Morse theory of critical points

Jessica Zhang

April 3, 2021

1 Introduction

The overarching goal of Morse theory is to use certain functions in order to understand what a manifold looks like. One example of a so-called "Morse function" is a "height function" on a manifold. The height function $h : M \to \mathbb{R}$ on the torus $M = S^1 \times S^1$ is shown in Figure 1.

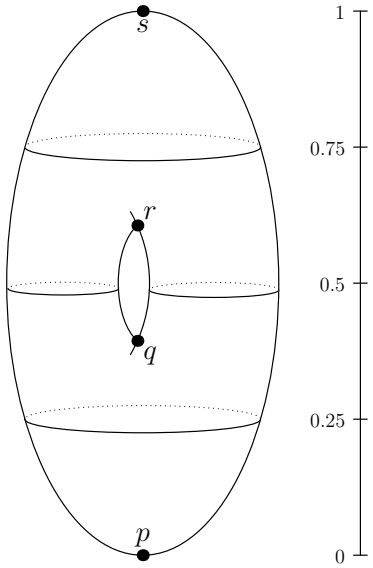

Figure 1: The height function on a torus.

Consider the four indicated points. Each of these are at a height where the *sublevel set* $M^a = h^{-1}((-\infty, a])$ changes. For example, as h passes through p, the level set changes from being empty to containing a single loop. As such the sublevel set changes from being the empty set to being a point or a disk (which is "basically" a point, since we can continuously squish it down into a point).

At q, the set M^a changes from containing a disk to a cylinder. The level set $h^{-1}(a)$ "pinches" itself and splits into two loops, but this pinching procedure isn't a continuous deformation. Indeed, a cylinder and a disk are fundamentally very different, and we can't get from one to the other simply by squishing them around.

The same thing happens as we pass r and s. In particular, at almost every point x, the sublevel sets $M^{x-\varepsilon}$ and $M^{x+\varepsilon}$ look the same. But at the "critical values" of $h(p)$, $h(q)$, $h(r)$, and $h(s)$, the sublevel sets suddenly change.

Since $M = M^\infty$, it follows that, if we can understand the critical points of h and if we can understand how M^a changes as it crosses each critical point, then we effectively "understand" the manifold M. This is guiding principle of Morse theory: *The critical points of a well-chosen function $f : M \to \mathbb{R}$ help us understand the topology of M.*

In this writeup, we will first discuss what a "well-chosen function" is. Then we will analyze in a bit more depth exactly what happens as we pass through each critical point of h. This will lead us to a definition of homotopy and CW complexes, and our guiding example of the height function on the torus will suggest a connection between any manifold M and a specific CW complex which can be defined using the critical points of a so-called Morse function on M.

2 Critical values and Morse functions

The points p, q, r, and s in Figure 1 are called critical points of h.

To explain why, let's briefly consider a smooth function $f : \mathbb{R} \to \mathbb{R}$. In this case, a *critical point* is just a point p where the derivative $f'(p) = 0$. At each critical point p, we can look at the sign of $f''(p)$ to see if p is a local minimum or maximum. In particular, if $f''(p) > 0$, then p is a local minimum. If $f''(p) < 0$, then p is a local maximum. And if $f''(p) = 0$, then this "second derivative test" is inconclusive, and p is usually (but not always) an inflection point.

In this last case, we could call p a *degenerate* critical point, since the second derivative doesn't give us any helpful information. In contrast, the points where $f''(p) \neq 0$ are *nondegenerate*, and we can understand their behavior by looking at their second derivative.

We can do the same thing in n dimensions. In particular, suppose M is an n-dimensional manifold, and suppose $f : M \to \mathbb{R}$ is smooth. (Here, as with elsewhere in this write-up, we use "smooth" to mean "infinitely differentiable.") Then the previous discussion suggests that we define a *critical point* of f as a point $p \in M$ if the differential $df : T_pM \to T_{f(p)}\mathbb{R}$ is zero. If we consider some local coordinates $(x^1, \ldots, x^n)$ of M, then T_pM can be given the coordinates $\frac{\partial}{\partial x^i}\big|_p$. Thus x is a critical point if and only if

$$\left.\frac{\partial f}{\partial x^1}\right|_p = \cdots = \left.\frac{\partial f}{\partial x^n}\right|_p = 0.$$

If p is a critical point, then we call $f(p) \in \mathbb{R}$ a *critical value* of f.

Similarly, we can use a "second derivative test" to differentiate between degenerate and nondegenerate critical points. In this case, we form a matrix

$$H := \left(\frac{\partial^2 f}{\partial x^i \partial x^j}(p)\right) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^1 \partial x^1}(p) & \frac{\partial^2 f}{\partial x^1 \partial x^2}(p) & \cdots & \frac{\partial^2 f}{\partial x^1 \partial x^n}(p) \\ \frac{\partial^2 f}{\partial x^2 \partial x^1}(p) & \frac{\partial^2 f}{\partial x^2 \partial x^2}(p) & \cdots & \frac{\partial^2 f}{\partial x^2 \partial x^n}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x^n \partial x^1}(p) & \frac{\partial^2 f}{\partial x^n \partial x^2}(p) & \cdots & \frac{\partial^2 f}{\partial x^n \partial x^n}(p) \end{pmatrix}.$$

It turns out that if this matrix, known as the *Hessian*, is invertible, then its eigenvalues determine whether or not p is a minimum, maximum, or saddle point. Thus we call it nondegenerate, since its behavior can be understood by looking at the second partial derivatives. If H is noninvertible, however, we call p a *degenerate* critical point.

Note that this coincides with our definition when $M = \mathbb{R}$. After all, for $f : \mathbb{R} \to \mathbb{R}$, the Hessian is singular exactly when the second derivative $\frac{d^2 f}{dx^2} = 0$.

One nice property of the Hessian is that it is symmetric, since

$$\frac{\partial^2 f}{\partial x^i \partial x^j} = \frac{\partial^2 f}{\partial x^j \partial x^i}.$$

Symmetric matrices are particularly useful because they give rise to symmetric bilinear forms. In particular, if A is symmetric, then we can define

$$Q(v, w) = v^T A w.$$

We denote the bilinear form corresponding to the Hessian H of f as f_{**}. (Note that this is a slight abuse of notation, since H also depends on p.) To put this all into coordinates, suppose that $v = \sum a_i \left.\frac{\partial}{\partial x^i}\right|_p$ and $w = \sum b_j \left.\frac{\partial}{\partial x^j}\right|_p$. Then we have

$$f_{**}(v, w) = v^T H w = \sum_{ij} a_i b_j \frac{\partial^2 f}{\partial x^i \partial x^j}(p).$$

The bilinear form f_{**} has an index and a nullity: We define its *index* to be the dimension of the subspace of T_pM on which f_{**} is negative definite. Its *nullity* is the dimension of its null space, that is, the subspace of vectors v so that $f_{**}(v, w) = 0$ for all $w \in T_pM$.

At a critical point p, note that the nullity is nonzero if and only if p is degenerate. After all, a tangent vector v is in the null space if and only if $Hv = 0$. If H is not invertible, i.e., if p is degenerate, then we can find a nontrivial vector v satisfying $Hv = 0$; otherwise, the only vector v with $Hv = 0$ is $H^{-1}(0) = 0$.

It turns out that the behavior of f near a nondegenerate critical point is entirely determined by its index at that point (or, more accurately, by the index of f_{**} at that point). In particular, we have the following theorem.

Theorem 2.1 (Morse Lemma)**.** *Let p be a nondegenerate critical point for f. Then there is a local coordinate system $(y^1, \dots, y^n)$ in a neighborhood U of p with $y^i(p) = 0$ for all i and such that the identity*

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

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

To prove this, we will first need a technical lemma.

Lemma 2.2. *Let f be a smooth function in a convex neighborhood V of 0 in $\mathbb{R}^n$, with $f(0) = 0$. Then there exist some smooth functions g_i defined on V such that*

$$f(x^1, \dots, x^n) = \sum_{i=1}^{n} x^i g_i(x^1, \dots, x^n) \tag{2}$$

and $g_i(0) = \frac{\partial f}{\partial x^i}(0)$.

Proof. Notice that

$$f(x^1, \dots, x^n) = \int_0^1 \frac{df(tx^1, \dots, tx^n)}{dt} dt.$$

But we know by the chain rule that

$$\frac{df(tx^1, \dots, tx^n)}{dt} = \sum_{i=1}^{n} \frac{\partial f}{\partial x^i}(tx^1, \dots, tx^n) \cdot x^i \, dt.$$

This implies that we can simply define

$$g_i(x^1, \dots, x^n) = \int_0^1 \frac{\partial f}{\partial x^i}(tx^1, \dots, tx^n) \cdot x^i \, dt. \qquad \square$$

With this lemma, we can now prove the Morse Lemma.

Proof of Theorem 2.1. We will first show that if there are coordinates $(z^1, \dots, z^n)$ which satisfy Equation (1), then λ must be the index of f at p.

Intuitively, this makes sense: Thus, the first λ coordinates $z^1, \dots, z^\lambda$ are "responsible" for all the "negative behavior" (which corresponds to the index) of the Hessian.

Indeed, taking the second derivative $\frac{\partial^2 f}{\partial z^i \partial z^j}$, we see that it is negative if and only if $i = j \leq \lambda$. In particular, we have

$$\frac{\partial^2 f}{\partial z^i \partial z^j} \begin{cases} -2 & \text{if } i = j \leq \lambda, \\ 2 & \text{if } i = j > \lambda, \\ 0 & \text{if } i \neq j. \end{cases}$$

Hence the Hessian with respect to the basis $\frac{\partial f}{\partial z^i}$ is exactly

$$\begin{pmatrix} -2 & & & & & \\ & \ddots & & & 0 & \\ & & -2 & & & \\ & & & 2 & & \\ & 0 & & & \ddots & \\ & & & & & 2 \end{pmatrix}.$$

This gives a subspace of dimension λ on which the Hessian is negative definite, since there are exactly λ negative eigenvalues (with multiplicity). Hence the index is at least λ.

Similarly, there is a subspace of dimension $n - \lambda$ on which the Hessian is *positive* definite. Thus λ is as large as possible, and so the index is exactly λ.

Now we must show that there does, in fact, exist a coordinate system $(y^1, \dots, y^n)$ satisfying the equation. We will begin with a coordinate system $(x^1, \dots, x^n)$ and show that $(y^1, \dots, y^n)$ can be obtained via a nonsingular linear transformation.

Translation implies that we can suppose without loss of generality that p is the origin of $\mathbb{R}^n = (x^1, \dots, x^n)$ and that $f(p) = 0$. We can thus apply Lemma 2.2 and rewrite f as

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

where $g_j(0) = \frac{\partial f}{\partial x^j}(0)$. But since $0 = p$ is a critical point, we know that all of the partial derivatives are equal to 0, so $g_j(0) = 0$.

Thus we can apply Lemma 2.2 to each g_j, which gives us smooth functions h_{ij} such that

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

and $h_{ij}(0) = \frac{\partial g_j}{\partial x^i}(0)$. Returning to f, we can thus say that

$$f(x^1, \dots, x^n) = \sum_{i,j=1}^{n} x^i x^j h_{ij}(x^1, \dots, x^n),$$

where each h_{ij} satisfies

$$h_{ij}(0) = \frac{\partial g_j}{\partial x^i}(0) = \frac{\partial^2 f}{\partial x^i \partial x^j}(0).$$

Note that $h_{ij}(0) = h_{ji}(0)$ for all i, j. In fact, it turns out that we can assume more generally that $h_{ij} = h_{ji}$. In particular, if $h_{ij} \neq h_{ji}$, then let $\bar{h}_{ij} = \frac{1}{2}(h_{ij} + h_{ji}) = \bar{h}_{ji}$. Moreover, we can verify that $f = \sum x^i x^j \bar{h}_{ij}$. Finally, since

$$\bar{h}_{ij}(0) = \frac{1}{2}(h_{ij}(0) + h_{ji}(0)) = \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^i \partial x^j}(0) + \frac{\partial^2 f}{\partial x^j \partial x^i}(0)\right) = \frac{\partial^2 f}{\partial x^i \partial x^j}(0),$$

we know that we could've just chosen $\bar{h}_{ij}$ for h_{ij} in the first place. Thus we can assume that $h_{ij} = h_{ji}$ for all i and j.

To prove that we can express the region near 0 as specified by Equation (1), we will perform induction. In particular, we will assume that we have some neighborhood U of 0 on which we have

$$f = \pm(u^1)^2 \pm \cdots \pm (u^{r-1})^2 + \sum_{i,j \geq r} u^i u^j H_{ij}(u^1, \ldots, u^n),$$

where $H_{ij} = H_{ji}$. This hypothesis holds when $r = 1$ thanks to our symmetry assumption on h_{ij}.

Now the goal is to find some new coordinates which do not affect $u^1, \ldots, u^{r-1}$, but will help us isolate $(u^r)^2$. To do so, we must first check that we can ensure that H_{rr} is nonzero near 0.

Indeed, there exists a change of the last $n - r + 1$ coordinates so that $H_{rr}(0) \neq 0$. After all, if this weren't the case, then we would have $H_{ij}(0) = 0$ for all $i, j \geq r$. But this would imply that 0 were a *degenerate* critical point of f. This is because the entries of its Hessian would be zero at every coordinate (i, j) if $i, j \geq r$, and so the Hessian would be singular.

Thus we without loss of generality suppose that the coordinates are such that $H_{rr} \neq 0$. The only H_{ij}'s that we want to change are where at least one of i, j is r. Thus we'll try to only change the u_r coordinate, say to v_r. Then our goal is to write f as

$$f = \sum_{i \leq r} \pm(v^i)^2 + \sum_{i,j > r} v^i v^j H'_{ij}(v^1, \ldots, v^n)$$

for some functions $v^i = v^i(u^1, \ldots, u^n)$.

Since we want the coefficient of $(v^r)^2$, it makes sense to define the function $g(u^1, \ldots, u^n)$ as

$$g(u^1, \ldots, u^n) = \sqrt{|H_{rr}(u^1, \ldots, u^n)|}.$$

Then we might want to define $v^r = gu^r$, so that $(v^r)^2 = \pm(u^r)^2 H_{rr}(u^1, \ldots, u^n)$.

But such a definition does not allow us to somehow get rid of the coefficients of $v^i v^r$ (or, symmetrically, of $v^r v^j$). Thus we would actually like to find some function $h(u^1, \ldots, u^n)$ so that $v^r = gu^r + h$ works by allowing us to cancel the extra terms obtained in $(v^r)^2$ with the terms in $v^i v^r = u^i v^r$ for $i > r$. In fact, because we know that h should probably include a term for each u^j, we can write it as $\sum u^j h_j$, where $j > r$.

In this case, we have

$$\begin{aligned}(v^r)^2 &= \pm(u^r)^2 H_{rr} + 2\sqrt{|H_{rr}|} u^r \sum_{i>r} u^i h_i + \left(\sum_{i>r} u^i h_i\right)^2, \\ v^i v^r &= u^i u^r \sqrt{|H_{rr}|} + \sum_{j>r} u^i u^j h_j.\end{aligned}$$

To get the $u^r u^i$ terms to cancel out, we need

$$H_{rr}\sqrt{|H_{rr}|} h_i = H_{ir}\sqrt{|H_{rr}|},$$

and so we simply need $h_i = \frac{H_{ir}}{H_{rr}}$. (Note that the factor of 2 drops out because $v^i v^r = v^r v^i$.)

Indeed, we can check that these coordinates work, and so induction implies that we can write $f = \sum \pm (v^i)^2$. □

The Morse Lemma tells us, in effect, that nondegenerate critical points locally look quadratic and are entirely determined by the index.

In the next couple sections, we will show that the topology of a space M is defined by the critical values of a so-called Morse function $f : M \to \mathbb{R}$. A *Morse function* on M is a smooth real-valued function on M such that none of its critical points are degenerate and if p, q are critical, then $f(p) \neq f(q)$.

The first condition means that we can use the Morse lemma to fully understand the behavior near each critical point. And since, as we will see, the topology of M only ever "changes" when crossing a critical point, the second condition lets us understand these changes without having them potentially influence one another.

Before we continue, it is worth noting that any smooth manifold admits a Morse function (and, indeed, uncountably many Morse functions!). We won't show this here, but the idea is to embed M into $\mathbb{R}^n$. This can always be done, as long as n is sufficiently large. Then let $L_p : M \to \mathbb{R}$ be defined as $L_p(q) = |p - q|^2$. (Note that the height function of the torus is just a particular example of such a function!) It turns out that, for almost every p, this function L_p has no degenerate critical points.

3 Morse functions and topology

In this section, we will discuss a particular kind of topological equivalence, namely homotopic equivalence. Intuitively, two objects are homotopic if there is some "continuous deformation" taking one to the other.

Formally, we say that two functions $f, g : X \to Y$ are *homotopic*, and write $f \simeq g$, if there is a continuous function $H : X \times [0,1] \to Y$ such that $H(x,0) = f(x)$ and $H(x,1) = g(x)$. We usually think of the first parameter of H as representing "space," while the second parameter represents"time." In other words, we think of H as a continuous way to start at f and end at g.

One example of a homotopy is the *straight-line homotopy*, which is defined by letting $H(x,t) = (1-t)f(x) + tg(x)$. If $X = [0,1]$, then we can visualize f and g as curves. Figure 2 shows an example of the straight-line homotopy in this case.

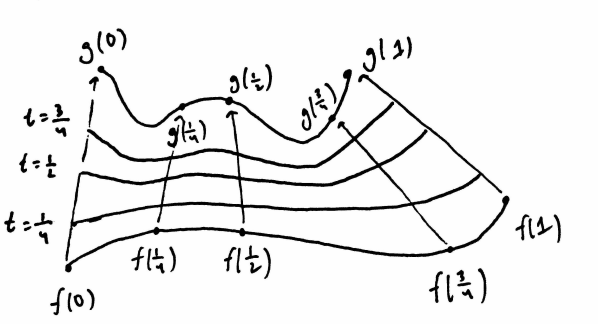

Figure 2: A straight-line homotopy between $f : [0,1] \to \mathbb{R}^2$ and $g : [0,1] \to \mathbb{R}^2$ effectively *deforms* f into g by pulling each $f(x)$ along the straight line toward $g(x)$.

Two spaces X and Y are *homotopy equivalent*, and have the same *homotopy type*, if there exist functions $f : X \to Y$ and $g : Y \to X$ such that $g \circ f \simeq \mathrm{id}_X$ and $f \circ g \simeq \mathrm{id}_Y$. If X is homotopy equivalent to a point, then we call it *contractible*, because we are basically able to contract it into a point.

One special case of homotopy equivalent spaces occurs when one is a deformation retract of another. If $Y \subseteq X$, then a deformation retraction of X onto Y is a function $F : X \times [0, 1] \to X$ such that

$$F(x, 0) = x, \quad F(x, 1) \in Y, \quad F(y, 1) = y,$$

where $x \in X$ and $y \in Y$. If such a function exists, then we call Y a *deformation retract* of X.

The types of homotopy equivalences which we are concerned with are mostly all deformation retracts, which should simply be thought of as continuous contractions of X onto a subspace Y. In general, we won't need to use the formal definitions of homotopies and deformation retracts in this write-up, and we'll just rely on this intuitive definition.

The reason we care about homotopy equivalence is because critical points turn out to tell us a lot about the homotopy type of a shape. Using the height function of the torus again, we start out with the empty set.

As we cross the point p, the sublevel set $f^{-1}((-\infty, a])$ becomes a disk, or, when $a = h(p)$, a single point. In fact, a disk is contractible. This can either be seen by by simply observing that a disk can be shrunk down to its center point, or by explicitly constructing the straight-line homotopy between a disk D and its center point 0.

The straight-line homotopy is achieved by letting $f : D \to \{0\}$ be the function taking every point to 0, and $g : \{0\} \to D$ be the inclusion map. Obviously $f \circ g$ is the identity, and thus is homotopic to the identity on $\{0\}$. But $g \circ f$ is homotopic to the identity (on D) as well, which we can see by considering the map $H(x, t) = (1 - t)x$. We have $H(x, 0) = x = \mathrm{id}_D(x)$ and $H(x, 1) = 0 = g \circ f(x)$.

Thus we see that we have effectively changed from the empty set to a point. In fact, we should think of this as *attaching* a point, or *adding* a point to the previous homotopy type, which happened to be the empty set.

When we pass q, the sublevel set becomes a cylinder. Note that a cylinder can be continuously compressed into a circle, but a circle cannot contract to a point, so the cylinder is not homotopy equivalent to the disk. In fact, in this case, the best way to think of the cylinder is as attaching a "handle" to the disk we already have, as illustrated in the second row of Figure 3.

Indeed, whenever we pass a critical point, the homotopy type of the sublevel set M^a changes by attaching some shape. In fact, each time, we seem to attach some *k-cell*, that is, some shape homeomorphic to a k-dimensional rectangle $[a_1, b_1] \times \cdots \times [a_k, b_k]$. This dimension k turns out to be exactly equal to the index λ.

We will first show that the homotopy type of M^a change when a does not pass through a critical value of f. We will then show that passing through a critical value of index λ has the effect of attaching a cell of dimension λ.

Proposition 3.1. *Let $f : M \to \mathbb{R}$ be smooth. Suppose now that $a < b$. Suppose moreover that $f^{-1}([a, b])$ is compact, and contains no critical points of f. Then M^a is diffeomorphic to M^b.*

Proof. The idea behind this proof is to construct a family of diffeomorphisms $\varphi_t : M \to M$ so that φ_{b-a} induces a diffeomorphism from M^a to M^b. Intuitively, the diffeomorphism φ_t will push M^b down onto M^{b-t} by moving perpendicular to the level sets of f.

To construct the family of diffeomorphisms, we use the *gradient* ∇f of f. The gradient is a vector field on M, that is, it attaches a vector $\nabla f(p)$ to each point p of M. It acts very similarly

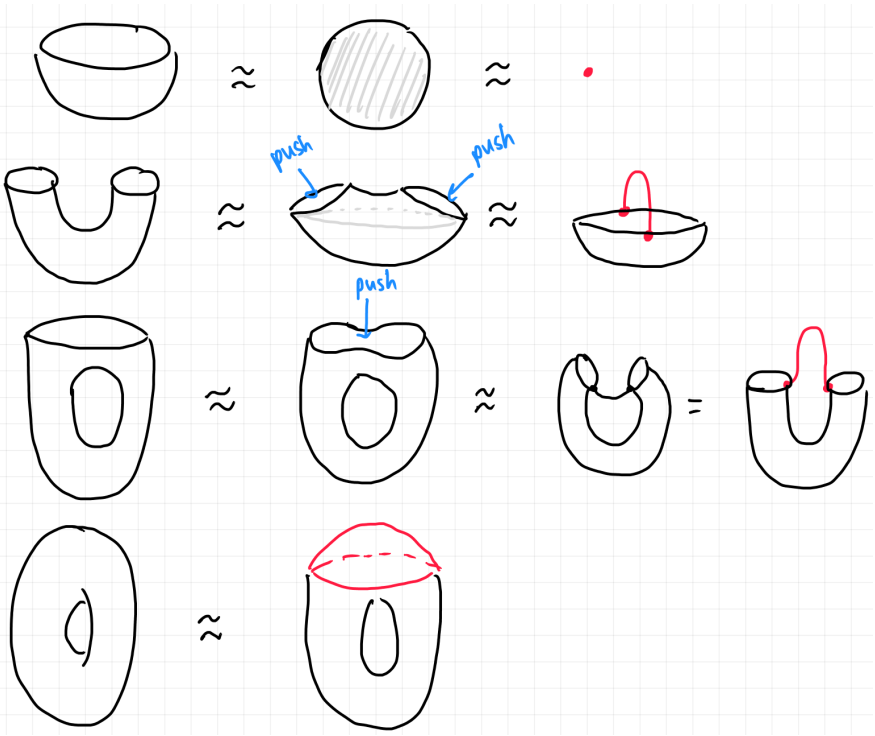

Figure 3: The homotopy type changes as we move across each critical point.

to the gradient in regular multivariable calculus. In particular, the gradient at p is orthogonal to the level set $f(x) = f(p)$ at noncritical points p, and vanishes at critical points.

Thus it makes sense to consider the vector field ∇f, and then to consider diffeomorphisms which follows the trajectories of ∇f at each point p. But these "diffeomorphisms which follow the trajectories of a vector field" can only be defined if the vector field vanishes outside of some compact set.

We can get around this difficulty by scaling ∇f. In particular, define $\lambda : M \to \mathbb{R}$ to be equal to $1/\langle \nabla f_p, \nabla f_p \rangle$ for $p \in f^{-1}([a,b])$, and to vanish outside a compact neighborhood K of this set. This is possible because $f^{-1}([a,b])$ is compact. (Note that the definition of λ on $f^{-1}([a,b])$ arises from the fact that we want the vectors of our scaled gradient to all be unit vectors. Note further that we never divide by zero because of the hypothesis that $f^{-1}([a,b])$ contains no critical points.)

Hence define the vector field $X : M \to T_pM$ by

$$X : p \mapsto X_p := \lambda(p) \nabla f_p.$$

By definition, this vanishes outside of the compact set K.

The diffeomorphisms in question are created by looking at all t for any given p. For each $p \in M$, let $\varphi(t,p)$ be the function in t defined by

$$\frac{\partial \varphi(t,p)}{\partial t} = X_{\varphi(t,p)} \tag{3}$$

with initial condition $\varphi(0,p) = p$. This can be done because it is known that ordinary differential equations locally have unique solutions, say for $t \in U_p = (-\varepsilon_p, \varepsilon_p)$.

We can do this for each p. Since K is compact, there exist finitely many U_p which cover K. Thus let ε_0 be the smallest of these ε_p's. Then by setting $\varphi(t,p) = 0$ for all $p \notin K$, it follows that we have a unique solution to Equation (3), at least for $|t| < \varepsilon$.

Suppose $|t| \geq \varepsilon$. If t is positive, then write $t = k \cdot \frac{\varepsilon}{2} + r$, where $k \in \mathbb{N}$ and $|r| < \frac{\varepsilon}{2}$. Then define

$$\varphi(t, p) = \varphi\left(\frac{\varepsilon}{2}, p\right) \circ \cdots \circ \varphi\left(\frac{\varepsilon}{2}, p\right) \circ \varphi(r, p),$$

where $\varphi(\frac{\varepsilon}{2}, p)$ is repeated k times. If t is negative, do the same thing with $-\frac{\varepsilon}{2}$.

Note that $\varphi(t, p)$ has thus been defined so that $\varphi(t, p) \circ \varphi(s, p) = \varphi(t + s, p)$ for every $t, s \in \mathbb{R}$. Thus we now have a family of diffeomorphisms

$$\varphi_t : p \mapsto \varphi(t, p)$$

where $t \in \mathbb{R}$ which satisfies the property that

$$\varphi_t \circ \varphi_s = \varphi_{t+s}.$$

Such a family is known as a *1-parameter group of diffeomorphisms*.

We claim that φ_{b-a} is a diffeomorphism which takes M^b to M^a. To do so, we will show that the map $t \mapsto f(\varphi_t(p))$ is a linear map with derivative 1 for any fixed p, as long as $f(\varphi_t(p)) \in [a, b]$. This follows from the fact that the dot product of ∇f_p and some tangent vector v is exactly equal to the directional derivative of f along v. In particular, this gives us

$$\frac{df(\varphi_t(p))}{dt} = \left\langle \frac{d\varphi_t(p)}{dt}, \nabla f_p \right\rangle = \langle X_p, \nabla f_p \rangle = 1,$$

as long as $\varphi_t(p) \in f^{-1}([a, b])$. This shows that M^a and M^b are diffeomorphic. □

Remark 3.2. It turns out that, by "moving through" the diffeomorphisms φ_t, we can actually construct a deformation retraction from M^b to M^a.

This means that the only points at which the manifold M^a can change are critical points. In fact, the following proposition characterizes how M^a changes across a critical point.

Proposition 3.3. *As before, let $f : M \to \mathbb{R}$ be smooth. Suppose q is a nondegenerate critical point of f with index λ, and define $c = f(q)$. Suppose there exists some $\varepsilon_0 > 0$ so that $f^{-1}([c-\varepsilon_0, c+\varepsilon_0])$ is compact and contains no critical point besides p. Then, for all sufficiently small ε, the sublevel set $M^{c+\varepsilon}$ has the homotopy type as $M^{c-\varepsilon}$ with a λ-cell attached.*

Note that if Proposition 3.3 can be applied to every critical point of a Morse function, then the critical points of f correspond to distinct critical values, and the critical values cannot have any accumulation points.

Proof. The goal will be to create an auxiliary function $F : M \to \mathbb{R}$ which differs from f only in a small neighborhood near q. It will be defined so that $F^{-1}((-\infty, c - \varepsilon])$ consists of $M^{c-\varepsilon}$, along with some region H containing p.

We will show that there is a λ-cell $e^\lambda \subseteq H$ so that $M^{c-\varepsilon} \cup e^\lambda$ can be attained by continuously deforming $M^{c-\varepsilon} \cup H$, and hence is homotopy equivalent to $M^{c-\varepsilon} \cup H$. Then we will apply Proposition 3.1 to the function F to show that $M^{c-\varepsilon} \cup H$ and $M^{c+\varepsilon}$ are homotopy equivalent, which will complete the proof. As an example on the solid torus, consider Figure 4.

Since we'll be working only near p, we can use $(u^1, \ldots, u^n)$ as the coordinate system of some neighborhood U of p from Theorem 2.1. Thus

$$f = c - (u^1)^2 - \cdots - (u^\lambda)^2 + (u^{\lambda+1})^2 + \cdots + (u^n)^2$$

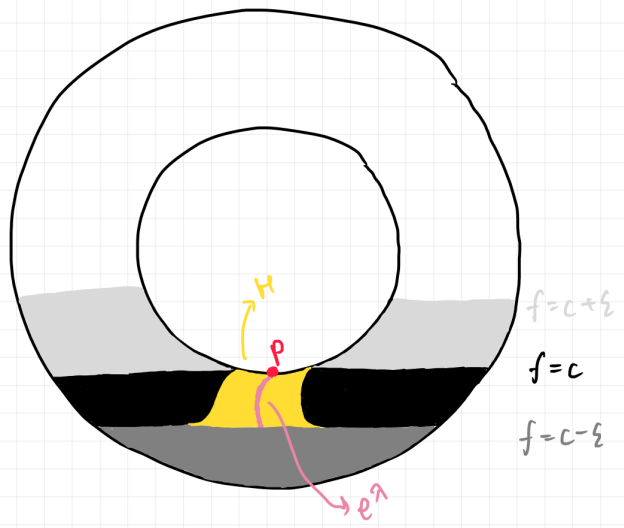

Figure 4: On the torus, the argument will roughly be to push $M^{c+\varepsilon}$ to $M^{c-\varepsilon} \cup H$, and then to squeeze the yellow area H into a single pink λ-cell e^λ.

on U. Observe that the critical point itself has coordinates $u^i = 0$.

It will be convenient to define the functions $\xi, \eta : U \to [0, \infty)$ as

$$\xi = (u^1)^2 + \cdots + (u^\lambda)^2$$
$$\eta = (u^{\lambda+1})^2 + \cdots + (u^n)^2.$$

With this definition, we can rewrite f as

$$f = c - \xi + \eta.$$

One way to visualize our current coordinate setup is to imagine having two axes, one representing the coordinates $u^1, \ldots, u^\lambda$, and the other representing the coordinates $u^{\lambda+1}, \ldots, u^n$. Then the level sets look like hyperbolas in this two-dimensional rendering thanks to the quadratic form of U. This is shown in Figure 5.

Now choose $\varepsilon < \varepsilon_0$ so that the image of U under the coordinate map

$$(u^1, \ldots, u^n) : U \to \mathbb{R}^n$$

contains the closed ball

$$\{(u^1, \ldots, u^n) : \sum (u^i)^2 \leq 2\varepsilon\}.$$

(The reasoning for this extra condition is simply that our definition of F will require a "bump" function μ which is flat everywhere outside this ball; thus we need this ball to be contained inside U so that we can use the coordinate system defined by the u^i's.) This is shown in Figure 6. We claim that, for any such ε, the conclusion of this proposition holds.

Now recall that we would like to find a smooth function F so that $F(p) < c-\varepsilon$, but $F^{-1}((-\infty, c+\varepsilon]) = M^{c+\varepsilon}$, where $\varepsilon > 0$ is suitably small. One way to do this is to define some "perturbation function" which is greater than ε at the point q and which gradually decreases away from q.

It turns out that, for each $\varepsilon > 0$, there exists a smooth function $\mu : \mathbb{R} \to \mathbb{R}$ such that

$$\mu(0) > \varepsilon$$
$$\mu(r) = 0 \quad \text{for } r \geq 2\varepsilon$$
$$\mu'(r) \in (-1, 0] \quad \text{for all } r.$$

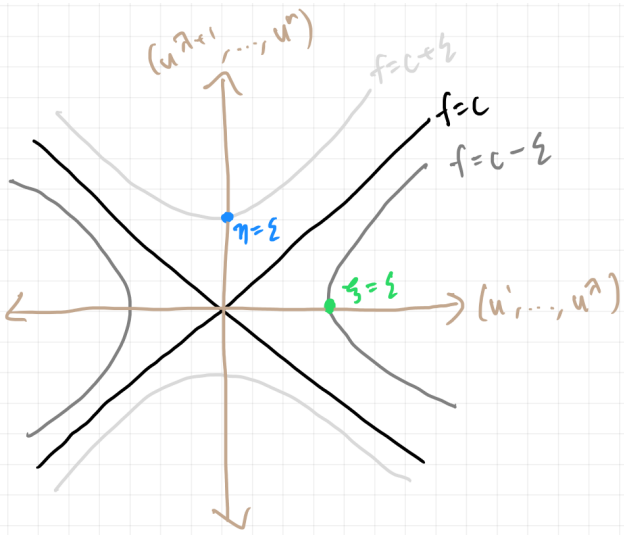

Figure 5: A schematic of the coordinate setup. Note that $f = c - \xi + \eta$ implies, for example, that the intersection of the $f = c - \varepsilon$ hyperbola and the $(u^1, \dots, u^\lambda)$-axis is the set of points with $\xi = \varepsilon$, as indicated by the green point.

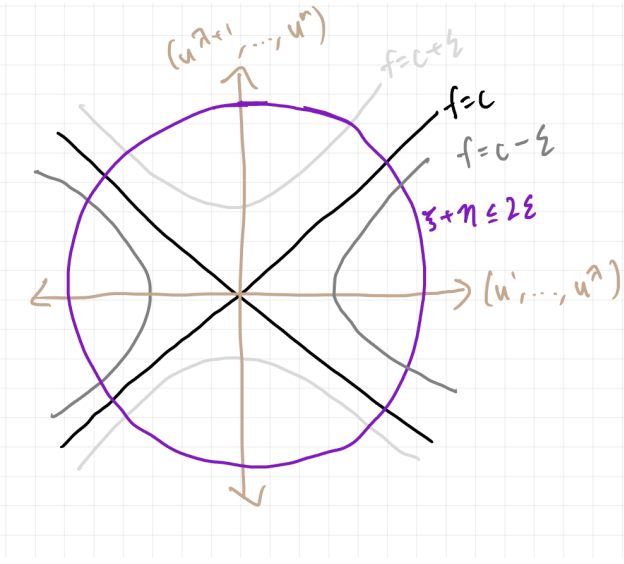

Figure 6: The purple ball intersects the $(u^1, \dots, u^\lambda)$-axis where $\xi = 2\varepsilon$. Our condition simply says that the ball is contained in our coordinate system.

The existence of such a function is nontrivial, but since it relies on some other machinery, we will simply take this fact for granted. Note that the 2ε in the second condition can't be replaced by ε, for example, since that would violate the third condition, which is what forces μ to decrease relatively gradually.

Now define F to coincide with f outside of U, and to be equal to

$$\begin{aligned} F(p) &= f(p) - \mu((u^1)^2 + \cdots + (u^\lambda)^2 + 2(u^{\lambda+1})^2 + \cdots + 2(u^n)^2) \\ &= c - \xi + \eta - \mu(\xi + 2\eta). \end{aligned}$$

The reason, intuitively, as to why we choose the input of μ to be $\xi + 2\eta$ is twofold: On one hand, we need the input of μ to grow when the distance from the origin grows, since we need the perturbation factor to be small *anywhere* far away from the origin (which is p). And on the other hand, we still want to maintain some kind of "imbalance" between ξ and μ in the input, since this will help achieve the imbalance between $F^{-1}((-\infty, c-\varepsilon])$ and $F^{-1}((-\infty, c+\varepsilon])$.

We will show that F has the necessary properties to make it behave "like f." The important properties of F, in particular, will be that $F^{-1}((-\infty, c+\varepsilon]) = M^{c+\varepsilon}$ and that F has exactly the same critical points as f.

We begin by showing that $F^{-1}((-\infty, c+\varepsilon])$ is equal to $M^{c+\varepsilon} = f^{-1}((-\infty, c+\varepsilon])$. If $p \notin U$, then f and F are defined to be equal. Otherwise, if $\xi + 2\eta \geq 2\varepsilon$, then we know that the perturbation factor μ is equal to 0, and so $F = f$ in this case. Otherwise, suppose $\xi + 2\eta < 2\varepsilon$ at some point p. We claim that $f(p), F(p) \leq c + \varepsilon$, so that p is contained in both $F^{-1}((-\infty, c+\varepsilon])$ and $M^{c+\varepsilon}$. To see this, recall that $F(p) \leq f(p)$ since μ is nonnegative when $\xi + 2\eta < 2\varepsilon$. Hence we have

$$F(p) \leq f(p) = c - \xi + \eta \leq c + \frac{1}{2}\xi + \eta < c + \varepsilon,$$

which proves that the sublevel sets of f and F are the same at $c + \varepsilon$.

To show that F and f have the same critical points, note that, as before, we can assume that p has $\xi + 2\eta < 2\varepsilon$. Note that p is a critical point if and only if the total derivative $dF_p = 0$. But we also know that

$$dF = \frac{\partial F}{\partial \xi} d\xi + \frac{\partial F}{\partial \eta} d\eta.$$

Recalling that $F = c - \xi + \eta - \mu(\xi + 2\eta)$, we can evaluate the partial derivatives as

$$\begin{aligned} \frac{\partial F}{\partial \xi} &= -1 - \mu'(\xi + 2\eta) \in [-1, 0) \\ \frac{\partial F}{\partial \eta} &= 1 - 2\mu'(\xi + 2\eta) \in [1, 3). \end{aligned}$$

In particular, neither partial derivative is ever 0. Thus $dF = 0$ implies that $d\xi = d\eta = 0$ at p. But this is only possible if $u^i = 0$ for each i. Since the origin of the $(u^1, \ldots, u^n)$-coordinate system is just q, this means that q is the only critical point of F with $\xi + 2\eta < 2\varepsilon$. This proves that F and f have the same critical points.

This means that Proposition 3.1 is applicable to the region $F^{-1}([c-\varepsilon, c+\varepsilon])$. To see this, note that $F \leq f$ and the fact that $F^{-1}((-\infty, c+\varepsilon]) = M^{c+\varepsilon}$ imply that

$$F^{-1}([c-\varepsilon, c+\varepsilon]) \subseteq f^{-1}([c-\varepsilon, c+\varepsilon]).$$

But the only critical point in $f^{-1}([c-\varepsilon, c+\varepsilon])$ is p, and $F(p) < c - \varepsilon$, and so it follows that $F^{-1}([c-\varepsilon, c+\varepsilon])$ cannot contain any critical points. Not furthermore that $f^{-1}([c-\varepsilon, c+\varepsilon])$ being

compact and $F^{-1}([c-\varepsilon, c+\varepsilon])$ being a closed subset implies that the latter set is also compact. Thus we can apply Proposition 3.1.

In other words, we conclude that $M^{c+\varepsilon}$ and $F^{-1}((-\infty, c-\varepsilon])$ are diffeomorphic, hence homotopy equivalent. For convenience, we write $F^{-1}((-\infty, c-\varepsilon]) = M^{c-\varepsilon} \cup H$, where H is the closure of $F^{-1}((-\infty, c-\varepsilon]) \setminus M^{c-\varepsilon}$. This H is exactly the H mentioned at the beginning of this proof.

As such, we now will show that there is a λ-cell $e^\lambda \subset H$ such that $M^{c-\varepsilon} \cup e^\lambda$ is homotopy equivalent to $M^{c-\varepsilon} \cup H = F^{-1}((-\infty, c-\varepsilon])$. Since we already showed that this last expression is homotopy equivalent to $M^{c+\varepsilon}$, this will complete the proof.

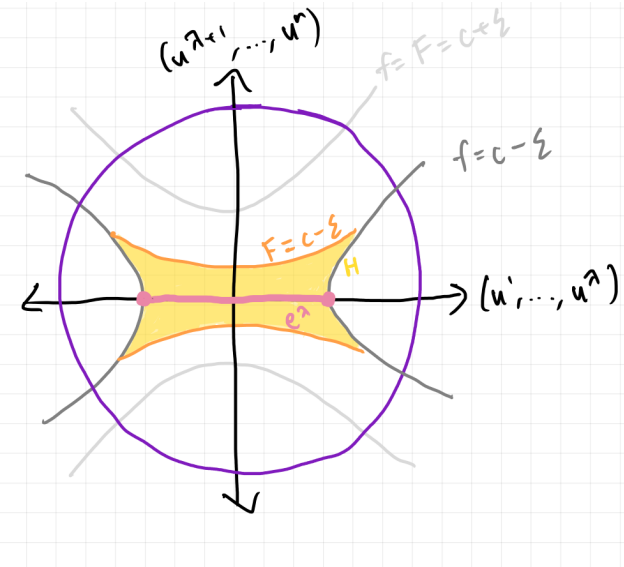

Figure 7: The handle H contains the cell e^λ. The set between the orange and light gray hyperbolas is $F^{-1}([c-\varepsilon, c+\varepsilon])$. Moreover, the two areas bounded by the dark gray $f = c-\varepsilon$ hyperbolas make up $M^{c-\varepsilon}$, so e^λ is connected to $M^{c-\varepsilon}$ as expected.

In fact, we can use the diagram shown in Figure 7 Note that the pink line e^λ is indeed λ-cell since the "x-axis" is really a λ-dimensional axis. Moreover, the diagram suggests that the cell we want is simply the set of points q with

$$\xi(q) \leq \varepsilon, \quad \eta(q) = 0.$$

Indeed, this cell is always contained in H. To see this, suppose $q \in e^\lambda$. Then we must show that $F(q) \leq c - \varepsilon$, but $f(q) \geq c - \varepsilon$. The latter inequality follows from the fact that

$$f(q) = c - \xi(q) + \eta(q) \geq c - \varepsilon.$$

The former follows from the fact that $\xi(q) \geq 0 = \xi(p)$. After all, since $\frac{\partial F}{\partial \xi} < 0$, it follows that

$$F(q) \leq F(p) < c - \varepsilon.$$

Thus $e^\lambda \subset H$.

This brings us to the final part of the proof: showing that e^λ is a suitable choice of λ-cell, i.e., that $M^{c-\varepsilon} \cup e^\lambda$ is a deformation retract of, and hence homotopy equivalent to, the set $M^{c-\varepsilon} \cup H = F^{-1}((-\infty, c-\varepsilon])$.

The deformation retraction is illustrated in Figure 8. Effectively, the red region gets pushed vertically onto the pink cell e^λ, the green region is pushed vertically onto a segment of the hyperbola

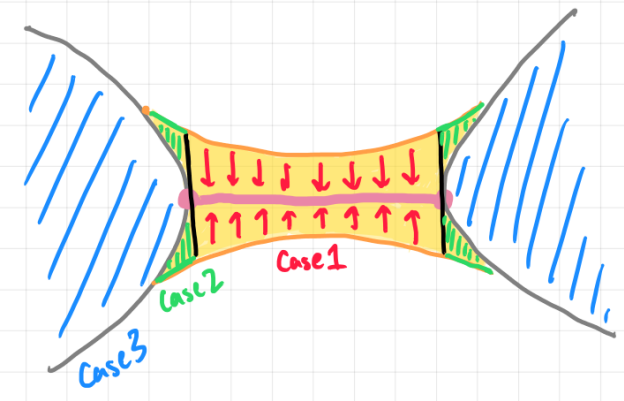

Figure 8: There are three cases for the deformation retraction, which will compress the shape to contain only the blue and pink areas.

$f = c - \varepsilon$, and the blue region (as well as any point of $M^{c-\varepsilon} \cup H$ which is not contained in U) doesn't move at all.

Note that the three regions correspond to $\xi \leq \varepsilon$, $\varepsilon \leq \xi \leq \varepsilon + \eta$, and $\varepsilon + \eta \leq \xi$, respectively. We will define functions r_t for $t \in [0,1]$ so that the function $r : M^{c-\varepsilon} \cup H \times [0,1] \to M^{c-\varepsilon} \cup H$ which takes (q,t) to $r_t(q)$ is the desired deformation retract.

In the first case, we will perform a straight-line homotopy, defined by

$$r_t : (u^1, \ldots, u^n) \mapsto (u^1, \ldots, u^\lambda, (1-t)u^{\lambda-1}, \ldots, (1-t)u^n).$$

In the second case, we will do another straight-line homotopy, but this time stopping at the hyperbola. It turns out that we can define r_t in this region to be

$$r_t : (u^1, \ldots, u^n) \mapsto (u^1, \ldots, u^\lambda, s_t u^{\lambda+1}, \ldots, s_t u^n),$$

where $s_t = (1-t) + t\sqrt{\frac{\xi - \varepsilon}{\eta}}$. This definition simply comes from the fact that we want s_1 to satisfy

$$f(u^1, \ldots, u^\lambda, s_1 u^{\lambda+1}, \ldots, s_1 u^n) = c - \varepsilon,$$

i.e., that, if we let $q = (u^1, \ldots, u^n)$, then we need

$$c - \xi(q) + s_1^2 \eta(q) = c - \varepsilon.$$

Finally, for the third case, we define r_t to be the identity. Since continuous maps can just be glued together, and since the definitions coincide at the intersections of any two cases, it follows that r gives a deformation retraction from $M^{c-\varepsilon} \cup H$ to $M^{c-\varepsilon} \cup e^\lambda$, thus concluding the proof. □

As might be supposed from Propositions 3.1 and 3.3, this lets us find the homotopy type of any sublevel set.

Theorem 3.4. *Suppose $f : M \to \mathbb{R}$ is a Morse function, that is, a smooth function with no nondegenerate critical points. If M^a is compact for each a, then M has the homotopy type of the manifold which has a cell of dimension λ for each critical point of index λ.*

We won't prove this theorem, since the proof is surprisingly tricky, since some work remains if we have infinitely many critical points. That being said, we've already proved the main ideas in the previous propositions!

As a final remark, note that Theorem 3.4 does not actually say how the cells are attached.