A GENTLE INTRODUCTION TO MORSE THEORY

ÁLVARO DEL PINO

Abstract. These are the notes for the Morse theory course within the Utrecht Summer School 2018 for undergrads. Their aim is to introduce the students to the study of manifolds through the use of Morse theory. As such, we assume no background on Differential Topology, but we expect familiarity with Calculus in several variables, point-set Topology, and basic Group Theory. A first course on Algebraic Topology is probably useful to put these results in context (but not really necessary).

The final result in the notes is the construction of Morse homology. We will hand-wave through some of the details, but the reader should get a reasonable idea of how the proofs work and how the theory can be applied to examples (surfaces and 3-manifolds).

1. Introduction

Today we are going to look at a particular area of Differential Topology called **Morse theory**. Morse theory will allow us to:

- Take a (smooth) manifold and decompose it into elementary pieces.
- Conversely, take these elementary pieces and use them to construct manifolds.
- Construct an invariant of the manifold called Morse homology.

If you do not know what a manifold is, do not worry. We will start with surfaces, which you know well after G. Cavalcanti's course last week. Then we will review what a manifold is, introducing the notions we need along the way. Since our focus will be on surfaces and 3-manifolds, we will be able to draw plenty of pictures!

This course is somewhat of a follow-up to Gil's course from last week. Let us briefly recall what Gil showed:

- Surfaces can be triangulated (i.e. decomposed into triangles that are glued in a reasonable way),
- A surface can be triangulated in many ways, but there is still a certain uniqueness: Any two triangulations have a common refinement.
- Given a triangulation, we can compute the sum

  $$\#\{\text{vertices}\} - \#\{\text{edges}\} + \#\{\text{faces}\},$$

  which we call the **Euler characteristic**.
- This number does not depend on the triangulation, so it is actually an invariant of the surface. It actually recovers a lot of information, because knowing orientability and the Euler characteristic is enough to recover the closed surface.

This approach is quite general in Topology: first we introduce some *auxiliary data* (the triangulation) that allows us to cook up a *number/group/chain complex* (the Euler characteristic) and then we show that the auxiliary data is irrelevant and what we computed is an *invariant* of our object of interest (the surface).

What we are going to do is very similar. To study M, we are going to use as auxiliary data a function $f : M \to \mathbb{R}$. Using the critical points of f, we will construct a certain algebraic object (the homology of a chain complex) that will turn out not to depend on f, but only on M. This invariant is called **Morse homology** and from it we can recover the Euler characteristic too.

If you do not know what a chain complex or a homology theory are, that is perfectly fine. We will explain what these are in the course. If you know what they are, here is a little spoiler: Morse homology is simply another way of computing singular/simplicial homology.

1.1. **Motivating examples.**

Instead of getting formal from the get-go, let us start by looking at some examples of functions on surfaces. Try to work them out yourself, see what they have in common, and try to make some reasonable conjectures as to what one might expect in general.

1.1.1. *The unit sphere.* Consider the unit sphere $\mathbb{S}^2 = \{(x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2\}$ in Euclidean 3-space. One of the simplest functions we can consider is simply the height:

$$h : \mathbb{S}^2 \subset \mathbb{R}^3 \to \mathbb{R}$$

$$h(x,y,z) = z.$$

Refer to Figure 1.

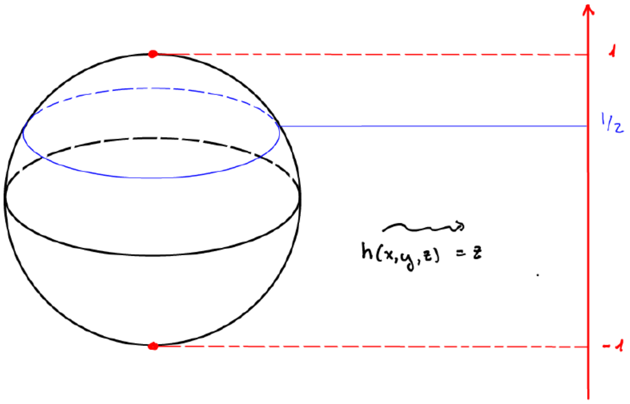


FIGURE 1. The height function on the unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$

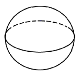
The poles are clearly critical points of h, since they are precisely the maximum and the minimum of the function. It is clear that there are no other critical points. For this, recall your Calculus in several variables: the gradient of h, as a function $\mathbb{R}^3 \to \mathbb{R}$, is the vertical vector $\nabla h = (0,0,1) = \partial_z$. A point of $\mathbb{S}^2$ is critical if the gradient ∇h is orthogonal to the sphere, which only happens at the poles.

Now, let us look at the level sets of the function h:

$$h^{-1}(z) = \begin{cases} \emptyset & \text{if } z < -1 \\ \{(0,0,-1)\} & \text{if } z = -1 \\ \text{a circle} & \text{if } -1 < z < 1 \\ \{(0,0,1)\} & \text{if } z = 1 \\ \emptyset & \text{if } z > 1 \end{cases}$$

And the sublevel sets:

$$h^{-1}((-\infty,z]) = \begin{cases} & \emptyset & \text{if } z < -1 \\ \bullet & \{(0,0,-1)\} & \text{if } z = -1 \\ & \text{a disc} & \text{if } -1 < z < 1 \\ & \mathbb{S}^2 & \text{if } 1 \leq z \end{cases}$$



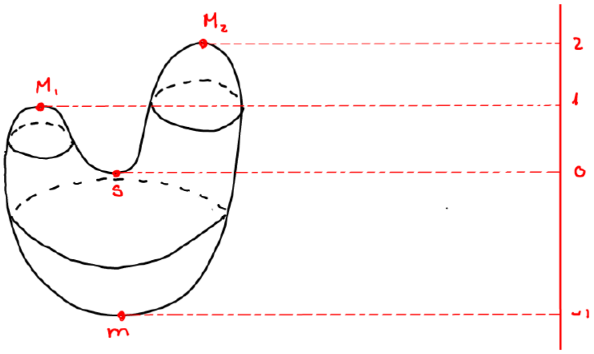
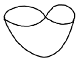
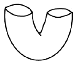


Figure 2. The height function on the "bean" in $\mathbb{R}^3$

1.1.2. *The bean.* We now consider the surface S in $\mathbb{R}^3$ shown in Figure 2. As a topological space, it is clearly homeomorphic to a sphere, but it is embedded in Euclidean 3-space in a strange way.

Let us consider the same function as before, the restriction to S of the function $h(x, y, z) = z$. There are a total of 4 critical points: a minimum m, two local maxima M_1 and M_2, and a saddle point s. Our analysis of sublevel sets yields:

$h^{-1}((-\infty, z]) =$

	$\emptyset$	if $z < -1$
•	$\{m\}$	if $z = -1$
	a disc	if $-1 < z < 0$
	a disc pinched at the boundary	if $z = 0$
	a cylinder	if $0 < z < 1$
	a disc	if $1 \leq z < 2$
	S	if $2 \leq z$

1.1.3. *The torus.* We consider the restriction of $h(x, y, z) = z$ to the torus $\mathbb{T}^2$ in $\mathbb{R}^3$ drawn as in Figure 3.

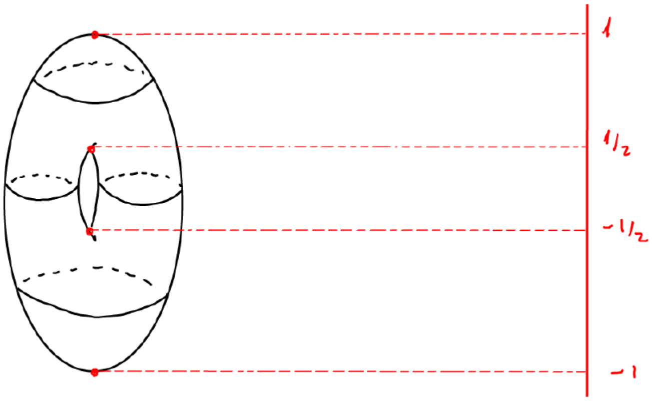
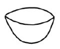
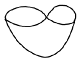
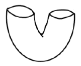
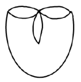
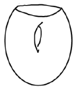


Figure 3. The height function on a 2-torus standing up in $\mathbb{R}^3$

There are again 4 critical points: a minimum m, a maximum M, and two saddle points s_1, s_2. Here are the sublevel sets:

$h^{-1}((-\infty, z]) =$

	$\emptyset$	if $z < -1$
	$\{m\}$	if $z = -1$
	a disc	if $-1 < z < -1/2$
	a disc pinched at the boundary	if $z = -1/2$
	a cylinder	if $-1/2 < z < 1/2$
	a cylinder with its boundaries pinched together	if $z = 1/2$
	$\mathbb{T}^2$ minus a disc at the top	if $1/2 < z < 1$
	$\mathbb{T}^2$	if $z \geq 1$

1.2. **Discussion.** Some phenomena should be apparent to you by examining these examples:

The function h provides a "movie" for the surface of interest by cutting it up in level sets $f^{-1}(z)$ that are collections of circles unless z is a critical value. As such, we need the function h to behave nicely if we want the movie to be useful; if the set of critical values is very complicated, there is no hope this will work. For instance, the constant zero function has a single level set, which is the whole surface. In Subsection 3.4 we will see that "most functions" have only isolated critical points, so this is not a problem.

The movie is only interesting at critical values. Indeed, the topology of the level and sublevel sets only changes at critical points. That is, if $[z_0, z_1]$ does not contain critical values:

$$\begin{aligned} f^{-1}(z_0) &\text{ is homeomorphic to } f^{-1}(z_1) \\ f^{-1}((-\infty, z_0]) &\text{ is homeomorphic to } f^{-1}((-\infty, z_1]) \\ f^{-1}([z_0, z_1]) &\text{ is homeomorphic to } f^{-1}(z_0) \times [z_0, z_1] \end{aligned}$$

The movie effectively decomposes the surface into a collection of elementary pieces (surfaces with boundary). Over each piece the height function has at most one critical point and is therefore easy to describe; see Figure 4.

I. The simplest piece is the cylinder. It contains no critical points.
II. The second piece is the upward disc corresponding to a minimum. The level sets of the function are concentric circles surrounding said minimum.
III. The third piece is the reflection of the second: the downward disc associated to a maximum.
IV. The fourth piece is the inverted pair of pants and its only critical point is the saddle. The level sets start as a circle that eventually gets pinched and separates into two.
V. The fifth piece is the pair of pants, also corresponding to the saddle. Here two circles eventually merge into one.

Remark 1. *The five pieces we just listed are the ones that appear in the examples above. Are there others?*

It is clear that a piece containing only a maximum or a minimum must be a disc. Similarly, a piece with no critical points must be a cylinder. Therefore any pieces we missed must contain a saddle. The key observation is the following: a pair of pants is obtained from the cylinder by attaching a band to

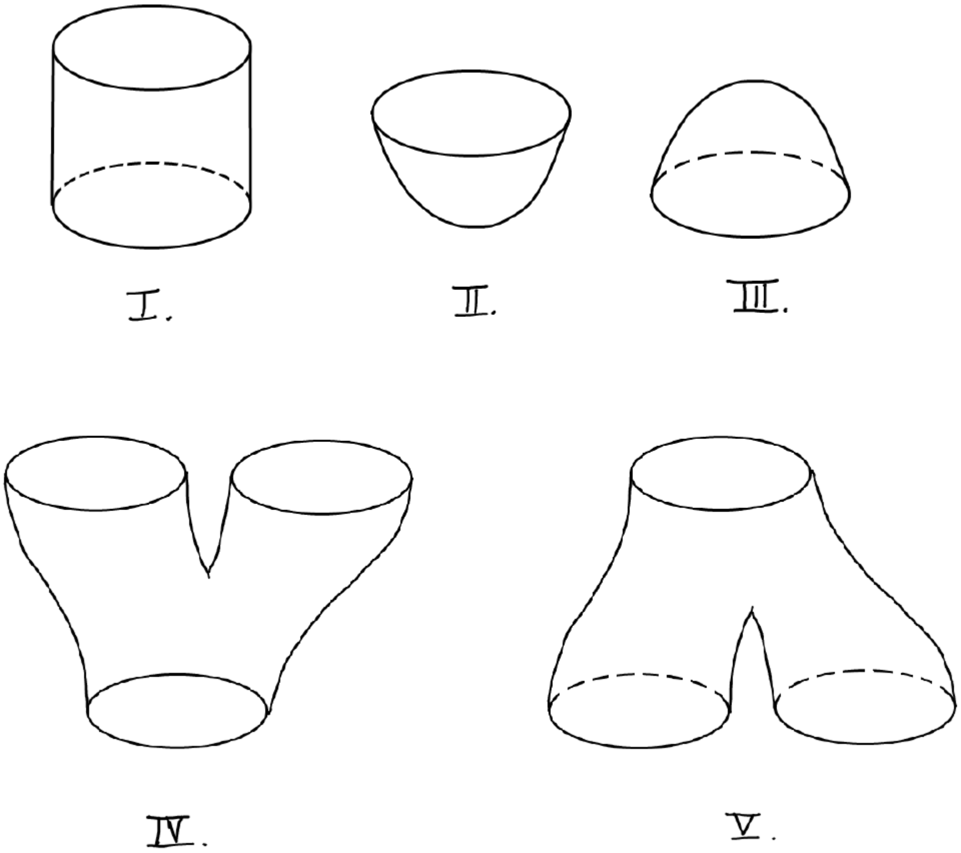


Figure 4. The elementary pieces used to construct orientable surfaces.

one of its boundary components. In both cases the band was attached in an orientable way*. Other elementary pieces can be obtained by glueing the band in a non-orientable manner, see Figure 5.*

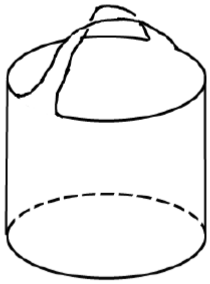


Figure 5. A non-orientable piece containing a single critical point for the height function. It is obtained from a cylinder by attaching a band at the top.

In this course we will focus on the orientable case (the five pieces from Figure 4), so you can safely forget these non-orientable ones.

Let us combine the elementary pieces: A disc glued to a pair of pants yields a cylinder, but the resulting function on the cylinder is not the standard one (the one with no critical points). Instead, it contains a pair of critical points: a saddle and either a maximum or a minimum; refer to Figure 6.

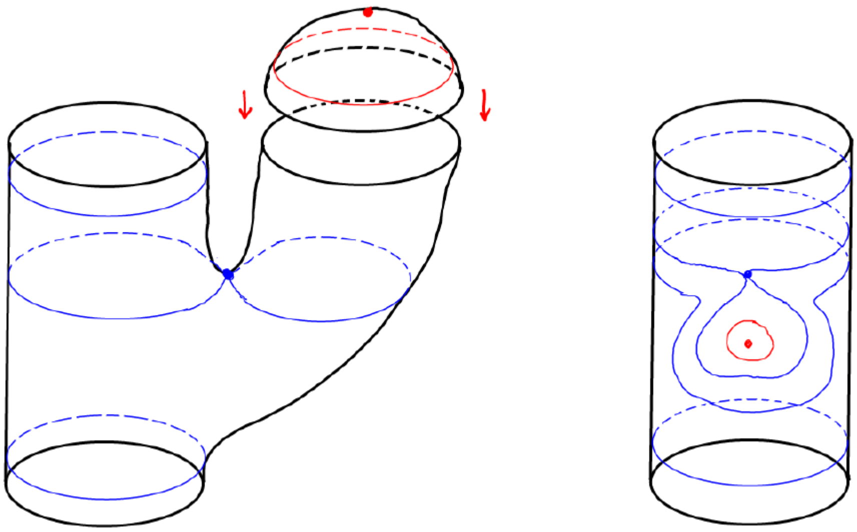


Figure 6. On the left we work with the height function; if we attach a disc (containing a maximum) to an inverted pair of pants, we obtain a cylinder, but the function on it is not the usual one. The figure on the right depicts exactly the same in a slightly different manner: the cylinder is put in a more standard form but the function is not the height anymore.

This suggests that a given function may be modified by adding/removing pairs of critical points; we call this a *birth/death*. We could try to prove that any two functions are related to one another by a sequence of birth/death events. This tells us that the number of critical points of each type is not an invariant but something like:

$$\#\{\text{minima}\} - \#\{\text{saddles}\} + \#\{\text{maxima}\}$$

may be. In fact, we shall see that this is nothing but the Euler characteristic you already know!

Our work now boils down to formalising all the statements we have just outlined. Afterwards, we will look at some examples of functions on 3-dimensional manifolds.

2. Manifolds

2.1. **The definition.** Depending on how you learned about surfaces, you might be a bit worried about what we are doing. If you studied them as subsets of $\mathbb{R}^3$ given as the zeroes of some function or parametrised by a subset of $\mathbb{R}^2$, probably you are familiar with the fact that we can take a smooth function on $\mathbb{R}^3$, restrict it to the surface, and then compute its critical points as we did in the examples above (by looking at the points where the gradient is orthogonal to the surface). If you have only seen surfaces as topological spaces (possibly constructed from a triangulation), then the idea that you can differentiate a function on them might seem a bit strange.

Repeating a bit what G. Cavalcanti did last week, I am going to define for you what a smooth manifold is; surfaces will be a particular example. What you should have in mind is that a smooth manifold is a topological space that locally looks like Euclidean space and in which we can differentiate functions. Notice that this is key, because we must be able to speak of critical points.

Here is a definition:

Definition 2. *A* ***(smooth) manifold*** *M of dimension n is a Haussdorff, second countable topological space that is modelled on Euclidean n–space. The last property means the following:*

For every point $p \in M$ there exists an open neighbourhood $p \in U_p \subset M$ endowed with a homeomorphism $\varphi_p : U_p \to \mathbb{D}^n$ onto the open n–ball satisfying that:

- *In each overlap $U_{p,p'} = U_p \cap U_{p'}$ the map*
$$\varphi_{p,p'} = \varphi_{p'} \circ \varphi_p^{-1} : \varphi_p(U_{p,p'}) \to \varphi_{p'}(U_{p,p'})$$
a diffeomorphism (a smooth bijective map).

The idea you should have in mind is that we have taken a collection of balls $\{U_p\}$ in $\mathbb{R}^n$ and glued them using the maps $\varphi_{p,p'}$. The functions $\{\varphi_p\}$ are called *charts*, because they identify pieces of our manifold with a piece of Euclidean space (like a chart/map of the Earth does). The functions $\{\varphi_{p,p'}\}$ are called *transition functions*. See Figure 7.

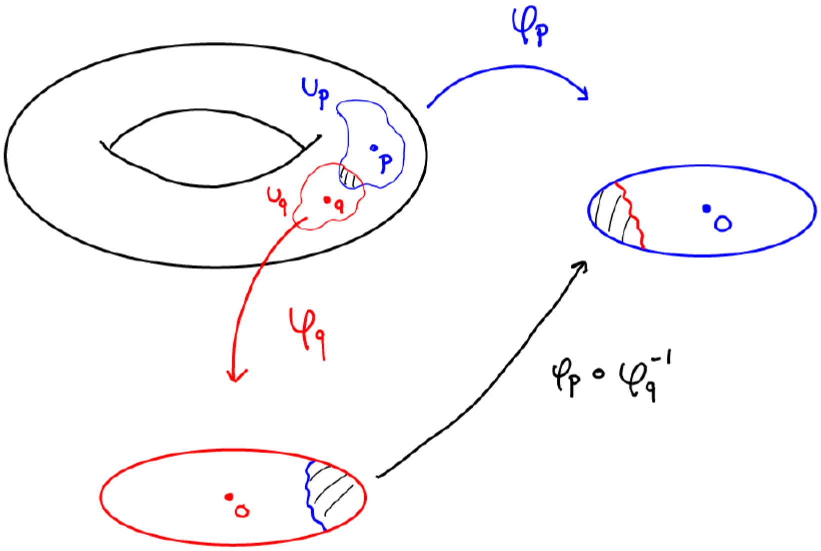


FIGURE 7. Two charts on the 2-torus. They overlap on a smaller disc. The transition function identifies the image of the smaller disc on the two different charts.

A manifold that is compact as a topological space is said to be **closed**. Otherwise it is said to be **open**. The sphere and the torus are closed, whereas the plane and the Möbius band are open. A manifold contained inside another manifold is said to be a **submanifold**. The unit sphere $\mathbb{S}^n$ is for instance a submanifold of $\mathbb{R}^{n+1}$.

2.2. **Little aside about orientability.** Later on we will need the notion of **orientable** smooth manifold. Recall your linear algebra: any basis of $\mathbb{R}^n$ can be written as a matrix with non-zero determinant (i.e. a matrix in $\mathrm{GL}(n)$). We can then say that a basis induces the *positive* orientation in $\mathbb{R}^n$ if its determinant is positive.

Remark 3. *Note that here we are essentially making a choice. We are saying that the standard basis of $\mathbb{R}^n$ corresponding to the identity matrix induces the "positive" orientation. However, given an abstract vector space V of dimension n, there is no standard basis. Fixing it is a choice that in particular determines what the "positive" orientation is.*

Given a manifold, each chart U_p is identified with a piece of $\mathbb{R}^n$ and therefore receives an orientation. More precisely, at every point of $\varphi_p(U_p) \subset \mathbb{R}^n$ the standard basis of $\mathbb{R}^n$ provides a basis of tangent vectors. We can now ask for the transition functions $\varphi_{p,p'}$ to preserve this orientation. I.e., for their differentials $[D(\varphi_{p,p'})]$ to take the standard basis to another positive basis. This means that the matrices $[D(\varphi_{p,p'})]$ should have positive determinant. In some manifolds we *can* pick all the transition functions to be like this; then we say that the manifold is **orientable**. If we actually pick transition functions with these properties, we say that it is **oriented**.

Exercise 4. *Let M be an orientable manifold. Show that M can be oriented in two distinct ways.*

Exercise 5. *Check that the sphere and the torus are orientable. Check that the Mobius band and the Klein bottle are not. Hint: look at Figure 8.*

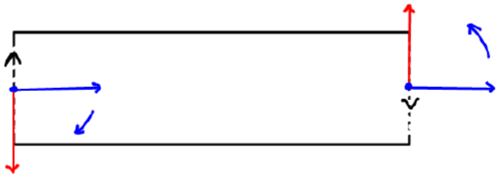


FIGURE 8. This is the usual depiction of the Möbius band (a square in which two of the opposing intervals are identified using opposite orientations). Observe that if we draw a basis for the tangent space and we move it along the band, its orientation changes.

2.3. **Smooth functions.** What is important in the definition of smooth manifold is that the transition functions are *smooth* maps, and not just homeomorphisms. This property is what allows us to talk about a function $f : M \to \mathbb{R}$ being smooth:

Definition 6. *A function $f : M \to \mathbb{R}$ is* ***smooth*** *if the compositions $f \circ \varphi_p^{-1} : \mathbb{D}^n \to \mathbb{R}$ are smooth for all p.*

I.e., it is smooth if it looks smooth over each chart. We have to show that this is a good definition. For this we have to compare what happens to the function as we look at different charts.

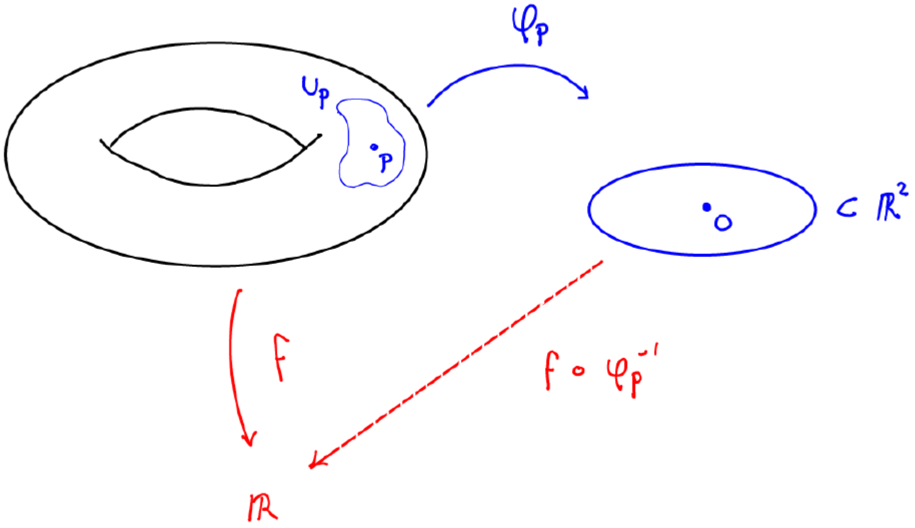


FIGURE 9. To check whether a function is smooth, we restrict it to the charts.

If we restrict $f \circ \varphi_p^{-1}$ to the overlap $U_{p,p'} \subset U_p$ of U_p and $U_{p'}$, we have that:

$$(f \circ \varphi_p^{-1})|_{\varphi_p(U_{p,p'})} = (f \circ \varphi_{p'}^{-1})|_{\varphi_{p'}(U_{p,p'})} \circ \varphi_{p,p'}.$$

Since $\phi_{p,p'}$ is a diffeomorphism, the left hand side is smooth if and only if the right hand side is smooth (this follows from the chain rule!). This shows that our definition is coherent.

Remark 7. *The general philosophy is this: a smooth function precomposed with a diffeomorphism is still smooth; as such, smooth functions make sense in manifolds with diffeomorphisms as transition functions.*

If you take the definition of manifold, we could impose other conditions on the transition functions. It is perfectly legitimate to require them to be just homeomorphisms. This yields the definition of topological manifold. *On them we cannot talk about smooth functions, only continuous ones!*

We could instead say our transition functions are biholomorphisms; such a manifold is said to be complex. *Note that holomorphic functions precomposed by biholomorphisms are still holomorphic; so one can consider holomorphic functions on complex manifolds.*

Exercise 8. *A good reality check you can do is the following: define a C^r-manifold to be a manifold given by C^r-transition functions (homeomorphisms admitting continuous derivatives up to order r). Check that one can consider C^r-functions on such a manifold, but C^{r+1}-functions do not make sense.*

Exercise 9. *Given two manifolds M and $\mathbb{N}$, a map $f : M \to N$ is said to be a* ***diffeomorphism*** *if it is a smooth bijection. Make sense of this statement. Hint: use the same philosophy as above.*

2.4. **Critical points and the differential.**

Definition 10. *A point $q \in M$ is a* ***critical point*** *of a smooth function $f : M \to \mathbb{R}$ if $\varphi_q(q)$ is a critical point of $f \circ \varphi_q^{-1}$.*

That is, a point q is critical for f if it is a critical point in terms of its chart U_q. It is easy to see that it will also be critical in other charts:

Lemma 11. *A point q is critical for $f : M \to \mathbb{R}$ if and only if $\varphi_p(q)$ is critical for $f \circ \varphi_p^{-1}$ (whenever $q \in U_p$).*

Proof. This follows from the chain rule:

$$[D(f \circ \varphi_p^{-1})(\varphi_p(q))] = [D(f \circ \varphi_q^{-1} \circ \varphi_{p,q})(\varphi_p(q))] =$$

$$[D(f \circ \varphi_q^{-1})(\varphi_q(q))][D\varphi_{p,q}(\varphi_p(q))]$$

Since $\varphi_{p,q}$ is a diffeomorphism, $[D\varphi_{p,q}(\varphi_p(q))]$ is an invertible matrix. Then $[D(f \circ \varphi_p^{-1})(\varphi_p(q))]$ is zero if and only if $[D(f \circ \varphi_q^{-1})(\varphi_q(q))]$ is zero. So q is either critical in all charts or in none. □

Let us inspect the matrix $[D(f \circ \varphi_p^{-1})(\varphi_p(q))]$ a bit. This is, by definition, the *gradient* of the function $f \circ \varphi_p^{-1}$ at the point $\varphi_p(q)$ in row form (as opposed to a column). Being in row form, you can multiply it with a vector in column form to produce a number. As such, vectors in row and column form are *dual* to each other (in the usual sense that the vectors at a point form a vector space and the covectors are elements of the dual vector space). We will henceforth say that a column vector is simply a vector and a row vector is a **covector**. You should think of vectors as directions and of covectors as duals of directions.

We would like to define this row gradient as a *global* object in our manifold. As such, we need to study how it changes when we apply a transition function. Recall your Calculus course:

Remark 12. *Let v be a vector based at a point $p \in \mathbb{R}^n$. Suppose we are given a diffeomorphism $F : \mathbb{R}^n \to \mathbb{R}^n$. Then, the differential $[DF(p)]$ maps v to the vector $[DF(p)]v$ based at the point $F(p)$. Suppose α is a covector based at p, so we can compute the quantity $\alpha(v)$. This number should not change if we change coordinates by applying F; i.e. if we evaluate the covector "whatever α is mapped to by F" to the vector $[DF(p)](v)$ we should obtain $\alpha(v)$ again. Therefore α should be mapped to $\alpha[DF(p)]^{-1}$ at $F(p)$, because then:*

$$\alpha[DF(p)]^{-1}([DF(p)](v)) = \alpha(v).$$

Indeed, this is what happens to the gradient in row form:

Lemma 13. *The gradient in row form transforms as:*

$$[D(f \circ \varphi_p^{-1})(\varphi_p(q))] = [D(f \circ \varphi_q^{-1})(\varphi_q(q))][D\varphi_{p,q}(\varphi_p^{-1}(q))].$$

Proof. This is simply the chain rule, where have applied $f \circ \varphi_p^{-1} = f \circ \varphi_q^{-1} \circ \varphi_{p,q}$. Multiplying by $[D\varphi_{p,q}(\varphi_p^{-1}(q))]^{-1}$ on the right we obtain the expression claimed in the previous remark. □

This shows that there exists a **covector field** in M (i.e. a choice of covector at every point of our manifold) which is defined on each chart U_p by $[D(f \circ \varphi_p^{-1})]$. We call it the **differential** of f (instead of the gradient, which should be a vector field), and we denote it by df.

Lemma 14. *The differential df evaluates to zero on all vectors tangent to level sets of f.*

Proof. In local coordinates, this is just the usual statement that the gradient of f is orthogonal to the level sets. □

What you should take away from this is that the "gradient of f" does not really exist in a manifold. However, the differential, which plays the role of its dual, does exist. Refer to Figure 10.

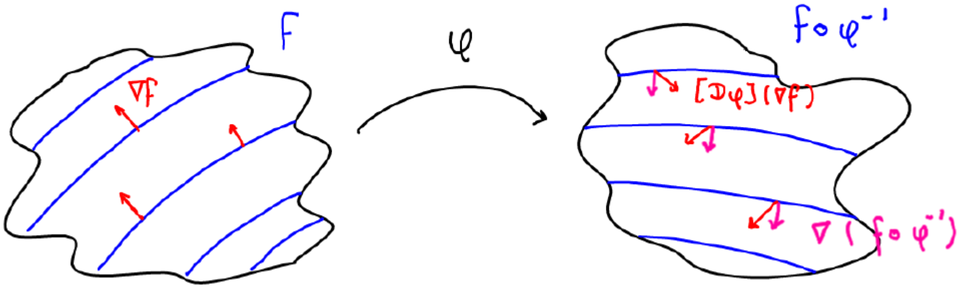


Figure 10. The gradient of a function $f \circ \varphi^{-1}$ is not $[D(\varphi)]$ applied to the gradient of f. Nonetheless, both are transverse to the level sets of $f \circ \varphi^{-1}$.

2.5. **Picking a gradient.** Even though there is not a uniquely defined gradient for $f : M \to \mathbb{R}$, we can nonetheless find a vector field (i.e. a choice of vector at each point of M) that more or less behaves like the gradient. What properties should such a vector field v have?

- v should be transverse to the level sets of f. This is the same as requiring $df(p)(v(p)) > 0$ whenever p is not a critical point.
- $v(p) = 0$ if and only if p is a critical point of f.

Note that for functions $\mathbb{R}^n \to \mathbb{R}$ the gradient satisfies these properties.

Without providing details, such a vector field can be found by covering M with a finite collection of charts, picking the actual gradient in each, and then interpolating between all these choices. This works because we can interpolate linearly between any two vectors v_1 and v_2 satisfying $df(v_i) > 0$ while preserving this positivity condition. We will often abuse notation and write ∇f for a vector field satisfying these properties; just recall that it is not unique!

Given a point $p \in M$ we can push it in the direction of ∇f. I.e. we define a map:

$$\phi_t : M \to M$$

such that $\phi_t(p)$ is the point obtained by pushing p using the vector field ∇f for a time t. More formally, the map ϕ_t solves the ordinary differential equation:

$$(\partial_t \phi_t)|_{t=t_0}(p) = \nabla f(\phi_{t_0}(p)).$$

The map ϕ_t is called the **flow** of ∇f. This map is smooth because the solution of an ODE with smooth coefficients is smooth.

The points q such that $\phi_t(p) = q$ for some $t \in \mathbb{R}$ are called the **flowline** associated to p; they are the points you can reach by pushing p forwards or backwards in time in the t variable. The following lemma states that as you push a point p along its flowline, you end up in a critical point.

Lemma 15. *Let M be a closed manifold and let $f : M \to \mathbb{R}$ be a function. Given any $p \in M$, the points in the limit set* $\lim_{t\to\infty} \phi_t(p)$ *are critical. The same applies as $t \to -\infty$.*

Here $\lim_{t\to\infty} \phi_t(p)$ should be understood as the set of points $q \in M$ such that there is a sequence of times $\{t_n\}_{n\in\mathbb{N}}$ satisfying that the limit $\lim_{t_n\to\infty} \phi_{t_n}(p)$ exists and is precisely the point q.

Proof. Let us argue for $t \to \infty$. The other case is the same.

The flowline containing p is the image of the map $\gamma : \mathbb{R} \to M$ given by $\gamma(t) = \phi_t(p)$. The composition $f \circ \gamma$ is an increasing function $\mathbb{R} \to \mathbb{R}$, since:

$$(f \circ \gamma)' = [df(\gamma(t))][D\gamma(t)](1) = [df(\gamma(t))](\nabla f(\gamma(t))) > 0,$$

by the definition of γ and the properties of ∇f. Since M is compact, f is bounded above and a unique supremum S exists. Using the fundamental theorem of calculus we deduce that $(f \circ \gamma)'$ must converge to zero as $t \to \infty$.

Compactness of M tells us that there is a convergent sequence of times t_n such that $\lim_{t_n \to \infty} \phi_{t_n}(p)$ is a point $q \in M$. Then:

$$f(q) = \lim_{t_n \to \infty} f(\phi_{t_n}(p)) = S$$
$$\nabla f(q) = \lim_{t_n \to \infty} \nabla f(\phi_{t_n}(p)) = 0.$$

If the set of critical points of f is very complicated, there may be more that one point q satisfying this conclusion. If for every critical value there is a single critical point, the point q we have found is unique. □

Exercise 16. *Find an example of M, f, p, and ∇f in which the limit set $\lim_{t \to \infty} \phi_t(p)$ is not just a point. Hint: find an example where the flowlines of ∇f spiral around the locus of critical points.*

2.6. **Some remarks to end this section.** Now that we know what a manifold is, we have two possible perspectives as a mathematician:

- We think of manifolds as a special type of topological space. Then we could ask things like: "When is a topological space homotopy equivalent to a manifold?"
- We think of manifolds as its own type of object. Then we need to define when two manifolds are "the same", and then we try to find invariants that help us distinguish manifolds up to this notion of "being the same".

Observe that homotopy equivalence is too weak of a notion, since it does not respect the Euclidean structure of the manifold (i.e. a homotopy equivalence does not map an n-dimensional ball to an n-dimensional ball necessarily). Two topological manifolds will be the same if there exists a **homeomorphism** (i.e. a continuous bijection) between them. Two smooth manifolds will be the same if there exists a **diffeomorphism** (i.e. a continuous bijection which is differentiable to all orders).

Remark 17. *The relationship between smooth and topological manifolds is quite subtle (last week Gil already commented a bit on this topic). It is clear that every smooth manifold is a topological manifold. Conversely, we could ask: is every topological manifold smoothable (i.e. homeomorphic to a smooth manifold)? If so, is this smooth manifold unique (up to diffeomorphism)?*

It turns out that the state of affairs is quite different in higher and lower dimensions. In dimensions 2 *and* 3 *every topological manifold has a unique smooth structure. In dimension* 4*, "most" topological manifolds do not admit smooth structures and* $\mathbb{R}^4$ *admits uncountably many distinct smooth structures; determining whether* $\mathbb{S}^4$ *admits more than one smooth structure is precisely the* Poincaré conjecture *in dimension 4. In dimensions larger than* 4*, a given topological manifold admits only finitely many distinct smooth structures.*

Remark 18. *We will henceforth work only with smooth manifolds, so I will just stop writing "smooth" every time. Additionally, we would like to work with* closed *manifolds. This is, however, not possible. Even though the manifolds we will start with will be closed, we will then cut them up into pieces that have a* ***boundary****. These pieces are not manifolds according to the definition above, but:*

Exercise 19. *Formalise the notion of "manifold with boundary"; its boundary should be a manifold in the standard sense. What does such an object look like locally?*

3. Morse theory

We have covered some of the theory of functions on manifolds. Our next objective is stating and proving some of the basic results in Morse theory. Before we do so, let us look at some examples.

3.1. **Functions with isolated critical points.** We want to use functions to cut our manifold into nice level sets. For this, we should consider functions having only isolated critical points. Let us see if we can come up with examples of functions $\mathbb{R}^n \to \mathbb{R}$ having an isolated critical point at zero:

3.1.1. *Dimension 1.* In dimension 1 we can consider the functions $f_+(x) = x^2$ and $f_-(x) = -x^2$. These are upward and downward parabolas, respectively, with the origin as their unique critical point. We can readily compute their Hessians: $H_0 f_+ = [1]$ and $H_0 f_- = [-1]$. It is easy to see that any other monomial x^k with $k > 2$ has an isolated critical point at 0 with vanishing Hessian.

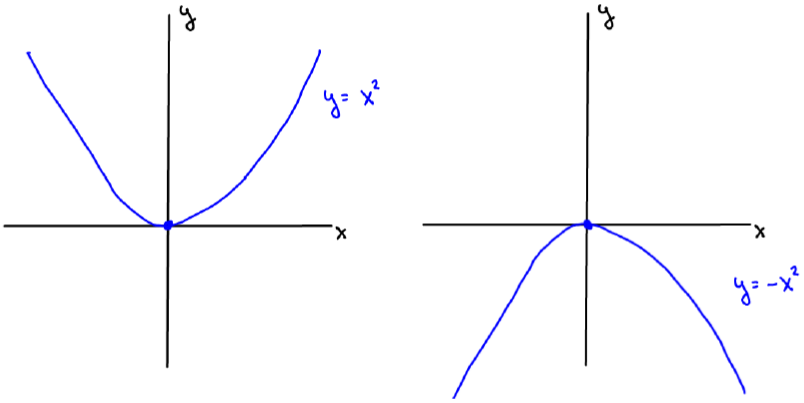


Figure 11. Quadratic functions from $\mathbb{R}$ to itself.

3.1.2. *Dimension 2.* In dimension 2 we can consider the functions $f_m(x,y) = x^2 + y^2$, $f_s(x,y) = x^2 - y^2$, and $f_M(x,y) = -x^2 - y^2$. The origin is their only critical point and their Hessians are:

$$H_0(f_m) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad H_0(f_s) = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad H_0(f_M) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}.$$

That is, they correspond to the minimum, the saddle, and the maximum, respectively.

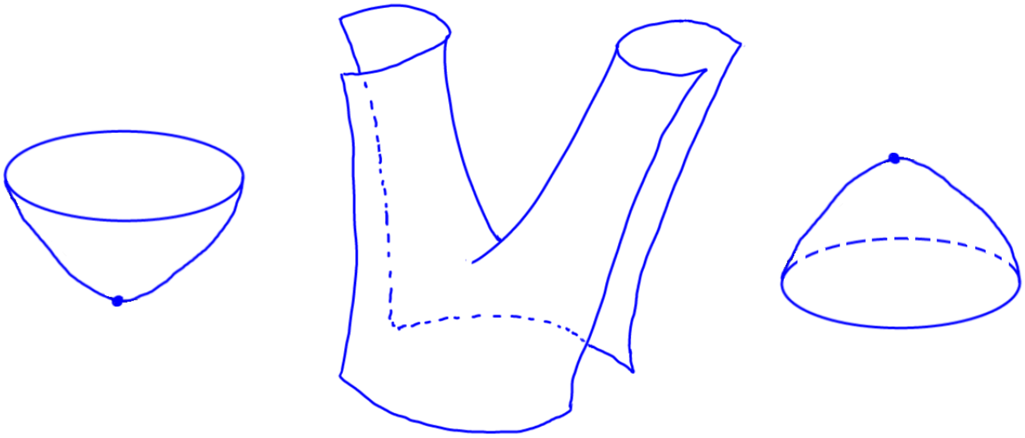


Figure 12. The minimum, the saddle, and the maximum.

Now observe that if we consider the function $f(x,y) = x^2$, the whole line $\{x = 0\}$ is formed by critical points. This has to do with the fact that the Hessian at each point $(0,y)$ is

$$H_{(0,y)}(f_m) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

which has zero determinant and vanishes in the direction of y (i.e. in the direction tangent to the family of critical points).

3.2. **The definition.** The last example suggests that we should focus on functions having Hessians with non-zero determinant at their critical points. A more serious motivation, that we will see later on, is that "most functions" have critical points satisfying this condition. Additionally, the gradient close to a critical point is very simple under this assumption.

In any case, we are interested in the following definition:

Definition 20. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth function. A critical point p of f is said to be **non-degenerate** or of **Morse type** if the Hessian*

$$H_p(f) = \begin{bmatrix} \partial^2_{x_1} f(p) & \partial_{x_2}\partial_{x_1} f(p) & \cdots & \ldots & \partial_{x_n}\partial_{x_1} f(p) \\ \partial_{x_1}\partial_{x_2} f(p) & \partial^2_{x_2} f(p) & \cdots & \ldots & \partial_{x_n}\partial_{x_2} f(p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \partial_{x_1}\partial_{x_n} f(p) & \partial_{x_2}\partial_{x_n} f(p) & \cdots & \ldots & \partial^2_{x_n} f(p) \end{bmatrix}$$

is non-degenerate (i.e. has non-zero determinant).

*A function $M \to \mathbb{R}$ is said to be of **Morse type** if all its critical points are of Morse type.*

We have to be a bit careful about the second definition. Is the Hessian well-defined in a manifold?

Lemma 21. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be the function (over a chart if you wish) and let $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ be a diffeomorphism (a transition function, for instance). Then the following formula holds:*

$$[H(f \circ \varphi)(p)] = [D(\varphi)(p)]^T [H(f)(\varphi(p))][D(\varphi)(p)]$$

if p is a critical point of $f \circ \varphi$.

Proof. Given a function $g : \mathbb{R}^n \to \mathbb{R}$, the transpose of its differential can be regarded as a function $Dg^T : \mathbb{R}^n \to \mathbb{R}^n$. The Hessian is precisely the differential of Dg^T. Applying this and the chain rule, we can compute:

$$[D(f \circ \varphi)(p)]^T = [D(\varphi)(p)]^T [D(f)(\varphi(p))]^T$$

and we apply the chain rule once more to $[D(f)(\varphi(p))]^T$:

$$[H(f \circ \varphi)(p)] = [D(\varphi)(p)]^T [H(f)(\varphi(p))][D(\varphi)(p)] + [D(D(\varphi)(p))^T][D(f)(\varphi(p))]^T.$$

Note that the second term involves the second derivatives of φ and the differential $[D(f)(\varphi(p))]^T$ of f. Since $\varphi(p)$ is a critical point of p, the second term vanishes and the claimed formula holds. □

The formula in the lemma should be familiar to you from Linear Algebra: this is the usual basis change for a quadratic form! In this particular case we are saying that the basis change for the Hessian, at a critical point, is given by the differential $[D(\varphi)]$ of the diffeomorphism φ. Sylvester's law of inertia says that some essential properties of the quadratic form survive under basis changes:

Lemma 22. *Let $f : M \to \mathbb{R}$ be a function and let $p \in M$ be a critical point of f. Then, the signature and determinant of the Hessian at p are well-defined (independently of the chart).*

Proof. Recall that the signature is the number of -1 and $+1$ when we diagonalise (as bilinear form) the matrix $H(f)$. □

Further, the signature is, up to a choice of chart, the only invariant for a Morse critical point:

Proposition 23. *Given a function $f : \mathbb{R}^n \to \mathbb{R}$ with a Morse critical point at zero, there exists a number s, called the **index**, and a change of coordinates $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ such that*

$$f \circ \varphi = -\sum_{i=1}^{s} x_i^2 + \sum_{i=s+1}^{n} x_i^2.$$

Proof. By diagonalisation, we know that there is some matrix A such that $A^T[H(f)(0)]A$ is diagonal with the first s diagonal entries -1 and the last $n-s$ entries $+1$. This implies that, if we consider A as a change of coordinates of $\mathbb{R}^n$, we have that the Taylor expansion of $f \circ A$ at zero is:

$$f \circ A = \sum_{i=1}^{s} x_i^2 + \sum_{i=s+1}^{n} x_i^2 + O(x^3),$$

that is, the expression we want plus some error term of higher order. Getting rid of this remainder is actually quite technical, if you are interested in seeing the rest of the proof, you should check out [1, Thm. 1.11] or [2, Lemma 2.2]. □

Probably you already knew this result in the particular case of surfaces: non-degenerate critical points either look like a minimum, like a maximum, or like a saddle.

3.3. **The Morse lemma.** Now we understand critical points a bit better. The following proposition proves that topology of the manifold changes only at critical points.

Proposition 24. *Let $f : M \to [0,1]$ be a function with no critical points and satisfying $f(\partial M) \subset \{0,1\}$. Then $M \cong f^{-1}(0) \times [0,1]$.*

Proof. We can pick a vector field v in M playing the role of the gradient of f, as in Subsection 2.5. Since f has no critical points by assumption, df is never vanishing so we can pick v satisfying $df(v) = 1$, this implies that v points outwards in $f^{-1}(1)$ and inwards in $f^{-1}(0)$. Refer to Figure 13.

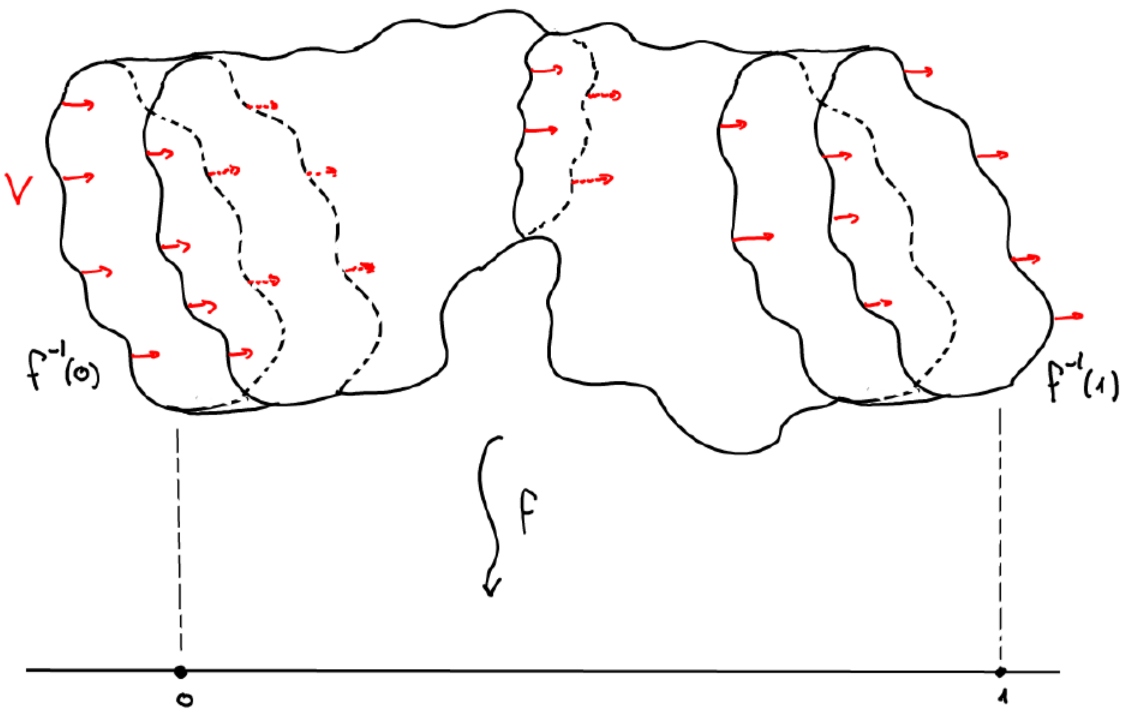


FIGURE 13. The function f divides our manifold M into level sets. Following the gradient v identifies the different level sets with one another.

The idea is that we should be able to take the level set $f^{-1}(0)$ and push it along v, effectively finding identifications between all the levels. $df(v) = 1$ means that when we push an amount of time t_0, we add t_0 to the function f. More formally: if we write ϕ_t for the flow of v, we have that ϕ_{t_0} is a diffeomorphism between $f^{-1}(t)$ and $f^{-1}(t+t_0)$. The map

$$f^{-1}(0) \times [0,1] \to M$$
$$(x,t) \to \phi_t(x)$$

is the desired diffeomorphism. □

3.4. **Genericity of Morse functions.**

One of the key lemmas in Morse theory says that being Morse is a *generic* property for functions, i.e. most functions are Morse. We will not prove this statement, but at least we shall give an intuition of why this is true.

Consider the space $C^\infty(M)$ of all smooth functions $M \to \mathbb{R}$. This is, of course, a set. It is, additionally, a vector space (although infinite dimensional). Furthermore, we can endow it with a topology. Let us explain how this is done when $M = [0,1]$ for simplicity. For each natural number k we can define a notion of distance between two functions f and g as follows:

$$d_k(f,g) = \max_{p\in[0,1]}\{|f^{(k)}(p) - g^{(k)}(p)|\}.$$

We consider the topology induced by the distance $d_0 + d_1 + d_2$. Thus, two functions are "close" to one another if all the derivatives up to order 2 are close at every point. This is called the C^2–**topology**.

Remark 25. *Although not necessary to us, it is true that one can use all the distances d_k to define a topology in $C^\infty(M)$.*

Exercise 26. *Find a sequence of functions $\{f_n : [0,1] \to \mathbb{R}\}_{n\in\mathbb{N}}$ converging to zero in the d_0 distance but not in the d_1 distance. Can you find a sequence that converges to zero in all the distances up to d_k, but not in d_{k+1}?*

Lemma 27. *In the C^2-topology, the Morse functions are an open set.*

Proof. Let $f : M \to \mathbb{R}$ be Morse function. Given another function g, a priori not Morse but close enough to f, we should show that each critical point of g is very close to a critical point of f and they are in 1 to 1 correspondence.

The Hessian $H(f)$ is precisely the differential of the map $p \to df^T(p)$ (which in a chart is just a map $\mathbb{R}^n \to \mathbb{R}^n$). Let p be a critical point of f. Since $H(f)(p)$ is non-degenerate, the map df is a diffeomorphism close to p, by the implicit function theorem. If g is sufficiently close to f, then dg is close to df and $H(g)$ is close to $H(f)$, so dg is a diffeomorphism close to p. This implies that there is a single point q such that $dg(q) = 0$. This is the unique critical point of g close to p. Since $H(g)(q)$ is close to $H(f)(p)$, it is non-degenerate. We deduce that all the critical points of g are Morse. □

Exercise 28. *Prove that Morse functions are not an open set in the topology defined by the distance d_0.*

It is not sufficient to show that Morse functions form an open set. We need to prove that the non-Morse functions form a small set. This is what the following result of M. Morse states:

Proposition 29. *In the C^2-topology, the non-Morse functions are a codimension-1 submanifold. In particular, Morse functions are dense in the C^2-topology.*

Proof. We cannot prove this statement without invoking some big machinery (*Sard's theorem*). We will explain roughly what this result says and how we can use it.

Suppose we are given two affine subspaces A and B in some affine space V. Suppose we leave A fixed, but we are allowed to slightly move B. Then, for most choices of B, the intersection $\dim(A \cap B)$ is as small as possible. For instance:

- two lines in $\mathbb{R}^3$ usually do not intersect,
- two planes in $\mathbb{R}^3$ usually intersect on a line,
- a line and a plane in $\mathbb{R}^3$ usually intersect at a point.

That is, for most choices of B:

- $\dim(A \cap B) = \dim(A) - \operatorname{codim}(B)$ if $\dim(A) + \dim(B) \geq \dim(V)$,
- $A \cap B = \emptyset$ if $\dim(A) + \dim(B) < \dim(V)$.

Sard's theorem (in one of its incarnations called *Thom's transversality*), tells us that the same is true for manifolds: Let V be a big manifold containing some smaller submanifolds A and B. Suppose we are allowed to slightly perturb B. Then, for most choices of B, their intersection will be a smaller manifold of dimension $\dim(B) - \operatorname{codim}(A)$; see Figure 14. Let us try to apply this result here here.

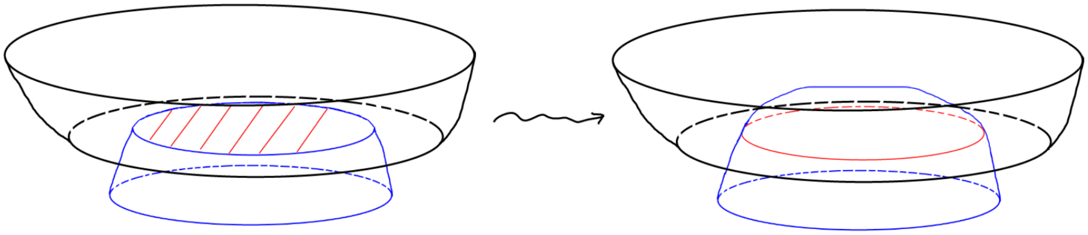


FIGURE 14. In the first figure we see two discs that intersect in a smaller disc. On the right hand side we have pushed the lower disc slightly upwards, effectively making them intersect in a circle.

We are interested in functions $f : \mathbb{R}^n \to \mathbb{R}$. Let us define the following infinite dimensional manifold:

$$J(\mathbb{R}^n, \mathbb{R}) = \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n \times \cdots \times \mathbb{R}^{c_k} \times \cdots$$

where the first term represents the domain of the function f, the second term corresponds to its value $y_0 = f(p)$, the third term is the gradient $y_1 = \nabla f(p)$, and in general the term $\mathbb{R}^{c_k}$ packages the k-order information of f, which we write as y_k. You should think of $J(\mathbb{R}^n, \mathbb{R})$ as a manifold packaging Taylor polynomials of functions at all points of $\mathbb{R}^n$.

A function $g : \mathbb{R}^n \to \mathbb{R}$ defines a submanifold of this big space by taking its graph:

$$\Gamma_g = \{(p, g(p), \nabla g(p), Hg(p), \cdots)\}.$$

This submanifold has dimension n. Within $J(\mathbb{R}^n, \mathbb{R})$, there exists a submanifold A of codimension n defined by the equation $y_1 = 0$. The critical points of f correspond exactly to the intersection $\Gamma_f \cap A$. Since $\dim(\Gamma_g) = \operatorname{codim}(A)$, this intersection should be zero dimensional, i.e. simply a collection of points.

Similarly, within the subspace A, points with degenerate Hessian are the zeroes of the equation $\det(y_2) = 0$; this defines a codimension–1 submanifold $B \subset A$. Therefore $\dim(\Gamma_g) = \operatorname{codim}(B) - 1$, so their intersection will be empty for most choices of g. See the upcoming remark. □

Remark 30. *We imagine the space of all functions as some infinite-dimensional manifold in which the non-Morse functions form (singular) hypersurfaces. It is then clear that if we have two different Morse functions and we interpolate between them, in general we will not be able to avoid the set of non-Morse functions.*

Consider for instance the function $f(x) = x^3$, which is non-Morse. We can then deform it slightly, yielding the family of functions:

$$f_\varepsilon : \mathbb{R} \to \mathbb{R}$$

$$f_\varepsilon(x) = x^3 - \varepsilon x.$$

These functions have critical points if and only if $\varepsilon \geq 0$. Then the critical points are at $\pm\sqrt{\varepsilon/3}$ and they are Morse when $\varepsilon \neq 0$. See Figure 15.

Exercise 31. *We claimed that if we fix an affine subspace A, for most choices of B (some other affine subspace), the intersection $A \cap B$ will have the smallest dimension possible. Try to make this claim rigorous.*

Note: Parallelism is a strange thing, because then intersections go to infinity. For most choices B is not parallel to A, so you can disregard this case (or even better, try to figure out what it means when you pass to projective space).

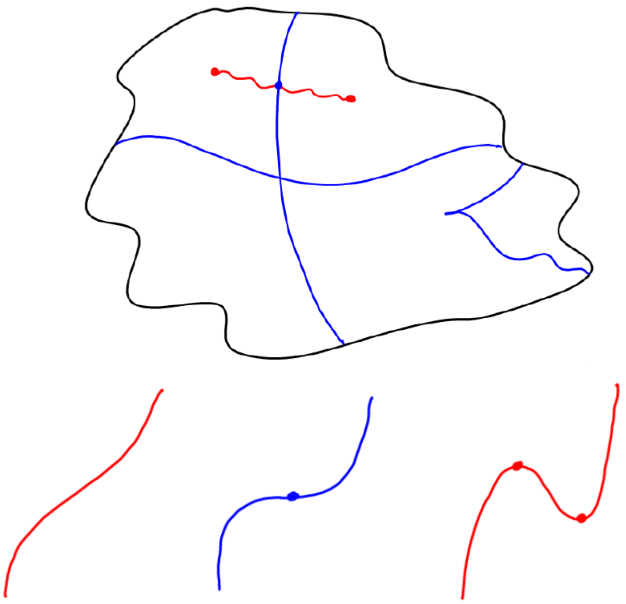


FIGURE 15. The bottom figure shows the family of functions f_ε. The map $\varepsilon \to f_\varepsilon$ is a path $\mathbb{R} \to C^\infty(\mathbb{R})$ that at $\varepsilon = 0$ crosses the hypersurface of non-Morse functions. The top figure depicts schematically the space $C^\infty(\mathbb{R})$ with the C^2-topology: the blue lines correspond to the hypersurface of non-Morse functions, the red path corresponds to $\varepsilon \to f_\varepsilon$.

3.5. **Morse theory in open manifolds.** These notes deal with Morse theory for closed manifolds. It turns out that one can apply Morse theory to open manifolds as well, but then one needs to control the behaviour of f at the infinity of M. Indeed, if we do not:

Exercise 32. *Let M be an open manifold, then M admits a function $f : M \to \mathbb{R}$ with* no *critical points.*

The idea is shown in Figure 16.

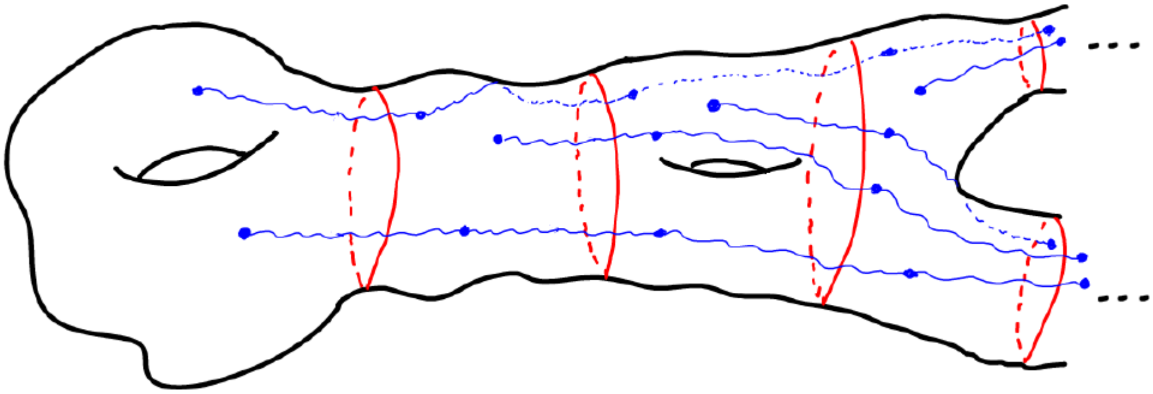


FIGURE 16. Let M be an open manifold. Then we can divide it in a countable number of compact pieces that go to infinity. Given any function $f : M \to \mathbb{R}$ we can proceed inductively taking its critical points and "pushing them" to the next piece. Taking the limit of this process yields a function $\tilde{f}$ with no critical points. Note that the level sets of the function $\tilde{f}$ go to infinity. This is a general phenomenon in Differential Geometry: problems are usually easier to solve in open manifolds.

The way in which one can control the behaviour of f at the infinity of M is by making the gradient "transverse to the boundary", i.e. pointing either fully inwards or fully outwards on each component at infinity. This forces the level sets not to go to infinity. Observe that when we cut a closed manifold into elementary pieces, these are not closed manifolds and the gradient is indeed transverse to their boundary.

4. Morse homology

Thus far we have shown some of the key lemmas in Morse theory:

- Most functions are Morse, so we can always pick one.
- We have a simple model close to a Morse critical point.
- The topology of the sublevel sets does not change in-between critical points.

Our aim now is to define this invariant of smooth manifolds called Morse homology. We shall see that this invariant describes in a fairly precise manner how the topology of the sublevel sets changes as we cross a critical point.

To see what we should expect, let us revisit surfaces once more.

4.1. Crossing a critical point (in dimension 2).

Let M be a closed surface. Let $f : M \to \mathbb{R}$ be a Morse function such that *the critical points correspond to distinct critical values.* Let us look at what happens every time we cross a critical point.

4.1.1. *Minima.* Each time a minimum appears, the number of connected components increases:

Lemma 33. *If z is a critical value of f corresponding to a point of index zero, then $f^{-1}(z+\varepsilon)$ has one more connected component than $f^{-1}(z-\varepsilon)$.*

Proof. The critical point associated to z is a minimum and thus gives birth to a new component. The function has no other critical points by assumption, so the Morse lemma (Proposition 24) tells us that the topology of the other connected components does not change. □

4.1.2. *Saddles.* When a saddle (a point of index 1) appears, two different phenomena can happen:

Lemma 34. *If z is a critical value of f corresponding to a point of index one, then either:*

- *there is a loop in $f^{-1}(z+\varepsilon)$ that cannot be deformed to lie in $f^{-1}(z-\varepsilon)$,*
- *$f^{-1}(z+\varepsilon)$ has one less connected component than $f^{-1}(z-\varepsilon)$.*

Proof. A point of index one appears when two points in a level set collapse together, as in Figure 17.

It can be readily seen that there are two distinct cases. In the first case, the two points belong to different connected components. When they collapse together, the connected components become the same, yielding the second conclusion.

When the points live in the same component, we can find a path connecting them inside the component. As they collapse together, this path becomes a loop. You should convince yourself that there is no way of moving this loop in a continuous way to put it within $f^{-1}(z-\varepsilon)$. If you know what the fundamental group is, we are effectively adding one generator to the fundamental group. □

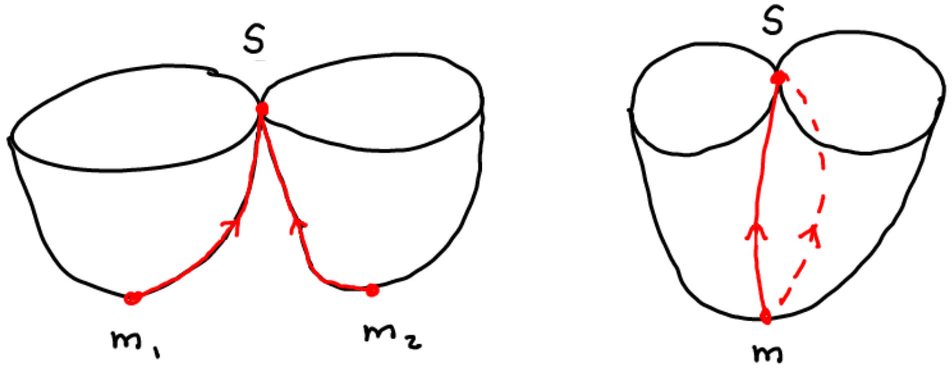


Figure 17. On the left hand side, two components come together. On the right hand side, two distinct points in the boundary of the same component come together. The red lines denote the gradient lines connecting the saddle point with the minima.

4.1.3. *What is the general pattern?* After studying these two cases, it might not be clear what we should expect in general. However, here is a little spoiler so that the case of maxima becomes clearer. When we cross a critical point p, either:

- We create a new closed manifold of dimension index(p). For instance, in index 0 we add a point, which corresponds to adding a connected component. In index 1 we add a circle (the new loop).
- We take a closed manifold of dimension index(p)-1 and we "eliminate it". When index(p) $= 0$ this cannot happen, but when index(p) $= 1$, we take two distinct components (i.e. two "points", which have dimension zero) and we join them together.

Remark 35. *It is not completely true that we create closed manifolds. We rather create closed* simplicial complexes *(if you do not know what these are, think about a space formed by gluing triangles/tetraheadra so that the resulting space has no boundary). In many cases these complexes are manifolds, but not always; this is known as the* Steenrod problem. *R. Thom showed that from dimension 7 onwards one cannot assume that these complexes are manifolds, i.e. "not every homology class can be realised by a submanifold".*

Remark 36. *By "eliminating" a closed manifold we mean making it* nullhomologous *in the sense of singular homology. If you do not know what this is, you should picture that when we cross the critical point we glue to our manifold a piece whose boundary is precisely the manifold of dimension* index(p) $- 1$. *This should become more clear once we discuss maxima.*

4.1.4. *Maxima.* Let us first explain what being nullhomologous means for a curve in our surface M.

Two oriented curves in a surface are **homologous** if they together bound a subsurface. A curve that bounds a subsurface by itself is called **nullhomologous** (because you can think of it as being homologous to the constant curve, which is just a point); see Figure 18 for examples. Being nullhomologous amounts to being zero in the abelianisation of the fundamental group.

Remark 37. *More in general, given a manifold M and a smaller manifold N inside of it, you could ask whether there exists a second manifold S inside of M such that the boundary of S is N. If this is the case, we can say N is nullhomologous. As before, one has to be careful and instead define this notion using simplicial complexes.*

Then:

Lemma 38. *If z is a critical value of f corresponding to a point of index two, then either:*

- *$f^{-1}(z+\varepsilon)$ has one more closed connected component than $f^{-1}(z-\varepsilon)$,*
- *a non-nullhomologous curve in $f^{-1}(z-\varepsilon)$ becomes nullhomologous in $f^{-1}(z+\varepsilon)$.*

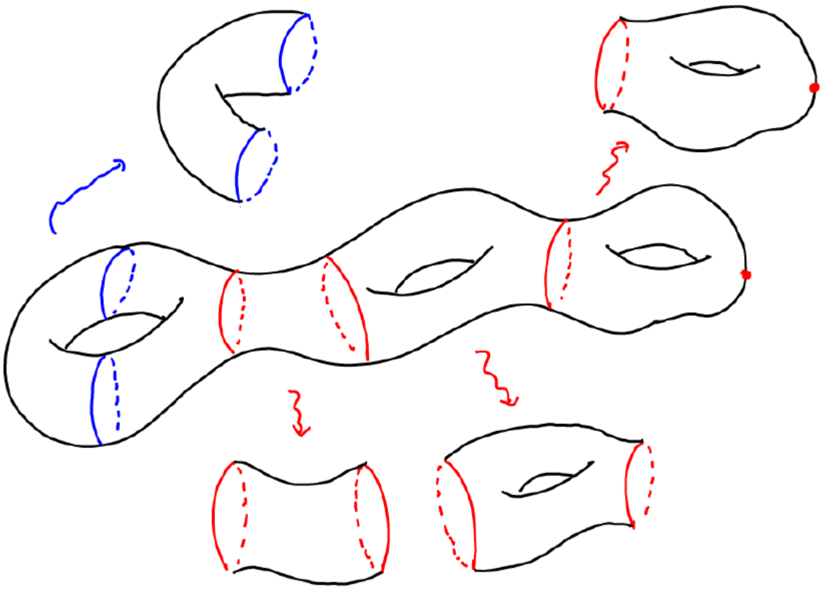


Figure 18. The figure shows a surface with 3 holes. The red curves are nullhomologous. The blue curves are not, but they are homologous to one another. For some pairs of homologous curves, we show the surface they bound together.

Proof. Take the little loop γ in the boundary of $f^{-1}(z-\varepsilon)$ that encircles the maximum of value z. There are two options: if it is nullhomologous, it is the only boundary of this connected component of $f^{-1}((-\infty, z-\varepsilon])$. This means that when we cross the critical point we create a new closed connected component (a new closed manifold of dimension 2). If the little loop is not nullhomologous, it certainly becomes so after the maximum appears, so we are in the second conclusion. See Figure 19. □

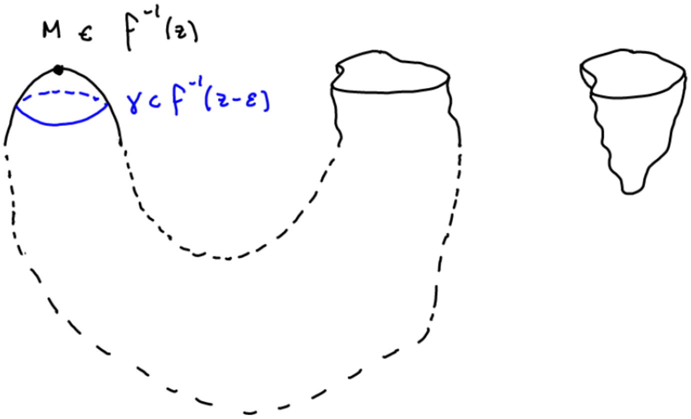


Figure 19. Here we see an example in which $f^{-1}(z)$ contains the maximum and two circles. One of the circles is in a different component, so we do not care. The other one is in our component, so the loop γ is not null-homologous until the maximum is reached.

In the introduction we discussed that as we modify a function, critical points appear and disappear in pairs. The idea in Morse homology is that when two such points appear, one of them creates a new closed manifold and the other point makes it nullhomologous. Therefore, counting the *"number of non-nullhomologous manifolds of a given dimension"* should provide an invariant of the manifold. See Figure 6.

4.2. **The Morse complex.** Let $f : M \to \mathbb{R}$ be a Morse function on a closed manifold M. We saw before (Proposition 24) that it is useful to find a vector field on M playing the role of the gradient of f. Let us fix a vector field ∇f satisfying the following properties:

- $\nabla f(p) = 0$ if and only if p is a critical point of f,
- $df(\nabla f) > 0$ if p is not a critical point,
- in the local model provided by Proposition 23, ∇f is the standard gradient.

This is what we did in Subsection 2.5, but now we add the third condition (it allows us to visualise and control the gradient close to the critical points easily).

Remark 39. *Let us look at Figure 17, particularly how the gradient behaves. Suppose we have a saddle that connects two different connected components. Then each of the two gradient lines going down from the saddle ends up in a different minimum. I.e. two different zero-dimensional manifolds have been connected by the gradient lines.*

If instead the gradient lines go to the same minimum, they generate a new non nullhomologous curve. Furthermore, if the two lines going upwards from the saddle end up in different maxima, we can see that the loop created by the saddle will divide the surface in two parts and is therefore nullhomologous. See Figure 20.

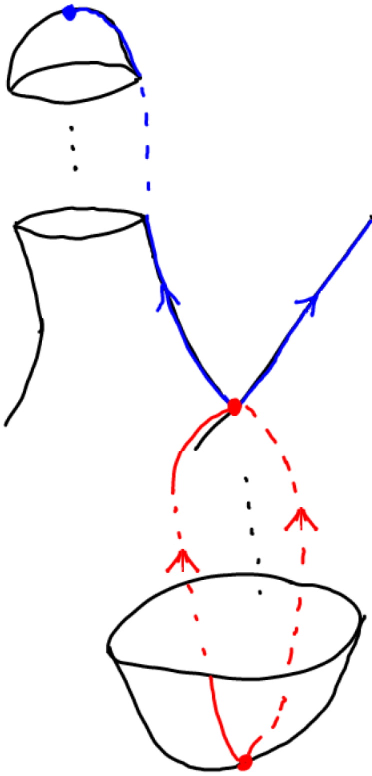


FIGURE 20. Since the two gradient lines going up from the saddle (in blue) go to different maxima, the two gradient lines going down to the minimum (in red) form a loop that divides the surface in two pieces.

The insight here is that looking at the gradient lines connecting the different critical points should recover the geometric phenomena of introducing and cancelling submanifolds.

Following this reasoning we are going to use ∇f to define an algebraic object, called the **Morse chain complex**. We define it as follows:

- $C^i(\nabla f)$ is the free abelian group $\mathbb{Z}[p^i_1, \cdots, p^i_{c_i}]$ generated by the critical points of f of index i. That is, an element in $C^i(\nabla f)$ is a *formal sum* of the form
$$\sum_{j=1}^{c_i} a_j p^i_j$$
where $\{p^i_1, \cdots, p^i_{c_i}\}$ is the collection of critical points of index i.
- There are group homomorphisms
$$\partial^i : C^i(\nabla f) \to C^{i-1}(\nabla f)$$
between these groups. It is sufficient for us to define what these homomorphisms do in terms of generators. We want $\partial^i(p^i_j)$ to be the collection of critical points that can be connected to p^i_j by a gradient flowline:
$$\partial^i(p^i_j) = \sum_{k=1}^{c_{i-1}} \#\{\text{gradient lines joining } p^i_j \text{ and } p^{i-1}_k\} p^{i-1}_k.$$
The point is that a particular point p^{i-1}_k can be connected to p^i_j by several gradient lines; if that happens we put a coefficient in front of it to remember this information.

Remark 40. *Observe that a gradient line is not simply an interval connecting two points. It is an interval with an orientation. When we say we count them, we must take this orientation into account, so some of them will count negatively. As an intuition, keep in mind the picture above: when two gradient lines going down from a saddle form a loop, they have opposite orientations, so when we pick an orientation for the loop one of them counts positively and the other one negatively.*

The groups $C^i(\nabla f)$, together with the maps ∂^i can be put together to form a **chain complex**:

$$C^0(\nabla f) \xleftarrow{\partial^1} C^1(\nabla f) \xleftarrow{\partial^2} \cdots \xleftarrow{\partial^{\dim(M)-1}} C^{\dim(M)-1}(\nabla f) \xleftarrow{\partial^{\dim(M)}} C^{\dim(M)}(\nabla f).$$

The map ∂^i is usually called the **differential**.

We reasoned above that a critical point either creates a new closed manifold or it cancels a manifold of dimension one less. This means that we should focus on those elements of $C^i(\nabla f)$ that do not cancel anything (because then they generate something new). This means, in terms of the map ∂^i, that we should look at its kernel $\ker(\partial^i) \subset C^i(\nabla f)$. Within this subgroup, not every element is good for us: we need to remove those elements that are cancelled by the points of index $i+1$. These are precisely the elements of $\text{image}(\partial^{i+1})$. This motivates us to look at the quotient:

$$H^i(\nabla f) = \frac{\ker(\partial^i)}{\text{image}(\partial^{i+1})}$$

which is called the $i^{\mathbf{th}}$ **Morse homology group**. Later on we shall show that under some reasonable assumptions $H^i(\nabla f)$ is well-defined and does not depend on ∇f. It will be an invariant of the manifold that we will denote by $H^i(M)$.

4.3. **2-dimensional examples.**

At this point there are many things to be checked, but first we will go through a few examples to get some intuition.

4.3.1. *The sphere.* Alright, so let us compute the Morse homology of both the unit sphere and the bean. Since they are actually homeomorphic, we should obtain the same result.

For the unit 2-sphere we see that there are a maximum and a minimum and no other critical points. Therefore our chain complex is:

$$C^0(\nabla f) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla f) = 0 \xleftarrow{\partial^2=0} C^2(\nabla f) = \mathbb{Z}[M].$$

And therefore:

$$H^0(\mathbb{S}^2) = \mathbb{Z}, \quad H^1(\mathbb{S}^2) = 0, \quad H^2(\mathbb{S}^2) = \mathbb{Z}.$$

Geometrically this corresponds to the fact that the sphere has one connected component and all the loops in it are nullhomologous.

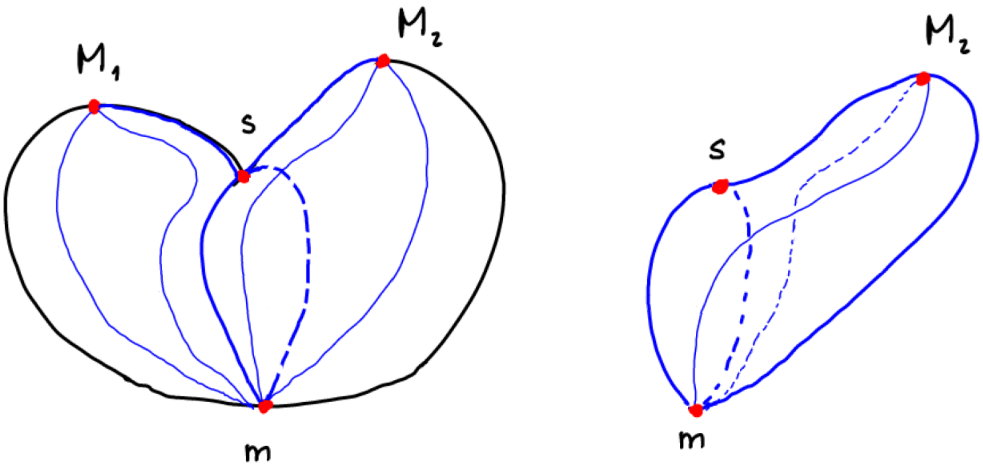


FIGURE 21. On the left hand side we see the gradient lines for the height function on the bean. On the right hand side we observe that the flowlines connecting M_2 with m form a disc whose boundary is given by those curves that are concatenations going from M_2 to s and then to m.

Similarly, if we compute the Morse chain complex of the bean (Figure 21) we have:

$$C^0(\nabla f) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla f) = \mathbb{Z}[s] \xleftarrow{\partial^2(M_i)=s} C^2(\nabla f) = \mathbb{Z}^2[M_1, M_2].$$

Note that $\partial^1(s)$ is indeed zero because the two gradient lines going down from s to m form a loop in which they have opposite orientations. There is only one gradient line going from M_i to s. Then:

$$H^0(\mathbb{S}^2) = \mathbb{Z}, \quad H^1(\mathbb{S}^2) = \mathbb{Z}/\mathbb{Z} = 0, \quad H^2(\mathbb{S}^2) = \mathbb{Z}[M_1 - M_2].$$

As before.

4.3.2. *The torus.* Let us draw the gradient lines in the torus as in Figure 22.

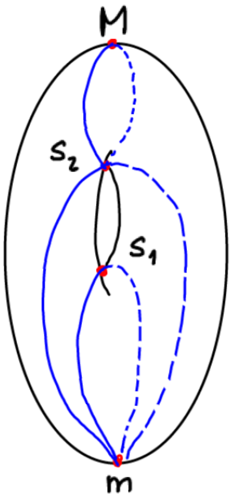


FIGURE 22. Gradient lines for the height function in the torus. Note that these are not the gradient lines that you would have chosen normally. Possibly you would have been tempted to draw them connecting s_2 and s_1. However, if ∇f is chosen generically, flowlines never connect saddle points to one another. This can be shown using Thom's transversality (see Proposition 29 and Proposition 45).

Then our chain complex is

$$C^0(\nabla f_2) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla f_2) = \mathbb{Z}^2[s_1, s_2] \xleftarrow{\partial^2=0} C^2(\nabla f_2) = \mathbb{Z}[M].$$

Reasoning as above we conclude that both differentials ∂^i are zero. Therefore:

$$H^0(\mathbb{T}^2) = \mathbb{Z}, \quad H^1(\mathbb{T}^2) = \mathbb{Z}^2, \quad H^2(\mathbb{T}^2) = \mathbb{Z}.$$

As we expected, since $\mathbb{T}^2$ has a single connected component and all the possible loops are generated by the parallel and the meridian.

4.3.3. *The surface of genus g.* Let us draw Σ_g, the surface of genus g (i.e. with g holes), as in Figure 23.

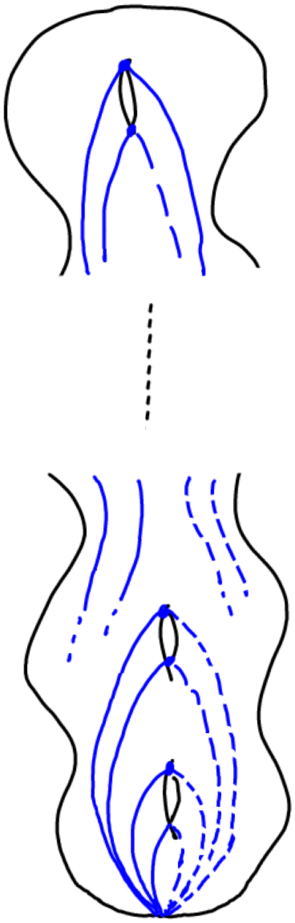


Figure 23. The height function on the genus g surface. We depict the descending manifolds of the index 1 points.

The gradient flowlines are very simple in this case, so we can readily see that our chain complex is

$$C^0(\nabla f_3) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla f_3) = \mathbb{Z}^{2g}[s_1, s_2, \cdots, s_{2g-1}, s_{2g}] \xleftarrow{\partial^2=0} C^2(\nabla f_3) = \mathbb{Z}[M].$$

and that both differentials ∂^i are zero. Therefore:

$$H^0(\mathbb{T}^2) = \mathbb{Z}, \quad H^1(\mathbb{T}^2) = \mathbb{Z}^{2g}, \quad H^2(\mathbb{T}^2) = \mathbb{Z}.$$

Indeed, all the loops in Σ_g are generated by a collection of $2g$ loops, as in Figure 24.

Remark 41. *A Morse function $f : M \to \mathbb{R}$ is said to be **perfect** if the corresponding differentials ∂^i are zero. In that case $H^i(M) = C^i(\nabla f)$. We have just shown that all closed surfaces admit perfect Morse functions. This is not true in general.*

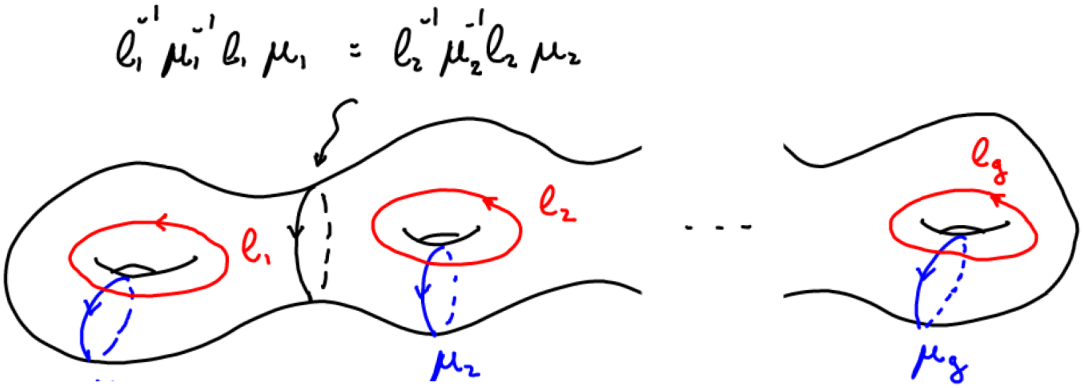


Figure 24. The curves in red and blue generate all the loops in the genus g surface. Note that the commutator of μ_i with l_i is independent of i, nullhomotopic, and not homotopically trivial.

4.4. **Ascending and descending manifolds.** Having gained some intuition through the examples, let us show that Morse homology is well-defined. The most pressing issue is whether the homomorphism ∂^i is well-defined at all. Indeed, this map counts the number of flowlines connecting the critical points of order $i+1$ with those of order i, and it is very much unclear whether there is a finite number of them. Let us look at the problem case by case.

Suppose we take a point of index 1. In the vicinity of such a point, our function looks like $-x_1^2+x_2^2+\cdots+x_n^2$. Refer to Figures 25 and 26.

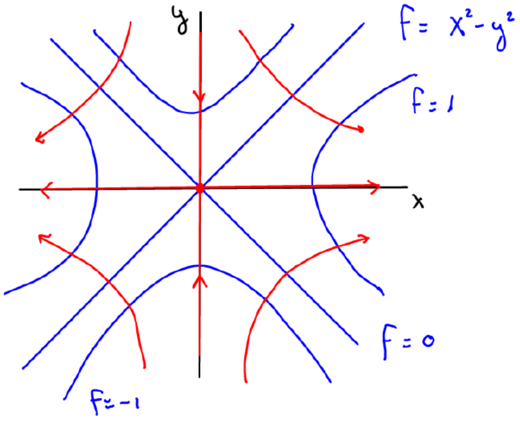


Figure 25. A critical point of index 1 in dimension 2.

The observation is that only two gradient flowlines go down from the point, so the number of flowlines connecting a minimum with an index 1 point is clearly finite (and at most 2!). If we consider a point of index $n-1$, we can reason similarly and conclude that it can only connect to a maximum by two flowlines.

The other cases are more involved. Let us take a point p of index s, so that our function looks like

$$f = -x_1^2 - x_2^2 + \cdots - x_s^2 + x_{s+1}^2 + \cdots + x_n^2.$$

For instance, if our point has index 2, the flowlines going down from p are arranged in a 2-dimensional disc, as in Figure 27.

More generally (recall that $\phi_t^{\nabla f}$ is the flow of ∇f):

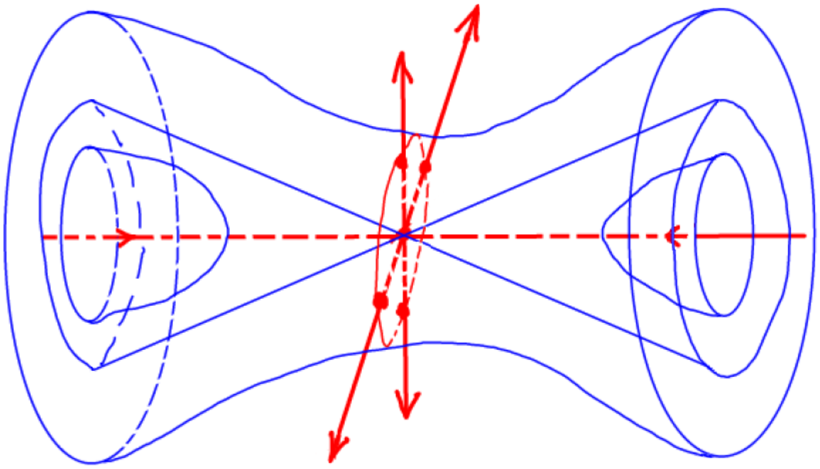


FIGURE 26. A critical point of index 1 in dimension 3.

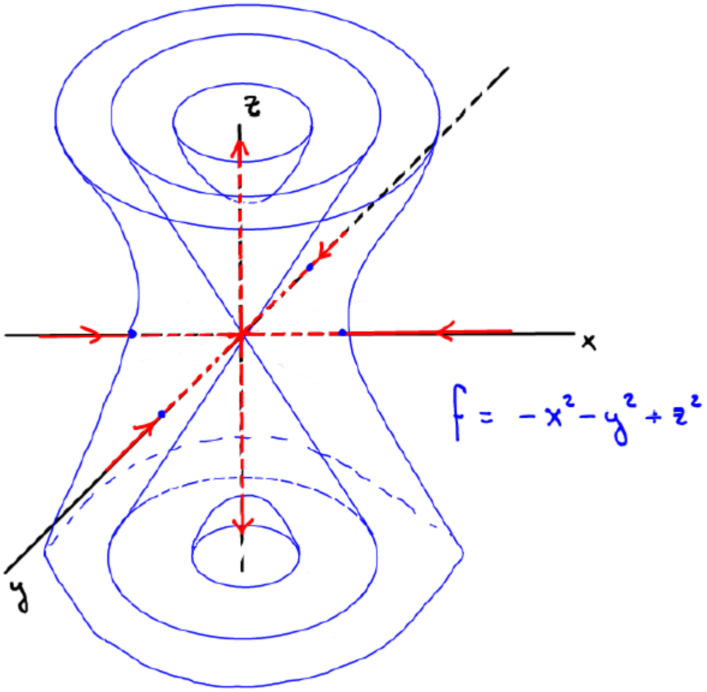


FIGURE 27. A critical point of index 2 in dimension 3. The flowlines going down correspond to the horizontal plane. The flowlines going up from the critical point correspond to the vertical axis.

Definition 42. *Let p be a critical point of index s. The **descending disc** of p is the set:*

$$\mathcal{D}(p) = \{q \in M | \lim_{t\to\infty} \phi_t^{\nabla f}(q) = p\}.$$

*The **ascending disc** of p is the set:*

$$\mathcal{A}(p) = \{q \in M | \lim_{t\to-\infty} \phi_t^{\nabla f}(q) = p\}.$$

I.e. the descending disc is formed by those flowlines that converge to the critical point as one moves positively in time. The ascending disc corresponds to those flowlines that converge to the critical point negatively in time.

Lemma 43. *Let M be an n-dimensional manifold. Let p be a critical point of index s. Then:*

- *$\mathcal{D}(p)$ is homeomorphic to an open disc of dimension s,*
- *$\mathcal{A}(p)$ is homeomorphic to an open disc of dimension $n - s$.*

Proof. In the local model the flowlines ending up at p are discs of the claimed dimension. Then we can reason as in the Morse lemma: following the flowlines identifies this small disc with the entirety of $\mathcal{D}(p)$ or $\mathcal{A}(p)$, respectively. □

The importance of the ascending and descending manifolds is the following trivial lemma:

Lemma 44. *Let p and q be critical points with $f(p) < f(q)$. The union of all flowlines connecting p with q is the set $\mathcal{A}(p) \cap \mathcal{D}(q)$.*

We want $\mathcal{A}(p) \cap \mathcal{D}(q)$ to be a finite collection of gradient flowlines under the assumption that p and q have indices i and $i+1$, respectively. The disc $\mathcal{A}(p)$ has dimension $n - i$ and $\mathcal{D}(q)$ has dimension $i + 1$; i.e. $\dim(\mathcal{D}(q)) - \operatorname{codim}(\mathcal{A}(p)) = 1$. We would like to reason as in Proposition 29 and say something like: "a little perturbation of $\mathcal{D}(q)$ makes it transverse to $\mathcal{A}(p)$ and therefore they intersect in a 1-dimensional set". Indeed:

Proposition 45. *Let p and q be critical points of indices i and $i + k$, respectively. For a generic choice of ∇f, the intersection $\mathcal{A}(p) \cap \mathcal{D}(q)$ is a finite collection of k–dimensional manifolds.*

In particular:

- *for $k = 0$, the intersection is empty,*
- *for $k = 1$ the intersection is a finite number of flowlines of ∇f.*

Proof. As in Proposition 29, Thom's transversality applies. However, do note that this is a slightly stronger version than before. We are claiming that the desired perturbation of $\mathcal{D}(q)$ can be achieved by perturbing ∇f itself. This is done by first moving $\mathcal{D}(q)$ as a submanifold and then modifying ∇f suitably. □

Remark 46. *We saw that most functions are Morse in a precise sense: non-Morse functions form a codimension-1 subspace inside the space of all functions. In a similar manner, within the space of all possible gradients, there is a codimension-1 subspace for which the conclusions of Proposition 45 fail. For instance: given a family of possible gradients $(\nabla^\varepsilon f)_{\varepsilon\in[0,1]}$ of f, usually there will be isolated values ε_0 for which $\nabla^{\varepsilon_0} f$ has two critical points of the same index connected by a flowline. This is called a* handleslide*.*

Remark 47. *To have a proper definition of ∂^i, we still have to describe explicitly which gradient lines count positively and which count negatively. In the examples below we will not need this, so feel free to skip this remark because it is quite technical.*

In any case, we proceed as follows: We select an orientation for our manifold M. Then we choose an orientation for each manifold $\mathcal{D}(p^i_j)$. This immediately provides an orientation for each ascending manifold $\mathcal{A}(p^i_j)$: Indeed, at the point p^i_j, we can choose a basis A of vectors tangent to $\mathcal{D}(p^i_j)$. We also pick a basis B for the vectors tangent to $\mathcal{A}(p^i_j)$. If we impose for A to be positive (for the orientation chosen) and $A + B$ to be positive (for the orientation chosen for M), then there is a unique choice for the orientation of B.

Now, we look at the intersection of $\mathcal{D}(p^i_j)$ with $\mathcal{A}(p^{i'}_{j'})$. Suppose we pick some orientation for $\mathcal{D}(p^i_j) \cap \mathcal{A}(p^{i'}_{j'})$, which at a point we represent by a basis C. Then there is a collection of linearly independent vectors A such that $C + A$ is a basis of $\mathcal{D}(p^i_j)$. Similarly, we extend C to a basis $C + B$ of $\mathcal{A}(p^{i'}_{j'})$. Then there is a unique choice of orientation of C making $C + A$, $C + B$, and $C + A + B$ positive (in $\mathcal{D}(p^i_j)$, $\mathcal{A}(p^{i'}_{j'})$, and M, respectively). This provides orientations for all intersections.

When $i' = i-1$, we can compare the orientation of $\mathcal{D}(p^i_j) \cap \mathcal{A}(p^{i'}_{j'})$ as a collection of gradient lines, with the orientation defined by the procedure we just described. If they agree, the gradient line counts positively and otherwise it counts negatively.

4.5. **The differential squares to zero.** We need one more ingredient to construct the homology groups above:

Proposition 48. *The composition $\partial^i \circ \partial^{i+1}$ is zero. In particular, the subgroup* $\mathrm{image}(\partial^{i+1})$ *is contained in* $\ker(\partial^i)$.

Let us look at the expression $\partial^i \circ \partial^{i+1}$. If we apply it to a critical point p^{i+1}_j of index $i+1$ we get:

$$\begin{aligned}
\partial^i \circ \partial^{i+1}(p^i_j) &= \partial^i(\sum_{k=1}^{c_i} \#\{\text{gradient lines joining } p^{i+1}_j \text{ and } p^i_k\} p^i_k) \\
&= \sum_{k=1}^{c_i} \sum_{l=1}^{c_{i-1}} \#\{\text{gradient lines joining } p^{i+1}_j \text{ and } p^i_k\} \#\{\text{gradient lines joining } p^i_k \text{ and } p^{i-1}_l\} p^{i-1}_l. \\
&= \sum_{l=1}^{c_{i-1}} \#\{\text{concatenation of two gradient lines joining } p^{i+1}_j \text{ with } p^{i-1}_l\} p^{i-1}_l
\end{aligned}$$

So we need to show that the numbers given by counting concatenations of two gradient flowlines should be zero. This motivates us to study the space of all gradient flowlines (not necessarily concatenations) between p^{i+1}_j and p^{i-1}_l. The descending manifold of p^{i+1}_j has dimension $i+1$ and the ascending manifold of p^{i-1}_l has dimension $n-i+1$. We might reason as in Proposition 45, and use the equality

$$(i+1) + (n-i+1) = n+2$$

to deduce that $\mathcal{D}(p^{i+1}_j) \cap \mathcal{A}(p^{i-1}_l)$ is 2-dimensional if we choose ∇f reasonably.

Let us look at the bean from Figure 21. The flowlines going down from each of the maxima to the minimum form a disc. The boundary of this disc is the union of the two curves given by concatenation and going from the maximum to the minimum. However, one of them is oriented properly (going counterclockwise around the disc) but the other one is not. We count one of them positively and the other one negatively, yielding zero.

This is the proof in general. The concatenations of two gradient lines joining p^{i+1}_j and p^{i-1}_l form the boundary of the 2-dimensional manifold $\mathcal{D}(p^{i+1}_j) \cap \mathcal{A}(p^{i-1}_l)$. This means that the union of all these gradient lines is nullhomologous and therefore, counted with appropriate orientations, they add to zero. □

4.6. **Invariance of Morse homology.** We have shown that there is this great algebraic gadget called Morse homology $H^i(\nabla f)$ that depends on the function f and on our choice of gradient ∇f. At this point, we would like to define:

Definition 49. *Let M be a closed manifold. The i^{th}* ***Morse homology group*** *of M is $H^i(\nabla f)$, where:*

- *$f : M \to \mathbb{R}$ is a Morse function,*
- *∇f is a choice of gradient for which $\mathcal{D}(p^i_k) \cap \mathcal{A}(p^{i'}_{k'})$ is a manifold of dimension $i - i'$.*

As we have remarked, most choices of f and ∇f satisfy these conditions.

So we must show that this definition does not really depend on ∇f nor on f. This is a bit involved, so let us just provide a picture and sketch the proof. We invite the reader to skip the rest of this section and jump to the examples.

Given two functions $f, g : M \to \mathbb{R}$, we can define the following function:

$$F : M \times [0,1] \to \mathbb{R}$$

$$F(p,t) = f(p)\chi(t) + g(p)(1-\chi(t)) + Ct^2(1-t^2).$$

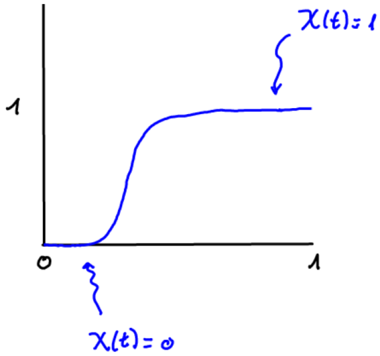


FIGURE 28. The function $\chi(t)$.

Here $\chi(t)$ is the function shown in Figure 28.

If C is sufficiently large, the critical points of F are exactly those of f (in $M \times \{0\}$) and those of g (in $M \times \{1\}$). Additionally, the points in $M \times \{1\}$ have index one more than previously. Now we can define a map Φ from $C^i(\nabla g)$ to $C^i(\nabla f)$ by counting flowlines. An argument as in Proposition 48 shows that this map descends to a map between $H^i(\nabla g)$ to $H^i(\nabla f)$; this boils down to showing that $\Phi \circ \partial^i = \partial^i \circ \Phi$.

By inverting the roles of f and g we can define a map Ψ from $C^i(\nabla f)$ to $C^i(\nabla g)$ as well. Proceeding as in Proposition 48 shows that $\Psi \circ \Phi$ induces the identity map in $H^i(\nabla g)$, showing that $H^i(\nabla g)$ and $H^i(\nabla f)$ are isomorphic. See Figure 29 for an example.

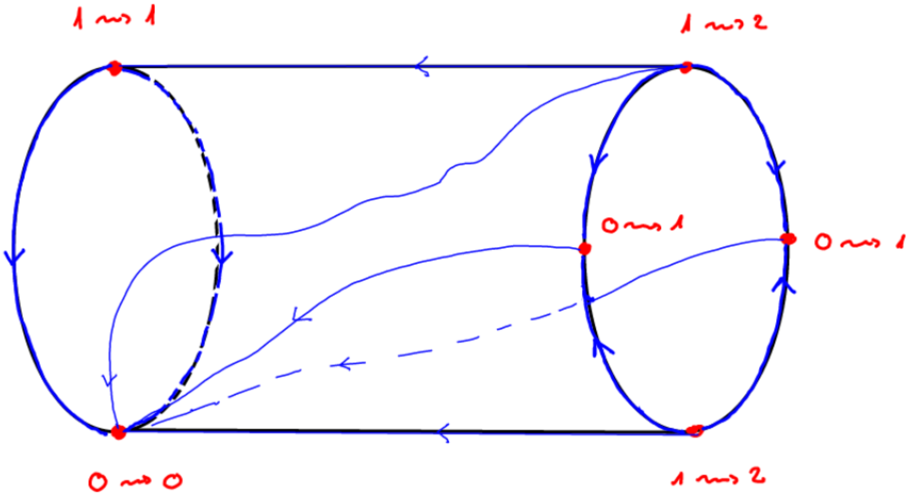


FIGURE 29. This is a cylinder, which we think of as $\mathbb{S}^1$ times the interval $[0, 1]$. The function f (left hand side) has a maximum and a minimum. The function g (right hand side) has two maxima and two minima. The gradient of the function F takes the upper maximum of g to the maximum of f, the two minima of g to the minimum of f, and the bottom maximum of g to zero. You should check that this defines a map between homologies that is the identity.

4.7. **3-dimensional examples.** With the tools we have now we can understand many examples of 3-dimensional manifolds. We will be able to tell them apart by computing their Morse homology.

4.7.1. *The standard sphere.* We can see the basic building blocks in dimension 3 by looking at the standard 3-sphere

$$\mathbb{S}^3 = \{(x_1, \cdots, x_4) \in \mathbb{R}^4 | \sum_i x_i^2 = 1\}.$$

You should probably imagine the 3-sphere as the 1-point compactification of $\mathbb{R}^3$ by adding the "point at infinity". Indeed, this identification can be obtained by looking at the stereographic projection of $\mathbb{S}^3$.

Let us look at the height function

$$h : \mathbb{S}^3 \to \mathbb{R}$$

$$h(x) = x_4.$$

It has a maximum and a minimum (the poles) and no other critical points. Its sublevel sets are:

$$h^{-1}((-\infty, t]) = \begin{cases} \emptyset & \text{if } t < -1 \\ \{(0,0,-1)\} & \text{if } t = -1 \\ \text{a 3–ball} & \text{if } -1 < t < 1 \\ \mathbb{S}^3 & \text{if } 1 \leq t \end{cases}$$

And all level sets inbetween -1 and 1 are 2-dimensional spheres. We can easily compute the Morse complex:

$$C^0(\nabla h) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla h) = 0 \xleftarrow{\partial^2=0} C^2(\nabla h) = 0 \xleftarrow{\partial^3=0} C^3(\nabla h) = \mathbb{Z}[M].$$

So h is a perfect Morse function and the homology agrees with the complex, i.e. $H^i(\mathbb{S}^3) = C^i(\nabla h)$. In particular, the fact that $H^1(\mathbb{S}^3) = 0$ tells us all loops are contractible and $H^2(\mathbb{S}^3) = 0$ tells us every closed surface divides $\mathbb{S}^3$ in several pieces.

4.7.2. *A characterisation of the sphere.* The following is a result of Reeb:

Lemma 50. *Let M be a closed 3-manifold endowed with a Morse function $f : M \to \mathbb{R}$ with only two critical points (a maximum and a minimum). Then M is diffeomorphic to $\mathbb{S}^3$.*

Proof. If f has only two critical points, M is obtained by gluing two copies D_1 and D_2 of the 3-disc $\mathbb{D}^3$ along their boundary $\mathbb{S}^2$; we call the gluing map $\psi : \partial D_1 \to \partial D_2$. Similarly, we can cut $\mathbb{S}^3$ into another two copies D_1' and D_2' of the 3-disc. Their boundaries are identified by the map $\psi' : \partial D_1 \to \partial D_2$ acting as $\psi'(x, y, z) = (x, y, -z)$. We identify D_1 with D_1'. Then we can identify D_2 with D_2' by the map

$$\Psi : (r, x, y, z) \to (r, \psi' \circ \psi^{-1}(x, y, z))$$

where r denotes the radius and (x, y, z) denote the coordinates in the boundary sphere. This is a homeomorphism but it is not differentiable at the origin of D_2 necessarily. See Figure 30. □

Remark 51. *One can modify the map Ψ to make it everywhere differentiable. From this we deduce that M is* diffeomorphic *to $\mathbb{S}^3$.*

In fact, the same statement can be proven for manifolds of dimension up to 6. However, from dimension 7 onwards one is unable to get rid of the non-differentiable point. Milnor gave examples of exotic 7-spheres (manifolds homeomorphic to the 7-sphere but not diffeomorphic to it) where this is precisely what happens. This says that one can obtain exotic n–spheres by considering "exotic" gluings between two $(n-1)$-spheres.

4.7.3. *Another function in the sphere.* The height function h on $\mathbb{S}^3$ had critical points of indices 0 and 3. However, as we discussed in the beginning, we should be able to introduce a pair of critical points s_1 and s_2 of indices 1 and 2 that cancel one another; i.e. $\partial^2(s_2) = s_1$.

Let us study how such a function f should look like. We start with the minimum m. When we cross it, a 3–ball appears. Then we cross the index 1 point s_1. The index tells us that only 2 gradient lines go down from s_1 towards m, so together they form a closed loop. I.e., after crossing the index 1 point our sublevel sets are solid tori. Let us look at the ascending manifold $\mathcal{A}(s_1)$ of s_1: it is the 2-dimensional disc whose boundary is the meridian of the torus. Refer to Figure 31.

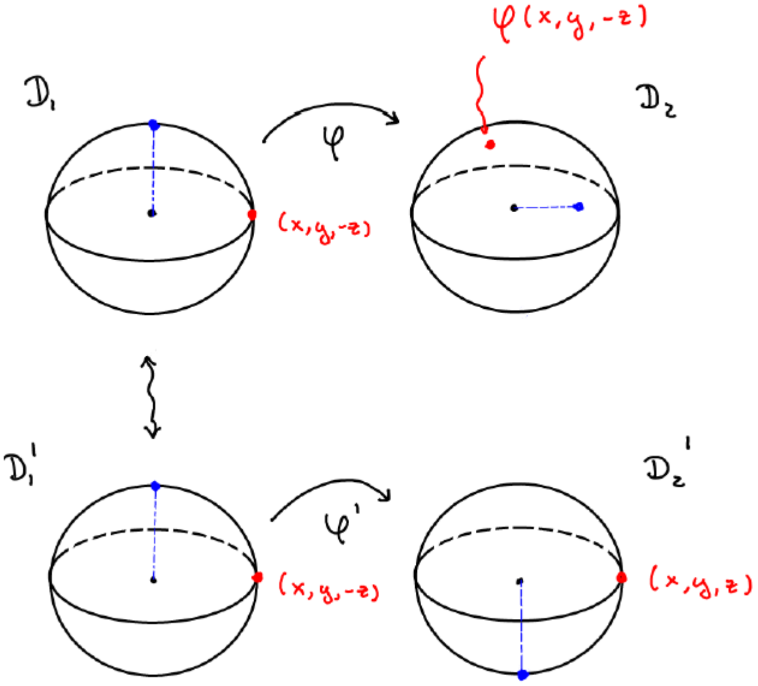


FIGURE 30. The gluing process from Lemma 50.

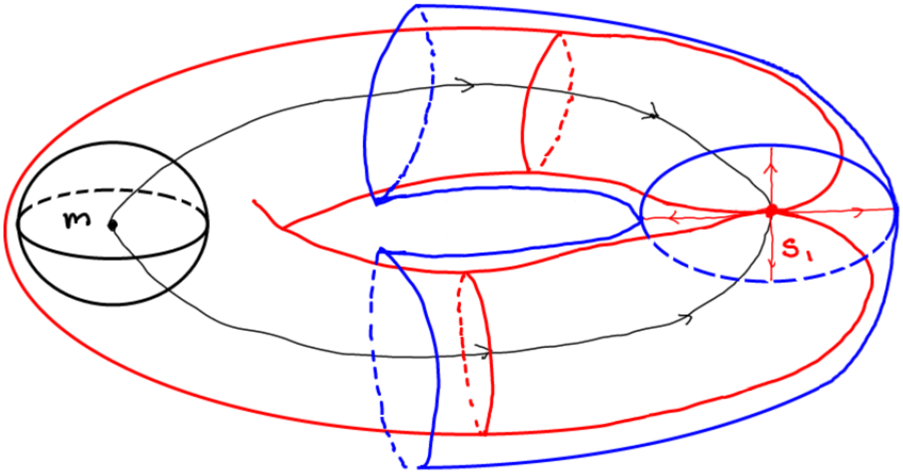


FIGURE 31. After crossing a minimum m and and index 1 point s_1 we obtain a solid torus. Its meridian is the ascending manifold of s_1.

Let us look at the point s_2 of index 2. Its descending manifold $\mathcal{D}(s_2)$ is a 2-dimensional disc. Since we want $\partial^2(s_2) = s_1$, this disc must cut $\mathcal{A}(s_1)$ in a single gradient flowline. This constrains how this disc must meet the boundary of our solid torus sublevel set. The easiest picture is shown in Figure 32.

After crossing s_2, the sublevel set becomes a 3–ball. We can see that the ascending manifold of s_2 is a pair of gradient flowlines (that of course end up in the only maximum M). The Morse complex reads:

$$C^0(\nabla f) = \mathbb{Z}[m] \xleftarrow{\partial^1=0} C^1(\nabla f) = \mathbb{Z}[s_1] \xleftarrow{\partial^2(s_2)=s_1} C^2(\nabla f) = \mathbb{Z}[s_2] \xleftarrow{\partial^3=0} C^3(\nabla f) = \mathbb{Z}[M].$$

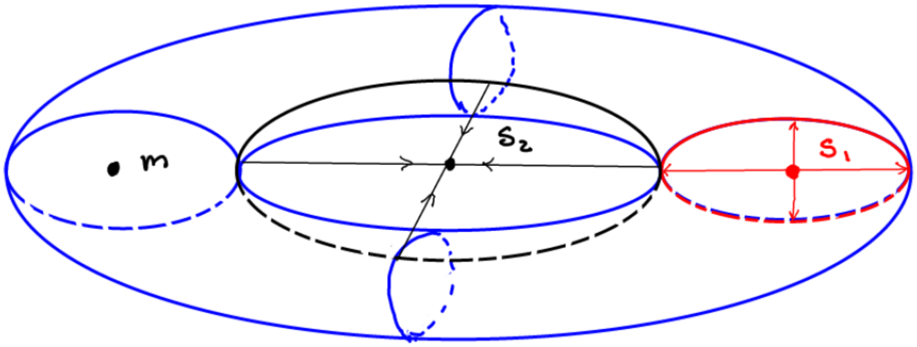


Figure 32. We cross and index 2 point s_2 whose descending manifold cuts the torus level set in its longitude. The longitude and the meridian cut in a single point so $\partial^2(s_2) = s_1$. After crossing s_2 the sublevel sets are 3-balls.

Which naturally yields the same result.

However, note that the function f has provided us with something interesting. If z is a critical value inbetween the critical values corresponding to s_1 and s_2, we have just shown that $f^{-1}((-\infty, z])$ is a solid torus. Look at $f^{-1}([z, \infty)$; it is also a solid torus! This is because it is obtained from the 3–ball at infinity by adding a tube (corresponding to s_2). I.e. the 3-sphere can be obtained by gluing 2-solid tori using the formula

$$\mu_1 \to l_2 \qquad l_1 \to -\mu_2$$

where μ_i is the meridian and l_i is the longitude in each of the 2–tori. See Figure 33.

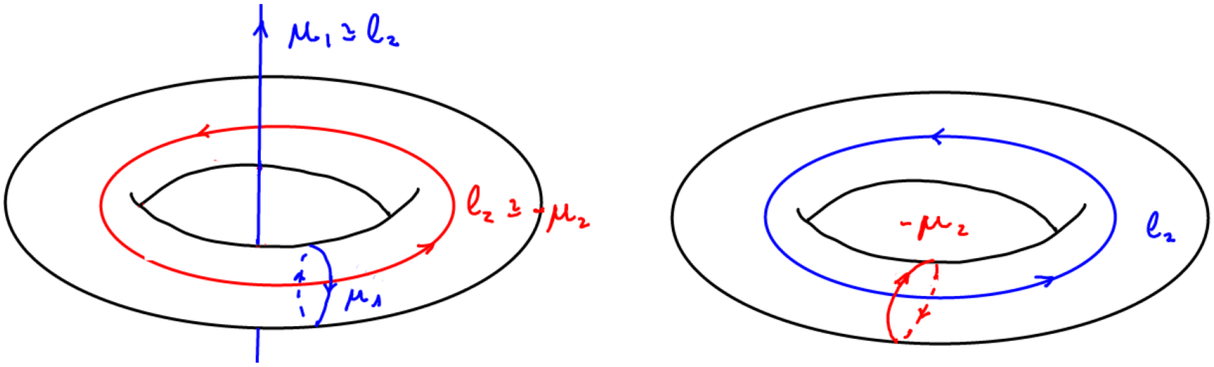


Figure 33. The 3-sphere can be obtained by gluing two tori. The torus on the right corresponds to the complement of the first in $\mathbb{S}^3$ (it goes across its hole).

4.7.4. *Lens spaces.* Having understood the 3-sphere, we can try to study other 3-manifolds. We will construct them by gluing elementary pieces associated to critical points. Then it will be very easy to compute their Morse homology, effectively showing that they are not the 3-sphere!

We should start by the easiest manifolds possible. We have seen that a 3-manifold with 2 critical points is just $\mathbb{S}^3$. Additionally:

Exercise 52. *A closed* 3*-manifold does not admit a Morse function with* 3 *critical points.*

Proof. Two of the points are the maximum M and the minumum m. Let p be the third critical point. p cannot be a minimum nor a maximum, because the manifold is closed. Suppose p has index 1: then we can cut the manifold in two pieces: one of them containing m and p, and therefore being a solid torus (with boundary a torus) and the other one containing M and therefore being a ball (with boundary the 2-sphere). We cannot glue the torus to the 2-sphere by a diffeomorphism. The reversed situation takes place when p has index 2. □

So we should look at Morse functions with 4 critical points; the interesting case is where there is one of each index. Let us denote them by m, s_1, s_2, and M. As before, once we cross m and s_1, our sublevel sets look like the solid torus. We additionally see the descending manifold $\mathcal{D}(s_2)$, which is a disc, approach the boundary of our sublevel set and cut it in some curve γ. The most convenient way of writing this curve is as $pl + q\mu$; this says that γ winds around the longitude p times and q times around the meridian. Refer to Figures 34 and 35.

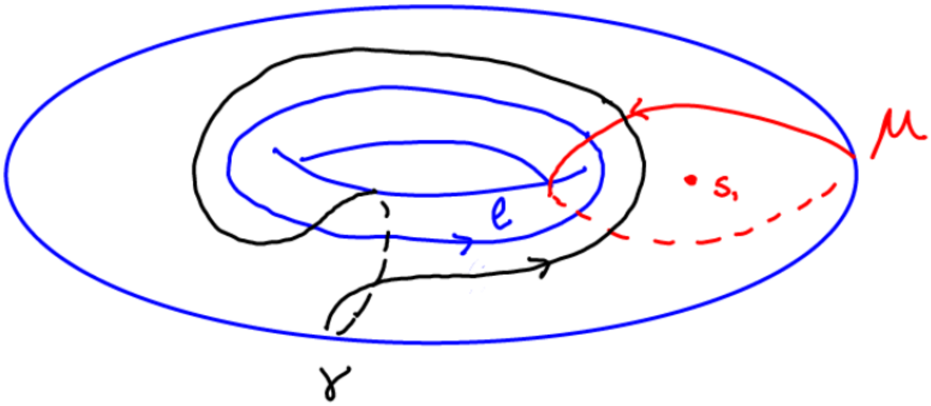


FIGURE 34. Here the curve γ is homologous to $l + \mu$, so $p = q = 1$.

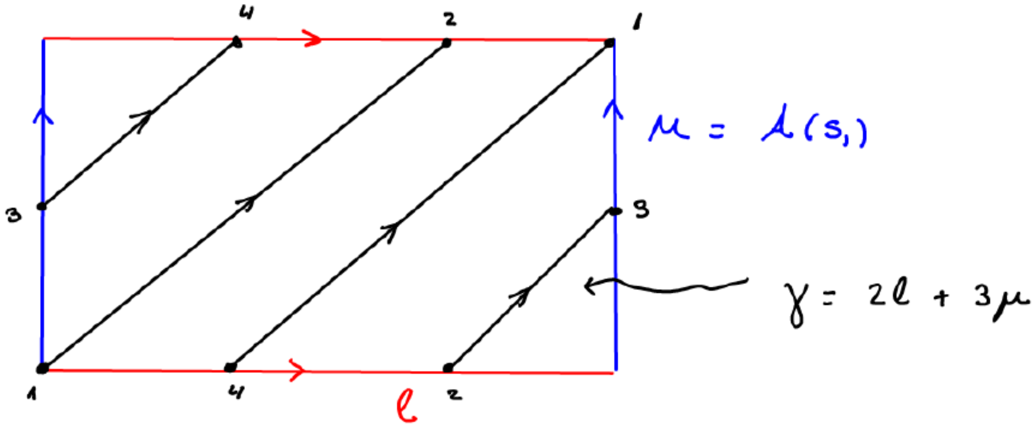


FIGURE 35. Here we depict the torus as a square with opposite sides identified. The curve γ is homologous to $2l + 3\mu$, so $p = 2$ and $q = 3$.

Definition 53. *The closed 3-manifold obtained by considering $\gamma \to pl + q\mu$ is said to be the* ***lens space*** *$L(p, q)$.*

Exercise 54. *All the lens spaces $L(1, q)$ are just $\mathbb{S}^3$.*

Proof. There exists a change of coordinates in our sublevel set that takes μ to μ and γ to l, effectively putting us into the situation above where we decomposed $\mathbb{S}^3$ into two solid tori. This change of coordinates is given by cutting the solid torus along a disc bounding the meridian μ and gluing it back introducing $-q$ turns along μ; look at Figure 36. □

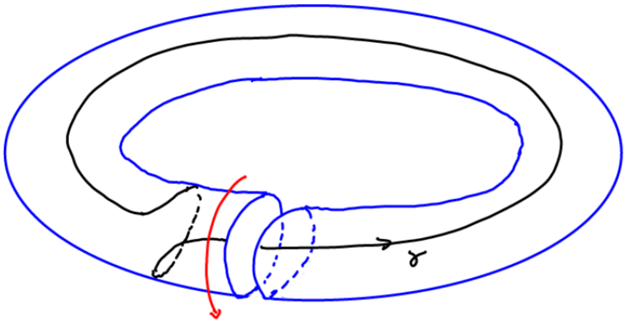


FIGURE 36. The curve γ corresponds to $l + \mu$. By cutting along the disc bounding μ, twisting a full turn, and gluing back again we map γ to l.

Alright, so we should look at pairs of numbers (p, q) where p is not 1. The key point is that this is enough information to compute the Morse complex associated to the function f. Indeed, if γ loops p times in the direction of l, it will cut μ p times (if we choose γ nicely). Since μ is the boundary of the ascending disc of s_1, we deduce:

$$\begin{tikzcd} C^0(\nabla f) = \mathbb{Z}[m] & C^1(\nabla f) = \mathbb{Z}[s_1] \arrow[l, "\partial^1=0"'] & C^2(\nabla f) = \mathbb{Z}[s_2] \arrow[l, "\partial^2(s_2)=ps_1"'] & C^3(\nabla f) = \mathbb{Z}[M]. \arrow[l, "\partial^3=0"'] \end{tikzcd}$$

And the homology of $L(p, q)$:

$$H^0(L(p,q)) = \mathbb{Z}, \quad H^1(L(p,q)) = \mathbb{Z}/p\mathbb{Z}, \quad H^2(L(p,q)) = 0, \quad H^3(L(p,q)) = \mathbb{Z}.$$

So $L(p, q)$ is not diffeomorphic to $L(p', q')$ if p and p' are different!

Remark 55. *There are a couple of interesting questions at this point. When are $L(p, q)$ and $L(p, q')$ homeomorphic? When are they homotopy equivalent? We have just shown that their homologies are the same and you can in fact prove easily, using Van Kampen, that their fundamental group is $\mathbb{Z}/p\mathbb{Z}$.*

To actually answer these questions you need to define another two invariants called the Reidemeister torsion *the* torsion linking form*, and it turns out that:*

- *they are homotopy equivalent if and only if qq' is a square mod p,*
- *they are homeomorphic if and only if q' agrees with $\pm q^{\pm 1}$ mod p.*

In particular, you have 3-manifolds for which the homotopy classification is not the same as the classification as manifolds! Recall that the Poincaré conjecture (now proven by Perelman) states that a manifold homotopy equivalent to $\mathbb{S}^3$ is homeomorphic to it.

4.8. **Euler characteristic.** The final result in these notes is:

Proposition 56. *The following alternated sums all compute the usual* ***Euler characteristic****:*

- $\sum_i (-1)^i \#\{\textit{critical points of index } i\} = \sum_i (-1)^i \dim(C^i(\nabla f))$,
- $\sum_i (-1)^i \dim(H^i(M))$,
- $\sum_i (-1)^i \#\{\textit{simplices of dimension } i \textit{ in a given triangulation}\}$.

Here by dim we mean the number of $\mathbb{Z}$ terms in the group H^i. As we saw before, sometimes H^i has a *torsion* part (for instance, $\mathbb{Z}/p\mathbb{Z}$). We ignore such terms.

Proof. The equivalence between the first two is a purely algebraic fact that follows by recalling that the kernel of ∂_i contains the image of ∂_{i+1}:

$$\begin{aligned}\sum_i (-1)^i \dim(C^i(\nabla f)) &= \sum_i (-1)^i [\dim(\ker(\partial_i)) + \dim(C_i/\ker(\partial_i))] \\ &= \sum_i (-1)^i [\dim(H^i(M)) + \dim(\text{image}(\partial_{i+1})) + \dim(\text{image}(\partial_i))] \\ &= \sum_i (-1)^i \dim(H^i(M)).\end{aligned}$$

The equivalence between the first and the last follows by choosing f suitably. Let us prove it just for surfaces (the general case is the same). Imagine a function that has a minimum for every vertex, a saddle for every edge, and a maximum for every face. This is easy to construct. The level sets start as circles around each vertex. Then they grow tentacles along the edge until they meet in the central point of the edge. Then the level sets grow inward into the faces until they disappear in the central point of the face. Then it is clear that both counts are the same. □

If you are familiar with other homology theories, in fact we may prove more:

Exercise 57. *For the particular choice of f given in the proof, show that the Morse complex is precisely the singular chain complex of the triangulation. As such, both homology theories are the same.*

4.9. **The end.** If you have made it this far, congrats! I hope that these notes were not too terrible (in fact, if you found any mistakes along the way, let me know). If you are interested in this topic, I would recommend you to look further into the references below.

References

[1] Y. Matsumoto. *An introduction to Morse theory*. Translations of Mathematical Monographs. Iwanami Series in Modern Mathematics (2002) 219 pp.

[2] J. Milnor. *Morse Theory*. Princeton University Press (1963). ISBN 0-691-08008-9

Departement Wiskunde, Universiteit Utrecht, Budapestlaan 6, 3584 Utrecht, The Netherlands

Email address: a.delpinogomez@uu.nl