A Concise Introduction to Morse Theory

Brandon Li

May 2023

The basic idea of Morse theory is that a well chosen smooth function on a manifold can be used to extract information about the topology of that manifold. In this writeup, we will explore what this statement means by applying it to a few key examples and formalizing it into a set of definitions and theorems.

The first example we will encounter is the torus. Suppose we had a 2-torus T^2 embedded in $\mathbb{R}^3$, depicted in Fig. 1. Define a smooth function $f : T^2 \to \mathbb{R}$ which just gives the z-coordinate of the embedding. We will refer to f as the *height function*, as the level sets of f resemble the contour lines of a topographic height map. It turns out that the level sets actually tell us practically everything we might want to know about its topology.

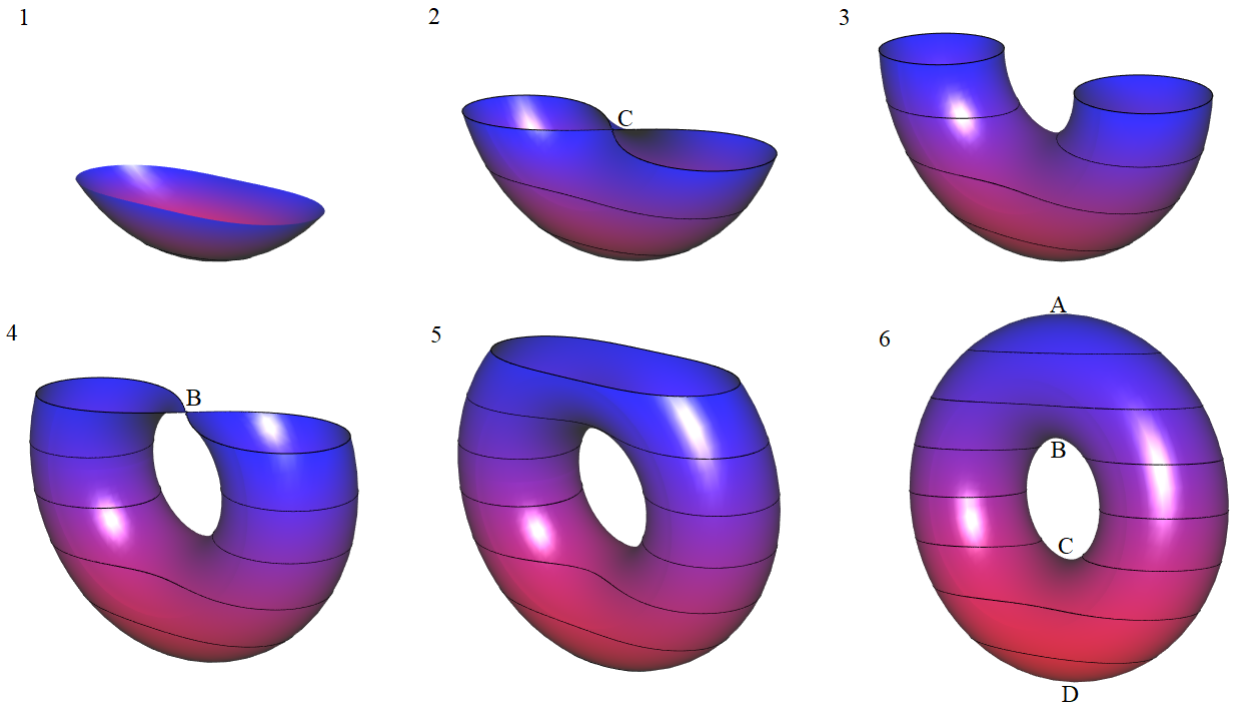

Figure 1: Contour lines of torus

In this diagram, we have labelled the four critical points of the torus A through D. The six consecutive images show level sets of the height function as they progress from low to high. We see that the topology of the sublevel sets (everything below a contour line) progresses from a disk (1) to a tube (3), then turning into a punctured torus (5) and finally becoming a full torus (6). The points at which the topology changes are precisely the critical points, which tells us that critical

points are where the interesting things happen. Outside the critical points, everything stays the same. Moving on from this example, let us start to introduce some definitions.

Definition 1.1. The Hessian d^2 of a smooth function $f : M \to \mathbb{R}$ at a critical point p is defined by the formula $d^2 f_p(X, Y) = X\tilde{Y} f(p)$. On critical points, the Hessian is a well defined symmetric bilinear form.

Proof. Let $\tilde{Y}$ and $\tilde{X}$ be extensions of Y and X respectively. Comparing $d^2 f_p(X, Y)$ and $d^2 f_p(Y, X)$, we find $X\tilde{Y} f - Y\tilde{X} f = [\tilde{X}, \tilde{Y}] f = df[\tilde{X}, \tilde{Y}] = 0$. This immediately implies $X\tilde{Y} f = Y\tilde{X} f = X\hat{Y} f$ where $\hat{Y}$ is another extension of Y. We see that the Hessian is both symmetric and well defined. □

Remark 1.1. The Hessian is well defined at critical points of functions on any smooth manifold. If the manifold is Riemannian, then the Hessian can be extended everywhere using the covariant derivative.

Next, we will introduce a class of functions that generalize the height function we saw before, and that are the central objects of study in Morse theory.

Definition 1.2. A function $f : M \to \mathbb{R}$ is called *Morse* if its critical points are non-degenerate. Equivalently, for all critical points, the eigenvalues of the Hessian are all non-zero. Intuitively, this makes all level sets near critical points look like either hyperboloids or ellipses.

Now, just to prove the point Morse functions are not rare at all, we state the following fact.

Fact 1.2. *Almost all functions are Morse. In other words, the set of Morse functions is dense in $C^\infty(M, \mathbb{R})$ under the $C^2(M, \mathbb{R})$ topology.*

For interested readers, the proof of this statement can be found in [2]. It is actually quite easy to come up with specific Morse functions. For example, if M is a submanifold of $\mathbb{R}^n$, then the square distance function $f(p) = ||p - p_0||^2$ is Morse for almost every p_0. Now that we have a sense of how common these functions are, let us explore some of their properties.

Lemma 1.3. *(Morse Lemma). Let f be a Morse function. If c is a critical point then there exists a chart parameterized by coordinates $(x_1, \dots x_n)$ on which $f(x) = f(c) - x_1^2 - \cdots - x_k^2 + x_{k+1}^2 + \cdots + x_n^2$. We call k the index of the point c. We see from this lemma that k is equal to the maximal dimension of any subspace on which $d^2 f_p$ is negative definite. This means k is well defined and thus independent of choice of chart.*

The proof of this lemma involves some tricky manipulation of coordinate charts, it is omitted in favor of emphasizing its consequences rather than the content of the proof itself. The proof can be found in Minor's book [3]. One implication of this lemma is that it tells us about the nature of critical points.

Corollary 1.4. *Critical points of Morse functions are isolated.*

Proof. Around any critical point there is a Morse chart. In this chart, the differential of f is non-zero everywhere except 0. It follows that Morse functions on compact manifolds have a finite number of critical points. □

Now we will formulate a precise notion of what it means for the topology of level sets to change at critical points. Before that, let us introduce a bit of notation.

Notation 1.5. Define M^a to be the set $f^{-1}((-\infty, a])$. If a is a regular value of f, then M^a is a submanifold of M with boundary $f^{-1}(a)$. This simply follows from the Regular Value Theorem.

This next theorem shows that the topology of a manifold doesn't change in between critical points.

Theorem 1.6. *If $a, b \in \mathbb{R}$ and $[a, b]$ contains no critical values ($f^{-1}[a, b]$ has no critical points), then M^a is diffeomorphic to M^b.*

Proof. The idea of the proof is to basically flow from $f^{-1}(b)$ to $f^{-1}(a)$ along the gradient of f, and use the resulting flow map to give the diffeomorphism between the two manifolds. First, let U be a compact neighborhood of $f^{-1}([a, b])$. Next, find a smooth function ρ that satisfies the following properties: The first is that

$$\rho(x) = \begin{cases} -\frac{1}{|\operatorname{grad} f|^2}, & x \in f^{-1}([a, b]) \\ 0, & x \in M \setminus U. \end{cases} \tag{1}$$

Secondly, ρ can vary in the region between $f^{-1}([a, b])$ and $M \setminus U$, but it must do so in a way that smoothly connects them together. Once we have such a function, we consider the vector field $\rho \cdot \nabla f$ and let $\theta : M \times \mathbb{R} \to M$ be the flow of this vector field. If we compute the rate of change of f along a trajectory of the flow, we see that f decreases at a constant speed since $\frac{d}{dt} f(\theta(x_0, t)) = df \frac{-\operatorname{grad} f}{|\operatorname{grad} f|^2} = -\frac{\langle \nabla f, \nabla f \rangle}{|\nabla f|^2} = -1$. This shows that the rate of change of f is always -1 regardless of the choice of flow line. Thus, if we flow for a time $b - a$, then the resulting flow map sends $f^{-1}(b)$ to $f^{-1}(a)$ and thus M^b to M^a. Finally, since flows are diffeomorphisms, we have the desired result. □

Remark 1.7. In fact, M^a is a deformation retract of M^b and the inclusion map is a homotopy equivalence.

Theorem 1.8. *Suppose $a, b \in \mathbb{R}$ and $f^{-1}([a, b])$ is compact and contains exactly 1 critical point with index k. Then, M^b is homotopy equivalent to M^a with a k-cell attached. As the proof is quite long, we will only give an outline.*

Proof outline. The basic idea is to define another function F that agrees with f outside of a neighborhood of the critical point p and compare level sets of F with that of f. For full details, see [3].

Now, as an example, let us apply this theorem to the torus. In the top part of this Figure 2, we go from a disk to a topological cylinder passing through a critical point of index 1. We see that if we attach a line segment (1-cell) to a disk, we do indeed get a cylinder. Similarly, if we attach a 1-cell to a cylinder we get a punctured torus. This demonstrates that the theorem does work in the case of our example. Moving on, let us describe another easy application of the preceding theorems.

Theorem 1.9 (Reeb's theorem)**.** *If a Morse function f on a compact manifold M has exactly two critical points, then M is a topological sphere.*

Proof. Compactness implies that f achieves its minimum and maximum at the critical points. The Morse Lemma implies for some $\epsilon > 0$, $f^{-1}([a, a + \epsilon])$ and $f^{-1}([b - \epsilon, b])$ are homeomorphic to disks with the same dimension as M. Now, $M^{a+\epsilon}$ and $M^{b-\epsilon}$ are diffeomorphic by Theorem 1.6. So, $M = M^{b-\epsilon} \cup f^{-1}([b-\epsilon, b])$ is the gluing of two disks along their boundaries, hence M is a sphere. □

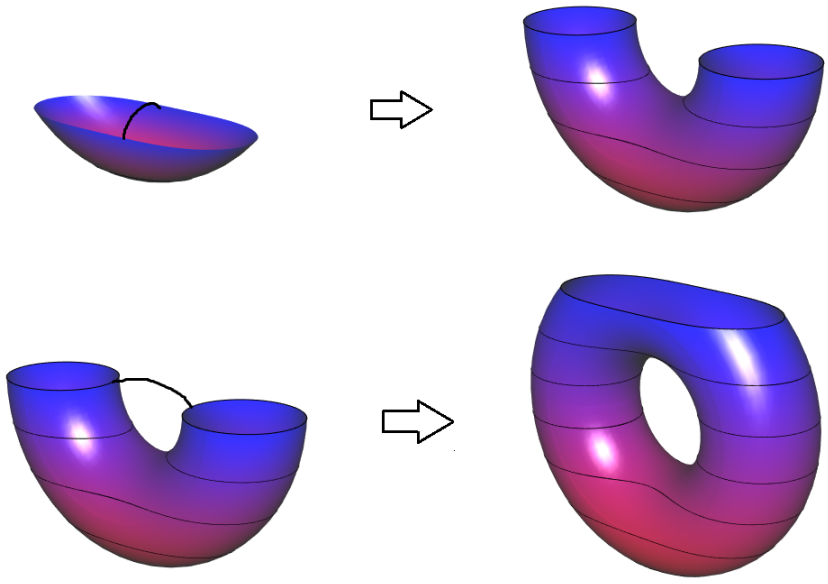

Figure 2: Attaching cells to sublevel sets

Let us now move on to the seemingly unrelated topic of infinite dimensional topological space. We will study one particular example of such a space: The set of paths between two points on a manifold. Given two points $p, q \in M$, let us denote the set of piecewise smooth paths joining p to q as $\Omega(M; p, q)$. It turns out that the ideas of Morse theory in a finite dimensional setting can be generalized to give us information about these infinite dimensional spaces.

The generalization of a function on a manifold is a functional on the space of paths, ie. a mapping that takes a path and returns a real number. One obvious example is the functional that gives the length of a path along a Riemannian manifold. Another related one, which we will be looking at, has its origins in the concept of kinetic energy from physics. As we will see, it is possible to study these functions in the same way we study Morse functions. The generalization of the Morse index for the energy functional will give us topological information about the path space. It is a useful tool for those who wish to study such spaces.

Definition 1.3. The energy functional, also sometimes called the action, is defined as $E[\gamma] = \frac{1}{2}\int_a^b ||\dot{\gamma}||^2\, dt$ where γ is a (piecewise) smooth path from a to b.

The energy functional is actually related to the length functional since it satisfies the following equivalence:

Lemma 1.10. *Consider a path $\gamma : [a, b] \to M$. Then γ is minimizes the energy functional if and only if it minimizes the length functional and has unit speed.*

Proof. We can show this by applying the Cauchy-Schwarz inequality to the functions 1 and $||\dot{\gamma}||$, and we find

$$L[\gamma] := \int_a^b 1 \cdot ||\dot{\gamma}||\, dt \leq \int_a^b 1\, dt \int_a^b ||\dot{\gamma}||^2\, dt = 2(b-a)E[\gamma] \tag{2}$$

where equality holds if and only if $||\dot{\gamma}||$ is constant. Suppose η minimizes the energy functional. It must be unit speed, otherwise the inequality $L[\eta] \leq 2(b-a)E[\eta]$ is strict meaning a unit speed

reparameterization of η would have smaller energy. We therefore must have $[\eta] = 2(b-a)E[\eta]$. Let us take an arbitrary curve γ with the same endpoints as η. The length of γ could not be smaller than the length of η because if so, we could reparameterize γ to get a curve γ' such that $L[\gamma'] = 2(b-a)E[\gamma']$ but that would mean $E[\gamma'] < E[\eta]$ which is a contradiction. For the other proof direction, suppose η was a unit speed length minimizing curve. Then for any other curve γ, we have $E[\eta] = \frac{L[\eta]}{2(b-a)} \leq \frac{L[\gamma]}{2(b-a)} \leq E[\gamma]$ hence η is energy minimizing. □

On a similar note, the following three statements are all equivalent for a path γ: (i) γ is a critical point of energy functional. (ii) γ is a critical point of length functional, and (iii) γ is a geodesic.

The proof requires the first variation formula and is left as an exercise. The point of this is that there is not really a fundamental difference between working with the length functional versus working with the energy functional. We choose the energy functional because the computations are a bit more convenient.

We can now define an infinite dimensional version of the hessian for paths. This is also called the index form, and it will turn out to be the same index form we defined in class.

Definition 1.4. Let γ be a geodesic from $[a,b] \to M$. Given two vector fields $X(t)$ and $Y(t)$ along γ that vanish at the endpoints, define the Hessian $I(X,Y)$ as follows: Choose a two-parameter proper variation of paths $\gamma(x,y,t)$ such that $\partial_x\gamma(0,0,t) = X(t)$ and $\partial_y\gamma(0,0,t) = Y(t)$. Then $I(X,Y) := \frac{\partial^2 E[\gamma]}{\partial x \partial y}|_{(0,0)}$. We will show that I is a well defined, symmetric bilinear form just like the Hessian from before.

Proof. Define $V = \partial_t\gamma$, $\tilde{X} = \partial_x\gamma$ and $\tilde{Y} = \partial_y\gamma$ such that $\tilde{X}(0,0,t) = X(t)$ and $\tilde{Y}(0,0,t) = Y(t)$. Using Lemma 6.2 and Proposition 7.3 in the textbook, we see

$$\frac{\partial^2 E[\gamma]}{\partial x \partial y}|_{(0,0)} = \frac{1}{2}\partial_x\partial_y \int_a^b \langle V, V\rangle \, dt = \partial_y \int_a^b \langle D_x V, V\rangle \, dt \tag{3}$$

$$= \partial_y \int_a^b \langle D_t\tilde{X}, V\rangle \, dt = \int_a^b \langle D_y D_t\tilde{X}, V\rangle + \langle D_t\tilde{X}, D_y V\rangle \, dt \tag{4}$$

$$= \int_a^b \langle D_t D_y\tilde{X}, V\rangle + \langle R(\tilde{Y}, V)\tilde{X}, V\rangle + \langle D_t\tilde{X}, D_t\tilde{Y}\rangle \, dt \tag{5}$$

The first term can be rewritten using integration of parts as $\langle D_y\tilde{X}, V\rangle|_a^b - \int_a^b \langle D_y\tilde{X}, D_t V\rangle \, dt$, now $D_t V$ is zero since $\gamma|_{0,0}$ is a geodesic and $\langle D_y\tilde{X}, V\rangle|_a^b$ is zero since the variation is proper. Finally equating $V = \dot{\gamma}$ shows

$$\frac{\partial^2 E[\gamma]}{\partial x \partial y}|_{(0,0)} = \int_a^b \langle \mathrm{Rm}(\tilde{Y}, \dot{\gamma}, \tilde{X}, \dot{\gamma}\rangle + \langle D_t\tilde{X}, D_t\tilde{Y}\rangle \, dt \tag{6}$$

From this formula, we see that I is a well defined form since its expression only depends on the vector fields evaluated at $x = y = 0$, where $\tilde{X} = X$ and $\tilde{Y} = Y$. We also see immediately that I is symmetric and bilinear. □

Note that this is the same index form as the one we defined in class. By Proposition 10.24, we may express this formula slightly differently.

Fact 1.11. *$I(X,Y) = -\int_a^b \langle D_t^2 X + R(X,\dot{\gamma})\dot{\gamma}, Y\rangle dt - \sum_i \langle \Delta_i D_t X, Y\rangle$ where $\Delta i D_t$ is the jump in D_t at discontinuities. The steps of the proof are the same as in Proposition 10.24.*

We can define the *null space* of I as the dimension of the maximal subspace on which I evaluates to zero, or equivalently, the set of vector fields X such that $I(X, Y) = 0$ for every Y. This leads to the following corollary:

Corollary 1.12. *X belongs to the null space of I if and only if X is a Jacobi field.*

Proof. See Corollary 10.25 in Lee.

Theorem 1.13 (Morse)**.** *The index λ of I is equal to the number of points $t \in (a, b)$ so that $\gamma(a)$ is conjugate to $\gamma(t)$ counted with multiplicity. λ is always finite.*

Proof outline. We first want to split space of vector fields into two orthogonal subspaces, where on one of them I is positive definite. Working with the subspace that is not positive definite, we can define the index k as a function of the affine parameter τ along the geodesic. If we prove that $k(\tau)$ is monotone increasing, equal to 0 for small τ, continuous from below, and jumps by the nullity of I at conjugate points, then we will have shown the result.

Finally, we will discuss one (non-trivial) consequence of this theorem.

Theorem 1.14 (Fundamental Theorem of Morse Theory)**.** *Let M be a geodesic and p and q be points that are not conjugate along any geodesic. Then the space of paths from p to q is homotopy equivalent to a countable CW-complex containing one cell of dimension k for each geodesic from p to q with index k.*

Remark 1.15. This theorem does not directly follow from the preceding results. Showing it involves constructing finite dimensional approximations of the path space which is a fair amount of work. The reason it is mentioned here is that this theorem is the culmination of all the ideas we have developed so far, and is interesting in its own right.

References

[1] Shintaro Fushida-Hardy. Morse theory.

[2] Marco Gualtier. Morse theory, 2010.

[3] J.W. Milnor. *Morse Theory.* Annals of mathematics studies. Princeton University Press, 1963.