Morse Theory and the Morse-Smale-Witten Complex

1 Introduction and Basic Definitions

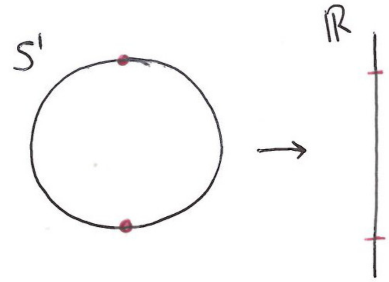
We give a brief introduction to Morse theory via the Witten approach of unstable manifolds. Throughout these notes, M will denote a smooth manifold, compact and without boundary. The basic idea of Morse theory is as follows: we fix a smooth function $f : M \to \mathbb{R}$ which we think of as a "height function" on M, and then use differential information about f to study the manifold M.

The incredible fact is that this differential data is enough to *completely determine* the topology of M (in a very explicit way), at least as long as we choose f sufficiently "generic". This leads to a number of extremely deep results (see for instance Theorem 2.5).

1.1 Critical Points

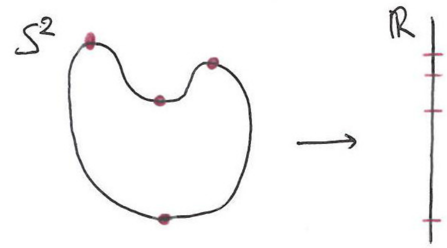
We begin with some basic definitions. We say that a point $p \in M$ is a **critical point** for f if $(\mathrm{d}f)_p = 0$. In co-ordinates this just means that all the first partial derivatives of f vanish.

Example 1.1. Consider the height function on the circle embedded in $\mathbb{R}^2$, given by the map $(x, y) \mapsto y$:



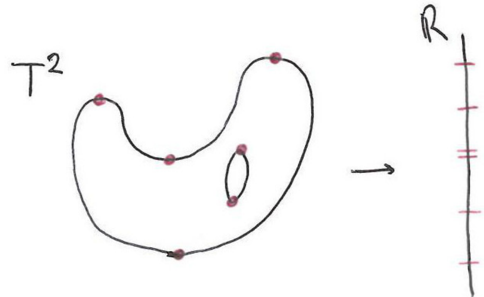
This has two critical points, corresponding to the minimum and maximum of the function.

Example 1.2. Similarly, consider the height function on the following embedded submanifold $S^2 \subseteq \mathbb{R}^3$:



This has four critical points: two local maxima, one saddle point and one minimum.

Example 1.3. Finally consider the height function on the following torus embedded in $\mathbb{R}^3$:



This has six critical points: one minimum, three saddle points and two local maxima.

We aim to carry out a systematic analysis of these critical points, and to be able to do this, we need to introduce some more vocabulary.

A critical point p is called **nondegenerate** if $\mathrm{d}f$ has a simple zero at p (meaning that f vanishes to first order but not higher). In co-ordinates $(x_1, \ldots, x_n)$, this means that the Hessian matrix of partial second derivatives is nonsingular:

$$\det \begin{pmatrix} \dfrac{\mathrm{d}^2 f}{\mathrm{d}x_1 \mathrm{d}x_1} & \cdots & \dfrac{\mathrm{d}^2 f}{\mathrm{d}x_1 \mathrm{d}x_n} \\ \vdots & & \vdots \\ \dfrac{\mathrm{d}^2 f}{\mathrm{d}x_n \mathrm{d}x_1} & \cdots & \dfrac{\mathrm{d}^2 f}{\mathrm{d}x_n \mathrm{d}x_n} \end{pmatrix} \neq 0$$

Geometrically, it means that the graph of $\mathrm{d}f$ near p (viewed as a submanifold of the total space of T^*M) intersects the zero section $M \subseteq T^*M$ transversely.

A function f is called a **Morse function** if all its critical points are nondegenerate. From now on we will assume all our functions are Morse. We do not lose much generality here: it can be shown that, in some well-defined sense, *almost all* functions are Morse (see §1.2 of [**?**]).

1.2 Morse Lemma and the Index of a Critical Point

Yet another characterisaton of nondegenerate critical points is given by the following result.

Proposition 1.4 (Morse Lemma)**.** Let p be a critical point of f. Then p is nondegenerate if and only if there exist local co-ordinates $x_1, \ldots, x_n$ centred at p with respect to which we have

$$f(x) = f(p) - x_1^2 - \ldots - x_k^2 + x_{k+1}^2 + \ldots + x_n^2$$

where k is some fixed integer, independent of the choice of charts, which we call the **index** of the critical point p.

The index should be thought of as the number of independent directions in which the function "descends" away from p. For instance, the index of a maximum point is n (the dimension of the manifold) and the index of a minimum point is 0. On a surface a saddle point has index 1.

More invariantly, if we view the Hessian matrix as a symmetric bilinear form on T_pM, then the index is the dimension of the maximal subspace on which the Hessian is negative-definite (see [**?**] §2 for more details).

A straightforward consequence of the previous proposition is:

Lemma 1.5. Nondegenerate critical points are always isolated.

Hence if f is a Morse function its set of critital points is discrete; and in fact since M is compact there are only finitely many. This will turn out to be important later on.

2 Topology of Sublevel Sets

With these preliminaries taken care of, we turn our attention to the so-called **sublevel sets** $f^{-1}(-\infty, r]$. In order to compare these for different values of r we will make use of the gradient flow of f, which we now explain.

Let us fix a Riemannian metric $\langle -, - \rangle$ on M (it turns out that the precise choice of metric does not affect the resulting theory). We can then define a vector field $\mathrm{grad} f$ on M, related to $\mathrm{d}f$ via the identity

$$\langle \mathrm{grad} f, V \rangle = \mathrm{d}f(V) = V(f)$$

for all vector fields V.

Remark 2.1. Intuitively this is just the ordinary vector field $\mathrm{grad} f = \nabla f$ that applied mathematicians and physicists know and love. In particular, $(\mathrm{grad} f)_p$ points in the direction of greatest increase of f at p.

We can then define the **gradient flow** associated to $-\mathrm{grad} f$. This exists for all time because M is compact (the reader unfamiliar with flows can find a full treatment in §9 of [**?**]).

Remark 2.2. Continuing to think of f as a height function on M, the flow associated to $-\text{grad}f$ moves "downwards" while that associated to $\text{grad}f$ moves "upwards". This perhaps goes some way to explaining the convention of taking the *negative* gradient flow: our gravity-shackled minds are accustomed to seeing things flow down, not up.

By pushing down along flow lines, we obtain our first big result.

Theorem 2.3. Suppose that $f^{-1}[r,s]$ contains no critical points of f. Then the sublevel sets $f^{-1}(-\infty,r]$ and $f^{-1}(-\infty,s]$ are homeomorphic.

Proof sketch. Consider the diffeomorphisms ϕ_t of M, given by translating along the flow lines for time t. The restriction of ϕ_t to $f^{-1}(-\infty,s]$ is certainly a homeomorphism onto its image. If we take $t = s - r$, then we conclude from the information we know about ϕ_t' that im $\phi_t = f^{-1}(-\infty,r]$, and this completes the proof (we use the fact that ϕ_t' is nonvanishing in this region, which of course only holds if the region does not contain a critical point). □

Aside. In fact, each sublevel set carries naturally the structure of a smooth manifold with boundary, and then the above homemorphism turns out to be a diffeomorphism.

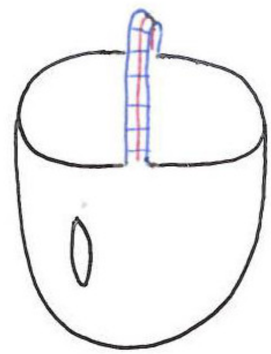
Consider now what happens to the topology of $f^{-1}(-\infty,r]$ as we gradually increase r. If r is small enough we have the empty set, and if r is big enough we have the whole manifold M. We're interested in what happens in between. From the previous result, we know that the topology can only change if r crosses a critical value of f.

So suppose that p is a critical point of f, with index k. Write $r = f(p)$, and consider the sets $f^{-1}(-\infty,r-\epsilon]$ and $f^{-1}(-\infty,r+\epsilon]$ where ϵ is sufficiently small. We want to know how to obtain the second space from the first.

Using the Morse Lemma, we know that f can be written locally as

$$f(x) = r - x_1^2 - \ldots - x_k^2 + x_{k+1}^2 + \ldots + x_n^2$$

and so near p the manifold M has a k-dimensional space of "downward directions" and a complimentary $(n-k)$-dimemsional space of "upward directions". These first k directions give a k-disc which is attached to $f^{-1}(\infty,r-\epsilon]$ by its boundary; the other $(n-k)$ directions then contribute to a "thickening" of this disk into a "handle" attached to $f^{-1}(-\infty,r-\epsilon]$:



In the above figure $f^{-1}(-\infty, r-\epsilon]$ is drawn in black, the k-dimensional cell corresponding to the downward directions is drawn in red and the $(n-k)$-dimensional "thickening" is drawn in blue.

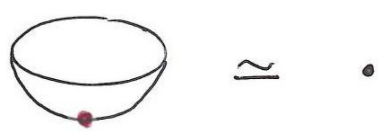
Theorem 2.4. Let p be a nondegenerate critical point of f of index k and write $f(p) = r$. Then $f^{-1}(-\infty, r+\epsilon]$ is homeomorphic to $f^{-1}(-\infty, r-\epsilon]$ together with a handle $D^k \times D^{n-k}$ attached along $S^{k-1} \times D^{n-k}$.

Note: If we only care about homotopy type, then

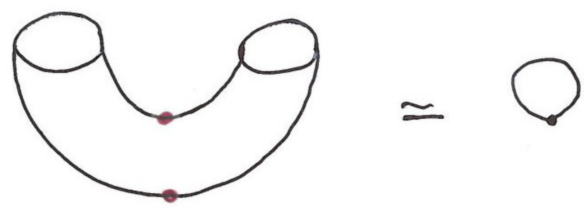
$$f^{-1}(-\infty, r+\epsilon] = f^{-1}(-\infty, r-\epsilon] \cup e^k$$

where e^k is a k-cell glued along its boundary.

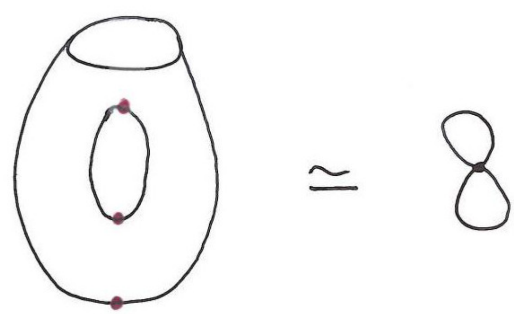
We can now build up the manifold M in stages by increasing r and considering $S_r = f^{-1}(-\infty, r]$. We begin with the empty set, and add in a k-cell for every index k critical point that we pass, until we end up with all of M. The example of the torus "standing on end" is illusrated below. We begin with the empty set, which becomes a 0-cell once we cross the minimum of the height function, d:



So when we consider S_r for $d \leq r < c$ we still have something homotopic to a 0-cell. The next critical point we pass is the lower saddle point, c. Beyond this point, S_r looks looks like a cylinder for $c \leq r < b$. Homotopically this is just a circle, and indeed a circle is obtained by attaching a 1-cell to a 0-cell. This is exactly what we expect as the index of point c is exactly one.



Passing the upper saddle point, b, we get a vase, or more formally a Riemann surface of genus one with connected (nonempty) boundary. Homotopically we have just added another 1-cell, the index of b, so we get the wedge sum of two circles. S_r is homotopic to this wedge sum of two circles for $b \leq r < a$.



Finally we pass the maximum, a, and S_a gives us the whole torus. This is obtained by gluing in a 2-cell along the boundary of the wedge sum of circles. To see more clearly how this works in detail, consider the representation of the torus as a quotient of a square.

These constructions can be carried out for *any* compact manifold equipped with a Morse function. This proves the following deep result:

Theorem 2.5. Every compact manifold is homotopy equivalent to a cell complex.

And given any Morse function f on M we have a quite explicit method for constructing the associated cellular decomposition. However, this is not the cell complex we are most interested in: there is an alternative cellular decomposition of M, called the *Morse–Smale–Witten complex*, which is more useful in modern applications.

3 The Witten approach

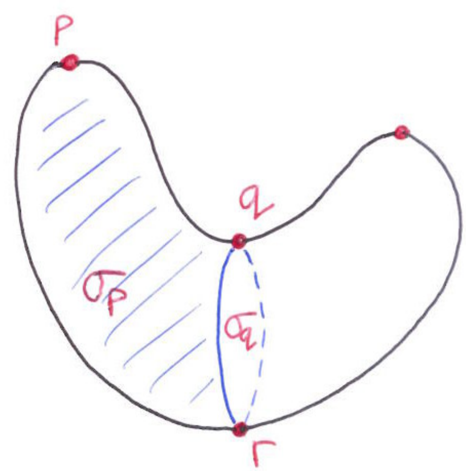
The Morse–Smale–Witten complex associates a cell to each critical point of f; intuitively this "downward cell" consists of the flow lines which "flow out" of the critical point. As we will see, these patch together to give a cell complex structure on M.

3.1 Unstable Manifolds and Stable Manifolds

If p is a critical point, we define the associated **unstable manifold** $W^u(p)$ by

$$W^u(p) = \{x \in M : \lim_{t \to -\infty} x_t = p\}$$

where $x_t = \phi(x, t)$ is the flow associated to $-\text{grad} f$ (as in the previous section). Thus $W^u(p)$ is the union of the flow lines which "flow down" out of p. For instance, on the S^2 which we saw in Example 1.2, we have:



This example illustrates a general fact about unstable manifolds, namely that each is homeomorphic to a disc with dimension equal to the index of the critical point (intuitively this makes sense, because the index counts the number of independent "downward directions" moving out of the point).

Theorem 3.1. $W^u(p) \cong \overset{\circ}{D}^k$ where $k = \mathrm{Ind}(p)$.

We also define for a critical point p, the **stable manifold** $W^s(p)$ to be

$$W^s(p) = \{x \in M : \lim_{t\to\infty} x_t = p\}$$

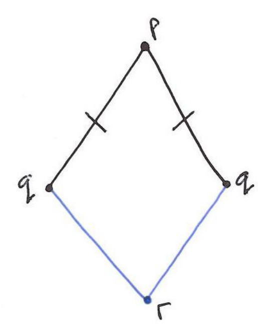
i.e the union of the flow lines which "flow down" into p. We see also that $W^s(p) \cong \overset{\circ}{D}^{n-k}$ where $k = \mathrm{Ind}(p)$ and $n = \dim(M)$.

We say that the data of our manifold M, the Morse function f and the metric $\langle\,,\,\rangle$ is **Morse–Smale** if for every pair of critical points $x, y \in \mathrm{Crit}(f)$, $W^u(x)$ is transverse to $W^s(y)$. This can always be achieved by small perturbations of the starting data.

Once we are in this situation, We consider the intersection $M(x, y) = W^u(x) \cap W^s(y)$, which by transversality is a submanifold of M. If we quotient by reparametrisation, we obtain the moduli space $M(x, y)/\mathbb{R}$ of flow lines from x to y. In order to understand how unstable manifolds glue to give a cell complex structure on M, we need to in particular understand the boundaries of these cells.

3.2 Boundaries of Unstable Manifolds

Coming back to Example 1.2 (illustrated in the above figure), consider in more detail $W^u(p)$. In order to see what happens at the boundary of this cell, we take a flow line from p to r and deform it towards the edge. What happens is that it eventually "snaps" to give a so-called **broken flow line** from p to r, which passes through the index 1 critical point q (see figure). (This isn't so much a flow line as it is two *separate* flow lines, appended together.) Depending on which direction we choose to deform in, we get two different broken flow lines, corresponding to the "front" and "back" of the manifold.



Taken together, the "tail ends" of these broken flow lines (obtained by throwing away those segments joining p and q) give $W^u(q)$. On the other hand, note that $\bar{W}^u(q) = \partial W^u(p)$.

This provides us with a nice way of thinking about the boundary of an unstable manifold $W^u(x)$: it consists of the "tail ends" of all the broken flow lines which begin at a critical point x and pass through a point of index $\mathrm{Ind}(x) - 1$.

Theorem 3.2. The boundary of an unstable manifold is given by

$$\partial W^u(p) = \sum_q n_{pq} \bar{W}^u(q)$$

where q runs over all critical points with $\mathrm{Ind}(q) = \mathrm{Ind}(p) - 1$ and n_{pq} is the number of flow lines from p to q, i.e the number of flowlines in $M(x, y)$ counted with orientations.

To see that the values n_{pq} are always finite, note that in general the space of flow lines from p to q has dimension

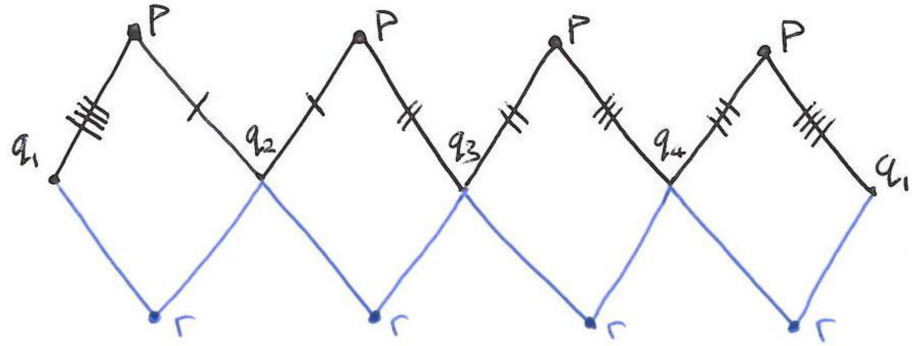
$$\mathrm{Ind}(p) - \mathrm{Ind}(q) - 1$$

and so in particular if $\mathrm{Ind}(q) = \mathrm{Ind}(p) - 1$ then the space is zero-dimensional, so we have a finite number of flow lines (for more details see §2.2 of [**?**]; you will see that we have to make an assumption here about the genericity of the metric). Of course there are only finite many critical points (see §1), so this sum is certainly well-defined.

3.3 The Morse-Smale-Witten Complex

We want the boundary maps we have just defined to fit into a chain complex, and so we must check that $\partial^2 = 0$.

Consider first the case $\mathrm{Ind}(p) = 2$. From our earlier discussion we know that $\partial W^u(p)$ consists of the "tail ends" of broken flow lines which start at p and pass through a critical point q of index $2 - 1 = 1$. Therefore schematically $W^u(p)$ looks like:



The boundary $\partial W^u(p)$ consists of all the lines in the bottom half of the diagram. The boundary of any two lines in the same diamond is 0 because the orientations cancel out; hence $\partial^2 = 0$ in each diamond seperately, and so $\partial^2 W^u(p) = 0$ as required. (Here we are using the fact that a compact 1-manifold with boundary is a union of circles and intervals, and hence its oriented boundary has degree 0.)

The general case is not so different, except we must replace the lines in the bottom half of the above diagram by $(k-1)$-discs. We won't say any more about this here: the interested reader is referred to §3.1 of [**?**].

Thus we get a chain complex, which we call the **Morse-Smale-Witten complex**. Its chain groups are

$$C_k = \bigoplus_p \mathbb{Z}$$

where p runs over all critical points of index k. The boundary maps are just given by the honest-to-god boundary maps described above (see Theorem 3.2), extending $\mathbb{Z}$-linearly.

The fact that $\partial^2 W^u(p) = 0$ for each generator $W^u(p)$ means that $\partial^2 = 0$, so we have a chain complex. Its homology is by definition the **Morse homology** of M with respect to f. In fact we have:

Theorem 3.3. Morse homology is isomorphic to singular homology.

In paticular, the Morse homology groups do not depend on the choice of f (although the chain groups definitely do).

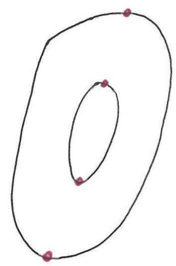
Remark 3.4. As we alluded to earlier, we can use the cells $W^u(p)$ to give M the structure of a cell complex. Viewed in this way, the chain complex for Morse homology is nothing but the chain complex for cellular homology.

Remark 3.5. Note that we could have defined the Morse-Smale-Witten complex without any reference to the unstable manifolds $W^u(p)$, by defining our chain groups

$$C_k = \bigoplus_p \mathbb{Z}_p$$

where p runs over all critical points of index k, and then *defining* the boundary map ∂ using the formula in Theorem 3.2. There is nothing mathematically incorrect about this approach, but using unstable manifolds gives a great deal more geometric intuition.

Exercise 1. Work out the Morse-Smale-Witten complex for the following example of the tilted torus:



Exercise 2. Work out the Witten complex for Example 1.3 seen earlier.

Exercise 3. Given a cycle in M, think about how to express this in terms of the Witten complex. In other words, construct the isomorphim between Morse homology and singular homology.

Remark 3.6. A lovely application of the Witten complex is to give an alternative proof of Poincaré duality: simply replace the functon f by $-f$. This replaces a k-dimensional downward cell by an $(n-k)$-dimensional transverse upward cell.