorphism.

Proof: The first step is to check that the second set of relations are the only relations which can occur if all the s_i's are non-zero. This follows from the fact that if two flow lines have a point in common then they are equal together with the previous theorem. If one of the s_i's is zero then we are dealing with a piecewise flow line. If two piecewise flow lines have a point in common then this point must be one of the joining points or else they have a common segment. An elementary analysis leads to the conclusion that the only identifications which can take place if one of the s_i's is zero are consequences of the first set of relations. □

Since the spaces $\mathcal{K}(a_{i-1}, a_i)$ are diffeomorphic to the compactified spaces $\overline{\mathcal{M}}(a_{i-1}, a_i)$ in such a way that the composition in the category corresponds to $\circ_1$, this shows us how to recover the manifold M from the category $\mathcal{C}_f$. To prove Theorem 12.1 we are therefore reduced to the combinatorial exercise necessary to identify the space $\mathcal{R}_f$ with the classifying space $B\mathcal{C}_f$. Recall from chapter 13 the definition of the classifying space of the category $\mathcal{C}_f$. Comparing it with Definition 12.5 of the space $\mathcal{R}_f$ we see that these spaces are very similar but they are not obviously the same. The essential difference is that $\mathcal{R}_f$ is built up out of cubes whereas the classifying space is built out of simplices. Nonetheless these two spaces are homeomorphic.

Theorem 12.7 *There is a homeomorphism*

$$\mathcal{R}_f \cong B\mathcal{C}_f.$$

Proof: The main point of this argument is that the equivalence relations used to build $\mathcal{R}_f$ can be imposed in two steps; the first step turns the cubes into simplices and the second step then imposes the gluing relations among the simplices that make up $B\mathcal{C}_f$. To see this precisely we first look at the image of a single cube

$$J_{\mathbf{a}} \times I^l = J_{\mathbf{a}} \times I^l \times (\gamma_0, \ldots, \gamma_l)$$

in the quotient space $\mathcal{R}_f$. Define

$$\mathbf{a}_i = (a_0, \ldots a_{i-1}, a_{i+1}, \ldots a_{l+1}) \quad \text{where } 0 \leq i \leq l+1$$

and maps

$$\delta_i : J_{\mathbf{a}_i} \times I^{l-1} \longrightarrow J_{\mathbf{a}} \times I^l$$

by the formulas

$$\delta_i(t; s_1, \ldots, s_{l-1}) = \begin{cases} (t; 0, s_1, \ldots, s_{l-1}), & \text{if } i = 0 \\ (t; s_1, \ldots, s_{i-1}, 1, s_i, \ldots, s_{l-1}), & \text{for } 1 \leq i \leq l \\ (t; s_1, \ldots, s_{l-1}, 0), & \text{if } i = l+1. \end{cases}$$

Now consider the spaces

$$J_{\mathbf{a}} \times I^{l-1}/\sim$$

where we make the following list of identifications: If $1 \leq i \leq l$ so that $J_{\mathbf{a}} = J_{\mathbf{a}_i}$ then

$$(t; s_1, \ldots s_{i-1}, 0, s_{i+1}, \ldots, s_l) \sim \begin{cases} (t; s_1, \ldots, s_{i-1}, 0, s'_{i+1}, \ldots s'_l) & \text{if } f \in [f(a_i), f(a_0)] \\ (t; s'_1, \ldots, s'_{i-1}, 0, s_{i+1}, \ldots, s_l), & \text{if } t \in [f(a_{l+1}, f(a_i)]; \end{cases}$$

and in the case $i = 0, 1$ then

$$\begin{aligned} (t; 0, s_2, \ldots, s_l) &\sim (t; 0, s'_2, \ldots s'_l) \quad \text{if } t \in [f(a_1), f(a_0)] \\ (t; s_1, \ldots, s_{l-1}, 0) &\sim (t; s'_1, \ldots, s'_{l-1}, 0) \quad \text{if } t \in [f(a_{l+1}, a_l)]; \end{aligned}$$

finally

$$\begin{aligned} (f(a_{l+1}); s_1, \ldots, s_l) &\sim (f(a_{l+1}); s'_1, \ldots, s'_l) \\ (f(a_0); s_1, \ldots, s_l) &\sim (f(a_0); s'_1, \ldots, s'_l). \end{aligned}$$

Note that if two points in $J_{\mathbf{a}} \times I^l$ are identified then they have the same image in $\mathcal{R}_f$, so we can equally well build the space $\mathcal{R}_f$ from the spaces $J_{\mathbf{a}} \times I^l/\sim$. The main point is the space $J_{\mathbf{a}} \times I^l/\sim$ is naturally homeomorphic to a $l+1$-simplex, with the δ_i map corresponding to the inclusions of the i^{th} face. That is, we have the following combinatorial result, whose verification is straightforward.

Lemma 12.8 *There are homeomorphisms*

$$h_{\mathbf{a}} : J_{\mathbf{a}} \times I^l/\sim \longrightarrow \Delta^{l+1}$$

which make the following diagrams commute

$$\begin{array}{ccc} J_{\mathbf{a}} \times I^l/\sim & \xrightarrow{h_{\mathbf{a}}} & \Delta^{l+1} \\ {\scriptstyle \delta_i}\uparrow & & \uparrow{\scriptstyle \delta_i} \\ J_{\mathbf{a}_i} \times I^{l-1}/\sim & \xrightarrow[h_{\mathbf{a}_i}]{} & \Delta^l \end{array}$$

where the left hand δ_i the map just constructed, and the right hand δ_i is the inclusion of the i^{th} face as described in chapter 13.

At this stage we have used up the first relations (12.1). Now we impose the relations (12.2) to get the following result.

Proposition 12.9 *There is a homeomorphism*

$$\mathcal{R}_f \cong \bigsqcup_{\mathbf{a}} \Delta^{l(\mathbf{a})+1} \times \mathcal{K}(\mathbf{a})/\sim$$

where

$$(s_0,\dots,s_{i-1},0,s_{i+1},\dots,s_l;\gamma_0,\dots,\gamma_l)\sim$$
$$\begin{cases}(s_1,\dots s_l;\gamma_1,\dots,\gamma_l) & if i=0\\ (s_0,\dots,s_{i-1},s_{i+1},\dots,s_l;\gamma_0,\dots,\gamma_{i-1}\circ_1\gamma_i,\dots\gamma_l) & if 1\le i\le l-1\\ (s_0,\dots,s_{l-1};\gamma_0,\dots,\gamma_{l-1}) & if i=l\end{cases}$$

Proof: This follows from the previous lemma by imposing the second list of relations (12.2). □

Notice that Theorem 12.7 now follows from this theorem by the definition of the classifying space of the category $\mathcal{C}_f$. As noted before, Theorem 12.7 implies Theorem 12.1.

□

Chapter 13

Simplicial Sets and Classifying Spaces

[Add more pictures

In the last chapter we saw how the topology of the framed moduli spaces of gradient flows of a Morse function determine the relative attaching maps in a CW complex homotopy equivalent to the manifold. In the next chapter we will strengthen this to show how these moduli spaces determine a simplicial structure of the manifold itself. That is, the topology of the moduli spaces determine the full topology of the manifold as a simplicial space; not only the homotopy type. In order to describe how this works we need the language of simplicial sets and spaces as well as the notion of a simplicial classifying space. We discuss these notions in this chapter. Good references for this theory are [?][?][?].

13.1 Simplicial Sets and Spaces

The idea of simplicial sets is to provide a combinatorial technique to study cell complexes built out of simplices; i.e. simplicial complexes. A simplicial complex X is built out of a union of simplices, glued together along faces. Thus if X_n denotes the indexing set for the n-dimensional simplices of X, then we can write

$$X = \left(\bigcup_{n \geq 0} \Delta^n \times X_n \right) \Big/ \sim$$

where Δ^n is the standard n-simplex in $\mathbb{R}^n$;

$$\Delta^n = \{(t_1, \ldots, t_n) \in \mathbb{R}^n \mid 0 \leq t_j \leq 1, \text{ and } \sum_{i=1}^{n} t_i \leq 1\}.$$

The gluing relation in this union can be encoded by set maps among the X_n's that would tell us for example how to identify an $n-l$-simplex indexed by an element of X_{n-1} with a particular face of an n-simplex indexed by an element of X_n. In particular, for every $x \in X_n$, we have $n+1$ "faces" $\partial_0, \ldots, \partial_n$ in X_{n-1}, and we have $n+1$ "degenerate simplices" $\sigma_0, \ldots, \sigma_n$ in X_{n+1} that are $n+1$-simplices that are "degenerate" in the sense that they are projected onto one of their faces (geometrically, these will all appear to be the same as x; but the parameterization of them as simplices is $n+1$-dimensional instead of n-dimensional, and one of the directions is degenerate).

Thus in principle, simplicial complexes can be studied purely combinatorially in terms of the sets X_n and set maps between them. The notion of a *simplicial set* makes this idea precise.

Definition 13.1 *A* simplicial set X_* *is a collection of sets*

$$X_n, \quad n \geq 0$$

together with set maps

$$\partial_i : X_n \longrightarrow X_{n-1} \quad \textit{and} \quad s_j : X_n \longrightarrow X_{n+1}$$

for $0 \leq i, j \leq n$ *called* face *and* degeneracy *maps respectively. These maps are required to satisfy the following compatibility conditions*

$$\begin{aligned} \partial_i \partial_j &= \partial_{j-1} \partial_i \quad \textit{for } i < j \\ s_i s_j &= s_{j+1} s_i \quad \textit{for } i < j \end{aligned}$$

and

$$\partial_i s_j = \begin{cases} s_{j-1}\partial_i & \textit{for } i < j \\ 1 & \textit{for } i = j,\ j+1 \\ s_j \partial_{i-1} & \textit{for } i > j+1 \end{cases}$$

As mentioned above, the maps ∂_i and s_j encode the combinatorial information necessary for gluing the simplices together. To say precisely how this works, consider the following maps between the standard simplices:

$$\delta_i : \Delta^{n-1} \longrightarrow \Delta^n \quad \text{and} \quad \sigma_j : \Delta^{n+1} \longrightarrow \Delta^n$$

for $0 \leq i, j \leq n$ defined by the formulae

$$\delta_i(t_1, \ldots, t_{n-1}) = \begin{cases} (t_1, \ldots, t_{i-1}, 0, t_i, \ldots, t_{n-1}) & \text{for } i \geq 1 \\ (1 - \sum_{q=1}^{n-1} t_q, t_1, \ldots, t_{n-1}) & \text{for } i = 0 \end{cases}$$

and

$$\sigma_j(t_1, \ldots, t_{n+1}) = \begin{cases} (t_1, \ldots, t_{i-1}, t_i + t_{i+1}, t_{i+2}, \ldots, t_{n+1}) & \text{for } i \geq 1 \\ (t_2, \ldots, t_{n+1}) & \text{for } i = 0\ . \end{cases}$$

δ_i includes Δ^{n-1} in Δ^n as the i^{th} face, and σ_j projects, in a linear fashion, Δ^{n+1} onto its j^{th} face.

We can now define the space associated to the simplicial set X_* as follows.

Definition 13.2 *The* geometric realization *of a simplicial set X_* is the space*

$$\|X_*\| = \left(\bigcup_{n \geq 0} \Delta^n \times X_n \right) \Big/ \sim$$

where if $t \in \Delta^{n-1}$ and $x \in X_n$, then

$$(t, \partial_i(x)) \sim (\delta_i(t), x)$$

and if $t \in \Delta^{n+1}$ and $x \in X_n$ then

$$(t, s_j(x)) \sim (\sigma_j(t), x).$$

In the topology of $\|X_*\|$, each X_n is assumed to have the discrete topology, so that $\Delta^n \times X_n$ is a discrete set of n-simplices.

Thus $\|X_*\|$ has one n-simplex for every element of X_n, glued together in a way determined by the face and degeneracy maps.

Example 13.1 *Consider the simplicial set $\mathbf{S}_*$ defined as follows. The set of n-simplices is given by*

$$\mathbf{S}_n = \mathbb{Z}/(n+1)\mathbb{Z}, \text{ generated by an element } \tau_n.$$

The face maps are given by

$$\partial_i(\tau_n^r) = \begin{cases} \tau_{n-1}^r & \text{if } r \leq i \leq n \\ \tau_{n-1}^{r-1} & \text{if } 0 \leq i \leq r-1. \end{cases}$$

The degeneracies are given by

$$s_i(\tau_n^r) = \begin{cases} \tau_{n+1}^r & \text{if } r \leq i \leq n \\ \tau_{n+1}^{r+1} & \text{if } 0 \leq i \leq r-1. \end{cases}$$

Notice that there is one zero simplex, two one simplices, one of them the image of the degeneracy $s_0 : \mathbf{S}_0 \longrightarrow \mathbf{S}_1$, and the other nondegenerate (i.e. not in the image of a degeneracy map). Notice also that all simplices in dimensions larger than one are in the image of a degeneracy map. Hence we have that the geometric realization

$$\|\mathbf{S}_*\| = \Delta^1/(0 \sim 1) = S^1.$$

Let X_* be any simplicial set. There is a particularly nice and explicit way for computing the homology of the geometric realization, $H_*(\|X_*\|)$.

Consider the following chain complex. Define $C_n(X_*)$ to be the free abelian group generated by the set of n-simplices X_n. Define the homomorphism

$$d_n : C_n(X_*) \longrightarrow C_{n-1}(X_*)$$

by the formula

$$d_n([x]) = \sum_{i=0}^{n} (-1)^i \partial_i([x])$$

where $x \in X_n$.

Proposition 13.3 *The homology of the geometric realization $H_*(\|X_*\|)$ is the homology of the chain complex*

$$\longrightarrow \ldots \xrightarrow{d_{n+1}} C_n(X_*) \xrightarrow{d_n} C_{n-1}(X_*) \xrightarrow{d_{n-1}} \ldots \xrightarrow{d_0} C_0(X_*).$$

Proof: It is straightforward to check that the geometric realization $\|X_*\|$ is a CW complex and that this is the associated cellular chain complex. □

Besides being useful computationally, the following result establishes the fact that all CW complexes can be studied simplicially:

Theorem 13.4 *Every CW complex has the homotopy type of the geometric realization of a simplicial set.*

Proof: Let X be a CW complex. Define the singular simplicial set of X , $\mathcal{S}(X)_*$ as follows. The n-simplices $\mathcal{S}(X)_n$ is the set of singular n-simplices,

$$\mathcal{S}(X)_n = \{c : \Delta^n \longrightarrow X\}.$$

The face and degeneracy maps are defined by

$$\partial_i(c) = c \circ \delta_i : \Delta^{n-1} \longrightarrow \Delta^n \longrightarrow X$$

and

$$s_j(c) = c \circ \sigma_i : \Delta^{n+1} \longrightarrow \Delta^n \longrightarrow X.$$

Notice that the associated chain complex to $\mathcal{S}(X)_*$ as in Proposition 13.3 is the singular chain complex of the space X. Hence by Proposition 13.3 we have that

$$H_*(\|\mathcal{S}(X)\|) \cong H_*(X).$$

This isomorphism is actually realized by a map of spaces

$$E : \|\mathcal{S}(X)_*\| \longrightarrow X$$

defined by the natural evaluation maps

$$\Delta^n \times \mathcal{S}(X)_n \longrightarrow X$$

given by

$$(t, c) \longrightarrow c(t).$$

It is straightforward to check that the map E does induce an isomorphism in homology. In fact it induces an isomorphism in homotopy groups. We will not prove this here; it is more technical and we refer the reader to [**?**] for details. Note that in the case where X is simply-connected, this would follow from the homological isomorphism by the Hurewicz theorem. A map between spaces that induces an isomorphism in homotopy groups is called a *weak homotopy equivalence*. Thus *any* space is weakly homotopy equivalent to a CW complex (i.e. the geometric realization of its singular simplicial set). But by the Whitehead theorem, two CW complexes that are weakly homotopy equivalent are homotopy equivalent. Hence X and $\|\mathcal{S}(X)_*\|$ are homotopy equivalent. □

We end this section by observing that the notion of simplicial set can be generalized as follows. We say that X_* is a *simplicial space¡* if it is a simplicial set (i.e. it satisfies Definition 13.1) where the sets X_n are topological spaces and the face and degeneracy maps

$$\partial_i : X_n \longrightarrow X_{n-1} \quad \text{and} \quad s_j : X_n \longrightarrow X_{n+1}$$

are continuous maps. The definition of the geometric realization of a simplicial space X_*, $\|X_*\|$, is the same as in Definition 13.2 with the proviso that the topology of each $\Delta^n \times X_n$ is the product topology. Notice that since the "set of n-simplices" X_n is actually a space, it is not necessarily true that $\|X_*\|$ is a CW complex, and in particular Proposition 13.3 does not hold. However if in fact each X_n is a CW complex and the face and degeneracy maps are cellular, then $\|X_*\|$ does have a natural CW structure induced by the product CW structures on $\Delta^n \times X_n$.

Notice that this simplicial notion generalizes even further. For example a *simplicial group* would be defined similarly, where each X_n would be a group and the face and degeneracy maps are group homomorphisms. Simplicial vector spaces, modules, etc. are defined similarly. The categorical nature of these definitions should by now be coming clear. Indeed most generally one can define a *simplicial object in a category* $\mathcal{C}$ using definition 13.1 where now the X_n's are assumed to be objects in the category and the face and degeneracies are assumed to be morphisms. If the category $\mathcal{C}$ is a subcategory of the category of sets then geometric realizations can be defined as in Definition 13.2. Notice for example that the geometric realization of a simplicial (abelian) group is a topological (abelian) group.

13.2 Categories and Classifying Spaces

We will now use simplicial techniques to construct classifying spaces of groups and of categories.

Let G be a topological group, and suppose that EG is a contractible space with a free, right G action

$$EG \times G \longrightarrow EG.$$

Let BG be the orbit space $BG = EG/G$. The following are standard facts from algebraic topology (bundle theory).

Theorem 13.5 *The quotient map $EG \longrightarrow BG$ is a principal G-bundle satisfying the following properties.*

1. *The homotopy type of BG is well defined. That is, it is independent of the choice of EG (so long as EG is contractible with a free G action.)*
2. *$\pi_q(BG) \cong \pi_{q-1}(G)$ so in particular if G is discrete BG is an Eilenberg–MacLane space $K(G, 1)$.*
3. *(Steenrod's classification theorem) Let X be any CW complex with a basepoint. Give BG a basepoint, and let $[X, BG]$ denote the (based) homotopy classes of basepoint preserving maps from X to BG. Then there is a bijective correspondence*

$$[X, BG] \xrightarrow[\psi]{\cong} Prin_G(X)$$

where $Prin_G(X)$ is the set of isomorphism classes of principal G-bundles over X. The correspondence ψ is given by

$$\psi([f]) = f^*(EG).$$

Because of this theorem the bundle $EG \longrightarrow BG$ is uniquely determined up to homotopy and is referred to as the universal principal G bundle.

We now describe a simplicial space model for the universal principal G - bundle $EG \longrightarrow BG$. Let G be a topological group and let $\mathcal{E}G_*$ be the simplicial space defined as follows. The space of n-simplices is given by the $n+1$-fold cartesian product

$$\mathcal{E}G_n = G^{n+1}.$$

The face maps $\partial_i : G^{n+1} \longrightarrow G^n$ are given by the formula

$$\partial_i(g_0, \ldots, g_n) = (g_0, \ldots, \hat{g}_i, \ldots, g_n).$$

The degeneracy maps $s_j : G^{n+1} \longrightarrow G^{n+2}$ are given by the formula

$$s_j(g_0, \ldots, g_n) = (g_0, \ldots, g_j, g_j, \ldots, g_n).$$

Exercise 13.1 *Show that the geometric realization $\|\mathcal{E}G_*\|$ is contractible. Hint: Let $\|\mathcal{E}G_*\|^{(n)}$ be the n-skeleton,*

$$\|\mathcal{E}G_*\|^{(n)} = \bigcup_{p=0}^{n} \Delta^p \times G^{p+1}.$$

Then show that the inclusion of one skeleton in the next $\|\mathcal{E}G_*\|^{(n)} \hookrightarrow \|\mathcal{E}G_*\|^{(n+1)}$ *is canonically null-homotopic. One way of doing this is to establish a homeomorphism between* $\|\mathcal{E}G_*\|^{(n)}$ *and* n*-fold join* $G * \cdots * G$*. See [?] for details.*

Notice that the group G acts freely on the right of $\|\mathcal{E}G_*\|$ by the rule

$$\|\mathcal{E}G_*\| \times G = \left(\bigcup_{p\geq 0} \Delta^p \times G^{p+1} \right) \times G \longrightarrow \|\mathcal{E}G_*\|$$
$$(t; (g_0, \ldots, g_p)) \times g \longrightarrow (t; (g_0 g, \ldots, g_p g)) .$$

Thus we can define $EG = \|\mathcal{E}G_*\|$. The orbit space $BG = EG/G$ has a similar simplicial structure defined as follows.

Let $\mathcal{B}G_*$ be the simplicial space whose n-simplices are the cartesian product

$$\mathcal{B}G_n = G^n. \tag{13.1}$$

The face and degeneracy maps are given by

$$\partial_i(g_1, \ldots, g_n) = \begin{cases} (g_2, \ldots, g_n) & \text{for } i = 0 \\ (g_1, \ldots, g_i g_{i+1}, \ldots g_n) & \text{for } 1 \leq i \leq n-1 \\ (g_1, \ldots, g_{n-1}) & \text{for } i = n. \end{cases}$$

The degeneracy maps are given by

$$s_j(g_1, \ldots, g_n) = \begin{cases} (1, g_1, \ldots, g_n) & \text{for } j = 0 \\ (g_1, \ldots g_j, 1, g_{j+1}, \ldots, g_n) & \text{for } j \geq 1. \end{cases}$$

The simplicial projection map

$$\pi : \mathcal{E}G_* \longrightarrow \mathcal{B}G_*$$

defined on the level of n-simplicies by

$$\pi(g_0, \ldots, g_n) = (g_0 g_1^{-1}, g_1 g_2^{-1}, \ldots, g_{n-1} g_n^{-1})$$

is easily checked to commute with face and degeneracy maps and so induces a map on the level of geometric realizations

$$\pi : EG = \|\mathcal{E}G_*\| \longrightarrow \|\mathcal{B}G_*\|$$

which induces a homemorphism

$$BG = EG/G \xrightarrow{\cong} \|\mathcal{B}G_*\|.$$

Thus for any topological group the construction in (13.1) gives a simplicial space model for its classifying space. This is referred to as the *simplicial bar construction.* Notice that when G is discrete the bar construction is a CW

complex for the classifying space $BG = K(G,1)$ and Proposition 13.3 gives a particularly nice complex for computing its homology. (The homology of a $K(G,1)$ is referred to as the homology of the group G.) The n - chains are the group ring

$$C_n(\mathcal{B}G_*) = \mathbb{Z}[G^n] \cong \mathbb{Z}[G]^{\otimes n}$$

and the boundary homomorphisms

$$d_n : \mathbb{Z}[G]^{\otimes n} \longrightarrow \mathbb{Z}[G]^{\otimes n-1}$$

are given by

$$d_n(a_1 \otimes \cdots \otimes a_n) = \quad (a_2 \otimes \cdots \otimes a_n) + \sum_{i=1}^{n-1} (-1)^i (a_1 \otimes \cdots \otimes a_i a_{i+1} \otimes \cdots \otimes a_n) + (-1)^n (a_1 \otimes \cdots \otimes a_{n-1}).$$

This complex is called the *bar complex* for computing the homology of a group and was discovered by Eilenberg and MacLane in the mid 1950's.

We end this chapter by observing that the bar construction of the classifying space of a group did not use the full group structure. It only used the existence of an associative multiplication with unit. That is, it did not use the existence of inverse. This allows the generalization of this construction to general categories. The main reference for this construction is [**?**].

Let $\mathcal{C}$ be a category. Let $Mor(A,B)$ be the set of morphisms between objects A and B. $\mathcal{C}$ is a *topological category* if the sets of morphisms are topologized and the composition pairings

$$Mor(A,B) \times Mor(B,C) \longrightarrow Mor(A,C)$$

are continuous. If no topology is specified, the morphisms are assumed to have the discrete topology. For $\gamma \in Mor(A,B)$ we say that A is the source of γ and B is the target.

The *simplicial nerve* of a category $\mathcal{C}$, is the simplicial space $\mathcal{NC}_*$ defined similarly to the simplicial bar construction of a group. More precisely, the space of 0-simplices $\mathcal{NC}_0$ is the space of objects, and for $n > 0$ the space of n-simplices $\mathcal{NC}_n$ is the space of n-tuples of composable morphisms:

$$\mathcal{NC}_n = \{(\gamma_1, \ldots, \gamma_n) \mid \text{ the target of } \gamma_i = \text{ the source of } \gamma_{i+1}, i = 1, \ldots, n-1\} \tag{13.2}$$

Following the lead of (13.1) the face maps are defined by the formula

$$\partial_i(\gamma_1, \ldots, \gamma_n) = \begin{cases} (\gamma_2, \ldots, \gamma_n) & \text{for } i = 0 \\ (\gamma_1, \ldots, \gamma_i\gamma_{i+1}, \ldots \gamma_n) & \text{for } 1 \le i \le n-1 \\ (\gamma_1, \ldots, \gamma_{n-1}) & \text{for } i = n. \end{cases}$$

If $n = 1$, $\partial_0(\gamma)$ is the source of γ and $\partial_1(\gamma)$ is its target. The degeneracy maps are given by

$$s_j(\gamma_1, \ldots, \gamma_n) = \begin{cases} (1, \gamma_1, \ldots, \gamma_n) & \text{for } j = 0 \\ (\gamma_1, \ldots \gamma_j, 1, \gamma_{j+1}, \ldots, \gamma_n) & \text{for } j \ge 1. \end{cases}$$

If $n = 0$, $s_j(o) = 1_o$, the identity on the object o.

Example 13.2 *Let G be a group, $\mathcal{C}_G$ be the category with one object (say $*$) and $Mor(*,*) = G$. The composition law in the category is given by group multiplication. We then see that the nerve of this category is the simplicial bar construction*

$$(\mathcal{N}\mathcal{C}_G)_* = \mathcal{B}G_*.$$

Motivated by this example we make the following definition.

Definition 13.6 *The classifying space of a category $\mathcal{C}$ is the geometric realization of its simplicial nerve:*

$$B\mathcal{C} = \|\mathcal{N}\mathcal{C}_*\|.$$

We will apply this construction in Morse theory by studying the category whose objects are the critical points of a Morse function and whose morphisms are the moduli spaces of flow lines. We will discuss in great detail in the next chapter. We end this chapter with two examples.

Example 13.3 *Let $\mathcal{C}_{\mathbb{R}}$ be the category whose objects are finite dimensional vector spaces over $\mathbb{R}$ and whose morphisms are isomorphisms* between them. (So if two vector spaces have different dimensions the space of morphisms between them is empty.) We then have

$$B\mathcal{C}_{\mathbb{R}} = \coprod_{n\geq 0} BGL_n(\mathbb{R}).$$

Example 13.4 *Let $\mathcal{C}$ denote the category whose objects are finite sets and whose morphisms are set bijections. (Again if two sets have different cardinality there are no morphisms between them.) We then have*

$$B\mathcal{C} = \coprod_{n\geq 0} B\Sigma_n$$

where Σ_n is the symmetric group on n letters.

Chapter 14

Spectral Sequences and the Filtered Nerve of a Morse Function

In this chapter we discuss spectral sequences in some generality and then define and study a spectral sequence associated to a Morse function on a compact manifold. This spectral sequence arises from a natural filtration of the category $\mathcal{C}_f$ of a Morse function $f : M \longrightarrow \mathbb{R}$. This filtration is given on the level of objects by the index of the critical points. It induces a filtration on the associated simplicial nerve, $\mathcal{N}_f$, and hence on the level of classifying spaces. This filtration will induce a spectral sequence converging to any generalized homology of the manifold (e.g. homology, stable homotopy, K-theory, cobordism theory). In the case of homology we will show how the spectral sequence degenerates to give the Morse–Smale chain complex for computing the homology of the manifold.

14.1 Spectral Sequences

A spectral sequence is the algebraic machinery for studying sequences of long exact sequences that are interelated in a particular way. We begin by illustrating this with the example of a filtered complex.

Let C_* be a chain complex, and let $A_* \subset C_*$ be a subcomplex. The short exact sequence of chain complexes

$$0 \longrightarrow A_* \hookrightarrow C_* \longrightarrow C_*/A_* \longrightarrow 0$$

leads to a long exact sequence in homology:

$$\begin{aligned} \longrightarrow \ldots \longrightarrow H_{q+1}(C_*, A_*) \longrightarrow H_q(A_*) \longrightarrow H_q(C_*) \\ \longrightarrow H_q(C_*, A_*) \longrightarrow H_{q-1}(A_*) \longrightarrow \ldots \end{aligned}$$

This is useful in computing the homology of the big chain complex, $H_*(C_*)$ in terms of the homology of the subcomplex $H_*(A_*)$ and the homology of the quotient complex $H_*(C_*, A_*)$. A spectral sequence is the machinery used to study the more general situation when one has a *filtration* of a chain complex C_* by subcomplexes

$$0 = F_0(C_*) \hookrightarrow F_1(C_*) \hookrightarrow \ldots \hookrightarrow F_k(C_*) \hookrightarrow F_{k+1}(C_*) \hookrightarrow \ldots \hookrightarrow C_* = \bigcup_k F_k(C_*).$$

Let D_*^k be the subquotient complex $D_*^k = F_k(C_*)/F_{k-1}(C_*)$ and so for each k there is a long exact sequence in homology

$$\longrightarrow H_{q+1}(D_*^k) \longrightarrow H_q(F_{k-1}(C_*)) \longrightarrow H_q(F_k(C_*)) \longrightarrow H_q(D_*^k) \longrightarrow \ldots$$

By putting these long exact sequences together, in principle one should be able to use information about $\oplus_k H_*(D_*^k)$ in order to obtain information about

$$H_*(C_*) = \varinjlim_k H_*(F_k(C_*)).$$

A spectral sequence is the bookkeeping device that allows one to do this. To be more specific, consider the following diagram.

$$
\begin{array}{ccccccc}
0 & & & & 0 & & \\
\downarrow i & & & & \downarrow i & & \\
H_q(F_1(C_*)) & & & & H_{q-1}(F_1(C_*)) & = & H_{q-1}(D^1_*) \\
\downarrow i & & & & \downarrow i & & \\
\vdots & & & & \vdots & & \\
\downarrow i & & & & \downarrow i & & \\
H_q(F_{k-p}(C_*)) & \xrightarrow{j} & H_q(D^{k-p}_*) & \xrightarrow{\partial} & H_{q-1}(F_{k-p-1}(C_*)) & \xrightarrow{j} & H_{q-1}(D^{k-p-1}_*) \\
\downarrow i & & & & \downarrow i & & \\
\vdots & & & & H_{q-1}(F_{k-p}(C_*)) & \xrightarrow{j} & H_{q-1}(D^{k-p}_*) \\
\downarrow i & & & & \downarrow i & & \\
\vdots & & & & \vdots & & \\
\downarrow i & & & & \downarrow i & & \\
H_q(F_{k-2}(C_*)) & & & & H_{q-1}(F_{k-3}(C_*)) & & \\
\downarrow i & & & & \downarrow i & & \\
H_q(F_{k-1}(C_*)) & \xrightarrow{j} & H_q(D^{k-1}_*) & \xrightarrow{\partial} & H_{q-1}(F_{k-2}(C_*)) & \xrightarrow{j} & H_{q-1}(D^{k-2}_*)) \\
\downarrow i & & & & \downarrow i & & \\
H_q(F_k(C_*)) & \xrightarrow{j} & H_q(D^k_*) & \xrightarrow{\partial} & H_{q-1}(F_{k-1}(C_*)) & \xrightarrow{j} & H_{q-1}(D^{k-1}_*)) \\
\downarrow i & & & & \downarrow i & & \\
\vdots & & & & \vdots & & \\
\downarrow i & & & & \downarrow i & & \\
H_q(C_*) & & & & H_{q-1}(C_*) & &
\end{array}
\tag{14.1}
$$

The columns represent the homology filtration of $H_*(C_*)$ and the three maps ∂, j, and i combine to give long exact seqences at every level.

Let $\alpha \in H_q(C_*)$. We say that α has *algebraic filtration* k, if α is in the image of a class $\alpha_k \in H_q(F_k(C_*))$ but is not in the image of $H_q(F_{k-1}(C_*))$. In such a case we say that the image $j(\alpha_k) \in H_q(D^k_*)$ is a *representative* of α. Notice that this representative is not unique. In particular we can add any class in the

image of

$$d_1 = j \circ \partial : H_{q+1}(D_*^{k+1}) \longrightarrow H_q(D_*^k)$$

to $j(\alpha_k)$ and we would still have a representative of $\alpha \in H_q(C_*)$ under the above definition.

Conversely, let us consider when an arbitrary class $\beta \in H_q(D_*^k)$ represents a class in $H_q(C_*)$. By the exact sequence this occurs if and only if the image $\partial(\beta) = 0$, for this is the obstruction to β being in the image of $j : H_q(F_k(C_*)) \to H_q(D_*^k)$ and if $j(\tilde{\beta}) = \beta$ then β represents the image

$$i \circ \cdots \circ i(\tilde{\beta}) \in H_q(C_*).$$

Now $\partial(\beta) = 0$ if and only if it lifts all the way up the second vertical tower in diagram 14.1. The first obstruction to this lifting, (i.e. the obstruction to lifting $\partial(\beta)$ to $H_{q-1}(F_{k-2}(C_*))$ is that the composition

$$d_1 = j \circ \partial : H_q(D^k) \longrightarrow H_{q-1}(D_*^{k-1})$$

maps β to zero. That is elements of $H_q(C_*)$ are represented by elements in the subquotient

$$\ker(d_1)/\operatorname{Im}(d_1)$$

of $H_q(D_*^k)$. We use the following notation to express this. We define

$$E_1^{r,s} = H_{r+s}(D_*^r)$$

and define

$$d_1 = j \circ \partial : E_1^{r,s} \longrightarrow E_1^{r-1,s}.$$

r is said to be the *algebraic filtration* of elements in $E_1^{r,s}$ and $r + s$ is the *total degree* of elements in $E_1^{r,s}$. Since $\partial \circ j = 0$, we have that

$$d_1 \circ d_1 = 0$$

and we let

$$E_2^{r,s} = \ker(d_1 : E_1^{r,s} \to E_1^{r-1,s})/\operatorname{Im}(d_1 : E_1^{r+1,s} \to E_1^{r,s})$$

be the resulting homology group. We can then say that the class $\alpha \in H_q(C_*)$ has as its representative, the class $\alpha_k \in E_2^{k,q-k}$.

Now let us go back and consider further obstructions to an arbitrary class $\beta \in E_2^{k,q-k}$ representing a class in $H_q(C_*)$. Represent β as a cycle in E_1: $\beta \in \ker(d_1 = j \circ \partial \in H_q(D_*^k))$. Again, β represents a class in $H_q(C_*)$ if and only if $\partial(\beta) = 0$. Now since $j \circ \partial(\beta) = 0$, $\partial(\beta) \in H_{q-1}(F_{k-1}(C_*))$ lifts to a class, say $\tilde{\beta} \in H_{q-1}F_{k-2}(C_*)$. Remember that the goal was to lift $\partial(\beta)$ all the way up the vertical tower (so that it is zero). The obstruction to lifting it the next stage, i.e. to $H_{q-1}(F_{k-3}(C_*))$ is that $j(\tilde{\beta}) \in H_{q-1}(D_*^{k-2})$ is zero. Now the fact that a d_1 cycle β has the property that $\partial(\beta)$ lifts to $H_{q-1}F_{k-2}(C_*))$ allows to define a map

$$d_2 : E_2^{k,q-k} \longrightarrow E_2^{k-2,q-k+1}$$

and more generally,

$$d_2 : E_2^{r,s} \longrightarrow E_2^{r-2,s+1}$$

by composing this lifting with

$$j : H_{s+r-1}(F_{r-2}(C_*)) \longrightarrow H_{s+r-1}(D_*^{r-2}).$$

That is, $d_2 = j \circ i^{-1} \circ \partial$. It is straightforward to check that $d_2 : E_2^{r,s} \longrightarrow E_2^{r-2,s+1}$ is well defined, and that elements of $H_q(C_*)$ are actually represented by elements in the subquotient homology groups of $E_2^{*,*}$:

$$E_3^{r,s} = \ker(d_2 : E_2^{r,s} \to E_2^{r-2,s+1}) / \operatorname{Im}(d_2 : E_2^{r+2,s-1} \to E_1^{r,s})$$

Inductively, assume the subquotient homology groups $E_j^{r,s}$ have been defined for $j \leq p-1$ and differentials

$$d_j : E_j^{r,s} \longrightarrow E_j^{r-j,s+j-1}$$

defined on representative classes in $H_{r+s}(D_*^r)$ to be the composition

$$d_j = j \circ (i^{j-1} = i \circ \cdots \circ i)^{-1} \circ \partial$$

so that $E_{j+1}^{*,*}$ is the homology $\ker(d_j)/\operatorname{Im}(d_j)$. We then define

$$E_p^{r,s} = \ker(d_{p-1} : E_{p-1}^{r,s} \to E_{p-1}^{r-p+1,s+p-2}) / \operatorname{Im}(d_{p-1} : E_{p-1}^{r+p-1,s-p+2} \to E_{p-1}^{r,s}).$$

Thus $E_p^{k,q-k}$ is a subquotient of $H_q(D_*^k)$, represented by elements β so that $\partial(\beta)$ lifts to $H_q(F_{k-p}(C_*))$. That is, there is an element $\tilde{\beta} \in H_q(F_{k-p}(C_*))$ so that

$$i^{p-1}(\tilde{\beta}) = \partial(\beta) \in H_{q-1}(F_{k-1}(C_*)).$$

The obstruction to $\tilde{\beta}$ lifting to $H_{q-1}(F_{k-p-1}(C_*))$ is $j(\beta) \in H_q(D_*^{k-p})$. This procedure yields a well defined map

$$d_p : E_p^{r,s} \longrightarrow E_p^{r-p,s+p-1}$$

given by $j \circ (i^{p-1})^{-1} \circ \partial$ on representative classes in $H_q(D_*^k)$. This completes the inductive step. Notice that if we let

$$E_\infty^{r,s} = \varinjlim_p E_p^{r,s}$$

then $E_\infty^{k,q-k}$ is a subquotient of $H_q(D_*^k)$ consisting of precisely those classes represented by elements $\beta \in H_q(D_*^k)$ so that $\partial(\beta)$ lifts all the way up the vertical tower i.e. $\partial(\beta)$ is in the image of i^p for all p. This is equivalent to the condition that $\partial(\beta) = 0$ which as observed above is precisely the condition necessary for β to represent a class in $H_q(C_*)$. These observations can be made more precise as follows.

Theorem 14.1 *Let $I^{r,s} = \mathrm{Im}(H_{r+s}(F_r(C_*)) \longrightarrow H_{r+s}(C_*))$. Then $E_\infty^{r,s}$ is isomorphic to the quotient group*

$$E_\infty^{r,s} \cong I^{r,s}/I^{r-1,s+1}.$$

Thus the $E_\infty^{*,*}$ determines $H_*(C_*)$ up to extensions. In particular, if all homology groups are taken with field coefficients we have

$$H_q(C_*) \cong \bigoplus_{r+s=q} E_\infty^{r,s}.$$

In this case we say that $\{E_p^{r,s}, d_p\}$ is a *spectral sequence* starting at $E_1^{r,s} = H_{r+s}(D_*^r)$, and converging to $H_{r+s}(C_*)$.

Often times a filtration of this type occurs when one has a topological space X filtered by subspaces,

$$* = X_0 \hookrightarrow X_1 \hookrightarrow \ldots \hookrightarrow X_k \hookrightarrow X_{k+1} \hookrightarrow \ldots \hookrightarrow X.$$

We get a spectral sequence as above by applying the homology of the chain complexes to this topological filtration. This spectral sequence converges to $H_*(X)$ with E_1 term $E_1^{r,s} = H_{r+s}(X_r, X_{r-1})$. From the construction of this spectral sequence one notices that chain complexes are irrelevant in this case; indeed all one needs is the fact that each inclusion $X_{k-1} \hookrightarrow X_k$ induces a long exact sequence in homology. Hence if $h_*(-)$ is any *generalized homology theory* (that is, a functor that obeys all the Eilenberg–Steenrod axioms but dimension) then the inclusions $X_{k-1} \hookrightarrow X_k$ induce long exact sequences in $h_*(-)$, and one gets, by a procedure completely analogous to the above, a spectral sequence converging to $h_*(X)$ with E_1 term

$$E_1^{r,s} = h_{r+s}(X_r, X_{r-1}).$$

Particularly important examples of such generalized homology theories include stable homotopy ($\cong$ framed bordism), other bordism theories, and K-homology theory. Similar spectral sequences also exist for cohomology theories. The reader is referred to [**?**] for a good general reference on spectral sequences with many examples of those most relevant in algebraic topology. A quick description of spectral sequences with complete details can be found in [**?**].

14.2 A Filtration of the Category of a Morse Function

Let $f : M \longrightarrow \mathbb{R}$ be a Morse function on a compact manifold M, and suppose that it satisfies the Morse–Smale transversality conditions. As in the last chapter, let $\mathcal{C}_f$ be the category associated to f. In this section we will see how $\mathcal{C}_f$ is naturally a filtered category, which induces a filtration of $B\mathcal{C}_f \cong M$. We will examine the resulting spectral sequence converging to $H_*(M)$.

For $k \geq 0$ let $\mathcal{C}_f^k$ be the subcategory of $\mathcal{C}_f$ whose objects are critical points of f of index $\leq k$. This is referred to as a *full* subcategory because the space of morphisms between any two objects a and b in the subcategory $\mathcal{C}_f^k$ is the full space of morphisms (namely the compactified moduli space $\overline{\mathcal{M}}(a,b)$) in the big category $\mathcal{C}_f$. On the level of classifying spaces we therefore have a filtration of simplicial spaces (nerves)

$$* \hookrightarrow B\mathcal{C}_f^0 \hookrightarrow B\mathcal{C}_f^1 \hookrightarrow \ldots \hookrightarrow B\mathcal{C}_f^k \hookrightarrow B\mathcal{C}_f^{k+1} \hookrightarrow \ldots \hookrightarrow B\mathcal{C}_f \cong M.$$

Thus there is a resulting spectral sequence, which we refer to as the *index spectral sequence.*

Theorem 14.2 *Let $Crit_k$ denote the set of critical points of f having index k. Then there is a natural homotopy equivalence*

$$B\mathcal{C}_f^k / B\mathcal{C}_f^{k-1} \simeq \bigvee_{Crit_k} S^k.$$

Thus in the index spectral sequence we have that

$$E_1^{r,s} = H_{r+s}(B\mathcal{C}_f^r, B\mathcal{C}_f^{r-1}) = \begin{cases} \bigoplus_{Crit_r} \mathbb{Z} & \text{if } s = 0 \\ 0 & \text{if } s > 0. \end{cases}$$

Furthermore the E_1 chain complex

$$\longrightarrow \ldots \longrightarrow E_1^{r,0} \xrightarrow{d_1} E_1^{r-1,0} \xrightarrow{d_1} \ldots$$

is precisely the Morse–Smale chain complex

$$\longrightarrow \ldots \longrightarrow \bigoplus_{Crit_r} \mathbb{Z} \xrightarrow{\partial} \bigoplus_{Crit_{r-1}} \mathbb{Z} \xrightarrow{\partial} \ldots$$

described in chapter 5. Hence

$$E_2^{r,s} = \begin{cases} H_r(M) & \text{if } s = 0 \\ 0 & \text{if } s > 0 \end{cases}$$

and the spectral sequence collapses at E_2 (that is $E_2^{r,s} = E_p^{r,s}$ for all $p \geq 2$.)

Proof: Recall that the homeomorphism $B\mathcal{C}_f \xrightarrow{\cong} M$ given in chapter 12, section 9.2 was defined via two homeomorphisms:

$$\phi : \mathcal{R}_f \xrightarrow{\cong} M \quad \text{and} \quad \mathcal{R}_f \xrightarrow{\cong} B\mathcal{C}_f$$

where

$$\mathcal{R}_f = \bigsqcup_{\mathbf{a}} J_{\mathbf{a}} \times I^{l(\mathbf{a})} \times \mathcal{K}(\mathbf{a}) / \sim .$$

Notice that $\mathcal{R}_f$ has an index filtration, by letting $\mathcal{R}_f^k$ be the above union, except restricting to sequences $(\mathbf{a})$ that start at critical points $s(\mathbf{a}) = a_0$ of index $\leq k$. Now the proof of Theorem 12.1 actually proves that

$$\mathcal{R}_f^k \cong B\mathcal{C}_f^k.$$

Thus it is enough to prove this theorem with $\mathcal{R}_f^k$ replacing $B\mathcal{C}_f^k$ everywhere. The reason we choose to do this is because the homeomorphism

$$\phi : \mathcal{R}_f \longrightarrow M$$

is explicitly given. In particular the restriction of ϕ to $J_{\mathbf{a}} \times I^{l(\mathbf{a})} \times \mathcal{K}(\mathbf{a})$ is given by

$$\phi(t; s_1, \ldots, s_l; \gamma_0 \ldots, \gamma_l) = (\gamma_0 \circ_{s_1} \cdots \circ_{s_l} \gamma_l)(t)$$

where recall if $\mathbf{a} = (a_0, \ldots, a_{l+1})$ then $J_{\mathbf{a}} = [f(a_{l+1}), f(a_0)]$.

To make things clearer, assume that f has exactly one critical point a_0 of index k. Notice then that the space $\mathcal{R}_f^k/\mathcal{R}_f^{k-1}$ is a quotient space of the disjoint union

$$\bigsqcup_{s(\mathbf{a})=a_0} J_{\mathbf{a}} \times I^{l(\mathbf{a})} \times \mathcal{K}(\mathbf{a})$$

where the disjoint union is taken over all ordered sequences of critical points $(\mathbf{a})$ with starting point $s(\mathbf{a})$ equal to the critical point a_0, and where the quotient space consists of identifying many things in this space that we will explore later. (In the general setting the disjoint union would be taken over all sequences that start at any critical point of index k.)

Now $\phi(\mathcal{R}_f^k - \mathcal{R}_f^{k-1})$ is the set of points that lie on a piecewise flow line that starts at a_0 that do not also lie on a piecewise flow line that starts anywhere else. In other words, it is the set of points that line on a flow line (not piecewise) that starts at a_0; i.e. the unstable manifold $W^u(a_0)$, which is a disk of dimension k.

Now $\mathcal{R}_f^k$ is compact, and therefore so is $\mathcal{R}_f^k/\mathcal{R}_f^{k-1}$. Therefore $\mathcal{R}_f^k/\mathcal{R}_f^{k-1}$ is the one-point compactification of $\mathcal{R}_f^k - \mathcal{R}_f^{k-1} \cong W^u(a_0) \cong D^k$. By the uniqueness of one-point compactifications, $\mathcal{R}_f^k/\mathcal{R}_f^{k-1} \cong S^k$.

It is also helpful to prove this step a little more explicitly: we first extend ϕ to a map

$$\hat{\phi} : \mathcal{R}_f^k/\mathcal{R}_f^{k-1} \longrightarrow \widehat{W}^u(a_0)$$

where $\widehat{W}^u(a_0) = W^u(a_0) \cup \{*\} \cong S^k$ is the one-point compactification of $W^u(a_0)$.

Exercise 14.1 *Show that $\hat{\phi}$ is a homeomorphism.*

This proves the first statement of Theorem 14.2. The d_1 differential can be computed by studying the homotopy class of the attaching map

$$S^{k-1} = \partial\overline{W}^u(a_0) \hookrightarrow \mathcal{R}_f^{k-1} \xrightarrow{proj} \mathcal{R}_f^{k-1}/\mathcal{R}_f^{k-2} \simeq \bigvee_{Crit_{k-1}} S^{k-1}.$$

But the above analysis of identifying the strata $\mathcal{R}_f^q - \mathcal{R}_f^{q-1}$ with the union of the unstable disks of the critical points of index q, shows that this attaching map is precisely the relative attaching map in the CW complex associated to the Morse function f (see chapter 6). The rest of Theorem 14.2 now follows. □

We end this chapter by remarking that since the index spectral sequence arose from a topological filtration of $M = B\mathcal{C}_f$, one gets corresponding spectral sequences with arbitrary generalized homology applied to this filtration. However in the general setting one would *not* expect the spectral sequence to collapse at E_2 (or any fixed E_p for that matter). The spectral sequence for stable homotopy is an interesting example, where Franks' theorem (Theorem 11.3) can be interpreted as saying that the higher differentials can be computed in terms of stable homotopy classes corresponding under the Thom–Pontryagin construction to the framed moduli spaces of flows.

Part IV

Generalizations

Chapter 15

Morse–Bott theory: Functions with Nondegenerate Critical Submanifolds

[Draw more pictures to illustrate [group action: more explicit, problems with transversality

In many of the most important cases, we have a group G acting on the manifold M, and the function $f : M \longrightarrow \mathbb{R}$ is G-invariant in the sense that $f(x) = f(g \cdot x)$ for all $x \in M$ and $g \in G$. In many cases we might hope to study the topology of M/G using f.

As always, we can again perturb f so that it is Morse or Morse–Smale, but this might not be desirable since doing so may destroy the G invariance. For instance, consider the torus T^2 embedded in $\mathbb{R}^3$ as in Figure 15.1, lying on its side. Rotating around the z axis provides an action of the group S^1 on the torus. The height (z coordinate) is a function from T^2 to $\mathbb{R}$, and it is invariant under the S^1 rotation.

Clearly, the height function is not Morse, since the critical points form two continuous famillies: one circle's worth on the top and one circle's worth on the bottom. This is to be expected since if p is a critical point of f and f is invariant under G, then for any $g \in G$, $g \cdot p$ is a critical point of f as well. So as long as the group G is more than zero-dimensional we do not expect f to be Morse (though it might be if, for instance, p is a fixed point for G).

In this situation, it is clear that no small perturbation will both preserve the S^1 invariance and satisfy the Morse condition. In fact, the most we can do is have critical submanifolds.

It turns out we can generalize Morse theory to allow for critical submanifolds

Figure 15.1: A torus embedded in $\mathbb{R}^3$, lying on its side. The height function is an S^1-invariant Morse–Bott function.

of M (instead of just critical points). We need to define what it means for such a critical submanifold to be nondegenerate, in analogy with nondegenerate critical points, and we will see that much of the same theory generalizes to this setting. The origins of this theory and its analogy to Morse theory are due to Bott [**?**][**?**]. Another good reference is [**?**].

15.1 The General Theory

Let $f : M^m \longrightarrow \mathbb{R}$ be a smooth function on an m-dimensional manifold. An n-dimensional connected submanifold

$$N^n \hookrightarrow M^m$$

is said to be a *critical submanifold* if every point of N is a critical point. That is,

$$df(x) = 0 \quad \text{for all} \quad x \in N.$$

Let $\nu(N) \longrightarrow N$ be the normal bundle of the submanifold N. This is the $m - n$ dimensional orthogonal complement to the tangent bundle of N inside the tangent bundle of M. That is, for any $x \in N$ we have a splitting of the tangent space

$$T_x M \cong T_x N \oplus \nu_x(N).$$

The Hessian at f at x, viewed as a symmetric bilinear form

$$Hess_x(f) : T_x M \times T_x M \longrightarrow \mathbb{R}$$

has the obvious property that it is zero when either of the coordinates lie in $T_x N \subset T_x M$. Thus the Hessian determines a symmetric bilinear form on the normal bundle

$$Hess_x^N(f) : \nu_x(N) \times \nu_x(N) \longrightarrow \mathbb{R}$$

The critical submanifold N is said to be *nondegenerate* if $Hess^N_x(f)$ is a nonsingular form at every $x \in N$. The function $f : M \longrightarrow \mathbb{R}$ is said to be *Morse–Bott* if all of its critical points lie in a disjoint union of nondegenerate critical submanifolds. Notice that a Morse function is a Morse–Bott function, all of whose critical submanifolds are zero-dimensional (i.e. points).

Again, let N be a nondegenerate critical submanifold of $f : M \longrightarrow \mathbb{R}$. Using the induced Riemannian metric on the normal bundle $\nu(N)$ of N, the Hessian $Hess^N$ defines a self adjoint endomorphism

$$A_N : \nu(N) \longrightarrow \nu(N)$$

by

$$\langle A_N(x), y\rangle = Hess^N(f)(x, y).$$

The nondegeneracy of N implies that all the eigenvalues of A_N are nonzero. Hence we can decompose each normal space $\nu_x(N)$ into its positive and negative eigenspace

$$\nu_x(N) = \nu^+_x(N) \oplus \nu^-_x(N)$$

which fit together to give a splitting of bundles

$$\nu(N) = \nu^+(N) \oplus \nu^-(N).$$

The fibre dimension of the negative normal bundle $\nu^-(N)$ is called the *index* of the critical submanifold N.

The following is the analogue of the Morse lemma in this context.

Theorem 15.1 *Let $f : M \longrightarrow \mathbb{R}$ be a smooth function on an m dimensional manifold that has a nondegenerate critical submanifold $N \hookrightarrow M$ of dimension n and index λ. Let*

$$p^+ : \nu^+(N) \longrightarrow N \quad \textit{and} \quad p^- : \nu^-(N) \longrightarrow N$$

be the positive and negative normal bundles as above. Then for any $x \in N$ there is a neighborhood U of x in N, a neighborhood V of x in M, local trivializations

$$\psi^- : (p^-)^{-1}(U) \xrightarrow{\cong} U \times \mathbb{R}^\lambda$$
$$\psi^+ : (p^+)^{-1}(U) \xrightarrow{\cong} U \times \mathbb{R}^{m-n-\lambda},$$

and a diffeomorphism

$$\phi : U \times \mathbb{R}^\lambda \times \mathbb{R}^{m-n-\lambda} \xrightarrow{\cong} V$$

so that

$$f \circ \phi : U \times \mathbb{R}^\lambda \times \mathbb{R}^{m-n-\lambda} \longrightarrow \mathbb{R}$$

is given by the formula

$$f \circ \phi(u; v_1, \ldots, v_\lambda; w_1, \ldots, w_{m-n-\lambda}) = f(x) - \sum_{i=1}^{\lambda} v_i^2 + \sum_{j=1}^{m-n-\lambda} w_j^2.$$

This theorem is simply a parameterized form of the classical Morse lemma (Theorem 5.3) and we leave its proof to the reader.

We call a function $f : M \longrightarrow \mathbb{R}$ *Morse–Bott* if all of its critical points lie in a disjoint union of connected, nondegenerate critical submanifolds. An important property of Morse–Bott functions that the more stringent condition of being a Morse function does not satisfy, is that this property is preserved under fibre bundles. That is, we have the following straightforward result.

Proposition 15.2 *Let $\pi : E \longrightarrow M$ be a smooth fibre bundle. Then a smooth function $f : M \longrightarrow \mathbb{R}$ is Morse–Bott if and only if the composition $f \circ \pi : E \longrightarrow \mathbb{R}$ is Morse–Bott. Moreover if $N \hookrightarrow M$ is a nondegenrate critical submanifold of f, then its index is equal to the index of $\pi^{-1}(N) \hookrightarrow E$ as a critical submanifold of $f \circ \pi$.*

As was proved in chapter 4 for Morse functions, a Morse–Bott function $f : M \longrightarrow \mathbb{R}$ has the property that every point $x \in M$ has unique flow line

$$\gamma_x : \mathbb{R} \longrightarrow M$$

satisfying the flow equations

$$\frac{d\gamma_x}{dt}(s) + \nabla_{\gamma_x(s)}(f) = 0$$

and the initial condition

$$\gamma_x(0) = x.$$

Furthermore all flow lines begin and end at points in critical submanifolds. These results are all consequences of the existence and uniqueness of solutions of ordinary differential equations and the compactness of the manifold M.

Of course, as before, if the point x lies on a critical submanifold then the flow line γ_x is the constant function at x.

Let N be an n dimensional critical submanifold of index λ, of a Morse–Bott function $f : M \longrightarrow \mathbb{R}$, where M is m-dimensional. Analogous to the case of nondegenerate critical points we define the stable and unstable manifolds of N as follows.

$$W^s(N) = \{x \in M : \lim_{t \to \infty} \gamma_x(t) \in N\}$$
$$W^u(N) = \{x \in M : \lim_{t \to -\infty} \gamma_x(t) \in N\}.$$

Define

$$\pi^u : W^u(N) \longrightarrow N \quad \text{and} \quad \pi^s : W^s(N) \longrightarrow N$$

by

$$\pi^u(x) = \lim_{t \to -\infty} \gamma_x(t) \quad \text{and} \quad \pi^s(x) = \lim_{t \to \infty} \gamma_x(t)$$

respectively. A parameterized version of Theorem 2 (stating that the stable and unstable manifolds of Morse functions are disks) is the following.

Theorem 15.3 *The maps*

$$\pi^u : W^u(N) \longrightarrow N \quad \textit{and} \quad \pi^s : W^s(N) \longrightarrow N$$

are smooth fibre bundles with fibers diffeomorphic to the disks D^λ and $D^{m-n-\lambda}$ respectively. Moreover as bundles they are isomorphic to the negative and positive normal bundles

$$p^- : \nu^-(N) \longrightarrow N \quad \textit{and} \quad p^+ : \nu^+(N) \longrightarrow N$$

respectively.

Continuing the analogy with Morse functions we say that the Morse–Bott function $f : M \longrightarrow \mathbb{R}$ satisfies the *Morse–Smale transversality condition* if the intersections of the unstable and stable manifolds of all of the critical submanifolds,

$$W^u(N_1) \cap W^s(N_2)$$

are transverse. In this case we set

$$W(N_1, N_2) = W^u(N_1) \cap W^s(N_2).$$

This is the space of points that lie on flow lines that start in N_1 and end in N_2. We note that by transversality this space is a manifold of dimension

$$\begin{aligned} dim(W^u(N_1)) + dim(W^s(N_2)) - dim(M) & \\ &= (n_1 + \lambda_1) + (n_2 + (m - n_2 - \lambda_2)) - m \\ &= n_1 + \lambda_1 - \lambda_2. \end{aligned}$$

Here $n_i = dim(N_i)$, $\lambda_i = index(N_i)$.

Notice that as in the critical point situation, the real numbers $\mathbb{R}$ acts on these spaces:

$$\begin{aligned} W(N_1, N_2) \times \mathbb{R} &\longrightarrow W(N_1, N_2) \\ (x, t) &\longrightarrow \gamma_x(t) \end{aligned}$$

When $N_1 \neq N_2$ this action is free, and we write

$$\mathcal{M}(N_1, N_2) = W(N_1, N_2)/\mathbb{R}.$$

$\mathcal{M}(N_1, N_2)$ is the moduli space of flows beginning in N_1 and ending in N_2. It is a manifold of dimension $n_1 + \lambda_1 - \lambda_2 - 1$. In particular when $n_1 + \lambda_1 < \lambda_2 - 1$ then this space is empty; that is there are no flow lines from N_1 to N_2. Notice that when $N_1 = N_2$, we have

$$W(N_1, N_1) = N_1$$

and the action of $\mathbb{R}$ on this space is trivial. Hence $\mathcal{M}(N_1, N_1) = N_1$.

The next step in developing Morse theory for Morse–Bott functions would be to see how the topology of the manifold M is determined by the topology of the critical submanifolds and the spaces of flow lines between them. In analogy with Theorem 5.6 we have the following:

Theorem 15.4 *Let $f : M \longrightarrow \mathbb{R}$ be a Morse–Bott function on a compact manifold. Let $c \in \mathbb{R}$ be a critical value and $\epsilon > 0$ be such that c is the only critical value in the closed interval $[c-\epsilon, c+\epsilon]$. Let $N_1, N_2, \ldots, N_k$ be the set of connected critical submanifolds with critical value c. Then the inclusion*

$$M^{c-\epsilon} \cup W^u(N_1) \cup W^u(N_2) \cup \cdots \cup W^u(N_k) \hookrightarrow M^{c+\epsilon}$$

is a homotopy equivalence (in fact a deformation retract).

This theorem has as its a consequence that a Morse–Bott function $f : M \longrightarrow \mathbb{R}$ defines defines a complex built out of the unstable manifolds associated to critical submanifolds, $W^u(N_i)$, that is homotopy equivalent to M. This is the analogue of the CW complex associated to a Morse function discussed in chapter 2. To make this complex precise we introduce the notion of a *disk bundle* complex.

Definition 15.5 *A finite disk bundle complex is a space X of the form*

$$X = \bigcup_i \zeta(K_i)$$

where this is a finite union of the total spaces of finite dimensional closed disk bundles

$$\zeta(K_i) \longrightarrow K_i$$

where K_i is a finite CW complex of dimension k_i and ζ_i is fiber dimension d_i. This union is required to satisfy the following property:

Let n_i be the total dimension, $n_i = k_i + d_i$. For $m \geq 0$, let

$$X^{(m)} = \bigcup_{n_i \leq m} \zeta(K_i)$$

Suppose $\zeta(K_{j_1}), \ldots, \zeta(K_{j_r})$ is the set of disk bundles of total dimension $n_{j_1} = \cdots = n_{j_r} = m+1$. Then there are attaching maps from the boundary sphere bundles

$$\phi_{j_i} : \partial(\zeta(K_{j_i})) \longrightarrow X^{(m)}$$

so that the inclusion $X^{(m)} \hookrightarrow X^{(m+1)}$ extends to a homeomorphism

$$X^{(m)} \cup \textstyle\bigcup_{\phi_{j_i}} \zeta(K_{j_i}) \xrightarrow{\cong} X^{(m+1)}$$

where the complex on the left is the disjoint union of $X^{(m)}$ with the disk bundles

$$\zeta(K_{j_1}), \ldots, \zeta(K_{j_r})$$

glued along the boundary sphere bundles via the maps ϕ_{j_i}.

Notice that a disk bundle complex with the property that each of the K_i's is a point is simply a finite CW complex.

Given a connected disk bundle complex X, notice that there is a natural filtration analogous to the filtration by skeleta of a CW complex:

$$X^{(0)} \hookrightarrow X^{(1)} \hookrightarrow \ldots \hookrightarrow X^{(m-1)} \hookrightarrow X^{(m)} \hookrightarrow \ldots \hookrightarrow X. \tag{15.1}$$

By construction this is a filtration of finite length. The subquotients of this filtration is a wedge of Thom-spaces:

$$X^{(m)}/X^{(m-1)} = \bigvee_{n_j=m} \zeta(K_j)/\partial\zeta(K_j).$$

In the example of a CW complex this subquotient is a wedge of spheres of dimension m. The associated cellular chain complex is replaced in the general disk bundle complex setting by the spectral sequence in homology associated to this filtration.

Theorem 15.6 *Given a disk bundle complex X as above, then filtration (15.1) induces a spectral sequence converging to the homology $H_*(X)$ with E_2 term*

$$E_2^{p,q} = \bigoplus_{n_j=p} H_{p+q}(\zeta(K_j); \partial\zeta(K_j))$$

converging to $H_{p+q}(X)$. Moreover, if X is an oriented disk bundle complex (that is, all of the disk bundles $\zeta(K_i) \longrightarrow K_i$ are oriented bundles), then the E_2 term of this spectral sequence is isomorphic to

$$E_2^{p,q} \cong \bigoplus_{n_j=p} H_{p+q-d_j}(K_j).$$

Proof: The spectral sequence itself is simply the spectral sequence in homology associated to filtration 15.1. The identification of the E_2-term of this spectral sequence comes from the identification of the subquotients in the filtration 15.1 of a disk bundle complex. The second statement in the theorem follows from the Thom isomorphism theorem. Notice then that no orientability requirements are needed if one considers homology with $\mathbb{Z}_2$-coefficients. □

We now apply this general theory to the context of a nondenerate function on a compact manifold.

Theorem 15.7 *Let $f : M \longrightarrow \mathbb{R}$ be a Morse–Bott function on M, a compact manifold. Let $N_1, \ldots, N_k$ be the critical submanifolds. Let $m_i = dim(N_i)$, $\lambda_i = index(N_i)$, and $n_i = m_i + \lambda_i$. Then M is homotopy equivalent to a disk bundle complex $X(f)$ with disk bundles $W^u(N_i) \longrightarrow N_i$ (or equivalently $\nu^-(N_i) \longrightarrow N_i$). In particular there is the associated filtration and spectral sequence converging to $H_*(M)$ as in theorem 15.6. When all the negative normal bundles are oriented, or if we take $\mathbb{Z}_2$-coefficients, the E_2 term of this spectral sequence is given by*

$$E_2^{p,q} = \bigoplus_{n_j=p} H_{p+q-\lambda_j}(N_j) = H_{m_j+q}(N_j)$$

converging to $H_{p+q}(M)$. When f is a Morse function so that all of N_j's are points, the E_2 chain complex is the Morse–Smale chain complex and so the spectral sequence collapses.

Proof: The disk bundle complex is constructed via Theorem 15.4 and induction (this is analogous to Theorem 5.7). The associated filtration and spectral sequence comes from Theorem 15.6. The relationship with the Morse–Smale chain complex follows from the fact that the filtration in the case of a Morse function is simply the cellular skeletal filtration whose associated chain complex is, by definition the Morse–Smale chain complex. Because of this analogy we refer to this spectral sequence as the *Morse–Smale* spectral sequence. We note that essentially this same spectral sequence appeared in the early work of Bott on degenerate functions [**?**]. □

We end this section by observing that the next natural question to ask is whether one can recover stronger information about the topology of M directly in terms of the topology of the critical submanifolds of a Morse–Bott function and in terms of the spaces of flows between them. For example, is the analogue of Theorem 12.1 true? Is there a category $\mathcal{C}_f$ whose objects are the critical points of f (topologized as a disjoint union of critical submanifolds) and whose morphisms are an appropriate compactification of the spaces of flows, so that its classifying space is homeomorphic to M? These and related questions are currently being investigated by M. Betz.

15.2 Equivariant Morse Functions

In this section we discuss an important class of examples of Morse–Bott functions. These occur when there is a smooth group action on a compact manifold,

$$G \times M \longrightarrow M$$

and a function $f : M \longrightarrow \mathbb{R}$ that is invariant under the action. That is,

$$f(gx) = f(x)$$

for all $g \in G$ and $x \in M$. In particular f defines a function on the orbit space

$$f : M/G \longrightarrow \mathbb{R}.$$

If G is a compact group and the action is free, then the orbit space M/G inherits a manifold structure and the projection

$$M \longrightarrow M/G$$

is a principal bundle. If the map on the orbit space $f : M/G \longrightarrow \mathbb{R}$ is a Morse function, or even more generally, a Morse–Bott function, then the map on the total space $f : M \longrightarrow \mathbb{R}$ will be Morse–Bott. (Recall that the action is *free* if it has no fixed points; that is, $gx = x$ if and only if $g = 1 \in G$.) Now even if

the action is not free, it still makes sense to consider Morse–Bott equivariant functions $f : M \longrightarrow \mathbb{R}$, and our goal in this section is to study how the general theory of Morse–Bott functions can be applied to recover information about the *equivariant* topology of M in terms of the equivariant topology of the critical submanifolds and the spaces of flows.

Let X be a space acted upon by a group G. The *equivariant (co)homology* of X is the homology of the homotopy orbit space:

$$H_*^G(X) = H_*(EG \times_G X)$$
$$H_G^*(X) = H^*(EG \times_G X).$$

Here, as usual, EG denotes a contractible space with a free G-action. The homotopy orbit space $EG \times_G X$ is the orbit space of the diagonal action of G on $EG \times X$. The reason for studying the homotopy orbit space rather than the honest orbit space X/G, is the following. If

$$h : X \longrightarrow Y$$

is a G-equivariant map that is a homotopy equivalence, then h induces a homotopy equivalence on the homotopy orbit space

$$EG \times_G X \xrightarrow{h \times 1} EG \times_G Y.$$

This is proved by studying the induced map of principal G-bundles

$$\begin{array}{ccc} EG \times X & \xrightarrow{1 \times h} & EG \times Y \\ \downarrow & & \downarrow \\ EG \times_G X & \xrightarrow{1 \times h} & EG \times_G Y \end{array}$$

and observing that $1 \times h$ is a homotopy equivalence on the total spaces, and a homeomorphism on the fibers. Hence it induces a homotopy equivalence of the base spaces. In this case we say that h is a *weak equivariant homotopy equivalence,* and this observation says that such equivalences preserve the homotopy type of the homotopy orbit spaces. Hence equivariant homology and cohomology is an invariant of the weak equivariant homotopy type.

Notice that the homology of the honest orbit space X/G is *not* an invariant of the weak equivariant homotopy type. Consider the following example.

Example 15.1 *Consider the unique map*

$$p : EG \longrightarrow *$$

where $*$ denotes the one point space with the trivial G-action. Obviously p is equivariant, and since EG is contractible it is a homotopy equivalence. The induced map on orbit spaces is the unique map

$$p : BG = EG/G \longrightarrow *$$

which is rarely a homotopy equivalence. Notice, however, that on the level of equivariant homology we see that

$$H_*^G(*) = H_*(BG).$$

Now let M be a compact manifold with a smooth action by a group G, and let $f : M \longrightarrow \mathbb{R}$ be a Morse–Bott, smooth function which is invariant under this group action. Notice that the space of critical points is a G-invariant subspace of M. That is, if $x \in M$ is a critical point and $g \in G$, then $gx \in M$ is also a critical point. Thus given any critical point a of f, the orbit a of the the action of G consists entirely of critical points. Such an orbit is of the form G/H where H is the isotropy subgroup of a. That is

$$H = \{g \in G : ga = a\}.$$

An interesting special case of an equivariant Morse–Bott function f is when the critical points of f consist of a disjoint union of isolated orbits: $G/H_1 \sqcup \cdots \sqcup G/H_k$. This is the equivariant analogue of a Morse function, where the critical points are isolated.

In the general setting of an equivariant Morse–Bott function $f : M \longrightarrow \mathbb{R}$ we may write the space of critical points of f as a disjoint union $N_1 \sqcup \cdots \sqcup N_k$ where each N_i is a G-invariant critical submanifold of M. Notice that if G is not a connected group then each N_i may itself have several connected components, each of which is a critical submanifold of M. Notice that since the elements of G act as diffeomorphisms, both of M and the submanifolds N_i, each of the connected components of N_i have the same dimension and the same index. As in the last section we write the dimension of N_i as m_i, the index as λ_i, and the sum $n_i = m_i + \lambda_i$. $\nu_i^- \longrightarrow N_i$ is the negative normal bundle.

Now consider the induced function

$$\bar{f} : EG \times M \xrightarrow{proj} M \xrightarrow{f} \mathbb{R}.$$

By the equivariance of f this descends to a well defined map on the homotopy orbit space

$$\bar{f} : EG \times_G M \longrightarrow \mathbb{R}.$$

This is still a Morse–Bott function, with critical submanifolds $EG \times_G N_i$. As in the last section this gives a filtration of the homotopy type of $EG \times_G M$ and we get the following spectral sequence.

Theorem 15.8 *There is a spectral sequence converging to the equivariant homology $H_*^G(M)$ with E_2 term given by*

$$E_2^{p,q} \cong \bigoplus_{n_j = p} H_{p+q}^G(\nu_j^-, \partial\nu_j^-)$$

converging to $H_{p+q}^G(M)$. If the negative normal bundles are all orientable (or if we take $\mathbb{Z}_2$ coefficients) the E_2 term is given by

$$E_2^{p,q} \cong \bigoplus_{n_j = p} H_{p+q-\lambda_j}^G(N_j).$$

Remark 15.1 *1. This theorem is an application of Theorem 15.7. However one must observe that although $\bar{f} : EG \times_G M \longrightarrow \mathbb{R}$ is indeed a Morse–Bott function, the space EG is often times infinite dimensional. This can be dealt with in one of two ways. First, one may typically filter EG by finite dimensional compact G-equivariant submanifolds*

$$* \hookrightarrow EG^{(1)} \hookrightarrow \ldots \hookrightarrow EG^{(k)} \hookrightarrow EG^{(k+1)} \hookrightarrow \ldots \hookrightarrow EG,$$

apply Theorem 15.7 to each $\bar{f} : EG^{(k)} \times_G M \longrightarrow \mathbb{R}$ and take the limit. The alternative is to verify that the disk bundle complex $X(f)$ can be actually made to be equivariantly homotopy equivalent to M, and apply $EG \times_G -$ to filtration 15.1.

2. When the critical points of f are isolated orbits $G/H_1, \ldots, G/H_k$, then notice that since

$$EG \times_G G/H \cong EG \times_H * = BH$$

we have that the E_2 term of the spectral sequence is

$$\bigoplus_{i=i}^{k} H_*(BH_i).$$

We end this chapter with an example calculation. Let $M = S^2$ be the unit sphere in $\mathbb{R}^3$, and let $G = S^1$ be the circle group which acts on S^2 by rotation around the axis through the north and south poles (see Figure 15.2). We call the north pole N and the south pole S. These are the only fixed points of this action. In Atiyah and Bott [**?**] the Poincare polynomials of $ES^1 \times_{S^1} S^2$ were computed using equivariant Morse theory and an implicit appeal to the Serre spectal sequence. We observe that an alternative way to compute the equivariant homology $H_*^{S^1}(S^2)$ is by using Theorem 15.8:

Theorem 15.9

$$\begin{aligned} H_n^{S^1}(S^2) &\cong H_n(\mathbb{CP}^\infty) \oplus H_{n-2}(\mathbb{CP}^\infty) \\ &\cong \begin{cases} \mathbb{Z} & \text{if } n = 0 \\ \mathbb{Z} \oplus \mathbb{Z} & \text{if } n > 0 \end{cases} \end{aligned}$$

Proof: Let $f : S^2 \longrightarrow \mathbb{R}$ be the height function. f is a Morse function with two critical points: N having index 2, and S having index 0. f is clearly invariant under the S^1 action. Thus

$$\bar{f} : ES^1 \times_{S^1} S^2 \longrightarrow \mathbb{R}$$

has two critical submanifolds:

1. $ES^1 \times_{S^1} N \cong BS^1 \simeq \mathbb{CP}^\infty$ and

Figure 15.2: S^2 with S^1 action as rotation around the z axis. As usual, height is the Morse–Bott function, which in this case has two critical points, each of which is a fixed point for the S^1 action.

2. $ES^1 \times_{S^1} S \cong BS^1 \simeq \mathbb{CP}^\infty$.

Thus by Theorem 15.8 there is a spectral sequence with

$$E_2^{p,q} \cong \begin{cases} H_q(\mathbb{CP}^\infty) & \text{if } p = 0 \\ H_{p+q-2}(\mathbb{CP}^\infty) = H_q(\mathbb{CP}^\infty) & \text{if } p = 2 \\ 0 & \text{otherwise} \end{cases}$$

which converges to $H^{S^1}_{p+q}(S^2)$. Thus all the nonzero elements in the E_2 term lie in even total degree $(p+q)$. Since the differentials all lower total degree by one, they must be all zero. Hence the spectral sequence collapses and we have

$$H^{S^1}_m(S^2) \cong \bigoplus_{p+q=m} E_2^{p,q} = H_m(\mathbb{CP}^\infty) \oplus H_{m-2}(\mathbb{CP}^\infty).$$

□

15.3 Transversality in Equivariant Morse theory

One major consideration we have ignored is whether or not we can perturb equivariant functions $f : M \longrightarrow \mathbb{R}$ to a Morse–Bott function, or whether such equivariant Morse–Bott functions even exist at all. This is, in general, a difficult question

[say more here about weeping and gnashing of teeth [Redo CJS for Morse–

Bott, possibly by using Francesco's perturbed Morse stuff

Chapter 16

Morse Theory on Hilbert Manifolds: The Palais–Smale Condition (C)

In this chapter we outline the generalization of Morse theory due to Palais and Smale [**?**][**?**] to Hilbert manifolds. The reader is referred to [**?**] for details. The authors are grateful to M. Sanders for preparing a summary of this work. This chapter is an expanded version of that summary.

Throughout this chapter we assume that M is a smooth, complete, Riemannian, Hilbert manifold with no boundary; that is there is a Hilbert space $\mathcal{H}$ so that every point in M has a neighborhood diffeomorphic to $\mathcal{H}$. Let

$$f : M \longrightarrow \mathbb{R}$$

be a smooth function and $p \in M$ a critical point. As in the compact manifold case one can define the Hessian

$$Hess_p(f) : T_pM \times T_pM \longrightarrow \mathbb{R}.$$

The critical point p is nondegenerate if $Hess_p(f)$ is nonsingular. Equivalently this means that the adjoint map

$$A(f) : T_pM \longrightarrow T_pM$$

defined by

$$\langle A(f)(x), y\rangle = Hess_p(f)(x)(y)$$

is an isomorphism. In this case the eigenvalues of $A(f)$ are bounded away from zero, and the *index* of p is defined to be the supremum of the dimensions of the subspaces on which $A(f)$ is negative definite. (Notice that the index may very well be infinite.) The *coindex* of p is the supremum of the dimensions of the subspaces on which $A(f)$ is positive definite. $f : M \longrightarrow \mathbb{R}$ is a *Morse* function

if all of its critical points are nondegenerate. The following version of the Morse lemma was proved in [**?**][**?**].

Theorem 16.1 *Let $\mathcal{H}$ be a Hilbert space and $f : \mathcal{H} \longrightarrow \mathbb{R}$ a smooth function satisfying*

1. *$f(0) = 0$,*

2. *0 is a nondegenerate critical point of f.*

Then there is an origin-preserving smooth diffeomorphism ϕ of a neighborhood of the origin into $\mathcal{H}$ so that

$$f(\phi(v)) = |Pv|^2 - |(1-P)(v)|^2$$

where P is an orthogonal projection in $\mathcal{H}$.

Let $f : M \longrightarrow \mathbb{R}$ be a Morse function on a Hilbert manifold. This theorem describes the local dynamics of f near critical points. However in this situation as in previous ones, our goal is to recover as much about the topology of M as is possible from the critical points and the topology of the flows between them. Among the main difficulties in the infinite dimensional situation include the fact that the index of a critical point may be infinite, and also the fact that the flow line through a point $x \in M$, that is maximal solution curves to the flow equations

$$\frac{d\gamma}{dt} + \nabla_\gamma(f) = 0$$

with the initial condition

$$\gamma(0) = x,$$

may not be defined on the entire real line. Moreover it may not begin and end at critical points.

Exercise 16.1 *Review the arguments in chapters 4 and 5 that prove these assertions in the finite dimensional setting and see where compactness is used.*

This problem strongly affects the argument used to describe the homotopy type of M in terms of a cell complex with cells corresponding to the critical points. In order to recover this type of theorem in the general Hilbert manifold setting one must make certain further assumptions about the gradient vector field $\nabla(f)$.

Definition 16.2 *A smooth vector field X on a Hilbert manifold M is* strongly transverse *to a function $f : M \longrightarrow \mathbb{R}$ on a closed interval $[a, b]$, if for some $\delta > 0$ the following two conditions hold for $V = f^{-1}(a - \delta, b + \delta)$:*

1. *X(f) is nonvanishing on V*

2. *If $x \in V$ and γ_x is the maximal solution curve of the equation*

$$\frac{d\gamma}{dt} + X_\gamma = 0$$

subject to the initial condition

$$\gamma_x(0) = x,$$

then $\gamma_x(t)$ is defined and not in V for some $t > 0$ and also for some $t < 0$.

Observation 16.3 *If $f : M \longrightarrow \mathbb{R}$ is a smooth function on a* compact *manifold M, then the gradient vector field $\nabla(f)$ is strongly transverse to f on any closed interval $[a, b]$ which contains no critical values.*

In the presence of strong transversality, deforming along flow lines works as it does in the finite dimensional setting and one obtains the following analogue of the *regular interval theorem.* (Compare Theorem 5.1; see [**?**] for details.)

Theorem 16.4 *Let M be a Hilbert manifold without boundary and let $f : M \longrightarrow \mathbb{R}$ be a smooth function. Suppose there exists a smooth vector field X which is strongly transverse to f on the interval $[a, b]$. Then $N = f^{-1}(a)$ is a closed submanifold of M and for some $\delta > 0$ there is a diffeomorphism onto its image*

$$F : N \times (a - \delta, b + \delta) \longrightarrow M$$

that maps $N \times \{c\}$ diffeomorphically onto $f^{-1}(c)$ for all $c \in (a - \delta, b + \delta)$. In particular F restricts to give a diffeomorphism

$$F : N \times [a, b] \xrightarrow{\cong} f^{-1}[a, b].$$

Corollary 16.5 *If $f : M \longrightarrow \mathbb{R}$ and X is a vector field on M strongly transverse to f on $[a, b]$ as in Theorem 16.4, then there is a diffeomorphism*

$$F : M^a = f^{-1}(-\infty, a] \xrightarrow{\cong} f^{-1}(-\infty, b] = M^b.$$

When M is compact, the gradient vector field $\nabla(f)$ is strongly transverse to f on intervals containing no critical values, and it was this property that was essential in the proof of the regular neighborhood theorem (Theorem 5.1). The proof of Theorem 16.4 proceeds similarly. Now in order to insure that a function $f : M \longrightarrow \mathbb{R}$ on an infinite dimensional Hilbert manifold satisfies this strong transversality condition, a somewhat more verifiable condition is assumed. This is the Palais–Smale condition (C).

Theorem 16.6 (Palais–Smale Condition (C)) *Let S be any subset of M satisfying the following conditions:*

1. *f is bounded on S, and*

2. *$|\nabla(f)|$ gets arbitrarily close to zero on S.*

Then there is a critical point of f in the closure of S.

Again, we remark that condition (C) is obviously satisfied if M is a compact manifold. In the general case of Hilbert manifolds, condition (C) is used in a straightforward way to prove the following results, which describe how condition (C) yields the strong transversality of $\nabla(f)$.

Proposition 16.7 *If $f : M \longrightarrow \mathbb{R}$ satisfies condition (C), then for any $a < b \in \mathbb{R}$, there are at most finitely many critical points with critical values in $[a, b]$. In particular the critical values of f are isolated.*

Proposition 16.8 *Let $f : M \longrightarrow \mathbb{R}$ satisfy condition (C). Then if $\gamma : (\alpha, \beta) \longrightarrow M$ is a flow line (i.e. a maximal solution curve to the flow equation) then one of the following hold:*

1. *$\lim_{t\to\beta} f(\gamma(t)) = \infty$ or*

2. *$\beta = \infty$ and $\lim_{t\to\infty} \gamma(t)$ exists and is a critical point of f.*

A similar statement holds for the limit as t tends to α.

Corollary 16.9 *Propositions 16.7 and 16.8 imply that if f has no critical values in a closed interval $[a, b]$, then the gradient vector field $\nabla(f)$ is strongly transverse to f in this interval. In particular this implies that*

$$M^a \cong M^b.$$

Thus like in the compact manifold setting if $f : M \longrightarrow \mathbb{R}$ satisfies condition (C) then the topology of the level sets does not change between critical values. The following theorem describes how the homotopy type changes when one passes a critical value. This is the analogue of Theorem 5.6. We refer the reader to [**?**] for a proof.

Theorem 16.10 *Let $f : M \longrightarrow \mathbb{R}$ be a Morse function satisfying condition (C). Let c be a critical value of f, $a_1, \ldots, a_n$ the critical points at level c of which $a_1, \ldots, a_p$ are those with finite indices, say $\lambda_1, \ldots, \lambda_p$. As before, let*

$$W^u(a_i) = \{x \in M : \lim_{t\to-\infty} \gamma_x(t) \text{ exists and equals } a_i\}$$

be the unstable manifold of the critical point a_i. We then have the following.

1. *$W^u(a_i)$ is diffeomorphic to the disk D^{λ_i}, and*

2. *The inclusion*

$$M^a \cup D^{\lambda_1} \cup \cdots \cup D^{\lambda_p} \cong M^a \cup W^u(a_1) \cup \cdots \cup W^u(a_p) \hookrightarrow M^b$$

is a deformation retract, and in particular, a homotopy equivalence.

Corollary 16.11 *If $f : M \longrightarrow \mathbb{R}$ is a Morse function satisfying condition (C), then M is homotopy equivalent to a CW complex $X(f)$ having one cell in dimension $\lambda < \infty$ for every critical point of index λ.*

Remark 16.1 *1. The argument used to prove Corollary 16.11 from Theorem 16.10 is the same as in the finite dimensional (compact) case.*

2. By taking the cellular chain complex associated to the CW complex $X(f)$ as in Corollary 16.11, one obtains the analogue of the Morse–Smale complex for computing $H_*(M)$.

3. An interesting feature about these results is that the critical points of infinite index do not affect the homotopy type of the manifold. This is because the unstable manifold of such a critical point is an infinite dimensional disk with boundary an infinite dimensional sphere, which is contractible. Therefore in the cell attaching procedure the attaching map is null homotopic (any map from a contractible space is null homotopic). But attaching a contractible space via a null homotopic map does not change the homotopy type. Hence from the point of view of homotopy type, one may ignore the cells corresponding to infinite index critical points.

The next natural question would be to see if one could construct a category corresponding to a Morse function $f : M \longrightarrow \mathbb{R}$ satisfying condition (C) in such a way that its classifying space is homeomorphic to M. One obvious difficulty in doing so is how to account for flows that do not begin (or end) at critical points. This question and its applications are currently being investigated by M. Sanders.

We end by considering a classical example, studied in detail in [**?**] and in [**?**]. Let M^n be a closed, n - dimensional Riemannian manifold, and for p and q points in M, let

$$\Omega(M; p, q) = \{\alpha : I \longrightarrow M : \alpha(0) = p, \text{ and } \alpha(1) = q\}.$$

Here I is the closed interval $I = [0, 1]$, and the strict condition for $\alpha \in \Omega(M; p, q)$ is that it be absolutely continuous with square integrable first derivative. We refer the reader to [**?**] for details.

$\Omega(M^n; p, q)$ is a Hilbert manifold, modelled on the path space $(\mathbb{R}^n)^I$ which is a Hilbert space under the inner product

$$\langle \sigma, \rho \rangle = \int_0^1 \langle \sigma(t), \rho(t) \rangle dt.$$

The *action* or *energy* functional

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

is defined by

$$E(\phi) = \frac{1}{2} \int_0^1 |\frac{d\phi}{dt}|^2 dt.$$

This functional is of classical interest in geometry and the calculus of variations, in part because of the following theorem.

Theorem 16.12 *$\phi \in \Omega(M; p, q)$ is a critical point of the energy functional E if and only if ϕ is a geodesic parameterized proportionally to arclength. Furthermore, if the geodesic distance from p to q is d, then E takes on its minimum d^2 precisely on the set of minimal geodesics from p to q.*

We refer the reader to [[**?**], chapter III] for a proof of this theorem. We remark that the study of geodesics via the energy functional was one of the original motivations for Morse to develop his theory [**?**]. Milnor gives a full description of the relevant Morse theory in this setting. He uses finite dimensional approximations to the infinite dimensional manifold $\Omega(M; p, q)$ in order to use classical Morse theory to study its homotopy type. However Palais and Smale prove that the energy E satisfies condition (C) and hence their more general theory applies directly.

Part V

Morse field theory

Part VI

Role of S^1

Part VII

Miscellaneous

Chapter 17

Connections, Curvature, and the Yang–Mills Functional

During the last ten to fifteen years perhaps the area of the strongest impact of the techniques and ideas of Morse theory has been in Gauge theory. From the geometric and topological point of view, (as distinct from the physics viewpoint) this is the study of spaces of connections on a principal bundle over a finite dimensional (usually low dimensional) manifold. Certain functionals are defined on these spaces and their critical points (or critical submanifolds) often have intrinsic geometric interest. In particular several homotopy theoretic invariants of these critical spaces have been shown to yield invariants of the differential topological structure of the underlying manifold. Donaldson and others have achieved dramatic success by using these invariants to understand the differential topology of four dimensional manifolds. Floer has applied some of the homological constructions of Morse theory to obtain invariants of three dimensional manifolds.

In this chapter we outline some of the basic constructions of this theory. In later chapters we will discuss several examples of applications of pertaining to Morse theory, and in particular discuss the relationship between our classifying space constructions for compact manifolds and Floer's theory.

17.1 Connections and their Curvature

Let G be a compact, simply connected Lie group. Recall that the tangent bundle of any Lie group has a canonical trivialization

$$\begin{array}{lll} \psi : G \times T_eG & \xrightarrow{\cong} & TG \\ (g, v) & \longrightarrow D(\ell_g)(v) & \end{array}$$

where for $g \in G$, $\ell_g : G \longrightarrow G$ is the map given by left multiplication by g, and $D(\ell_g) : T_hG \longrightarrow T_{gh}G$ is the differential. r_g and $D(r_g)$ will denote the analogous maps corresponding to right multiplication.

The differential of right multiplication on G defines a right action of G on the tangent bundle TG. The trivialization ψ is equivariant with respect to this action, if we take as the right action of G on T_eG to be the *adjoint* action:

$$\begin{aligned} T_eG \times G &\longrightarrow T_eG \\ (v, g) &\longrightarrow D(\ell_{g^{-1}})(v)D(r_g). \end{aligned}$$

Under the identification of T_eG with the Lie algebra $\mathfrak{g}$ this is the usual adjoint action. Now let

$$G \longrightarrow P \xrightarrow{p} M$$

be a principal right G-bundle over a finite dimensional, Riemannian manifold M. So in particular G acts freely on the right of P and $M = P/G$. This adjoint representation of G on $\mathfrak{g}$ defines an induced vector bundle with fiber $\mathfrak{g} = T_eG$:

$$ad(P) : P \times_G \mathfrak{g} \longrightarrow M.$$

This bundle has the following relevance. Let $p^*(TM) \longrightarrow P$ be the pull-back bundle over P of the tangent bundle of M. We have a surjective map of bundles

$$TP \longrightarrow p^*(TM).$$

Define T_FP to be the kernel bundle of this map. That is, T_FP consists of those tangent vectors which are tangent to the fibers. Notice that the action of G on P defines an action of G on the tangent bundle TP, which restricts to an action of G on T_FP. Furthermore, by recognizing that the fibers are equivariantly homeomorphic to the Lie group G, the following is a direct consequence of the above considerations.

Proposition 17.1 *T_FP is naturally isomorphic to the pull-back of the adjoint bundle,*

$$T_FP \cong p^*(ad(P)).$$

Thus we have an exact sequence of G-equivariant vector bundles over P:

$$0 \longrightarrow p^*(ad(P)) \longrightarrow TP \xrightarrow{p_*} p^*(TM) \longrightarrow 0 \tag{17.1}$$

Definition 17.2 *A* connection *on the principal bundle P is an equivariant splitting*

$$\omega_A : TP \longrightarrow p^*(ad(P))$$

of the above exact sequence of vector bundles. That is, ω_A defines a G-equivariant isomorphism

$$\omega_A \oplus p_* : TP \longrightarrow p^*(ad(P)) \oplus p^*(TM).$$

The following is an important description of the space of connections on P, $\mathcal{A}(P)$.

Proposition 17.3 *The space of connections on the principal bundle P, $\mathcal{A}(P)$, is an affine space modelled on the infinite dimensional vector space of one-forms on M with coefficients in the bundle $ad(P)$, $\Omega^1(M; ad(P))$.*

Remark 17.1 *Recall that given a bundle over M, $\zeta \longrightarrow M$, the space of one-forms with coefficients in ζ, $\Omega^1(M;\zeta)$ is the space of sections of the tensor product bundle $T^*(M) \otimes \zeta$, where $T^*(M)$ is the cotangent bundle. This tensor product bundle is isomorphic to the bundle $Hom(T(M), \zeta)$ over M, the bundle whose fiber at $x \in M$ is the vector space of homomorphisms (linear transformations) from the tangent space T_xM to the fiber of ζ at x, ζ_x.*

Proof: Consider two connections $(\omega_A)_1$ and $(\omega_A)_2$,

$$(\omega_A)_1, (\omega_A)_2 : TP \longrightarrow p^*(ad(P)).$$

Since these are splittings of the exact sequence 17.1, they are both the identity when restricted to $p^*(ad(P)) \hookrightarrow TP$. Thus their difference $(\omega_A)_1 - (\omega_A)_2$ is zero when restricted to $p^*(ad(P)$. By the exact sequence it therefore factors as a composition

$$(\omega_A)_1 - (\omega_A)_2 : TP \longrightarrow p^*(TM) \xrightarrow{\alpha} p^*(ad(P))$$

for some bundle homomorphism $\alpha : p^*(TM) \longrightarrow p^*(ad(P))$. That is, for every $v \in P$, α defines a linear transformation

$$\alpha_v : p^*(TM)_v \longrightarrow p^*(ad(P))_v.$$

Now $p^*(TM)_v \cong T_{p(v)}M$ and $p^*(ad(P))_v \cong ad(P)_{p(v)}$. Hence for every $v \in P$, α defines (and is defined by) a linear transformation

$$\alpha_v : T_{p(v)}M \longrightarrow ad(P)_{p(v)}.$$

Furthermore, the fact that both $(\omega_A)_1$ and $(\omega_A)_2$ are *equivariant* splittings says that $(\omega_A)_1 - (\omega_A)_2$ is equivariant, which translates to the fact that α_v only depends on the orbit of v. That is,

$$\alpha_v = \alpha_{vg} : T_{p(v)}M \longrightarrow ad(P)_{p(v)}$$

for every $g \in G$. Thus α_v only depends on $p(v) \in M$. Hence for every $x \in M$, α defines, and is defined by, a linear transformation

$$\alpha_x : T_xM \longrightarrow ad(P)_x.$$

Thus α may be viewed as a section of the bundle of homomorphisms, $Hom(TM, ad(P))$, which, as remarked above, is a one-form,

$$\alpha \in \Omega^1(M; ad(P)).$$

Thus any two connections on P differ by an element in $\Omega^1(M; ad(P))$ in this sense.

Now reversing the procedure, any $\beta \in \Omega^1(M; ad(P))$ defines an equivariant homomorphism of bundles over P,

$$\beta : p^*(TM) \longrightarrow p^*(ad(P)).$$

By adding the composition

$$TP \longrightarrow p^*(TM) \xrightarrow{\beta} p^*(ad(P))$$

to any connection (equivariant splitting)

$$\omega_A : TP \longrightarrow p^*(ad(P))$$

one produces a new equivariant splitting of TP, and hence a new connection. The proposition follows. □

Remark 17.2 *Even though the space of connections $\mathcal{A}(P)$ is affine, it is not, in general a vector space. There is no "zero" in $\mathcal{A}(P)$ since there is no pre-chosen, canonical connection. The one exception to this, of course is when P is the trivial bundle*

$$P = M \times G \longrightarrow M.$$

In this case there is an obvious equivariant splitting of TP, which serves as the "zero" in $\mathcal{A}(P)$. Moreover in this case the adjoint bundle $ad(P)$ is also trivial,

$$ad(P) = M \times \mathfrak{g} \longrightarrow M.$$

Hence there is a canonical identification of the space of connections on the trivial bundle with $\Omega^1(M; \mathfrak{g}) = \Omega^1(M) \otimes \mathfrak{g}$, the space of Lie algebra valued one-forms on M. If one views the Lie algebra $\mathfrak{g}$ as a space of matrices, then a connection on the trivial bundle over M can be viewed as a matrix of one-forms on M.

Let $P \longrightarrow M$ be a principal G-bundle and let $\omega_A \in \mathcal{A}(P)$ be a connection. The *curvature* F_A is a two-form

$$F_A \in \Omega^2(M; ad(P))$$

which measures to what extent the splitting ω_A commutes with the bracket operation on vector fields. More precisely, let X and Y be vector fields on M. The connection ω_A defines an equivariant splitting of TP and hence defines a "horizontal" lifting of these vector fields, which we denote by $\tilde{X}$ and $\tilde{Y}$ respectively. One then defines

$$F_A(X, Y) = \omega_A[\tilde{X}, \tilde{Y}]$$

which is a section of $ad(P)$. The *covariant derivative* induced by the connection ω_A can also be defined in terms of the Lie bracket on $ad(P)$.

$$D_A : \Omega^0(M; ad(P)) \longrightarrow \Omega^1(M; ad(P))$$

is defined by

$$D_A(s)(X) = [\tilde{X}, s]$$

where X is a vector field on M.

The notion of covariant derivative, and hence connection extends to vector bundles as well. Let $\zeta \longrightarrow M$ be a finite dimensional vector field over M. A connection on ζ is a linear transformation

$$D_A : \Omega^0(M;\zeta) \longrightarrow \Omega^1(M;\zeta)$$

that satisfies the *Leibnitz rule*

$$D_A(f\phi) = df \otimes \phi + fD_A(\phi)$$

for any $f \in C^\infty(M)$ and any $\phi \in \Omega^0(M;\zeta)$.

Given two connections $(D_A)_1$ and $(D_A)_2$ on ζ and a function $f \in C^\infty(M)$ one can take the convex combination

$$f(D_A)_1 + (1-f)(D_A)_2$$

and obtain a new connection. From this it is not difficult to see that the space of connections on ζ is affine modelled on the vector space of one-forms $\Omega^1(M; End(\zeta))$, where $End(\zeta)$ is the bundle of endomorphisms of ζ.

Now let X be a vector field on M and D_A a connection on the vector bundle $\zeta \longrightarrow M$. The covariant derivative in the direction of X, which we denote by $(D_A)_X$ is an operator on the space of sections of ζ,

$$(D_A)_X : \Omega^0(M;\zeta) \longrightarrow \Omega^0(M;\zeta)$$

defined by

$$(D_A)_X(\phi) = \langle D_A(\phi); X\rangle.$$

One can then define the curvature $F_A \in \Omega^2(M; End(\zeta))$ by defining its action on a pair of vector fields X and Y to be

$$F_A(X,Y) = (D_A)_X(D_A)_Y - (D_A)_Y(D_A)_X - (D_A)_{[X,Y]}. \tag{17.2}$$

To interpret this formula notice that a priori $F_A(X,Y)$ is a second order differential operator on the space of sections of ζ. However, a direct calculation shows that for $f \in C^\infty(M)$ and $\phi \in \Omega^0(M;\zeta)$ then

$$F_A(X,Y)(f\phi) = fF_A(X,Y)(\phi)$$

and hence that $F_A(X,Y)$ is in fact a zero-order operator on $\Omega^0(M;\zeta)$. But a zero order operator on the space of sections of ζ is a section of the endomorphism bundle $End(\zeta)$. Thus F_A assigns to any pair of vector fields X and Y a section of $End(\zeta)$. Moreover it is straightforward to check that this assignment is tensorial in X and Y, (i.e. $F_A(fX,Y) = F_A(X,fY) = fF_A(X,Y)$). Thus F_A is an element of $\Omega^2(M; End(\zeta))$. The curvature measures the lack of commutativity in second order partial covariant derivatives.

Given a connection on a bundle $\zeta \longrightarrow M$, the linear mapping $D_A : \Omega^0(M;\zeta) \longrightarrow \Omega^1(M;\zeta)$ extends to a *de Rham* type sequence

$$\Omega^0(M;\zeta) \xrightarrow{D_A} \Omega^1(M;\zeta) \xrightarrow{D_A} \Omega^2(M;\zeta) \xrightarrow{D_A} \dots \tag{17.3}$$

where for $\phi \in \Omega^p(M;\zeta)$, $D_A(\phi)$ is the $p+1$-form defined by the formula

$$\begin{aligned} D_A(\phi)(X_0,\dots,X_p) = &\sum_{j=0}^{p}(-1)^j(D_A)_{X_j}(\phi(X_0,\dots,\hat{X}_j,\dots,X_p)) \\ &+\sum_{i<j}(-1)^{i+j}\phi([X_i,X_j],X_0,\dots,\hat{X}_i,\dots,\hat{X}_j,\dots,X_p). \end{aligned}$$

It is not generally true that $D_A \circ D_A = 0$. In fact we have the following

Proposition 17.4

$$D_A \circ D_A = F_A : \Omega^0(M;\zeta) \longrightarrow \Omega^2(M;\zeta)$$

where in this context the curvature F_A is interpreted as assigning to a section $\phi \in \Omega^0(M;\zeta)$ the two-form $F_A(\phi)$ which associates to vector fields X and Y the section $F_A(X,Y)(\phi)$ as defined in 17.2.

Proof: This is a straightforward verification using formulae 17.2 and 17.3 □.

Thus the curvature of a connection F_A can also be viewed as measuring the extent to which the covariant derivatives D_A fail to form a cochain complex on the space of differential forms with coefficients in the bundle ζ. However it is always true that the covariant derivative of the curvature tensor is zero. This is known as the *Bianchi identity*:

Theorem 17.5 *Let A be a connection on a vector bundle $\zeta \longrightarrow M$. Then*

$$D_A F_A = 0.$$

We end this section by observing that if $P \longrightarrow M$ is a principal G-bundle, with a connection ω_A, then any representation of G on a finite dimensional vector space V induces a connection on the corresponding vector bundle

$$P \times_G V \longrightarrow M.$$

We refer the reader to [**?**] and [**?**] for thorough discussions of the various ways of viewing connections. [**?**] has a nice, brief discussion of connections on principal bundles, and [**?**] and [**?**] have similarly concise discussions of connections on vector bundles.

17.2 The Gauge Group and its Classifying space

Let A be a connection on a principal bundle $P \longrightarrow M$ where M is a closed manifold equipped with a Riemannian metric. The Yang–Mills functional applied to A, $\mathcal{YM}(A)$ is the L^2 norm of the curvature,

$$\mathcal{YM}(A) = \frac{1}{2}\int_M |F_A|^2 \, d(vol).$$

We view $\mathcal{YM}$ as a mapping $\mathcal{YM} : \mathcal{A}(P) \longrightarrow \mathbb{R}$. Eventually $\mathcal{YM}$ will be treated as a Morse function, but first we investigate the group of symmetries of $\mathcal{A}$ that $\mathcal{YM}$ preserves. This is known as the *gauge group*, or *group of gauge transformations*, of P.

Definition 17.6 *The gauge group $\mathcal{G}(P)$ of the principal bundle P is the group of bundle automorphisms of $P \longrightarrow M$. That is, an element $g \in \mathcal{G}(P)$ is a bundle isomorphism of P with itself lying over the identity:*

$$\begin{array}{ccc} P & \xrightarrow[\cong]{\phi} & P \\ \downarrow & & \downarrow \\ M & \xrightarrow{=} & M. \end{array}$$

Equivalently, $\mathcal{G}(P)$ is the group $\mathcal{G}(P) = Aut_G(P)$ of G-equivariant diffeomorphisms of the space P.

The gauge group $\mathcal{G}(P)$ can be thought of in several equivalent ways. The following one is particularly useful.

Consider the conjugation action of the Lie group G on itself,

$$\begin{aligned} G \times G &\longrightarrow G \\ (g, h) &\longrightarrow ghg^{-1}. \end{aligned}$$

This left action defines a fiber bundle

$$Ad(P) = P \times_G G \longrightarrow P/G = M$$

with fiber G. We leave the following as an exercise for the reader.

Proposition 17.7 *The gauge group of a principal bundle $P \longrightarrow M$ is naturally isomorphic (as topological groups) the the group of sections of $Ad(P)$, $C^\infty(M; Ad(P))$.*

The gauge group $\mathcal{G}(P)$ acts on the space of connections $\mathcal{A}(P)$ by the pull-back construction. More generally, if $f : P \longrightarrow Q$ is any smooth map of principal G-bundles and A is a connection on Q, then there is a natural pull back connection $f^*(A)$ on Q, defined by pulling back the equivariant splitting of TQ to an equivariant splitting of TP in the obvious way. The pull-back construction for automorphisms $\phi : P \longrightarrow P$ defines an action of $\mathcal{G}(P)$ on $\mathcal{A}(P)$. The following is an exercise involving the definitions of the constructions of this chapter.

Proposition 17.8 *Let P be the trivial bundle $M \times G \longrightarrow M$. Then the gauge group $\mathcal{G}(P)$ is given by the function space from M to G,*

$$\mathcal{G}(P) \cong C^\infty(M; G).$$

Furthermore if $\phi : M \longrightarrow G$ is identified with an element of $\mathcal{G}(P)$, and $A \in \Omega^1(M; \mathfrak{g})$ is identified with an element of $\mathcal{A}(G)$, then the induced action of ϕ on G is given by

$$\phi^*(A) = \phi^{-1} A \phi + \phi^{-1} d\phi.$$

It is not difficult to see that in general the gauge group $\mathcal{G}(P)$ does not act freely on the space of connections $\mathcal{A}(P)$. However there is an important subgroup $\mathcal{G}_0(P) < \mathcal{G}(P)$ that does. This is the group of *based* gauge transformations. To define this group, let $x_0 \in M$ be a fixed basepoint, and let P_{x_0} be the fiber of P at x_0.

Definition 17.9 *The based gauge group $\mathcal{G}_0(P)$ is a subgroup of the group of bundle automorphisms $\mathcal{G}(P)$ which pointwise fix the fiber P_{x_0}. That is,*

$$\mathcal{G}_0(P) = \{\phi \in \mathcal{G}(P) : if v \in P_{x_0} then \, \phi(v) = v\}.$$

Theorem 17.10 *The based gauge group $\mathcal{G}_0(P)$ acts freely on the space of connections $\mathcal{A}(P)$.*

Proof: Suppose that $A \in \mathcal{A}(P)$ is a fixed point of $\phi \in \mathcal{G}_0(P)$. That is, $\phi^*(A) = A$. We need to show that $\phi = 1$.

The equivariant splitting ω_A given by a connection A defines a notion of parallel transport in P along curves in M (See [**?**] or [**?**].) It is not difficult to see that the statement $\phi^*(A) = A$ implies that application of the automorphism ϕ commutes with parallel transport. Now let $w \in P_x$ be a point in the fiber of an element $x \in M$. Given curve γ in M between the basepoint x_0 and x one sees that

$$\phi(w) = T_\gamma(\phi(T_{\gamma^{-1}}(w))$$

where $T\gamma$ is parallel transport along γ. But since $T_{\gamma^{-1}}(w) \in P_{x_0}$ and $\phi \in \mathcal{G}_0(P)$,

$$\phi(T_{\gamma^{-1}}(w)) = w.$$

Hence $\phi(w) = w$, that is, $\phi = 1$. □

Remark 17.3 *Notice that this argument actually says that if $A \in \mathcal{A}(P)$ is the fixed point of* any *gauge transformation $\phi \in \mathcal{G}(P)$, then ϕ is determined by its action on a single fiber.*

Let $\mathcal{B}(P)$ and $\mathcal{B}_0(P)$ be the orbit spaces of connections on P up to gauge and based gauge equivalence respectively,

$$\mathcal{B}(P) = \mathcal{A}(P)/\mathcal{G}(P) \quad \mathcal{B}_0(P) = \mathcal{A}(P)/\mathcal{G}_0(P).$$

Now it is straightforward to check directly that the Yang–Mills functional in invariant under gauge transformations. Thus it yields maps

$$\mathcal{YM} : \mathcal{B}(P) \longrightarrow \mathbb{R} \quad \text{and} \quad \mathcal{YM} : \mathcal{B}_0(P) \longrightarrow \mathbb{R}.$$

It is therefore important to understand the homotopy types of these orbit spaces. Because of the freeness of the action of $\mathcal{G}_0(P)$, the homotopy type of the orbit space $\mathcal{G}_0(P)$ is easier to understand. We end this section with a discussion of its homotopy type. Now since the space of connections $\mathcal{A}(P)$ is affine, it is contractible. Thus by the above theorem $\mathcal{B}_0(P) = \mathcal{A}(P)/\mathcal{G}_0(P)$ is the classifying space of the based gauge group,

$$\mathcal{B}_0(P) = B\mathcal{G}_0(P).$$

But the classifying spaces of the gauge groups are relatively easy to understand. The following describes their homotopy types (See [**?**].)

Theorem 17.11 *Let $G \longrightarrow EG \longrightarrow BG$ be a universal principal bundle for the Lie group G (so that EG is contractible). Let $y_0 \in BG$ be a fixed basepoint. Then there are homotopy equivalences*

$$B\mathcal{G}(P) \simeq Map^P(M, BG) \quad \textit{and} \quad \mathcal{B}_0(P) \simeq B\mathcal{G}_0(P) \simeq Map_0^P(M, BG)$$

where $Map(M, BG)$ is the space of all continuous maps from M to BG and $Map_0(M, BG)$ is the space of those maps that preserve the basepoints. The superscript P denotes the path component of these mapping spaces consisting of the homotopy class of maps that classify the principal G-bundle P.

Proof: Consider the space of all G-equivariant maps from P to EG, $Map^G(P, EG)$. The gauge group $\mathcal{G}(P) \cong Aut^G(P)$ acts freely on the left of this space by composition. It is easy to see that $Map^G(P, EG)$ is contractible, and its orbit space is given by the space of maps from the G-orbit space of P $(= M)$ to the G-orbit space of EG $(=$ BG$)$,

$$Map^G(P, EG)/\mathcal{G}(P) \cong Map^P(M, BG).$$

This proves that $Map(M, BG) = B\mathcal{G}(P)$. Similarly $Map_0^G(P, EG)$, the space of G-equivariant maps that send the fiber P_{x_0} to the fiber EG_{y_0}, is a contractible space with a free $\mathcal{G}_0(P)$ action, whose orbit space is $Map_0^P(M, BG)$. Hence $Map_0^P(M, BG) = B\mathcal{G}_0(P)$. □

17.3 The Critical Points of the Yang–Mills Functional

In this section we derive the *Yang–Mills* equations. These are the variational equations corresponding to the Yang–Mills functional

$$\mathcal{YM} : \mathcal{B}(P) \longrightarrow \mathbb{R}$$

where $P \longrightarrow M$ is a principal G-bundle over a Riemannian manifold M. That is, we derive the equations determining when a connection A on P is a critical point of $\mathcal{YM}$. We will then restrict to the important case when M is a compact 4-dimensional manifold and identify the space of absolute minima of $\mathcal{YM}$. We refer the reader to [**?**] for details and background for the arguments presented below.

Given a Riemannian metric on an n -manifold M^n and a vector bundle $\zeta \longrightarrow M$, let

$$* : \Omega^p(M;\zeta) \longrightarrow \Omega^{n-p}(M;\zeta)$$

be the Hodge star operator. Locally the operator can be described as follows. Let $I = (i_1, \ldots, i_p)$ be an increasing sequence of p-integers between 1 and n and let $dx_I = dx_{i_1} \wedge \cdots \wedge dx_{i_p}$. Then

$$*dx_I = dx_{n-I}$$

where $n - I$ is the ordered sequence of the $n - p$ integers that do not appear in I. The metric on M and on the bundle ζ is then used to extend this definition to globally defined forms.

Theorem 17.12 *A connection A on P is a critical point of the Yang–Mills functional $\mathcal{YM}$ if and only if it satisfies the Yang–Mills equations:*

$$D_A(*F_A) = 0.$$

where, like above, D_A is the covariant derivative induced by A, and F_A is the curvature two-form.

To prove this result we use the following straightforward calculation. Let A be a connection on P and $\eta \in \Omega^1(M; ad(P))$. Then by the results in section 1 imply that for $t \in \mathbb{R}$ we can define a new connection $A_t(\eta) = A + t\eta$.

Lemma 17.13 *The curvature of $A_t(\eta)$ is given by*

$$F_{A_t(\eta)} = F_A + tD_A\eta + \frac{1}{2}t^2[\eta,\eta] \in \Omega^2(M; ad(P)).$$

We now prove Theorem 17.12.

Proof: For forms α, $\beta \in \Omega^p(M; ad(P))$, write

$$\langle \alpha, \beta \rangle = \int_M \alpha \wedge *\beta \, d(vol).$$

So if A is a connection on P we have that the Yang–Mills functional is given by

$$\mathcal{YM}(A) = \frac{1}{2}\langle F_A, F_A \rangle = \frac{1}{2}|F_A|^2.$$

Now let $\eta \in \Omega^1(M; ad(P))$ and let $t \in \mathbb{R}$. Then by the above lemma we have that

$$|F_{A_t}|^2 = |F_A|^2 + 2t\langle D_A\eta, F_A \rangle + \mathcal{O}(t^2)$$

where $\mathcal{O}(s)$ denotes a term that is divisible by s. Hence

$$\frac{1}{t}(\mathcal{YM}(A+t\eta) - \mathcal{YM}(A)) = \langle D_A\eta, F_A\rangle + \mathcal{O}(t).$$

By taking the limit as $t \to 0$ and appealing to 17.3 we conclude that A is a critical point of $\mathcal{YM}$ if and only if for every $\eta \in \Omega^1(M; ad(P))$,

$$\langle D_A\eta, F_A\rangle = 0.$$

This is equivalent to the requirement that

$$\langle \eta, (D_A)^* F_A\rangle = 0$$

where $(D_A)^*$ denotes the adjoint of the operator D_A with respect to this inner product. Since this equation is true for *every* $\eta \in \Omega^1(M; ad(P))$, this is equivalent to the condition that

$$(D_A)^* F_A = 0.$$

Now a standard calculation (done for example in [**?**]) shows that

$$(D_A)^*\omega = *D_A(*\omega)$$

for any $\omega \in \Omega^2(M; ad(P))$. (On the right hand side of this equation $*$ denotes the Hodge $*$ operator.) This, together with the fact that the $*$ operation is an isomorphism implies that A is a critical point if and only if the Yang–Mills equations

$$D_A * F_A = 0$$

hold. □

We end this chapter by restricting to perhaps the most important case; when $P \longrightarrow M$ is a principal $\mathrm{SU}(n)$-bundle over a compact four dimensional manifold M.

Observe that the Bianchi identities (Theorem 17.5) say that any connection A satisfies $D_A F_A = 0$. Comparing this with the Yang–Mills equations, we see that a distinguished class of solutions to the Yang–Mills equations are the *self dual* and *anti- self dual* connections; that is, connections that satisfy

$$F_A = *F_A \quad \text{and} \quad F_A = - * F_A$$

respectively. (Note that dimension four is necessary for the curvature to satisfy self duality.)

Theorem 17.14 *Let $P \longrightarrow M$ be a principal $\mathrm{SU}(n)$-bundle over a closed, oriented four dimensional manifold M. Suppose the second Chern class of the bundle is given by*

$$c_2(P) = k \in H^4(M^4) \cong \mathbb{Z}.$$

Then a connection A is a global minimum of the Yang–Mills functional $\mathcal{YM}$ if and only if

$$F_A = \begin{cases} *F_A & \text{if } k \text{ is negative} \\ -*F_A & \text{if } k \text{ is positive} \\ 0 & \text{if } k = 0. \end{cases}$$

Proof: To ease notation let Ω^2 denote $\Omega^2(M; ad(P))$. The Hodge star operator,

$$* : \Omega^2 \longrightarrow \Omega^2$$

which is defined by the relation

$$\alpha \wedge *\beta = \langle \alpha, \beta \rangle \nu \in \Omega^4(M; \mathbb{R})$$

where ν is the volume form, satisfies

$$*^2 = 1.$$

We can therefore split Ω^2 into the positive and negative eigenspaces of $*$:

$$\Omega^2 = \Omega^2_+ \oplus \Omega^2_-.$$

Accordingly for any two-form ω we write $\omega = \omega_+ + \omega_-$. ω is self dual if and only if $\omega_- = 0$, and is anti-self-dual if and only if $\omega_+ = 0$. Using this splitting for the curvature form of a connection, we have

$$\mathcal{YM}(A) = \int_M |F_A|^2 \, d(vol) \tag{17.4}$$

$$= \int_M |(F_A)_+|^2 + |(F_A)_-|^2 \, d(vol). \tag{17.5}$$

On the other hand, the Chern class $c_2(P)$ can be computed in terms of the curvature of any connection via the following Chern–Weil formula (see [**?**] for a description of this):

$$k = c_2(P)[M] = \frac{-1}{4\pi^2} \int_M trace(F_A \wedge F_A) \, d(vol) \tag{17.6}$$

$$= \frac{-1}{4\pi^2} \int_M |(F_A)_+|^2 - |(F_A)_-|^2 \, d(vol) \tag{17.7}$$

Since equation 17.7 is independent of the connection A, the theorem follows by comparing it to 17.5. Notice that we can also see that the minimum value of $\mathcal{YM}$ is $4\pi^2|k|$. □

Chapter 18

Stable Holomorphic Bundles and the Yang–Mills Functional on Riemann Surfaces

In the final four chapters we describe some recent applications of the Morse theoretic aspects of Yang–Mills theory. In this chapter we sketch the work of Atiyah and Bott [**?**] on the use of the Yang–Mills equations on Riemann surfaces to study the cohomology of moduli spaces of holomorphic bundles over Riemann surfaces.

A quick summary of this approach is as follows: First they studied the Yang–Mills equations over the Riemann surface M, and showed that solutions correspond to conjugacy classes of a certain central extension of the fundamental group $\pi_1(M)$. It was then shown that for relatively prime integers n and k, such representations also parameterize stable holomorphic vector bundles over M having rank n and first Chern class $k \in H^2(M;\mathbb{Z}) = \mathbb{Z}$. (We will define the notion of "stable" later.) Thus the moduli space of stable holomorphic bundles can be studied Morse theoretically; as the critical submanifold of the Yang–Mills functional

$$\mathcal{YM} : \mathcal{B}_{P(n,k)}(M) \longrightarrow \mathbb{R}$$

where $P(n,k)$ denotes the principal $U(n)$ bundle over M with first Chern class $c_1(P(n,k) = k$. ($P(n,k)$ is well defined up to isomorphism.) It turns out that the space of connections up to gauge equivalence $\mathcal{B}_{P(n,k)}(M)$ is relatively easy to understand homotopy theoretically. Moreover Atiyah and Bott prove that $\mathcal{YM}$ can be viewed as a *perfect* nondegenerate function on $\mathcal{B}_{P(n,k)}$, which in our language says that the spectral sequence in homology going from the homology of the critical submanifolds to the homology of the ambient manifold collapses. That is, the homology of the ambient manifold is given directly in

terms of the homology of the critical submanifolds. Actually they prove that in this case a gauge-equivariant version of this property is true. Since in this case the ambient manifold $\mathcal{B}_{P(n,k)}(M)$ is easy to understand directly, one can use the theory "backwards" in order to obtain homological information about the critical submanifold, which in this case is the moduli space of stable holomorphic bundles.

18.1 The Homotopy Type of the Space of Connections on a Riemann Surface

We begin a more detailed description of the work in [**?**] by studying the space of connections on a principal bundle over a Riemann surface.

Let M_g be a closed Riemann surface of genus g. $U(n)$ bundles over M_g are classified up to isomorphism by homotopy classes of maps into the classifying space, $[M_g, BU(n)]$. Since $BU(n)$ is simply connected with $\pi_2(BU(n)) \cong \mathbb{Z}$, and since M_g is two dimensional it is not difficult to show that there is a bijective correspondence

$$[M_g, BU(n)] \cong \pi_2(BU(n)) \cong \mathbb{Z}.$$

This correspondence is given by sending a map $f : M_g \longrightarrow BU(n)$ to the cohomology class $f^*(c_1) \in H^2(M_g; \mathbb{Z}) \cong \mathbb{Z}$. Hence a $U(n)$ bundle over M_g is completely classified by its first Chern class. Now let $\mathcal{B}_{(n,k)} = \mathcal{B}_{(n,k)}(M_g)$ be the space of *based* gauge equivalence classes of connections on a $U(n)$-bundle of Chern classs k, $P(n,k) \longrightarrow M_g$. That is

$$\mathcal{B}_{(n,k)} = \mathcal{A}_{n,k}(M_g)/\mathcal{G}_0$$

where, as in the last chapter, $\mathcal{G}_0$ denotes the group of based gauge transformations. (The orbit space under the full gauge group is discussed in [**?**], but for technical reasons we will not deal with it here.) Now by Theorems 17.10 and 17.11 we know that $\mathcal{B}_{(n,k)}$ is homotopy equivalent to the classifying space of the based gauge group $\mathcal{G}_0$, and its homotopy type is given by

$$\mathcal{B}_{(n,k)} \simeq Map_0^k(M_g, BU(n))$$

where this mapping space consists of basepoint preserving maps that classify a bundle of first Chern class $k \in H^2(M_g)$.

We now study the homotopy type of this mapping space. The basic tool for this study is the following theorem of Dold and Thom [**?**] relating Eilenberg–MacLane spaces to topological abelian groups.

Theorem 18.1 *An Eilenberg–MacLane space of type $K(G,n)$ for $n \geq 2$ has the homotopy type of a topological abelian group. Conversely, any topological abelian group has the homotopy type of a product of Eilenberg–MacLane spaces.*

Remark 18.1 *The group structure of an Eilenberg–MacLane space $K(G,n)$ is given by a map (well defined up to homotopy)*

$$K(G,n) \times K(G,n) \longrightarrow K(G,n)$$

which can be thought of as a class in $H^n(K(G,n) \times K(G,n); G)$. In this context it is the sum of the two fundamental classes in $H^n(K(G,n);G)$ coming from the two factors.

Now consider the case of $U(1)$-bundles over the Riemann surface M_g. In this case $U(1) = S^1$ and $BU(1) = \mathbb{CP}^\infty$ which is an Eilenberg–MacLane space of type $K(\mathbb{Z},2)$. Thus $BU(1)$ has the homotopy type of a topological abelian group, and therefore so does any mapping space with target $BU(1)$. In particular $\mathcal{B}_{(1,k)} = Map_0^k(M_g, BU(1))$ has the homotopy type type of a topological abelian group, and so by the Dold–Thom theorem it is homotopy equivalent to a product of Eilenberg–MacLane spaces. Thus to completely determine its homotopy type we are reduced to computing its homotopy groups. We have that $\mathcal{B}_{(1,k)}$ is connected and for $q \geq 1$,

$$\begin{aligned}
\pi_q(\mathcal{B}_{(1,k)}) &= \pi_q(Map_0(M_g, K(\mathbb{Z},2)) \\
&= [S^q, Map_0(M_g, K(\mathbb{Z},2))] \\
&= \pi_0(Map_0(S^q, Map_0(M_g, K(\mathbb{Z},2))) \\
&= \pi_0(Map_0(S^q \wedge M_g, K(\mathbb{Z},2))) \\
&= [S^q \wedge M_g, K(\mathbb{Z},2)] \\
&= H^2(S^q \wedge M_g; \mathbb{Z})
\end{aligned}$$

where $S^q \wedge M_g$ is the "smash" product

$$S^q \wedge M_g = S^q \times M_g / S^q \vee M_g$$

which is homeomorphic to the q-fold suspension of the space M_g. Hence we have that for $q \geq 1$

$$\pi_q(\mathcal{B}_{(1,k)}) = H^2(S^q \wedge M_g) = \tilde{H}^{2-q}(M_g) = \begin{cases} \mathbb{Z}^{2g} & \text{if } q = 1 \\ 0 & \text{if } q > 1. \end{cases}$$

Now since the circle S^1 is a $K(\mathbb{Z},1)$ we've proven the following.

Theorem 18.2 *There is a homotopy equivalence*

$$\mathcal{B}_{(1,k)} \simeq Map_0^k(M_g, BU(1)) \simeq (S^1)^{2g}.$$

The homotopy type of $\mathcal{B}_{(n,k)} \simeq Map_0^k(M_g, BU(n))$ for $n \geq 2$ is more complicated because in this case $BU(n)$ is not an Eilenberg–MacLane space or a product of such. However one can study the *rational* homotopy type of these spaces in an analogous manner, as follows.

From classical homotopy theory one knows that the rational homotopy type of the group $U(n)$ is given as follows.

Proposition 18.3 *There is a rational homotopy equivalence*

$$U(n) \simeq_{\mathbb{Q}} S^1 \times S^3 \times \cdots \times S^{2n-1}.$$

Remark 18.2 *Spaces X and Y are said to have the same rational homotopy type if there is a third space Z and maps $X \to Z$ and $Y \to Z$ which both induce isomorphisms in rational homotopy groups, $\pi_*(-)\otimes\mathbb{Q}$, and equivalently, rational homology, $H_*(-;\mathbb{Q})$.*

This theorem is proved by induction, using the fibre bundles

$$U(n-1) \longrightarrow U(n) \longrightarrow S^{2n-1}.$$

Now another classical result in homotopy theory, due to Serre, asserts that the homotopy groups $\pi_q(S^{2k-1})$ are finite abelian groups for $q > 2k-1$. Hence the sphere S^{2k-1} has the same rational homotopy type as the Eilenberg–MacLane space $K(\mathbb{Z}, 2k-1)$. Thus this proposition implies that $U(n)$ has the rational homotopy type of a product of Eilenberg–MacLane spaces, and hence so does its classifying space. That is, the following holds.

Lemma 18.4 *The classifying space $BU(n)$ has the same rational homotopy type as the product of Eilenberg–MacLane spaces,*

$$\mathrm{B\,U}(n) \simeq_{\mathbb{Q}} \prod_{q=1}^{n} K(\mathbb{Z}, 2q).$$

Similarly, there is a rational homotopy equivalence

$$\mathrm{B\,SU}(n) \simeq_{\mathbb{Q}} \prod_{q=2}^{n} K(\mathbb{Z}, 2q).$$

Remark 18.3 *The q^{th} factor in the rational homotopy equivalence $BU(n) \longrightarrow \prod_{q=1}^{n} K(\mathbb{Q}, q)$ is given by the q^{th} Chern character rational characteristic class. This rational equivalence is very important in K-theory as it establishes an isomorphism between rational topological K-theory and a direct sum of copies of rational cohomology. See [**?**] or [**?**] for details.*

We can now repeat the argument used for Theorem 18.2 to prove the following.

Theorem 18.5 *There is a rational homotopy equivalence*

$$\coprod_{k\in\mathbb{Z}} \mathcal{B}_{(n,k)} \simeq_{\mathbb{Q}} \prod_{q=1}^{n} K(\mathbb{Z}, 2q-2) \times K(\mathbb{Z}^{2g}, 2q-1).$$

18.2 Yang–Mills Connections and Representations

The results of the first section imply that the Yang–Mills functional

$$\mathcal{YM} : \mathcal{B}_{(n,k)} \longrightarrow \mathbb{R}$$

is defined on a manifold (infinite dimensional) whose homotopy type is fairly well understood. The idea in [**?**] is to use Morse-theoretic techniques to deduce information about the topology of the space of critical points. As seen in the last chapter these are connections A which satisfy the Yang–Mills equations,

$$D_A * F_A = 0.$$

The next step in analyzing the moduli space of solutions is to identify it with a certain representation space.

Let A be a connection on $P(n,k) \longrightarrow M_g$. Given a loop $\gamma : S^1 \longrightarrow M_g$, then using the parallel transport defined by A, one can lift γ to a path in $P(n,k)$ that starts and ends in the same fiber. These points are related by the action of a unique element $g \in U(n)$, which is the *holonomy* of γ determined by the connection A. A is a flat connection (i.e. $F_A = 0$) if and only if the holonomy of any loop only depends on its homotopy class. Thus flat connections determine holonomy representations of the fundamental group

$$h_A : \pi_1(M_g) \longrightarrow U(n).$$

It is not difficult to see that two connections that are related by a gauge transformation yield holonomy representations that are conjugate in $U(n)$. Furthermore any representation $\rho : \pi_1(M_g) \longrightarrow U(n)$ defines a flat connection on the vector bundle

$$\tilde{M}_g \times_{(\pi_1(M_g))} \mathbb{C}^n \longrightarrow M_g$$

where $\tilde{M}_g$ is the universal cover of M_g and where $\pi_1(M_g)$ acts on $\mathbb{C}^n$ via ρ. These constructions establish a bijective correspondence between the space of conjugacy classes of representations of $\pi_1(M_g)$ and the space of gauge equivalence classes of flat connections on rank n complex vector bundles over M_g.

In [**?**] they prove that nonflat Yang–Mills connections (i.e. that satisfy $D_A * F_A = 0$ but $F_A \neq 0$) also can be described in terms of representations. This was done as follows.

The fundamental group $\pi_1(M_g)$ is generated by $2g$ generators $a_1, b_1, \ldots, a_g, b_g$ subject to the relation

$$\prod_{i=1}^{g} [a_i, b_i] = 1$$

where $[a,b] = aba^{-1}b^{-1}$ is the commutator. Therefore for positive genus this group has a central extension

$$1 \longrightarrow \mathbb{Z} \longrightarrow \Gamma \longrightarrow \pi_1(M_g) \longrightarrow 1$$

where Γ is the group generated by the elements $a_1, b_1, \ldots a_g, b_g$ and the central element J satisfying the relation

$$\prod_{i=1}^{g}[a_i, b_i] = J.$$

Let $\Gamma_{\mathbb{R}}$ be the group obtained from Γ by extending the center to $\mathbb{R}$, so that there is a central extension

$$1 \longrightarrow \mathbb{R} \longrightarrow \Gamma_{\mathbb{R}} \longrightarrow \pi_1(M_g) \longrightarrow 1.$$

Notice that by construction the quotient group $\Gamma_{\mathbb{R}}/\mathbb{Z}$ is isomorphic to $\mathbb{R}/\mathbb{Z} \times \pi_1(M_g) = U(1) \times \pi_1(M_g)$. Let $P(1,1) \longrightarrow M$ be a fixed U(1)-bundle having Chern class 1 having a fixed Yang–Mills connection. In the genus 0 case one can think of $P(1,1)$ as the Hopf bundle $S^3 \longrightarrow S^2 = \mathbb{CP}(1)$. The induced line bundle $E \longrightarrow \mathbb{CP}(1)$ is the canonical line bundle; That is,

$$E = \{(z, \ell) : \ell \,\text{is a (complex) line in}\, \mathbb{C}^2, \,\text{and}\, z \in \ell\}.$$

There is a natural inclusion of E into the two dimensional trivial bundle

$$E \subset \mathbb{CP}(1) \times \mathbb{C}^2.$$

Let $p : \mathbb{CP}(1) \times \mathbb{C}^2 \longrightarrow E$ be the splitting induced by fiberwise orthogonal projection using the standard metric on $\mathbb{C}^2$. This induces a covariant derivative D_A (and therefore a connection) on E defined to be the composition

$$D_A : \Omega^0(\mathbb{CP}(1), E) \hookrightarrow \Omega^0(\mathbb{CP}(1); \mathbb{C}^2) \xrightarrow{\quad d \quad} \Omega^1(\mathbb{CP}(1); \mathbb{C}^2) \xrightarrow{\quad p \quad} \Omega^1(\mathbb{CP}(1); E)$$

where d is the usual exterior derivative.

It is not difficult to check that D_A is a Yang–Mills connection on E. This induces a Yang–Mills connection on the underlying principal bundle (in this case the Hopf bundle) $P(1,1) \longrightarrow S^2$. A similar canoncial Yang–Mills connection A can be found on the Chern class 1 principal $U(1)$ bundle $P(1,1) \longrightarrow M_g$ for any closed Riemann surface M_g. If we normalize the metric on M_g so that it has volume equal to one, then an argument like the one given to prove Theorem 17.14 (to establish the minimum value of the Yang–Mills functional) shows that the curvature F_A is given by

$$F_A = -2\pi i \omega$$

where ω is the volume form. (Notice in this case since $U(1)$ is an abelian group the bundle $ad(P(1,1)) \longrightarrow M_g$ is trivial, and hence the curvature F_A is a two form with trivial (complex) coefficients as is the volume form ω.)

Consider the universal covering space $\tilde{M}_g \longrightarrow M_g$. It has a canonical flat $\pi_1(M_g)$-connection. Taking the Whitney sum of bundles and connections, we have a $U(1) \times \pi_1(M_g)$-bundle

$$P(1,1) \oplus \tilde{M}_g \longrightarrow M_g$$

with a Yang–Mills connection (which we still call A) with curvature $-2\pi i\omega$. The projection map

$$\Gamma_{\mathbb{R}} \longrightarrow \Gamma_{\mathbb{R}}/\mathbb{Z} \cong U(1) \times \pi_1(M_g)$$

allows us to lift this bundle and connection to a Yang–Mills connection on a principal $\Gamma_{\mathbb{R}}$ bundle. Finally given a homomorphism

$$\rho : \Gamma_{\mathbb{R}} \longrightarrow G$$

where G is a compact Lie group one gets a Yang–Mills connection A_ρ on the induced G-bundle. The following is an important characterization of Yang–Mills connections, proved in chapter 6 of [**?**].

Theorem 18.6 *The mapping $\rho \longrightarrow A_\rho$ induces a bijective correspondence between conjugacy classes of homomorphisms $\rho : \Gamma_{\mathbb{R}} \longrightarrow G$ and gauge-equivalence classes of Yang–Mills connections over M_g.*

18.3 The Moduli Space of Stable Holomorphic Bundles

The next step in Atiyah and Bott's analysis is to identify a certain moduli space of holomorphic bundles with the representation space of the group $\Gamma_{\mathbb{R}}$ described above, and hence with the space of Yang–Mills connections over M_g.

Let $E = P(n,k) \longrightarrow M_g$ be a fixed, smooth complex vector bundle of rank n and Chern class $k \in \mathbb{Z} = H^2(M_g)$. Let

$$\mathcal{C} = \mathcal{C}(n,k) = \mathcal{C}(E)$$

be the space of all holomorphic structures on E. A holomorphic structure on E is a local trivialization of E so that the structure maps

$$\phi_{i,j} : U_i \cap U_j \longrightarrow GL(n,\mathbb{C})$$

are holomorphic.

Holomorphic structures can be thought of as coming from unitary connections on E as follows (see Ch. 5 of [**?**] for details).

Given a Riemann surface M_g, the Hodge star operator acts on the space of ($\mathbb{C}$-valued) one forms $* : \Omega^1(M_g) \longrightarrow \Omega^1(M_g)$ with the property that $*^2 = -1$. This gives a splitting

$$\Omega^1(M_g) = \Omega^{(1,0)}(M_g) \oplus \Omega^{(0,1)}(M_g)$$

where

$$* = -i \quad \text{on} \quad \Omega^{(1,0)}, \quad * = i \quad \text{on} \quad \Omega^{(0,1)}.$$

More generally the holomorphic structure of M_g defines a splitting of the space of $\mathbb{C}$-valued m-forms,

$$\Omega^m(M_g) = \oplus_{p+q=m}\Omega^{p,q}(M_g)$$

where, in terms of local holomorphic coordinates z_i , if we write forms in terms of dz_i and $d\bar{z}_i$, then $\Omega^{p,q}$ consists of forms with p dz's and q $d\bar{z}$'s. This splitting induces a splitting of the exterior derivative d into

$$d' : \Omega^0 \longrightarrow \Omega^{(1,0)} \quad \text{and} \quad d'' : \Omega^0 \longrightarrow \Omega^{(0,1)}.$$

This structure identifies the holomorphic functions

$$\phi : M_g \longrightarrow \mathbb{C}$$

are those elements of $\Omega^0(M_g)$ which locally satisfy $d''(\phi) = 0$.

Now given a unitary connection A on the bundle $E \longrightarrow M_g$, there is a similar splitting of the E-valued forms

$$\Omega^1(M_g; E) = \Omega^{1,0}(M_g; E) \oplus \Omega^{0,1}(M_g; E)$$

with respect to which there is a corresponding decomposition of the covariant derivative $d_A = d'_A \oplus d''_A$,

$$d'_A : \Omega^0(M_g; E) \longrightarrow \Omega^{1,0}M(_g; E) \quad \text{and} \quad d''_A : \Omega^0(M_g; E) \longrightarrow \Omega^{0,1}(M_g; E).$$

Atiyah and Bott prove that this splitting is compatible with a holomorphic structure on E. In particular it identifies the holomorphic sections as those on which locally d''_A is zero. Moreover this construction defines a map

$$\mathcal{A} \longrightarrow \mathcal{C}$$

from the space of unitary connections on E to the space of holomorphic structures on E. In fact this map is a linear isomorphism of affine spaces. Locally this corresponds to the isomorphism

$$\Omega^1(\mathfrak{u}(n)) \cong \Omega^{0,1}(\mathfrak{gl}(n, \mathbb{C}))$$

(see [?], chapters 5 and 8 for details). The space of automorphisms $Aut(E)$ which locally are given by smooth maps of M_g into $GL(n, \mathbb{C})$ act naturally on the space of holomorphic structures, $\mathcal{C}$, and the orbit space $\mathcal{C}/Aut(E)$, the space of isomorphism classes of holomorphic structures, is the naturally studied object. Now the gauge group acting on the space of unitary connections $\mathcal{G} = \mathcal{G}(E)$ is given locally by smooth of maps of M_g into $U(n)$. Since $GL(n, \mathbb{C})$ is the complexification of $U(n)$ it is natural to think of $Aut(E)$ as the complexification $\mathcal{G}^c$ of the gauge group $\mathcal{G}$. It is not difficult to see that the above map descends to give a map of orbit spaces

$$\mathcal{B} = \mathcal{A}/\mathcal{G} \longrightarrow \mathcal{C}/\mathcal{G}^c. \tag{18.1}$$

The orbit space $\mathcal{C}/\mathcal{G}^c$ of isomorphism classes of holomorphic structures on E is not an appropriate space for classification theorems in algebraic geometry. This is essentially because it does not have a natural structure as an algebraic variety. According to Mumford [?] one must restrict to a certain subclass of

holomorphic structures $\mathcal{C}_s = \mathcal{C}_s(E)$ which he refers to as *stable.* This is an open subspace of $\mathcal{C}$ and indeed if the rank and the Chern class are relatively prime, $(n,k) = 1$, then the space $\mathcal{C}_s/\mathcal{G}^c$ is a compact manifold. This is the "moduli space" of fundamental interest in [**?**]. The definition of stability is given as follows.

Let $\mu(E)$ be the normalized Chern class, defined to be $\mu(E) =$ Chern class/ rank. A holomorphic bundle E is *stable* if for every proper subbundle D, $\mu(D) < \mu(E)$. E is said to be *semi-stable* if it satisfies the weaker condition that for every proper subbundle D, $\mu(D) \leq \mu(E)$. Harder and Narasimhan [**?**] showed that every holomorphic bundle has a canonical filtration

$$0 = E_0 \subset E_1 \subset \cdots \subset E_r = E$$

with $D_i = E_i/E_{i-1}$ semi-stable and

$$\mu(D_1) > \mu(D_2) > \cdots > \mu(D_r).$$

Clearly E is semi-stable if and only if $r = 1$. Given this filtration of E, we define the *type* of E to be the vector

$$\mu = (\mu_1, \ldots, \mu_n)$$

where if D_i has rank n_i and Chern class k_i (so that $n = \sum n_i$ and $k = \sum k_i$) then the coordinates of μ are the ratios k_i/n_i, each repeated n_i times, arranged in decreasing order. So the first n_1 coordinates of μ are $k_1/n_1 = \mu(D_1)$, the next n_2 are k_2/n_2, and so on. We have that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n$.

Let $\mathcal{C}_\mu \subset \mathcal{C}$ denote the space of all holomorphic bundles of a given type μ. So in particular if the coordinates of μ are all equal to k/n, then all bundles of type μ are semi-stable, and for this case we write $\mathcal{C}_\mu = \mathcal{C}_{ss}$.

By the naturality of the Harder–Narasimhan filtration of a holomorphic vector bundle, the symmetry group $Aut(E) = \mathcal{G}^c$ preserves the notion of type, which is to say that each space $\mathcal{C}_\mu$ is a $\mathcal{G}^c$-invariant subspace of $\mathcal{C}$.

Now consider the equivariant homology

$$H_*^{\mathcal{G}^c}(\mathcal{C}_\mu) = H_*(E\mathcal{G}^c \times_{\mathcal{G}^c} \mathcal{C}_\mu).$$

(See chapter 15, section 15.2 for a fuller discussion of equivariant homology.) In particular, since $\mathcal{C}$ is an affine space it is contractible, and hence

$$E\mathcal{G}^c \times_{\mathcal{G}^c} \mathcal{C} \simeq B\mathcal{G}^c \simeq Map^k(M_g, BGL(n,\mathbb{C})) \simeq Map^k(M_g, BU(n)).$$

Thus the calculations in section 18.1 of this chapter imply that the equivariant homology $H_*^{\mathcal{G}^c}(\mathcal{C}) \cong H_*(Map^k(M_g, BU(n))$ is well understood, at least with rational coefficients.

Now the subspaces $\mathcal{C}_\mu$ of $\mathcal{C}$ define a $\mathcal{G}^c$-equivariant filtration of $\mathcal{C}$ which, when we apply equivariant homology, defines a spectral sequence converging to $H_*^{\mathcal{G}^c}(\mathcal{C})$, with E_1-term given as a direct sum of the equivariant homologies $\oplus_\mu H_*^{\mathcal{G}^c}(\mathcal{C}_\mu)$. In particular a direct summand of the E_1-term is the equivariant homology of the space of semi-stable holomorphic bundles $\mathcal{C}_{ss}$.

Thus we are in the "backwards" situation where we have a spectral sequence that converges to something we know fairly well ($H_*^{\mathcal{G}^c}(\mathcal{C})$), but whose E_1-term we'd like to compute (in particular $H_*^{\mathcal{G}^c}(\mathcal{C}_{ss})$). One of the main results in [**?**] is that this spectral sequence collapses. ([**?**] did not phrase their results in terms of spectral sequences, but rather in terms of "equivariant perfect" filtrations, but it is equivalent.) The following are the main results of §7 of [**?**].

Theorem 18.7 *There is an isomorphism*

$$H_q^{\mathcal{G}^c}(\mathcal{C}) \cong \oplus_\mu H_{q-2d_\mu}^{\mathcal{G}^c}(\mathcal{C}_\mu)$$

where, in terms of the sequence

$$(n_1, k_1), \ldots, (n_r, k_r)$$

defining μ, we have

$$d_\mu = \sum_{i>j} \left((n_i k_j - n_j k_i) + n_i n_j (g-1)\right).$$

Alternatively, d_μ is the complex codimension of $\mathcal{C}_\mu$ in $\mathcal{C}$.

Furthermore [**?**] made the following calculation in rational cohomology:

Theorem 18.8 *The rational equivariant homology of a stratum, $H_*^{\mathcal{G}^c}(\mathcal{C}_\mu; \mathbb{Q})$ is the tensor product of the rational equivariant homologies of the semi-stable strata $\mathcal{C}_{ss}(D_i)$ of the subquotients D_i. Here, of course, the equivariant homologies are taken with respect to the appropriate symmetry groups $Aut(D_i)$.*

Since $H_*^{\mathcal{G}^c}(\mathcal{C}) = H_*(Map^k(M_g, BU(n))$ is understood, these theorems together describe these groups in terms of the equivariant homology of the semi-stable strata of E and its various Harder–Narasimhan subquotients D_i, which gives an inductive procedure for computing the equivariant homology of spaces of semi-stable holomorphic bundles. (The induction is taken over rank.)

One result coming from this procedure is the following. Since $H_*(BU(n))$ is torsion free, it is not difficult to see that $H_*(Map(M_g, BU(n))$ is torsion free. This is the E_∞ term of a collapsing spectral sequence, and hence one concludes that the E_1- term of the spectral sequence is torsion free. In particular we have the following.

Corollary 18.9 *The semi-stable stratum has no torsion in its equivariant homology.*

Let $\mathcal{C}_s \subset \mathcal{C}$ be the subspace of *stable* holomorphic structures on the underlying bundle E. The moduli space of interest in algebraic geometry is the space of isomorphism classes of stable holomorphic bundles, namely the space

$$N(n,k) = \mathcal{C}_s / Aut(E) = \mathcal{C}_s / \mathcal{G}^c.$$

Now when the rank n and the Chern class k are relatively prime, it is known that the semi-stable and stable strata agree,

$$\mathcal{C}_{ss} = \mathcal{C}_s$$

and so in this case

$$N(n,k) = \mathcal{C}_{ss}/\mathcal{G}^s.$$

The relationship with gauge theory is given via the following theorem of Narashiman and Seshadri [**?**] (see §8 of [**?**]).

Theorem 18.10 *A holomorphic vector bundle of rank n is stable if and only if it arises from an irreducible representation $\rho : \Gamma_{\mathbb{R}} \longrightarrow U(n)$. Moreover isomorphic bundles correspond to equivalent (conjugate) representations.*

In view of Theorem 18.6 this result can be stated more clearly as follows. Let $\mathcal{N} \subset \mathcal{A} = \mathcal{A}(E) = \mathcal{A}(n,k)$ be the set of connections giving minima of the Yang–Mills functional. As discussed earlier, these connections define representations $\rho : \Gamma_{\mathbb{R}} \longrightarrow U(n)$. Let $\mathcal{N}_s \subset \mathcal{N}$ be those Yang–Mills connections that give irreducible representations. Then Theorem 18.10 can be reformulated as follows.

Theorem 18.11 *Under the identification of $\mathcal{A}$ with $\mathcal{C}$ of 18.1, the induced map of quotient spaces,*

$$\mathcal{N}_s/\mathcal{G} \longrightarrow \mathcal{C}_s/\mathcal{G}^c$$

is a homeomorphism.

Atiyah and Bott then went on to prove that if $\mathcal{A}_\mu$ denotes the space of unitary connections corresponding to the space of holomorphic structure $\mathcal{C}_\mu$ of type μ, then the $\mathcal{A}_\mu$ play the role of equivariant Morse strata for the Yang–Mills functional. That is, if $\mathcal{N}_\mu = \mathcal{N} \cap \mathcal{A}_\mu$ then $\mathcal{A}_\mu$ is the stable manifold of the Yang–Mills functional for the critical submanifold $\mathcal{N}_\mu$. Moreover, they argued that Theorem 18.7 can be interpreted as saying that the Yang–Mills functional is equivariantly perfect. (They did not prove this explicitly, however.)

Now as mentioned above, when the rank n and the Chern class k are relatively prime, the semi-stable and stable strata agree,

$$\mathcal{C}_{ss} = \mathcal{C}_s$$

and so in this case the homological object of geometric interest is

$$H_*(N(n,k)) = H_*(\mathcal{C}_{ss}/\mathcal{G}^s)$$

as opposed to the equivariant homology,

$$H_*^{\mathcal{G}^s}(\mathcal{C}_{ss}) = H_*(E\mathcal{G}^s \times_{\mathcal{G}^s} \mathcal{C}_{ss})$$

which is computed using the above techniques. If the action of $\mathcal{G}^s$ on $\mathcal{C}_{ss}$ were free these homology groups would be the same. The action, however, is not

free, but in this case the only isotropy subgroup is the group of scalars in $Aut(E) = \mathcal{G}^c$. Said another way, let $\mathbb{C}^*$ be the nonzero complex numbers, viewed as a subgroup of $Aut(E) = \mathcal{G}^c$. Let $\bar{\mathcal{G}}^c$ denote the quotient group. Then $\overline{\mathcal{G}}^c$ acts freely on $\mathcal{C}_s$ with orbit space $N(n,k)$. Hence

$$H_*(N(n,k)) = H_*^{\overline{\mathcal{G}}^c}(\mathcal{C}_s) = H_*^{\overline{\mathcal{G}}^c}(\mathcal{C}_{ss})$$

if $(n,k) = 1$.

Thus to use the above techniques, what remains to do is study the relationship between $\mathcal{G}^s$ and $\overline{\mathcal{G}}^c$ equivariant homology. This is done by studying the exact sequence of groups

$$1 \hookrightarrow \mathbb{C}^* \hookrightarrow \mathcal{G}^s \longrightarrow \overline{\mathcal{G}}^c \longrightarrow 1$$

on the level of classifying spaces; that is, the induced fibration

$$B\mathbb{C}^* \longrightarrow B\mathcal{G}^c \longrightarrow B\overline{\mathcal{G}}^c.$$

Now $B\mathbb{C}^* \simeq BU(1) = \mathbb{CP}^\infty$ and Atiyah and Bott verify that this fibration behaves from the point of view of rational homology as though it were trivial. This is because by restrict an element $\phi \in Aut(E) = \mathcal{G}^c$ to a particular fiber and then taking determinants there is a homomorphism

$$\mathcal{G}^c \longrightarrow \mathbb{C}^*$$

so that the composition $\mathbb{C}^* \hookrightarrow \mathcal{G}^c \longrightarrow \mathbb{C}^*$ is of degree n. This implies that the induced product map on classifying spaces

$$B\mathcal{B}^c \longrightarrow B\mathbb{C}^* \times B\overline{\mathcal{G}}^c$$

induces an isomorphism in rational homology. In §9 of [**?**] this is used deduce information about $H_*(N(n,k)) = H_*^{\overline{\mathcal{G}}^c}(\mathcal{C}_{ss})$ from information about $H_*^{\mathcal{G}^s}(\mathcal{C}_{ss})$ when $(n,k) = 1$. In particular they prove the following.

Theorem 18.12 *If $(n,k) = 1$, the moduli space $N(n,k)$ of stable holomorphic bundles has torsion-free homology.*

Chapter 19

Instantons on Four-Manifolds

In this chapter we discuss another important topological aspect of Yang–Mills theory, the study of the space of *instantons* on a compact four-dimensional manifold. These are the minima of the Yang–Mills functional and so are described by self-dual or anti-self-dual connections, up to gauge equivalence. From a Morse theoretic point of view one would expect the topology of these spaces of minima to be related to the topology of the entire space of connections up to gauge equivalence. In this chapter we will outline what is known about this relationship.

We remark that the topology of these spaces of minima have been studied from many points of view. In particular the topology of the space of instantons on a simply connected closed four-manifold has been used by Donaldson to define invariants of the manifold that has had much success in studying the differential topology of four manifolds. We recommend the book by Donaldson and Kronheimer [**?**] for a discussion of this aspect of gauge theory. In this chapter we will limit ourselves to that aspect of the theory that fits most closely the general theme of these lecture notes; the relationship between the topology of the space of critical points to the topology of the ambient manifold.

19.1 The Atiyah–Jones conjecture, configuration spaces, and $\mathrm{SU}(2)$-instantons on S^4

Let $\mathrm{SU}(n) \longrightarrow P \longrightarrow M^4$ be a principal $\mathrm{SU}(n)$-bundle over a closed, oriented four dimensional manifold M^4. Much of what we say will apply well to principal G-bundles over M^4 where G is any compact, simple Lie group, but to make some of the analysis easier we restrict ourselves to $\mathrm{SU}(n)$ in this section. Let $\mathcal{B}_0(P)$ be the space of connections on P modulo the action of the based gauge group $\mathcal{G}_0(P)$. Then as seen in chapter 17 $\mathcal{B}_0(P)$ is homotopy equivalent to

the classifying space of the gauge group $B\mathcal{G}_0(P)$ which in turn is homotopy equivalent to the mapping space $Map_0^P(M, \mathrm{B\,SU}(n))$. Consider the Yang–Mills functional

$$\mathcal{YM} : \mathcal{B}_0(P) \longrightarrow \mathbb{R}.$$

$\mathcal{M}(P)$ will denote the space of minima of this functional. Elements of $\mathcal{M}(P)$ are referred to as *instantons* . As seen in chapter 17 instantons are gauge equivalence classes of self-dual or anti-self-dual connections.

Now if the functional $\mathcal{YM}$ were a nondegenerate function satisfying the Palais–Smale condition (C), then one would expect the homology of the manifold

$$\mathcal{B}_0(P) \simeq Map_0^P(M, \mathrm{B\,SU}(n)),$$

which is fairly well understood, to be induced (by way of a spectral sequence) by the homologies of critical submanifolds. However $\mathcal{YM}$ does not satisfy condition (C) and so the relationship between the homotopy type of $\mathcal{B}_0(P)$ and that of the critical submanifolds of $\mathcal{YM}$ are not so clear. Indeed it has only been in the very recent work of Sibner, Sibner, and Uhlenbeck [**?**] and of Sadun and Segert [**?**] that nonminimal critical points were found in the case of $M = S^4$ and $\mathrm{SU}(n) = \mathrm{SU}(2)$. Nonetheless understanding how much of the homotopy type of the mapping space $Map_0^P(M, \mathrm{B\,SU}(n))$ is seen by the homotopy type of the space of minima (instantons) $\mathcal{M}(P)$ is a very interesting question, which is as of now, far from having been answered. The basic conjecture on this topic was made by Atiyah and Jones [**?**] in the case of $M = S^4$, but applies equally well to all manifolds. To explain this conjecture and what is known about it, we first make a homotopy theoretic observation.

Proposition 19.1 *Isomorphism classes of* $\mathrm{SU}(n)$*-bundles over a closed, oriented, four dimensional manifold* M^4 *is in bijective correspondence with the integers. The correspondence is given by the second Chern class:*

$$P \longrightarrow c_2(P) \in H^4(M^4) \cong \mathbb{Z}.$$

Proof: Consider the natural inclusion $\mathrm{SU}(2) \hookrightarrow \mathrm{SU}(n)$. An inductive argument using the fibrations

$$\mathrm{SU}(k-1) \hookrightarrow \mathrm{SU}(k) \longrightarrow S^{2k-1}$$

shows that the inclusion $S^3 \cong \mathrm{SU}(2) \hookrightarrow \mathrm{SU}(n)$ induces an isomorphism in homotopy groups through dimension 3, and is surjective in dimension 4. Thus on the classifying space level the inclusion

$$\mathrm{B\,SU}(2) \hookrightarrow \mathrm{B\,SU}(n)$$

induces an isomorphism in homotopy groups through dimension 4 and is surjective through dimension 5. Hence since M^4 is 4-dimensional we have a bijection between the sets of homotopy classes of maps

$$[M^4, \mathrm{B\,SU}(2)] \cong [M^4, \mathrm{B\,SU}(n)].$$

Thus it suffices to prove this theorem when $n = 2$. But in this case the classifying space $\mathrm{B}\,\mathrm{SU}(2) = BSp(1)$ is given by the infinite quaternion projective space $\mathbb{HP}^\infty$. Now $\mathbb{HP}^\infty$ has a CW-complex structure with one cell in every dimension of the form $4k$. Hence the inclusion of $S^4 = \mathbb{HP}^1 \hookrightarrow \mathbb{HP}^\infty$ induces an isomorphism of homotopy groups through dimension 6 and is surjective in dimension 7. This implies that there is a bijection

$$[M^4, S^4] = [M^4, \mathbb{HP}^1] \xrightarrow{\cong} [M^4, \mathbb{HP}^\infty] = [M^4, \mathrm{B}\,\mathrm{SU}(2)].$$

Now since M^4 is a closed, orientable manifold, the homotopy classes of maps $[M^4, S^4] \cong \mathbb{Z}$ where the correspondence is given by the degree of the map. The proposition follows once it is recalled that the the universal second Chern class $c_2 \in H^4(\mathrm{B}\,\mathrm{SU}(2)) \cong \mathbb{Z}$ is the generator. □

Now suppose the Chern class $c_2(P) = k$. We then write $\mathcal{M}_k = \mathcal{M}_k(M)$ for $\mathcal{M}(P)$ and $\mathcal{B}_k = \mathcal{B}_k(M) \simeq Map_0^k(M, \mathrm{B}\,\mathrm{SU}(n))$ for $\mathcal{B}_0(P) \simeq Map_0^P(M, \mathrm{B}\,\mathrm{SU}(n))$. The following is referred to as the Atiyah–Jones conjecture [**?**].

Conjecture 19.2 (Atiyah–Jones conjecture) *There is an increasing sequence of integers* $\{q_k,\ k = 0, 1, \dots\}$ *so that the inclusion*

$$\mathcal{M}_k(M) \hookrightarrow \mathcal{B}_k(M)$$

induces an isomorphism in homotopy groups and homology groups through dimension q_k.

A proof of the homology isomorphism statement in this conjecture in the case $M = S^4$ and $G = \mathrm{SU}(2)$ has been recently announced by Boyer, Hurtubise, Mann, and Milgram. For general manifolds, the conjecture seems to be far from understood, although a certain "stable " version of it (which we will describe later) has been proved by Taubes [**?**] for general 4-manifolds and indeed for all compact, simple Lie groups; not only $\mathrm{SU}(n)$.

The original evidence for the Atiyah–Jones conjecture came in their original paper in which they studied the space of $\mathrm{SU}(2)$ instantons on S^4. In this case the space of connections up to gauge equivalence, $\mathcal{B}_k(S^4)$ is homotopy equivalent to the loop spaces,

$$\mathcal{B}_k(S^4) \simeq Map_0^k(S^4, \mathrm{B}\,\mathrm{SU}(2)) = \Omega_k^4\,\mathrm{B}\,\mathrm{SU}(2) \simeq \Omega_k^3\,\mathrm{SU}(2) \simeq \Omega_k^3 S^3.$$

Here we are using the fact that for any group G, $\Omega BG \simeq G$. The subscript k in $\Omega_k^3 S^3$ denotes the set of (basepoint preserving) self maps of the sphere S^3 of degree k. This loop space has been studied extensively in homotopy theory. In particular its homology and stable homotopy type are well understood (See for example [**?**][**?**]). In [**?**] Atiyah and Jones proved the following.

Theorem 19.3 *The inclusion of the instanton space into the space of all connections,*

$$\mathcal{M}_k(S^4) \hookrightarrow \mathcal{B}_k(S^4) \simeq \Omega_k^3 S^3$$

induces a surjection in homology in dimensions $\leq k$.

Atiyah and Jones proved this theorem by comparing a well known combinatorial approximation space for the homology of loop spaces due to Segal [**?**] to a similar combinatorial construction of instantons due to a theoretical physicist named 'tHooft. The building block of both these constructions is the configuration space of distinct points in a manifold. That is, for a manifold N, let

$$F(N,k) = \{(x_1,\ldots,x_k) \in N^k \,:\, x_i \neq x_j \text{ if } i \neq j\}.$$

The symmetric group Σ_k acts freely on $F(N,k)$ by permuting coordinates. We let $C_N(k)$ be the orbit space

$$C_N(k) = F(N,k)/\Sigma_k.$$

$C_N(k)$ is the configuration space of k *unordered* points in N.

Particularly important examples of these spaces are when $N = \mathbb{R}^n$. In this case we write $C_n(k) = C_{\mathbb{R}^n}(k)$. An important example of these spaces are when $n = 2$, where it is not difficult to see that

$$\pi_1(C_2(k)) = \beta_k,$$

where β_k is Artin's *braid group* on k-strings. (See chapter 20 for a more complete explanation). Another important example is $F(\mathbb{R}^\infty,k) = \lim_{k\to\infty} F(\mathbb{R}^n,k)$. It is easy to see that the natural inclusion $F(\mathbb{R}^n,k) \hookrightarrow F(\mathbb{R}^{n+1},k)$ is null homotopic and so the limit space $F(\mathbb{R}^\infty,k)$ is a contractible space with a free Σ_k action. Hence the orbit space is the classifying space,

$$C_\infty(k) = B\Sigma_k = K(\Sigma_k,1).$$

Now for each n, Segal [**?**] defined an important map from $C_n(k)$ to the loop space $\Omega^n_k S^n$, where again the subscript k denotes the component of self maps of S^n that have degree k. The map

$$\alpha_n : C_n(k) \longrightarrow \Omega^n_k S^n$$

is defined as follows. Let $(x_1,\ldots,x_k) \in F(\mathbb{R}^n,k)$. Let $B(x_i)$ denote a ball centered at x_i of radius $\epsilon_i/4$, where ϵ_i is the minimal distance between x_i and any of the other x_j's. Let $S(x_i) = B(x_i)/\partial B(x_i)$. Let D^n be the unit disk around the origin and consider the natural affine homeomorphism

$$\begin{aligned} h_i : B(x_i) &\longrightarrow D^n \\ x &\longrightarrow \frac{(x-x_i)}{\epsilon_i}. \end{aligned}$$

By identifying boundaries to a point h_i defines a homeomorphism

$$h_i : S(x_i) \longrightarrow S^n = D^n/\partial D^n.$$

Now consider the map

$$\alpha(x_1,\ldots,x_k) : S^n = \mathbb{R}^n \cup \infty \longrightarrow S^n = D^n/\partial D^n$$

defined to be the composition

$$\mathbb{R}^n \cup \infty \xrightarrow{p(x_1,\ldots,x_k)} S(x_1) \vee \cdots \vee S(x_n) \xrightarrow{h_1 \vee \cdots \vee h_n} S^n$$

where $p(x_1, \ldots x_n)$ is the map that sends every point in one of the $B(x_i)$'s to itself, and every other point (including ∞) to the basepoint. This is clearly a map of degree k, whose definition is invariant under the permutations of the coordinates in $(x_1, \ldots, x_k)$. Hence it defines a map

$$\alpha_n : C_n(k) \longrightarrow \Omega^n_k S^n.$$

Notice there is a natural "gluing" map

$$C_n(k) \times C_n(r) \longrightarrow C_n(k+r)$$

defined by sending a pair of configurations $\mathbf{x}$ and $\mathbf{y}$ of length k and r respectively to the configuration $\mathbf{x}' \cup \mathbf{y}''$, where $\mathbf{x}'$ is the configuration in the open upper half plane of $\mathbb{R}^n$ defined by the configuration $\mathbf{x}$ via a fixed homeomorphism of $\mathbb{R}^n$ to the upper half plane, and $\mathbf{y}''$ is the configuration in the lower half plane defined by $\mathbf{y}$ via a fixed homeomorphism of $\mathbb{R}^n$ to the lower half plane.

These gluing maps are compatible with the loop sum maps

$$\Omega^n_k S^n \times \Omega^n_r S^n \longrightarrow \Omega^n_{k+r} S^n$$

in these sense that the following diagram homotopy commutes:

$$\begin{array}{ccc} C_n(k) \times C_n(r) & \longrightarrow & C_n(k+r) \\ {\scriptstyle \alpha_n \times \alpha_r}\downarrow & & \downarrow{\scriptstyle \alpha_{k+r}} \\ \Omega^n_k S^n \times \Omega^n_r S^n & \longrightarrow & \Omega^n_{k+r} S^n. \end{array}$$

In particular there are "stabilization maps"

$$\iota_k : C_n(k) \longrightarrow C_n(k+1)$$

given by gluing on a fixed point, say the origin in $C_n(1) = \mathbb{R}^n$. By the above diagram this is homotopy compatible with the map

$$j_k : \Omega^n_k S^n \longrightarrow \Omega^n_{k+1} S^n$$

defined by taking the loop sum with a fixed map of degree one (say the identity map on S^n). Notice that each j_n is a homotopy equivalence; its homotopy inverse is given by taking the loop sum with a fixed map of degree -1.

Let $C_n(\infty) = \lim_{k\to\infty} C_n(k)$, where the limit is taken over the maps ι_k. The above remarks about the compatibility of the maps α_k say there is a map, well defined up to homotopy, from this limit to the corresponding limit space $\lim_{k\to\infty} \Omega^n_k S^n$ where the limit is taken over the maps j_k. Now since each j_k is a homotopy equivalence, this limit is homotopy equivalent to any one of the

components of $\Omega^n S^n$. By choosing the component of the basepoint, $\Omega_0^n S^n$, we get a map

$$\alpha : C_n(\infty) \longrightarrow \Omega_0^n S^n.$$

The following is a corollary of what is known as Segal's group completion Theorem [**?**],

Theorem 19.4 *The map*

$$\alpha : C_n(\infty) \longrightarrow \Omega_0^n S^n$$

induces an isomorphism in homology, and therefore these spaces are stably homotopy equivalent. Equivalently, there is a homology equivalence

$$\mathbb{Z} \times C_n(\infty) \longrightarrow \Omega^n S^n.$$

Remark 19.1 *For $n = \infty$ this gives a homology equivalence between $C_\infty(\infty) = B\Sigma_\infty = \lim_{n\to\infty} B\Sigma_n$ and $\Omega_0^\infty S^\infty = \lim_{n\to\infty} \Omega_0^n S^n$ which had been proved earlier by Barratt, Quillen, and Priddy. The resulting relationship between the symmetric groups and the stable homotopy groups of spheres has been very important in homotopy theory.*

Notice also that for $n = 2$ there is a similar homology equivalence between $C_2(\infty) = B\beta_\infty = \lim_{n\to\infty} B\beta_n$ and $\Omega_0^2 S^2$. The resulting relationship between braid groups and the homotopy of the sphere S^2 has had similar importance.

This theorem can be viewed has a stable theorem, in that it deals with the limit of the $C_n(k)$'s. The following says that the stabilization process is in some sense as nice as possible. It follows from calculations of F. Cohen [**?**].

Theorem 19.5 *The map*

$$\alpha_n : C_n(k) \longrightarrow \Omega_k^n S^n$$

induces an injection in homology, and is an isomorphism through dimension k.

These theorems were used by Atiyah and Jones to study SU(2)-instantons on S^4 by showing that the map

$$\alpha_3 : C_3(k) \longrightarrow \Omega_k^3 S^3$$

factors up to homotopy as a composition

$$C_3(k) \xrightarrow{\tau} \mathcal{M}_k(S^4) \hookrightarrow \mathcal{B}_k(S^4) \simeq \Omega_k^3 S^3.$$

Such a factorization together with Theorem 19.5 implies the Atiyah–Jones Theorem 19.3.

The map τ in this composition is the 'tHooft construction of instantons. It is actually given in terms of a map

$$\tau : C_4(k) \longrightarrow \mathcal{M}_k$$

defined as follows.

Think of $\mathbb{R}^4$ as the quaternions $\mathbb{H}$. Let $(a_1, \ldots, a_k)$ be a set of k-distinct quaternions. Also let $(q_1, q_2) \in \mathbb{H} \times \mathbb{H} - \{(0,0)\}$. Consider the $k+1 \times k$ dimensional matrix $\Phi((q_1, q_2); (a_1, \ldots, a_k)) =$

$$\begin{pmatrix} 1 & 1 & \ldots\ldots & 1 \\ q_1 - a_1 q_2 & 0 & \ldots\ldots & 0 \\ 0 & q_1 - a_2 q_2 & 0 \quad \ldots & 0 \\ \vdots & \ldots\ldots\ldots\ldots & \ddots & \vdots \\ 0 & \ldots\ldots\ldots\ldots & 0 & q_1 - a_k q_2 \end{pmatrix}$$

Notice that since the a_i's are distinct and (q_1, q_2) is not the origin in $\mathbb{H}^2$, this matrix has maximal rank $(= k)$, and hence can be thought of as a surjective left linear transformation from $\mathbb{H}^{k+1}$ to $\mathbb{H}^k$. This transformation is given by applying the $k+1 \times k$ matrix Φ on the right of a $(k+1)$-dimensional row vector. Notice that this defines a map of quaternionic bundles

$$\begin{array}{ccc} \mathbb{H}^2 - \{(0,0)\}/\mathbb{H}^* \times \mathbb{H}^{k+1} & \xrightarrow{\Phi} & \mathbb{H}^2 - \{(0,0)\} \times_{\mathbb{H}^*} \mathbb{H}^k \\ \downarrow & & \downarrow \\ S^4 = \mathbb{H}^2 - \{(0,0)\}/\mathbb{H}^* & \xrightarrow[=]{} & \mathbb{H}^2 - \{(0,0)\}/\mathbb{H}^* \end{array}$$

where $\mathbb{H}^* = \mathbb{H} - \{0\}$ is viewed as acting by scalar multiplication on the right of the relevant quaternionic vector space. The map Φ of total spaces of this bundle is given by

$$((q_1, q_2); v) \longrightarrow ((q_1, q_2); v\Phi)$$

where $v \in \mathbb{H}^{k+1}$ and where $\Phi = \Phi((q_1, q_2); (a_1, \ldots, a_k)$.

The left hand bundle over S^4 is the trivial $k+1$-dimensional quaternionic bundle, and the right hand bundle is the Whitney sum k times of the canonical (quaternionic) line bundle over $\mathbb{HP}^1 = S^4$. By choosing an orientation so that the line bundle has Chern class $c_2 = -1$ this bundle has Chern class $c_2 = -k$. Furthermore the map Φ is a surjection between these these bundles. Hence its kernel bundle is a one dimensional quaternionic bundle E (or 2-dimensional complex bundle) of Chern class k. Since this bundle comes embedded inside a trivial bundle, it comes equipped with a connection defined via orthogonal projection:

$$D : \Omega^0(S^4; E) \hookrightarrow \Omega^0(S^4; \mathbb{H}^{k+1}) \xrightarrow{d} \Omega^1(S^4; \mathbb{H}^{k+1}) \xrightarrow{p} \Omega^1(S^4; E)$$

where d is the exterior derivative and p is induced by the orthogonal projection of the trivial bundle $S^4 \times \mathbb{H}^{k+1}$ onto E.

It was shown in [**?**] that this connection is self-dual. Moreover it was shown that the association to the configuration $(a_1, \ldots, a_k)$ the bundle E with connection D induces a well defined map

$$\tau : C_4(k) \longrightarrow \mathcal{M}_k$$

which was first described in a somewhat different fashion by the physicist 'tHooft. This construction is a special case of what has become known as the ADHM construction [**?**] which gives a complete classification of instantons over S^4 in terms of quaternionic matrices. In any case Atiyah and Jones verified that the composition

$$C_3(k) \hookrightarrow C_4(k) \xrightarrow{\tau} \mathcal{M}_k \hookrightarrow \mathcal{B}_k \simeq \Omega^3_k S^3$$

is homotopic to α_3, and hence their theorem (Theorem 19.3) follows from Theorem 19.5.

19.2 Instantons on general four manifolds, gluing and the Taubes stability theorem

The generalization of the Atiyah–Jones Theorem 19.5 to arbitrary compact four-manifolds and arbitrary compact, simple Lie groups was obtained by Taubes [**?**]. More specifically, he proved the following:

Theorem 19.6 *Let M be a compact, connected, oriented, Riemannian four-manifold, and let G be a simple, connected Lie group. Fix an integer q. Then there is an integer $m(q)$ with the following significance: Let $P \longrightarrow M$ be a principal G-bundle whose corresponding adjoint vector bundle $ad(P)$ has Pontryagin class $p_1(ad(P)) \geq m(q)$. Then the inclusion of the moduli space into the full space of connections up to based gauge equivalence,*

$$\mathcal{M}(P) \hookrightarrow \mathcal{B}_0(P)$$

induces a surjection in homology and homotopy groups in dimensions $\leq q$.

In [**?**] Taubes also proved a stable analogue of the Atiyah–Jones conjecture in this generality (i.e. for all compact, Riemannian manifolds, and all compact, simple, Lie groups). For the sake of these notes we will limit our explanation of this result to the case $G = \mathrm{SU}(n)$.

Fix a four-manifold M. Let $P \longrightarrow M$ be a principal $\mathrm{SU}(n)$-bundle of Chern class $c_2(P) = k$ Recall that

$$\mathcal{B}_k \simeq Map^k_0(M, \mathrm{B\,SU}(n)).$$

As in the case when $M = S^4$, all of these path components of $Map_0(M, \mathrm{B\,SU}(n))$ are homotopy equivalent. Such a homotopy equivalence between Map^k_0 and Map^{k+1}_0 can be constructed as follows. Let

$$\gamma : S^4 \longrightarrow \mathrm{B\,SU}(n)$$

be a map having Chern class $c_2(\gamma) = 1$. Equivalently, this class generates $\pi_4(\mathrm{B\,SU}(n)) \cong \mathbb{Z}$. Let

$$p : M^4 \longrightarrow M^4 \vee S^4$$

be the map that pinches the boundary of a small disk in M^4 to a point. Then given $f \in Map^k(M, \mathrm{B\,SU}(n))$ the new map

$$j(f) : M^4 \xrightarrow{\ p\ } M^4 \vee S^4 \xrightarrow{\ f\vee\gamma\ } \mathrm{B\,SU}(n)$$

has degree $k+1$. This procedure describes a map

$$j : Map_0^k(M^4, \mathrm{B\,SU}(n)) \longrightarrow Map_0^{k+1}(M^4, \mathrm{B\,SU}(n)).$$

This map is clearly a homotopy equivalence; its homotopy inverse is given by the same procedure, replacing γ by $-\gamma$.

The main device used by Taubes in [**?**] is a "gluing map"

$$\tau : \mathcal{M}_k \longrightarrow \mathcal{M}_{k+N}$$

for N sufficiently large.

The gluing map τ was defined originally in [**?**] and it was proved in [**?**] to be is homotopy compatible with j. That is, the following diagram homotopy commutes:

$$\begin{array}{ccc} \mathcal{M}_k & \xrightarrow{\ \tau\ } & \mathcal{M}_{k+N} \\ \big\downarrow & & \big\downarrow \\ \mathcal{B}_k & \xrightarrow[j]{\ \simeq\ } & \mathcal{B}_{k+N}. \end{array}$$

This allows one to then define a map of the limiting spaces

$$h : \mathcal{M}_\infty \longrightarrow \mathcal{B}_\infty$$

where

$$\mathcal{M}_\infty = \lim_{k\to\infty} \mathcal{M}_k \quad \text{and} \quad \mathcal{B}_\infty = \lim_{k\to\infty} \mathcal{B}_k$$

where the limits are taken with respect to the maps τ and j respectively. Notice that since each map j is a homotopy equivalence, there is a natural homotopy equivalence

$$\mathcal{B}_\infty \simeq \mathcal{B}_0 \simeq Map_0^0(M^4, \mathrm{B\,SU}(n)).$$

The following, referred to as the "stable" version of the Atiyah–Jones conjecture, was also proved by Taubes in [**?**].

Theorem 19.7 *The induced map on the limit spaces*

$$h : \mathcal{M}_\infty \longrightarrow \mathcal{B}_\infty \simeq Map_0^0(M^4, \mathrm{B\,SU}(n))$$

is a homotopy equivalence.

We end this chapter with a rough description of Taubes' gluing construction and how he used it to prove these theorems. The reader is referred to [**?**] for details.

Let A be a connection on a bundle $P \longrightarrow M$ of Chern-class k. Now let $x \in M$. We can assume that x has a small neighborhood U homeomorphic to a disk on which the connection A is flat. The idea of the procedure is to "glue" in a connection over S^4 into this neighborhood. So let B be a connection on a bundle $Q \longrightarrow S^4$ of Chern class $c_2 = q$. We assume that B is flat in a neighborhood of $\infty \in S^4 = \mathbb{R}^4 \cup \infty$. Consider the pinch map

$$M^4 \longrightarrow U/\partial U \cong S^4$$

defined by mapping every point in U to itself, and every point outside U to the basepoint (which is identified with $\infty \in S^4 = \mathbb{R}^4 \cup \infty$.) Pulling back the bundle Q and connection B along this map defines a bundle $p^*(Q)$ over M, trivial outside of the neighborhood U, with a connection $p^*(B)$, which is flat outside of U. Given a parameter $\lambda \in (0, 1]$ one can then defined the "glued" connection $A' = A + \lambda^* B$ on the Whitney sum bundle $P \oplus p^*(Q)$. This procedure defines a "gluing" pairing

$$\mathcal{B}_k(M) \times \mathcal{B}_q(S^4) \longrightarrow \mathcal{B}_{k+q}(M).$$

Now let B be a fixed self-dual connection on the SU(2) bundle on S^4 of Chern class one. (This can be taken to be the Hopf bundle $S^3 = \mathrm{SU}(2) \longrightarrow S^7 \longrightarrow S^4$.) Using the canonical inclusion $\mathrm{SU}(2) \hookrightarrow \mathrm{SU}(n)$ this defines a self-dual connection (which we still call B) on an SU(2)-bundle $Q \longrightarrow S^4$ of Chern class one. Gluing with this connection defines a map

$$\mathcal{B}_k(M) \longrightarrow \mathcal{B}_{k+1}(M)$$

which is a homotopy equivalence, homotopic to the map

$$j : Map_0^k(M^4, \mathrm{B\,SU}(n)) \longrightarrow Map_0^{k+1}(M^4, \mathrm{B\,SU}(n))$$

described above.

The main part of the analysis in Taubes' work has to do with the studying the effect of this gluing procedure on the Yang–Mills energy. This can be expressed in the following way.

Normalize the Yang–Mills energy as follows. Let C be a connection on a bundle over M^4 of Chern class k. Let $\mathfrak{a}(C)$ be the number

$$\mathfrak{a}(C) = \int_M |P_- F_C|^2 d\,vol$$

where $P_- = (1 - *)/2$, where $*$ is the Hodge star operator. Notice that if we let $P_+ = (1 + *)/2$, then

$$F_C = P_+ F_C + P_- F_C$$

and we may think of $P_+ F_C$ and $P_- F_C$ as the self-dual and anti-self-dual parts of F_C respectively. The minima of $\mathfrak{a}$ are precisely $\mathfrak{a}^{-1}(0)$ and are realized by the self-dual connections.

Let $\mathcal{B}_\epsilon = \mathfrak{a}^{-1}[0,\epsilon) \subset \mathcal{B}(M)$. A *homotopy invariant* family of compact subsets $\mathcal{F}$ of $\mathcal{B}$ is a family of compact subsets with the property that if

$$G : \mathcal{B} \times I \longrightarrow \mathcal{B}$$

is a homotopy of the identity (i.e. $1 = G_0 = G_{|_{\mathcal{B}\times 0}}$), then if $K \subset \mathcal{F}$ then $G_1(K) \subset \mathcal{F}$. In [**?**] Taubes is concerned with families of compact sets, invariant under homotopies rel $\mathcal{B}_\epsilon$ (i.e. homotopies G as above so that each G_t maps $\mathcal{B}_\epsilon$ into itself). For example if

$$z \in H_m(\mathcal{B}, \mathcal{B}_\epsilon)$$

then z is represented by a singular m-dimensional chain in $\mathcal{B}$ whose boundary lies in $\mathcal{B}_\epsilon$. The set of such singular m-dimensional chains representing z is a family $\mathcal{F}_z$ of compact subsets of $\mathcal{B}$ which are homotopy invariant rel $\mathcal{B}_\epsilon$. As another example, consider a releative homotopy class $z \in \pi_m(\mathcal{B}, \mathcal{B}_\epsilon)$. z is represented by a map of an m- dimensional disk whose boundary is mapped to $\mathcal{B}_\epsilon$. The images of the set of such representatives of z is also a family $\mathcal{F}_z$ of compact subsets of $\mathcal{B}$ which homotopy invariant rel $\mathcal{B}_\epsilon$.

Taubes proves that when the above gluing is done carefully the gluing map $j : \mathcal{B}_k \longrightarrow \mathcal{B}_{k+1}$ can be chosen to have the following properties. Given a compact set $K \in \mathcal{B}_k$ then there is an integer N so that the induced homotopy equivalence

$$j : \mathcal{B}_k \longrightarrow \mathcal{B}_{k+N}$$

maps $(\mathcal{B}_k)_\epsilon$ into $(\mathcal{B}_{k+N})_{2\epsilon}$. Also

$$\sup_{j(K)} \mathfrak{a} < c \cdot \sup_K \mathfrak{a} + \epsilon$$

where c is a positive constant less than 1.

Taubes shows that this inequality implies the following property. Given $\epsilon > 0$ and a homotopy invariant (rel $(\mathcal{B}_k)_\epsilon$) family of compact subsets of $\mathcal{B}_k$, $\mathcal{F}$, there exists an $N > 0$ and a homotopy equivalence $j : \mathcal{B}_k \longrightarrow \mathcal{B}_{k+N}$ as above, which maps $(\mathcal{B}_k)_\epsilon$ to $(\mathcal{B}_{k+n})_{2\epsilon}$, and there exists a compact set $K \subset j(F)$ which lies in $(\mathcal{B}_{k+N})_{2\epsilon}$.

Taking the family $\mathcal{F}$ associated associated to a relative homotopy classe z as above, this property implies that under the map of relative homotopy groups

$$j : \pi_m(\mathcal{B}_k, (\mathcal{B}_k)_\epsilon) \longrightarrow \pi_m(\mathcal{B}_{k+N}, (\mathcal{B}_{k+N})_{2\epsilon}) \tag{19.1}$$

the class z gets mapped to zero.

On S^4 with its standard metric, Taubes shows that there exists an $\epsilon_0 > 0$ so that $(\mathcal{B}_k)_{\epsilon_0}$ has a strong deformation retraction ρ onto $\mathcal{M}_k$ for every k. This is a kind of tubular neighborhood of $\mathcal{M}_k$ in $\mathcal{B}_k$. The existence of this deformation retraction implies that the map j in 19.1 has can be interpreted as defining a direct limit system $\pi_m(\mathcal{B}_k, \mathcal{M}_k)$ with the property that

$$\lim_{k\to\infty} \pi_m(\mathcal{B}_k, \mathcal{M}_k) = 0.$$

The above theorems follow rather easily from this result in this case. In particular the map

$$\tau : \mathcal{M}_k \longrightarrow \mathcal{M}_{k+N}$$

is given by the composition

$$\mathcal{M}_k \subset (\mathcal{B}_k)_\epsilon \xrightarrow[\simeq]{j} (\mathcal{B}_{k+N})_{2\epsilon} \xrightarrow[\simeq]{\rho} \mathcal{M}_{k+N}.$$

In the general case ($M \neq S^4$) it is not necessarily true that $(\mathcal{B}_k)_\epsilon$ can be retracted onto $\mathcal{M}_k$ for any ϵ. In this case a more delicate min-max argument has to be used with the above energy estimates to prove Theorems 19.6 and 19.7. We urge the reader to consult [**?**] for details.

Chapter 20

Monopoles, Rational Functions, and Braids

In this chapter we consider the space of "time invariant" instantons, or *monopoles* in Euclidean space. This is the space of minima of a variant of the Yang–Mills energy functional called the "Yang–Mills–Higgs" functional. The ambient manifold on which this functional is defined is, via a theorem of Taubes, homotopy equivalent to the space of smooth self maps of the sphere S^2. The topology of the minima of this functional is, via a theorem of Donaldson, given by the space of *holomorphic* maps of the Riemann sphere to itself, or equivalently, the space rational functions of a complex variable. Thus the Morse theoretic question of how much of the topology of the ambient manifold is detected by the topology of the space of minima, is the question of how much of the topology of the entire space of smooth maps from S^2 to itself is detected by the topology of the space of holomorphic maps. The answer to this question was recently worked out by F. Cohen, R. Cohen, B. Mann, and R.J. Milgram in [**?**]. In this work a curious relationship between the space of monopoles and Artin's braid groups was deduced. We will discuss this work in this chapter. We refer the reader to [**?**] for details and to [**?**] for a more geometric description of the space of monopoles.

20.1 Time invariant connections, monopoles, and rational functions

Let $P = \mathbb{R}^4 \times \mathrm{SU}(2)$ be the trivial $\mathrm{SU}(2)$-bundle over $\mathbb{R}^4$. Clearly the vector bundle $ad(P)$ is also trivial, and so a connection C on P can be viewed as a one-form

$$C \in \Omega^1(\mathbb{R}^4, ad(P)) = \Omega^1(\mathbb{R}^4, \mathfrak{su}(2)) \cong \Omega^1(\mathbb{R}^4, \mathbb{R}^3).$$

Throughout this chapter we will be thinking of the last coordinate of $\mathbb{R}^4$ as "time" and so we will write an element of $\mathbb{R}^4$ as (v, t) where $v \in \mathbb{R}^3$ and $t \in \mathbb{R}$.

Then a connection C can be written in the form

$$C = A_1(v,t)dx_1 + A_2(v,t)dx_2 + A_3(v,t)dx_3 + \phi(v,t)dt$$

where A_1, A_2, A_3 and $\phi : \mathbb{R}^4 \longrightarrow \mathfrak{su}(2) \cong \mathbb{R}^3$ are smooth maps.

The connection C is said to be *time invariant* if each of the maps A_i and ϕ are time invariant. That is

$$A_i(v,t) = A_i(v,0) \quad \text{and} \quad \phi(v,t) = \phi(v,0)$$

for all $t \in \mathbb{R}$. We will only be considering time invariant connections in this chapter. Notice in this case the maps A_i and ϕ can be thought of as smooth maps $\mathbb{R}^3 \to \mathfrak{su}(2)$. In particular

$$A = A_1(v)dx_1 + A_2(v)dx_2 + A_3(v)dx_3$$

may be viewed as a connection on the trivial SU(2) bundle over $\mathbb{R}^3$. Thus time invariant connections on $\mathbb{R}^4$ are given by pairs (A, ϕ) where

1. A is a connection on the trivial SU(2)-bundle on $\mathbb{R}^3$,
2. ϕ is an $\mathfrak{su}(2)$ valued smooth function on $\mathbb{R}^3$.

The connection A is called the *gauge potential* and ϕ is called the *Higgs field*.

The pairs (A, ϕ) that we will consider will be required to satisfy several conditions (see [T, AH]). To describe the first condition, observe that given a time invariant connection C, the square of the norm of the curvature, $|F_C|^2$, at any given point $v \in \mathbb{R}^3$ can be expressed in terms of the associated pair (A, ϕ). It is given by $|F_A|^2 + |D_A\phi|^2$ where F_A is the curvature of the $\mathbb{R}^3$-connection A, and D_A is its covariant derivative, which operates on ϕ by considering ϕ as a zero-form with coefficients in $\mathfrak{su}(2)$. Hence $D_A\phi \in \Omega^1(\mathbb{R}^3, \mathfrak{su}(2))$. Thus the appropriate energy functional in this setting is the *Yang–Mills–Higgs* energy given by

$$\mathcal{U}(A,\phi) = \int_{\mathbb{R}^3} (|F_A|^2 + |D_A\phi|^2) d\,vol.$$

The first condition we will require is the finite energy condition:

$$\mathcal{U}(A,\phi) < \infty.$$

We also require a condition on the behavior of ϕ at infinity. There are several different conditions which may be imposed, but the usual normalizing condition is that

$$\lim_{|x|\to\infty} |\phi(x)| = 1$$

where we take the usual norms in $\mathbb{R}^3$. Actually the weakest convergence condition that seems to be sufficient is $1 - |\phi| \in L^6(\mathbb{R}^3)$. We do not assume any asymptotic conditions on the gauge potential A (other than that implied by

requiring that $\mathcal{U}(A,\phi) < \infty$). However we do impose an additional basepoint condition:

$$\lim_{t\to\infty} \phi(t,0,0) = (1,0,0).$$

Let $\mathcal{A}$ be the space of pairs (A,ϕ) which satisfy these three conditions. $\mathcal{A}$ is a space of time-invariant connections on $\mathbb{R}^4$, but unlike previous spaces of connections we have considered, $\mathcal{A}$ is *not* an affine space, indeed it is not contractible. The three conditions impose an interesting topology on $\mathcal{A}$ which was identified by Taubes [**?**] in the following manner.

Consider the following map

$$I : \Omega^2 S^2 \longrightarrow \mathcal{A},$$

where here $\Omega^2 S^2$ is the space of all smooth maps $S^2 \longrightarrow S^2$ which preserve the basepoint $(1,0,0) \in S^2$. Identify the unit sphere S^2 with the unit sphere in the Lie algebra $\mathfrak{su}(2)$. Given a map

$$\alpha : S^2 \longrightarrow S^2 \subset \mathfrak{su}(2)$$

define the pair $I(\alpha) = (A,\phi)$ by the formula

$$A = \beta(|x|)\left[\alpha\left(\frac{x}{|x|}\right), d\alpha\left(\frac{x}{|x|}\right)\right]$$
$$\phi = -\beta(|x|)\alpha\left(\frac{x}{|x|}\right).$$

In this formula $\beta : \mathbb{R} \longrightarrow [0,1]$ is a smooth cut-off function which is identically 0 if $t \leq 1/2$ and identically 1 if $t \geq 3/4$, and $[,]$ is the Lie bracket in $\mathfrak{su}(2)$. (See [**?**] for details). The following was proved in [**?**].

Theorem 20.1 *The map $I : \Omega^2 S^2 \longrightarrow \mathcal{A}$ is a homotopy equivalence.*

We remark that the homotopy inverse to the map I is given by sending a pair (A,ϕ) to the restriction of the Higgs field ϕ to the "sphere at ∞" which is to say the map $\bar{\phi} : S^2 \longrightarrow S^2$ given by

$$\bar{\phi}(x) = lim_{t\to\infty}\phi(tx).$$

Since the bundle over $\mathbb{R}^3$ is trivial the based gauge group of bundle automorphisms in this setting is given by

$$\mathcal{G}_0 = Map_0(\mathbb{R}^3, \mathrm{SU}(2))$$

where here Map_0 means smooth maps $g : \mathbb{R}^3 \longrightarrow \mathrm{SU}(2)$ which satisfy the basepoint condition

$$\lim_{t\to\infty} g(t,0,0) = 1 \in \mathrm{SU}(2).$$

The gauge group $\mathcal{G}_0$ acts on $\mathcal{A}$ by the formula

$$g(A,\phi) = (g^{-1}Ag + g^{-1}dg, g^{-1}\phi g).$$

This formula makes sense since g and ϕ are matrix valued functions and A is a matrix of one-forms. In this formula the expression dg refers to the differential of the function $g : \mathbb{R}^3 \longrightarrow \mathrm{SU}(2)$ which is a one-form with coefficients in the Lie algebra $\mathfrak{su}(2)$. (Here we are identifying the Lie algebra $\mathfrak{su}(2)$ with the tangent space at the identity of the Lie group SU(2).)

Like in previously studied examples, this action of the based gauge group $\mathcal{G}_0$ on $\mathcal{A}$ is free and it is not difficult to see that the quotient map the quotient map to the orbit space

$$\mathcal{A} \longrightarrow \mathcal{B} = \mathcal{A}/\mathcal{G}_0$$

is a principal $\mathcal{G}_0$ bundle. (Again we refer to [**?**] for details.) Moreover since $\mathbb{R}^3$ is contractible, the basepoint condition in $\mathcal{G}_0 = Map_0(\mathbb{R}^3, \mathrm{SU}(2))$ implies that $\mathcal{G}_0$ is a contractible topological group. Hence this bundle has contractible fibers. This implies the following.

Proposition 20.2 *The projection onto the orbit space*

$$\mathcal{A} \longrightarrow \mathcal{B} = \mathcal{A}/\mathcal{G}_0$$

is a homotopy equivalence.

Combining Theorem 20.1 and Proposition 20.2 we have the following theorem of Taubes proved in [**?**].

Theorem 20.3 *There is a natural homotopy equivalence*

$$\mathcal{B} \xrightarrow{\simeq} \Omega^2 S^2.$$

In view of this result the path components of $\mathcal{B}$ (and of $\mathcal{A}$) are labelled by the integers $\mathbb{Z}$ corresponding to the degree of the map in $\Omega^2 S^2$. We write $\mathcal{B}_k$ for the component corresponding to $\Omega^2_k S^2$. In the literature the integer k is referred to as the *charge* of the pair (A, ϕ).

The Yang–Mills–Higgs functional is invariant under the gauge group action and so defines a function

$$\mathcal{U} : \mathcal{B}_k \longrightarrow \mathbb{R}$$

for each $k \in \mathbb{Z}$. By the definition of $\mathcal{U}$ (actually its local relationship with the Yang–Mills functional on the associated time invariant connections on $\mathbb{R}^4$) it is not difficult to see that $(A, \phi) \in \mathcal{A}$ represents a minimum of $\mathcal{U}$ if and only if the associated four-dimensional connection

$$\alpha = A + \phi dt$$

is self-dual, or anti-self-dual:

$$*F_\alpha = \pm F_\alpha.$$

This duality equation, written directly in terms of the pair (A, ϕ) is given by the *Bogomolnyi equation*

$$*F_A = \pm D_A \phi.$$

Notice that in this equation, since the underlying manifold is $\mathbb{R}^3$, the Hodge star of the curvature two-form F_A is a one-form, as is the covariant derivative of the Higgs field ϕ.

Thus the space of minima $\mathcal{M}_k \subset \mathcal{B}_k$ of the Yang–Mills–Higgs functional $\mathcal{U}$, referred to as *monopoles,* is the space of gauge equivalence classes of pairs $(A, \phi) \in \mathcal{A}$ satisfying the Bogomolnyi equation. From the topological point of view, the basic question inspired by Morse theory is therefore the following.

Question 20.1 *How much of the homotopy type of $\mathcal{B}_k \simeq \Omega^2_k S^2$ is reflected in the homotopy type of of the space of monopoles $\mathcal{M}_k$?*

A major step toward the understanding of the spaces of monopoles was taken by Donaldson [**?**] who proved the following.

Theorem 20.4 *The space of monopoles of charge k $\mathcal{M}_k$, is homeomorphic to the space of basepoint point preserving holomorphic maps of the Riemann sphere S^2 to itself of degree k.*

We will denote the space of holomorphic maps in this theorem as Rat_k, as such holomorphic maps are given by rational functions

$\mathrm{Rat}_k = \{p/q : p \text{ and } q \text{ are monic, degree } k \text{ polynomials in } \mathbb{C} \text{ with no roots in common.}\}$

The basepoint condition in this description of the space of holomorphic maps is that they send $\infty \in \mathbb{C} \cup \infty = S^2$ to 1.

Thus the above question can be rephrased as follows.

Question. How much of the homotopy type of the full space of based, degree k self maps of the sphere, $\Omega^2_k S^2$, is reflected in the homotopy type of the space of holomorphic maps Rat_k?

The answer to this question is now fairly well understood by the results of [**?**]. We will discuss this work in the next section. We end this section with an observation about charge one monopoles $= \mathcal{M}_1 \cong \mathrm{Rat}_1$.

Notice that

$$\mathrm{Rat}_1 = \{\frac{z-u}{z-v} : u \neq v\}$$

and is therefore diffeomorphic to $\mathbb{C}^2 - \Delta$, where $\Delta \subset \mathbb{C}^2$ is the diagonal. But this space is diffemorphic to $\mathbb{C} \times \mathbb{C}^*$ under the map

$$(u, v) \longrightarrow (u, u - v).$$

Notice furthermore that $\mathbb{C} \times \mathbb{C}^* \cong \mathbb{R}^3 \times S^1$. Atiyah and Hitchin describe a homeomorphism $\mathcal{M}_1 \cong \mathbb{R}^3 \times S^1$ by associating to a monopole a notion of its center (in $\mathbb{R}^3$) and its phase angle (viewed as an element of S^1.) See [**?**] for details. In any case there is a natual homotopy equivalence between $\mathcal{M}_1 \cong \mathrm{Rat}_1$ and the circle S^1. Thus we may view the inclusion

$$\mathcal{M}_1 \hookrightarrow \mathcal{B}_1 \quad \text{or equivalently} \quad \mathrm{Rat}_1 \hookrightarrow \Omega^2_1 S^2$$

as an element of the homotopy group

$$\pi_1(\Omega^2 S^2) \cong \pi_3 S^2 \cong \mathbb{Z}.$$

It was verified in [**?**] that this element generates this group. Hence the inclusion map $\mathcal{M}_1 \hookrightarrow \mathcal{B}_1$ is an amusing geometric way of viewing the Hopf map $S^3 \longrightarrow S^2$.

20.2 Rational functions and braids

The first result concerning the homotopy type of the pair $(\mathcal{B}_k, \mathcal{M}_k)$, or equivalently the pair $(\Omega^2_k S^2, \mathrm{Rat}_k)$ is a result of Segal [**?**] analogous to the Taubes stability theorem for instantons (Theorem 19.7). Like in the instanton setting, in order to state the theorem we need to describe a notion of gluing of rational functions. It turns out that the Taubes gluing procedure for connections described in chapter X can be applied to time invariant connections and does define a notion of gluing of monopoles. However viewed in terms of rational functions the gluing procedure is much more conceptual. This can be stated in the following terms.

Lemma 20.5 *There exists a "gluing map"*

$$\mu : \mathrm{Rat}_k \times \mathrm{Rat}_r \longrightarrow \mathrm{Rat}_{k+r}$$

so that the following diagram homotopy commutes:

$$\begin{array}{ccc} \mathrm{Rat}_k \times \mathrm{Rat}_r & \xrightarrow{\mu} & \mathrm{Rat}_{k+r} \\ \cap\downarrow & & \downarrow\cap \\ \Omega^2_k S^2 \times \Omega^2_r S^2 & \xrightarrow[\sigma]{} & \Omega^2_{k+r} S^2. \end{array}$$

where σ is the loop sum operation.

Proof: Let $\alpha \in \mathrm{Rat}_k$ and $\beta \in \mathrm{Rat}_r$. Since rational functions are determined by their roots and poles, α and β can be viewed as being given by configurations of points in the complex plane $\mathbb{C}$ each labelled according to whether it is a root or a pole and by a positive integer determining the multiplicity. Let α' be the configuration of roots and poles in the upper half plane $\mathbb{C}_+ = \{z : Im(z) > 0\}$ given by the image of the roots and poles of α under a fixed diffeomorphism of $\mathbb{C}$ with $\mathbb{C}_+$. Similarly let β'' be the configuration of roots and poles in the lower half plane $\mathbb{C}_- = \{z : Im(z) < 0\}$ given by the image of the roots and poles of β under a fixed diffeomorphism of $\mathbb{C}$ with $\mathbb{C}_-$. Then $\alpha' \cup \beta''$ is a configuration of roots and poles that determines a rational function of degree $k + r$. This procedure defines the map

$$\mu : \mathrm{Rat}_k \times \mathrm{Rat}_r \longrightarrow \mathrm{Rat}_{k+r}$$

which is clearly homotopy compatible with the loop sum operation in the loop space. □

Gluing with a fixed element in Rat_1 defines an inclusion map

$$j : \mathrm{Rat}_k \hookrightarrow \mathrm{Rat}_{k+1}$$

compatible with the homotopy equivalence $j : \Omega^2_k S^2 \xrightarrow[\simeq]{} \Omega^2_{k+1} S^2$ given by taking the loop sum with the identity map (or any fixed map of degree one.) If

$$\mathrm{Rat}_\infty = \lim_{k\to\infty} \mathrm{Rat}_k$$

then there is an induced map (well defined up to homotopy) $j : \mathrm{Rat}_\infty \to \Omega^2_0 S^2$, or equivalently a map

$$j : \mathbb{Z} \times \mathrm{Rat}_\infty \longrightarrow \Omega^2 S^2.$$

In [**?**] Segal proved the following. Compare this theorem to the Taubes stability Theorem 19.5.

Theorem 20.6 *The map*

$$j : \mathbb{Z} \times \mathrm{Rat}_\infty \longrightarrow \Omega^2 S^2$$

is a homotopy equivalence.

This result says that the rational function spaces Rat_k define a filtration of the homotopy type of $\Omega^2 S^2$. But also recall Segal's previous result (Theorem 19.4) that the stable homotopy type of $\Omega^2 S^2$ is filtered by the configuration spaces $C_2(k)$. Now the homology and other homotopy invariants of these configuration spaces are well understood. Therefore the goal of the work in [**?**] was to compare the rational function filtration with the configuration space filtration of this loop space. Before we state the results of this work we pause to explain the relationship mentioned in chapter X between these configuration spaces and the braid groups.

Consider an element $b \in \pi_1(C_2(k))$. b is represented by a loop of configurations of k unordered points in the complex plane $\mathbb{C}$. That is, b can be viewed as a one parameter family $\{b_t : 0 \le t \le 1\}$, where each b_t is a set of k distinct points in the plane, with $b_0 = b_1 = \{1, 2, \ldots k\} \subset \mathbb{C}$. Pictorially, b can be viewed as follows.

For obvious reasons $\pi_1(C_2(k))$ is referred to as the *braid group* on k strings. We write this group as β_k. Observe that each β_k is an infinite group. Indeed if $\sigma_i \in \beta_k$ is the braid that crosses the i^{th} string over the $(i+1)^{st}$ string as pictured below, then β_k is is the group generated by the elements $\sigma_1, \ldots, \sigma_{k-1}$ subject to the relations

1. $\sigma_i\sigma_j = \sigma_j\sigma_i$ if $|i-j| \geq 2$
2. $\sigma_i\sigma_{i+1}\sigma_i = \sigma_{i+1}\sigma_i\sigma_{i+1}$

The reader is encouraged to verify these relations by drawing pictures. Notice in particular that $\beta_2 \cong \mathbb{Z}$ and that for any k there is a natural surjection onto the symmetric group $\beta_k \longrightarrow \Sigma_k$ obtained by sending a braid to the associated permutation of the endpoints.

The following is a well known result, which we already referred to in chapter 19.

Proposition 20.7 *The configuration space $C_2(k)$ is an Eilenberg–MacLane space of type $K(\beta_k, 1)$.*

Proof: Since $\pi_1 C_2(k) = \beta_k$ it remains to show that $C_2(k)$ is aspherical (that is, its higher homotopy groups are zero). Now consider the projection map from the configuration space of *ordered* points in $\mathbb{R}^2$

$$F(\mathbb{R}^2, k) \longrightarrow F(\mathbb{R}^2, k)/\Sigma_k = C_2(k).$$

Since the symmetric group Σ_k acts freely on $F(\mathbb{R}^2, k)$, this is a covering space. Hence it suffices to prove that $F(\mathbb{R}^2, k)$ is aspherical.

Let $R_q = \mathbb{R}^2 - \{q\}$ be the plane with q distinct points removed. We will prove that each of the spaces $F(R_q, k)$ is aspherical. We prove this by induction on k. For $k = 1$ $F(R_q, 1) = R_q$ which has the homotopy type of a bouquet (wedge) of q circles. Such spaces are aspherical with free fundamental group. Now inductively assume that for all $j < k$ $F(R_q, j)$ is aspherical for every q. We now prove that $F(R_q, k)$ is aspherical.

An element of $F(R_q, k)$ is a k-tuple of points in $R_q = \mathbb{R}^2 - \{q\}$. Let

$$p : F(R_q, k) \longrightarrow R_q$$

be the projection onto the first coordinate. This is easily seen to be a fibration with fiber $F(R_{q+1}, k-1)$. By the inductive assumption this fiber is aspherical. The base space R_q is also aspherical. Hence by the homotopy exact sequence of the fibration, the total space $F(R_q, k)$ is aspherical. □

This result, together with Theorem 19.4 implies that there is a map

$$\alpha : \mathbb{Z} \times K(\beta_\infty, 1) \longrightarrow \Omega^2 S^2$$

which induces an isomorphism in homology, and hence is a stable homotopy equivalence. Combining this with Theorem 20.6 we have the following.

Corollary 20.8 *The limiting spaces* Rat_∞ *and* $K(\beta_\infty, 1)$ *are stably homotopy equivalent.*

The following is the main result of [**?**].

Theorem 20.9 *The spaces* Rat_k *and* $K(\beta_{2k}, 1)$ *are stably homotopy equivalent. That is, there is a homotopy equivalence between the suspension spaces*

$$\Sigma^N \mathrm{Rat}_k \simeq \Sigma^N K(\beta_{2k}, 1)$$

for N *sufficiently large.*

This theorem gives an effective description of the stable homotopy type of the rational function spaces Rat_k and hence the monopole spaces $\mathcal{M}_k$ since quite a bit is known about the stable homotopy type of the braid groups. They have been used in many ways in topology. We recommend many of the articles of [**?**] and the survey article [**?**] for discussions about these applications.

The following consequence of Theorem 20.9 is the analogue of the Atiyah–Jones conjecture in the setting of monopoles and gives an answer to the question raised in section one.

Corollary 20.10 *The inclusion* $\mathcal{M}_k \hookrightarrow \mathcal{B}_k$ *induces a monomorphism in homology and is an isomorphism through dimension* $2k$*.*

Proof: By theorems 20.4 and 20.9 the the homomorphism

$$H_*(\mathcal{M}_k) \longrightarrow H_*(\mathcal{B}_k) \cong H_*(\Omega^2 S^2)$$

is given by the homomorphism

$$H_*(K(\beta_{2k}, 1)) \longrightarrow H_*(K(\beta_\infty, 1)) \cong H_*(\Omega^2 S^2)$$

which, by 20.7 is given by

$$\alpha : H_*(C_2(2k)) \longrightarrow H_*(\Omega^2 S^2).$$

The result now follows from Theorem 19.5. □

We remark that the *unstable* homotopy type of these rational functions has been studied in detail in [**?**]. Notice furthermore that Theorem 20.9 points to the possibility that many algebraic invariants of the braid groups will lead to interesting invariants of the spaces of monopoles. A particular example of this phenomenon was studied in [**?**] where it was seen that a certain class of representations of the braid groups give rise to families of elliptic operators coupled to monopoles. However the full extent of this relationship has yet to be understood.

We end this chapter by outlining the structure of the proof of Theorem 20.9. The reader is referred to [**?**] for details.

The proof revolves around the study of the subspace Rat_k^0 of Rat_k consisting of those rational functions with k distinct simple poles; this is the subspace of

generic rational functions. If $f = p/q \in \mathrm{Rat}_k^0$ then q is a monic, degree k polynomial with k distinct roots, and p is determined by its values at the roots of q. Notice that these values all must be nonzero since p and q have no roots in common. This gives a description of Rat_k^0 as

$$\mathrm{Rat}_k^0 \cong F(\mathbb{C}, k) \times_{\Sigma_k} (\mathbb{C}^*)^k$$

where the coordinates in $F(\mathbb{C}, k)$ correspond to the roots of the denominator and the coordinates of $(\mathbb{C}^*)^k$ are the values of the numerator at these poles. Now notice that there is a fibration

$$(\mathbb{C}^*)^k \longrightarrow \mathrm{Rat}_k^0 = F(\mathbb{C}, k) \times_{\Sigma_k} (\mathbb{C}^*)^k \longrightarrow F(\mathbb{C}, k)/\Sigma_k = C_2(k) \simeq K(\beta_k, 1).$$

Since the fiber $(\mathbb{C}^*)^k \simeq (S^1)^k$ is the classifying space of $\mathbb{Z}^k$, this fibration allows us to deduce that Rat_k^0 is the classifying space $B\beta_{2,k}$ of the semi-direct product

$$\beta_{2,k} = \beta_k \ltimes (\mathbb{Z})^k$$

where β_k acts on $(\mathbb{Z})^k$ through the homomorphism $\beta_k \longrightarrow \Sigma_k$ by permuting factors. This group $\beta_{2,k}$ can be thought of as the group of *framed braids* and it is naturally a subgroup of the braid group β_{2k}. The inclusion $\beta_{2,k} \hookrightarrow \beta_{2k}$ is given by the "cabling" process which can be described as follows. Start with k pairs of pieces of string and twist the i^{th} pair n_i times, where n_i is the integer in the i^{th} coordinate of an element of $(\mathbb{Z})^k$. Now braid the k pairs according to the braid $b \in \beta_k$. This gives a braid on $2k$ strings and the map which sends $(b; n_1, \ldots n_k)$ to this braid is the inclusion $\beta_{2,k} \hookrightarrow \beta_{2k}$.

We now have the following diagram.

$$\begin{array}{ccc} \mathrm{Rat}_k^0 = K(\beta_{2,k}, 1) & \xrightarrow{\phi} & K(\beta_{2k}, 1) \\ {\scriptstyle\psi}\big\downarrow & \cap & \\ \mathrm{Rat}_k\,. & & \end{array}$$

The following result, which implies Theorem 20.9, was proved in [**?**] using stable splitting theory in loop spaces, as well as homological calculations.

Proposition 20.11 *For suitably large N there is a map of suspension spaces*

$$\sigma : \Sigma^N \mathrm{Rat}_k \longrightarrow \Sigma^N \mathrm{Rat}_k^0$$

so that the composition

$$\Sigma^N \mathrm{Rat}_k \xrightarrow{\sigma} \Sigma^N \mathrm{Rat}_k^0 \xrightarrow{\phi} K(\beta_{2k}, 1)$$

is a homotopy equivalence.

Chapter 21

Floer's "Instanton Homology"

In this chapter we describe an application of gauge theory, studied from a Morse theoretic point of view, to the theory of three dimensional manifolds. This application, due to A. Floer [**?**] defines a homology theory for a 3-manifold Y that is computed via a chain complex whose chains are generated by irreducible representations of the fundamental group $\pi_1(Y)$ in the Lie group SU(2), or equivalently flat connections on the trivial SU(2) bundle over Y. Its boundary homomorphisms are computed by counting the self dual connections on $Y \times \mathbb{R}$ satisfying certain dimensional and asymptotic data. The chain complex has many of the same formal properties as the Morse–Smale complex of a Morse function on a compact manifold. However, unlike in the compact case its homology is not the homology of the underlying manifold (in this case the space of connections on Y up to gauge equivalence) but it does give an invariant of the topology of Y (i.e. it does not depend on any choices of metric used in its definition). In the first section we discuss work of Casson, which, from the topological viewpoint motivates Floer homology, and then discuss the Chern–Simons functional; the function that plays the role of the Morse function in this theory. In section two we outline the construction of Floer's instanton homology and then discuss how the classifying space model studied in chapter 12 might be used to understand a deeper relationship between the spaces of instantons on $Y \times \mathbb{R}$, viewed as moduli spaces of flows, and the topology of Y.

21.1 Representations of the fundamental group and the Chern–Simons functional

Let Y be a closed, oriented homology three sphere. That is,

$$H_*(Y;\mathbb{Z}) \cong H_*(S^3;\mathbb{Z}).$$

This isomorphism is realized by the map $Y \longrightarrow S^3$ defined by identifying all points outside a small closed disk in Y to a point. Notice that such a map is not necesarily a homotopy equivalence since $\pi_1(Y)$ is not necessarily zero. (All we are assuming is that $H_1(Y) = \pi_1(Y)/[\pi_1, \pi_1] = 0$.)

An important invariant of such a manifold is the *"Milnor–Rochlin invariant"* $\mu(Y) \in \mathbb{Z}_2$ defined as follows. By standard results of low dimensional topology Y is the boundary of a four-dimensional spin-manifold

$$Y = \partial W^4$$

whose signature $\sigma(W)$ is divisible by 8. (See [**?**] for a discussion of these points.) We then have

$$\mu(Y) = \sigma(Y)/8 \pmod 2.$$

It is a standard fact that this number is independent of the choice of spin-manifold W.

Casson defined another invariant $\lambda(Y) \in \mathbb{Z}$ which is defined by counting (with appropriate sign) the number of conjugacy classes of irreducible representations of $\pi_1(Y) \in \mathrm{SU}(2)$ and dividing by two. He then proved that $\lambda(Y) = \mu(Y)$ (mod 2), which implies the following.

Theorem 21.1 *If $\mu(Y) \neq 0$ then there exists a nontrivial representation*

$$\rho : \pi_1(Y) \longrightarrow \mathrm{SU}(2).$$

In particular if Y is homotopy equivalent to S^3 the $\mu(Y) = 0$.

Perhaps the most famous conjecture in Topology, the Poincare conjecture, states that any closed, simply connected 3-manifold is homeomorphic to S^3. It is clear that $\mu(S^3) = 0$ since in the definition of μ we could take $W = D^4$ whose signature is zero. Casson's theorem, stating that the μ-invariant of any homotopy 3-sphere is also zero, says that the μ-invariant will be of no use in trying to find a counter-example to the Poincare conjecture; a fact that had not been known until that time.

To make the definition of Casson's invariant precise one uses the notion of Heegard splittings of 3-manifolds and the theory of intersections of the resulting spaces of representations of the associated fundamental groups. Describing this theory will take us rather far afield, so we refer the reader to the exposition of this theory by Akbulut and McCarthy [**?**]. However there is an interesting relationship between Casson's invariant and a homology theory, developed by Floer [**?**], which is defined using Morse theory, based on a certain vector field on the space of connections on the trivial SU(2)-bundle on Y.

For now let Y be any closed, oriented 3-manifold and G any compact, simple Lie group. Let $\mathcal{A} = \mathcal{A}(Y)$ be the space of connections on the trivial bundle $Y \times G \longrightarrow Y$. Let $A \in \mathcal{A}(Y)$. Now $\mathcal{A}(Y)$ is affine based on the vector space of one-forms with coefficients in the Lie algebra $\mathfrak{g}$, $\Omega^1(Y; \mathfrak{g})$. Hence there is an natural isomorphism between the tangent space and this vector space,

$$T_A\mathcal{A}(Y) \cong \Omega^1(Y; \mathfrak{g}).$$

Now give Y a Riemannian metric. Since Y is a 3-manifold, the space of two-forms $\Omega^2(Y;\mathfrak{g})$ acts on the space of one-forms $\Omega^1(Y;\mathfrak{g})$ by the rule

$$\alpha(\beta) = \int_Y \text{trace}(\alpha \wedge \beta) d\,vol.$$

Thus we may think of Lie algebra valued two-forms as cotangent vectors:

$$\Omega^2(Y;\mathfrak{g}) \hookrightarrow (\Omega^1(Y;\mathfrak{g}))^* \cong T^*_A(\mathcal{A}(Y)).$$

In particular the curvature form

$$F_A \in T^*_A(\mathcal{A}(Y))$$

and so the mapping

$$A \longrightarrow F_A \in T^*_A(\mathcal{A}(Y))$$

is a section of the cotangent bundle $T^*(\mathcal{A}(Y))$ and hence is a one-form

$$F \in \Omega^1(\mathcal{A}(Y)).$$

Notice that using the metric and the induced Hodge star operator we have that the vector field dual to the one-form F is given by

$$A \longrightarrow *F_A \in \Omega^1(Y;\mathfrak{g}) \cong T_A(\mathcal{A}(Y)).$$

Clearly the zeros of this vector field are given by the flat connections on $Y \times G$.

It is not difficult to see that the curvature F is a closed one-form on $\mathcal{A}(Y)$ and since $\mathcal{A}(Y)$ is contractible, it must be the differential of a function. Equivalently, the vector field $*F_A$ is the gradient vector field of a function on $\mathcal{A}(Y)$. This function is the Chern–Simons functional

$$\psi : \mathcal{A}(Y) \longrightarrow \mathbb{R}$$

defined by

$$\psi(A) = \int_Y \text{trace}(\frac{1}{2} A \wedge dA + \frac{1}{3} A \wedge A \wedge A)$$

where in this formula A is viewed as a one-form (which can be done in a canonical way on a trivial bundle).

The idea in Floer homology is to use the Chern–Simons functional ψ in order to define a Morse–Smale type complex whose homology is an interesting invariant of the underlying manifold Y. In particular the chains in this complex are generated by the critical points of ψ, which is to say the zeros of the gradient vector field $*F$; that is, the flat connections. Thus the Euler characteristic of this homology theory would be a way of counting flat connections, and indeed is related to Casson's invariant. (The fact that Casson's invariant could be defined gauge-theoretically in this manner was first proved by Taubes [**?**].)

Now as we shall see there are many technical problems dealt with in [**?**] in defining this theory. The first one that is encountered is the fact that ψ is not

quite gauge invariant. This is important because as seen in chapter 18 conjugacy classes of irreducible representations correspond to gauge equivalence classes of flat connections. In this setting, since we are dealing with a trivial principal bundle the gauge group is given by

$$\mathcal{G} = Map(Y, G).$$

Notice that the set of path components

$$\pi_0(\mathcal{G}) = [Y, G] \cong \pi_3 G \cong \mathbb{Z}$$

and we can think of the path component that an element $g \in \mathcal{G}$ lies in as its *degree*. (Notice that in the case $G = \mathrm{SU}(2) \cong S^3$ then this notion is the actual degree of the smooth map $g : Y \longrightarrow S^3$.) As seen before the gauge group acts on $\mathcal{A}(Y) \cong \Omega^1(Y; \mathfrak{g})$ by the rule

$$g(A) = g^{-1}Ag + g^{-1}dg \in \Omega^1(Y; \mathfrak{g}). \tag{21.1}$$

The behavior of the Chern–Simons functional under a gauge transformation is given by

$$\psi(g(A)) = \psi(A) + \tau_G \deg(g)$$

where τ_G is a constant depending only on the Lie group G. ($\tau_{\mathrm{SU}(2)} = 2\pi$.)

Thus, although ψ is not fully gauge invariant, it is invariant under the action of $\mathcal{G}_0 \subset \mathcal{G}$, the subgroup of gauge transformations of degree zero. Since $\mathcal{G}/\mathcal{G}_0 \cong \mathbb{Z}$ (given by degree), ψ induces a map of infinite cyclic covering spaces

$$\begin{array}{ccc} \mathcal{B}_0 = \mathcal{A}(Y)/\mathcal{G}_0 & \xrightarrow{\psi} & \mathbb{R} \\ \downarrow & & \downarrow \\ \mathcal{B} = \mathcal{A}(Y)/\mathcal{G} & \xrightarrow[\psi]{} & \mathbb{R}/\tau_G\mathbb{Z} \cong S^1. \end{array}$$

Another way of viewing the Chern–Simons functional and its behavior under gauge transformation is as follows (see for example [**?**]).

Let $A \in \mathcal{A}(Y)$ and consider the path of connections

$$A_t = (1 - t)A + tA_0$$

where A_0 is the trivial connection on $Y \times G$ and $t \in [0, 1] = I$. This path of connections defines a connection $\bar{A}$ on the four-manifold with boundary, $Y \times I$. An alternative definition of the Chern–Simons functional is given by

$$\psi(A) = \frac{1}{8\pi^2} \int_{Y \times I} trace(F_{\bar{A}} \wedge F_{\bar{A}}).$$

Now let $g \in Map(Y, G) = \mathcal{G}$ be a gauge transformation. In order to compare $\psi(g(A))$ with $\psi(A)$ we study the closed manifold $Y \times S^1$. More precisely, let I_+ and I_- denote the upper and lower semi- circles of S^1 so that

$$I_+ \cap I_- = S^0 = \{-1, 1\} \subset S^1 \subset \mathbb{C}.$$

Consider the G-bundle $P \longrightarrow Y \times S^1$ that is trivial on $Y \times I_+$ and $Y \times I_-$ and is therefore defined by the clutching function

$$Y \times S^0 = Y \times \{-1\} \sqcup Y \times \{1\} \longrightarrow G$$

which maps $Y \times \{-1\}$ to $1 \in G$ and which when restricted to $Y \times \{1\}$ is given by $g : Y \longrightarrow G$. The degree of the map g is, up to multiplication by a constant (depending only on the Lie group G) the Chern class $c_2(ad(P))$ of the corresponding adjoint vector bundle. (If $G = \mathrm{SU}(n)$ the degree of G equals the Chern class of $P \longrightarrow Y \times S^1$)

Now the connections $\bar{A}$ on $Y \times I_+$ and $g(\bar{A})$ on $Y \times I_-$ fit together to give a connection A_g on $P \longrightarrow Y \times S^1$. Moreover, we have that

$$\psi(A) - \psi(g(A)) = \frac{1}{8\pi^2} \int_{Y \times S^1} trace(F_{A_g} \wedge F_{A_g}).$$

But this quantity is well known to give the second Chern class, (see [**?**]) and hence is determined by the degree of g.

Another technical issue dealt with in [**?**] in the attempt to view the Chern–Simons functional ψ as a Morse function, is the fact that the underlying space $\mathcal{B} = \mathcal{A}(Y)/\mathcal{G}$ is not a manifold. The reason for this is because the gauge group $\mathcal{G}$ does not act freely on $\mathcal{A}(Y)$. Actually by (21.1) it is apparent that the center of G always acts trivially on any connection. However the singularities $\mathcal{B}$ arise from *reducible* connections, that is those whose connections A whose isotropy subgroup

$$G_A = \{g \in \mathcal{G} : g(A) = A\}$$

is larger than the center. Alternatively, these are connections $A \in \mathcal{A}(Y)$ that arise from a connection on $Y \times H$ where H is a proper subgroup of G. This is a rather difficult problem to deal with in general, but in the case when $G = \mathrm{SU}(2)$ and Y a homology 3-sphere the situation is much easier. In particular the center of $\mathrm{SU}(2)$ is $\mathbb{Z}_2 = \{\pm 1\}$. Moreover the irreducible connections form an open and dense set $\mathcal{B}^*$ in $\mathcal{B}$. The reducible flat connections correspond to representations

$$\rho : \pi_1(Y) \longrightarrow \mathrm{SU}(2)$$

whose image lies in a proper subgroup of $\mathrm{SU}(2)$. But any such subgroup is a subgroup of $U(1) < \mathrm{SU}(2)$, and hence is abelian. Thus any such representation factors through a representation of the abelianization

$$\pi_1/[\pi_1, \pi_1] \cong H_1(Y)$$

which is zero precisely when Y is a homology 3-sphere. Hence in this case, up to gauge equivalence, there is only one reducible flat connection (representation), the trivial connection. This is one of the main reasons that as of this point in time the details of Floer homology have only been fully worked out in the case $G = \mathrm{SU}(2)$ and Y a homology 3 - sphere. We will assume we are in this situation in the rest of this chapter.

21.2 Instantons as flows

Perhaps the most difficult problem in trying to do classical Morse theory with the Chern–Simons functional is the fact that the indices at the critical points of ψ (i.e. the gauge equivalence classes of flat connections) have infinite index. That is, the Hessian of ψ at a flat connection has infinite dimensional negative (and positive) eigenspaces. Equivalently, the unstable and stable manifolds of the gradient vector field $\nabla\psi = *F$ are infinite dimensional. The saving fact, however, is that the *relative* indices between any two critical points is finite. To understand this phenomenon properly, we consider the flow equations.

Let a and b be flat connections and so represent critical points of ψ. A curve $\gamma : \mathbb{R} \longrightarrow \mathcal{B}$ is a gradient flow between a and b if

$$\lim_{t\to-\infty} \gamma(t) = a \quad \text{and} \quad \lim_{t\to\infty} \gamma(t) = b$$

and

$$\frac{d\gamma}{dt} = -\nabla_{\gamma(t)}(\psi) = - * F_{\gamma(t)}.$$

Now one can view a curve of gauge equivalence classes of connections $\gamma : \mathbb{R} \longrightarrow \mathcal{B}$ going between flat connections a and b as a gauge equivalence class of connection $\bar{\gamma}$ on the trivial bundle over $Y \times \mathbb{R}$ which, when viewed as a one-form is trivial in the $\mathbb{R}$ direction, and satisfies the asymptotic conditions that as $t \to \pm\infty$, $\bar{\gamma}$ approaches the flat connections b and a respectively. A direct calculation, comparing the curvatures of of the connections $\gamma(t)$ on Y at each t, with the curvature of the connection $\bar{\gamma}$ on the four manifold $Y \times \mathbb{R}$ one verifies the following. (See [**?**] for details.)

Theorem 21.2 *A curve $\gamma : \mathbb{R} \longrightarrow \mathcal{B}(Y)$ going between the flat connections a and b satisfies the flow equation*

$$\frac{d\gamma}{dt} = - * F_{\gamma(t)}$$

if and only if the connection $\bar{\gamma}$ on the 4-manifold $Y \times \mathbb{R}$ satisfies the self duality equation

$$F_{\bar{\gamma}} = *F_{\bar{\gamma}}.$$

Any connection on the trivial bundle $Y \times \mathbb{R} \times \mathrm{SU}(2)$ is gauge equivalent to one that is trivial in the $\mathbb{R}$ - direction. Hence, using the language of chapters 6–12 we have the following.

Corollary 21.3 *Let a and b be flat connections on $Y \times \mathrm{SU}(2)$ and so represent critical points of the Chern–Simons functional ψ. Then the "moduli space of flows" $\mathcal{M}(a, b)$ is equal to the moduli space of gauge equivalence classes of self-dual connections (instantons) on $Y \times \mathbb{R} \times \mathrm{SU}(2)$ which in the sense described above, asymptotically approach the flat connections a and b.*

Now recall from chapter 6 that in the case of a Morse–Smale function on a compact manifold, the dimension of the moduli space of flows between critical points is one less than the relative index

$$dim\, \mathcal{M}(a,b) = \lambda(a) - \lambda(b) - 1.$$

where λ denotes the index of the critical point.

In the case of the Chern–Simons functional pertubations in both the functional and the metric need to be performed before it satisfies the analogue of the Morse–Smale conditions (i.e. the nondegeneracy of the critical points and the transversality of the intersections of the unstable and stable manifolds). In any case, the dimensions of the instanton spaces $\mathcal{M}(a,b)$, where now a and b are flat connections give the notion of the relative index of a and b.

Now the dimension of $\mathcal{M}(a,b)$ can be described as the index of a certain elliptic differential operator. (Recall how in chapter 12 we saw that the dimension of the space of flows between critical points a and b of a Morse function on a compact manifold was given by the index of the operator we called $D\mathcal{S}$ defined on the tangent space to the space of paths $P_{a,b}$.) Furthermore this index can be computed by the Atiyah–Singer index theorem and was done so in [**?**]. It turns out that the instanton moduli space $\mathcal{M}(a,b)$ has many connected components, indexed by the integers, which is related to the "charge" or"degree" of an instanton on a compact four-manifold as described in chapter 17. However the dimensions of the various components of $\mathcal{M}(a,b)$ turn out to be congruent mod 8. Hence the "dimension" of $\mathcal{M}$ is well defined mod 8 and Floer defines the "relative index", $\lambda(a,b) \in \mathbb{Z}_8$ of two flat connections a and b by

$$\lambda(a,b) = dim\, \mathcal{M}(a,b) + 1 \pmod 8.$$

In particular given any flat connection a, Floer defines the "mod 8 - index" $\lambda(a) \in \mathbb{Z}_8$ to be the relative index of a and the trivial connection θ on $Y \times \mathrm{SU}(2)$,

$$\lambda(a) = \lambda(a,\theta) \in \mathbb{Z}_8.$$

Floer then uses these invariants to define the analogue of the Morse–Smale chain complex. For $p \in \mathbb{Z}_8$ he defines C_p to be the free abelian group generated by those gauge equivalence classes of flat connections (critical points of ψ) with $\lambda(a) = p$. Analogous to the Morse–Smale setting, he then defines a boundary homomorphism

$$\partial : C_p \longrightarrow C_{p-1}$$

by counting (with sign) numbers of flows. That is, if $a \in C_p$ and $b \in C_{p-1}$, then the relative index $\lambda(a,b) = 1 \pmod 8$, and so the dimension of the various components of $\mathcal{M}(a,b)$ are congruent to zero mod 8. The zero dimensional component is finite and so the coefficient $\langle \partial a, b \rangle$ is defined to be the number (counted with sign) of instantons in this component. Analytically, this number is the "spectral flow" from a to b, which is to say the number of negative eigenvalues of the Hessian of ψ at a that in a sense that can be made precise, "cross over" along a flow from a to b to become positive eigenvalues of the

Hessian of ψ at b. We refer the reader to [**?**] for details. The outcome of this analysis was the following.

Theorem 21.4 *The boundary homomorphisms $\partial : C_p \longrightarrow C_{p-1}$ satisfy*

$$\partial^2 = 0$$

and the induced $\mathbb{Z}_8$-graded chain complex

$$\ldots \xrightarrow{\partial} C_p \xrightarrow{\partial} C_{p-1} \xrightarrow{\partial} \ldots$$

has homology groups $IH_p(Y)$, for $p \in \mathbb{Z}_8$ which do not depend on the choices of metrics or pertubations made in the definitions. That is, the isomorphism classes of the groups $IH_*(Y)$ are invariants of the topology of Y.

We remark that unlike in the case of a Morse function on a compact manifold, where the homology of the Morse–Smale chain complex gives the homology of the manifold, these "instanton homology" groups do not reflect the homology of the manifold ($\mathcal{B}^* \subset \mathcal{B}$) upon which the Chern–Simons functional is defined. Nonetheless one might speculate that the classifying space theory developed in chapter 12 might lead to a way of understanding the relationship. This would be important because as it stands, the Floer homology groups are quite difficult to compute, and it would be helpful to understand these invariants better, either from a calculational or a qualitative point of view.

To make these ideas precise one needs to define a "Floer category" $\mathcal{C}(Y)$ whose objects are gauge equivalence classes of flat connections (or equivalently conjugacy classes of SU(2) representations of $\pi_1(Y)$) and whose morphisms between flat connections a and b is a suitable completion $\bar{\mathcal{M}}(a, b)$ of the moduli space of self dual connections on $Y \times \mathbb{R}$ which asymptotically approach a and b. One would need an appropriate refinement of Taubes gluing of instantons in order to define an associative composition pairing in this category. We remark that the following version of Taubes' gluing was used by Floer in his work, in particular to prove that $\partial^2 = 0$. (Compare Theorem 9.2.)

Proposition 21.5 *If $a_0, \ldots a_n$ are irreducible flat connections on $Y \times \mathrm{SU}(2)$ then for every compact subspace of the product of the moduli spaces*

$$K \subset \mathcal{M}(a_0, a_1) \times \cdots \times \mathcal{M}(a_{n-1}, a_n)$$

there is an $\epsilon > 0$ and a smooth map

$$\mu : K \times [0, \epsilon)^n \longrightarrow \mathcal{M}(a_0, a_n)$$

which is a diffeomorphism onto its image.

If one could refine this theorem so as to define an associative pairing and therefore a category along the lines done for compact manifolds in chapter 12, then one might expect to have sufficient data necessary for analyzing how the

moduli spaces build up the topology of the space of connections $\mathcal{B}$. In particular one might expect that an appropriate classifying space of this category is (at least) homotopy equivalent to $\mathcal{B}$ (or perhaps the homotopy orbit space $E\mathcal{G} \times_{\mathcal{G}} \mathcal{A} \simeq B\mathcal{G}$) that is filtered in a way analogous to the index filtration done in chapter 14, section 14.2. This should lead to a spectral sequence converging to $H_*(\mathcal{B})$ or $H_*(B\mathcal{G})$ which has E_2-term a variant of Floer homology (i.e. the homology of a Morse–Smale type complex. Compare Theorem 14.2.) Such a description would give a systematic way of understanding how the higher dimensional moduli spaces "bridge the gap" between the Floer type invariants (which only use the zero dimensional instanton spaces in their definition) and the full space of gauge equivalence classes on Y, a space whose homotopy type is relatively easy to understand. This would begin to lead to an understanding of a basic question this theory brings up. How well does the homotopy type of the (higher dimensional) moduli spaces of instantons on $Y \times \mathbb{R}$ that are asymptotically flat reflect the diffeomorphism type of the manifold Y?