V.1 Generic Smooth Functions

Many, perhaps most questions in the sciences and engineering can be posed in terms of real-valued functions. General such functions are a nightmare and continuous functions are not much better. Even smooth functions can be exceedingly complicated but when they are restricted to being generic they become intelligible.

The upright torus. We start with an example that foreshadows many of the results on generic smooth functions in an intuitive manner. Let $\mathbb{M}$ be the two-dimensional torus and $f(x)$ the height of the point $x \in \mathbb{M}$ above a horizontal plane on which the torus rests, as in Figure V.1. We call $f : \mathbb{M} \to \mathbb{R}$ a *height function*. Each real number a has a preimage, $f^{-1}(a)$, which we refer to as a *level set*. It consists of all points $x \in \mathbb{M}$ with height a. Accordingly, the *sublevel set* consists of all points with height at most a,

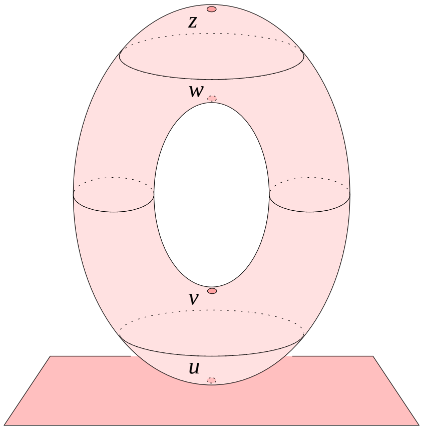

Figure V.1: The vertical height function on the torus with critical points u, v, w, z and level sets between their height values.

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

We are interested in the evolution of the sublevel set as we increase the threshold. Critical events occur when a passes the height values of the points u, v, w, z in Figure V.1. For $a < f(u)$ the sublevel set is empty. For $f(u) < a < f(v)$ it is a disk, which has the homotopy type of a point. For $f(v) < a < f(w)$ the sublevel set is a cylinder. It has the homotopy type of a circle which we imagine is obtained by gluing the two ends of an interval to the disk which is then shrunk to a point. For $f(w) < a < f(z)$ the sublevel set is a capped torus. It has the homotopy type of a figure-8 obtained by gluing the two ends of another interval to the cylinder which is then shrunk to a circle. Finally, for $f(z) < a$ we have the complete torus. It is obtained by gluing a disk to the capped torus. Figure V.2 illustrates the three intermediate stages of the evolution. We need background in differential topology to explain in what sense this evolution of the sublevel set is representative of the general situation.

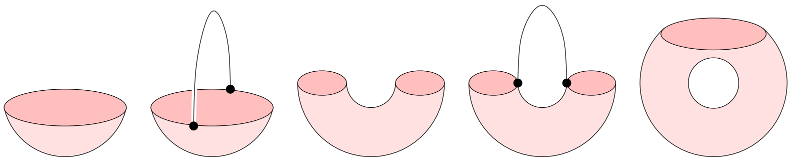

Figure V.2: Going from a disk to a cylinder is homotopically the same as attaching a 1-handle. Similarly, going from the cylinder to the capped torus is homotopically the same as attaching another 1-handle.

Smooth functions. Let $\mathbb{M}$ be a smooth d-manifold, that is, $\mathbb{M}$ has an atlas of coordinate charts each diffeomorphic to an open ball in $\mathbb{R}^d$. We recall that a diffeomorphism is a homeomorphism that is smooth in both directions. Denote the tangent space at a point $x \in \mathbb{M}$ by $\mathrm{T}\mathbb{M}_x$. It is the d-dimensional vector space consisting of all tangent vectors of $\mathbb{M}$ at x. A smooth mapping to another smooth manifold, $f : \mathbb{M} \to \mathbb{N}$, induces a linear map between the tangent spaces, the derivative $\mathrm{D}f_x : \mathrm{T}\mathbb{M}_x \to \mathrm{T}\mathbb{N}_{f(x)}$. We are primarily interested in real-valued functions for which $\mathbb{N} = \mathbb{R}$. Accordingly, we have linear maps $\mathrm{D}f_x : \mathrm{T}\mathbb{M} \to \mathrm{T}\mathbb{R}_{f(x)}$. The tangent space at a point of the real line is again a real line, so this is just a fancy way of saying that the derivatives are real-valued linear maps on the tangent spaces. Being linear, the image of such a map is either the entire line or just zero. We call $x \in \mathbb{M}$ a *regular point* of f if $\mathrm{D}f_x$ is surjective and we call x a *critical point* of f if $\mathrm{D}f_x$ is the zero map. If we have a local coordinate system $(x_1, x_2, \ldots, x_d)$ in a neighborhood of x then x is critical iff all its partial derivatives vanish,

$$\frac{\partial f}{\partial x_1}(x) \;=\; \frac{\partial f}{\partial x_2}(x) \;=\; \ldots \;=\; \frac{\partial f}{\partial x_d}(x) \;=\; 0.$$

The image of a critical point, $f(x)$, is called a *critical value* of f. We use second derivatives to further distinguish between different types of critical points. The

Hessian of f at the point x is the matrix of second derivatives,

$$H(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(x) & \frac{\partial^2 f}{\partial x_d \partial x_2}(x) & \ldots & \frac{\partial^2 f}{\partial x_d \partial x_d}(x) \end{bmatrix}.$$

A critical point x is *non-degenerate* if the Hessian is non-singular, that is, $\det H(x) \neq 0$. The points u, v, w, z in Figure V.1 are examples of non-degenerate critical points. Examples of degenerate critical points are $x_1 = 0$ of the function $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x_1) = x_1^3$ and $(x_1, x_2) = (0, 0)$ of $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1^3 - 3x_1x_2^2$. The degenerate critical point in that latter example is often referred to as a monkey saddle. Indeed, the graph of the function in a neighborhood goes up and down three times, providing convenient rest for the two legs as well as the tail of the monkey.

Morse functions. At a critical point all partial derivatives vanish. A local Taylor expansion has therefore no linear terms. If the critical point is non-degenerate then the behavior of the function in a small neighborhood is dominated by the quadratic terms. Even more, we can find local coordinates such that there are no higher-order terms.

Morse Lemma. Let u be a non-degenerate critical point of $f : \mathbb{M} \to \mathbb{R}$. There are local coordinates with $u = (0, 0, \ldots, 0)$ such that

$$f(x) = f(u) - x_1^2 - \ldots - x_p^2 + x_{p+1}^2 + \ldots + x_d^2$$

for every point $x = (x_1, x_2, \ldots, x_d)$ in a small neighborhood of u.

The number of minus signs in the quadratic polynomial is the *index* of the critical point, $\text{index}(u) = p$. The index classifies the non-degenerate critical points into $d + 1$ basic types. For a 2-manifold we have three types, *minima* with index 0, *saddles* with index 1, and *maxima* with index 2. Examples of all three types can be seen in Figure V.1. In Figure V.3 we display them by showing the local evolution of the sublevel set. A consequence of the Morse Lemma is that non-degenerate critical points are isolated. This implies that a Morse function on a compact manifold has at most a finite number of critical points. To contrast this with a function that is not Morse take the height function of a torus, similar to Figure V.1 but placing the torus sideways, the way it would naturally rest under the influence of gravity. This height function

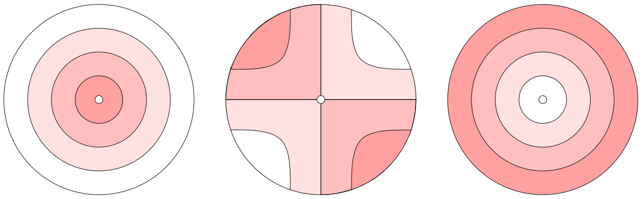

Figure V.3: From left to right: the local pictures of a minimum, a saddle, a maximum. Imagine looking from above with the shading getting darker as the function shrinks away from the viewpoint.

has an entire circle of minima and another circle of maxima. All these critical points are degenerate and their index is not defined.

Definition. A smooth function on a manifold, $f : \mathbb{M} \to \mathbb{R}$, is a *Morse function* if (i) all critical points are non-degenerate, and (ii) the critical points have distinct function values.

Sometimes the second condition is dropped but in this book we will always require both. For a geometrically perfect torus the height function satisfies condition (i) for all but two directions, the ones parallel to the symmetry axis of the torus. Condition (ii) is violated for another two circles of directions along which the two saddles have the same height. The height function of $\mathbb{S}^2$ is a Morse function for all directions. The distance from a point is a Morse function for almost all points. Exceptions for the torus are points on the symmetry axis and on the center circle, but there are others. The only exception for the 2-sphere is the center.

Gradient vector field. A *vector field* on a manifold is a function $X : \mathbb{M} \to \mathbb{TM}$ that maps every point $x \in \mathbb{M}$ to a vector $X(x)$ in the tangent space of $\mathbb{M}$ at x. Given $f : \mathbb{M} \to \mathbb{R}$ and X we denote the directional derivative of f along the vector field by $X[f]$. It maps every point $x \in \mathbb{M}$ to the derivative of f at x in the direction $X(x)$. A particularly useful vector field is the one that points in the direction of steepest increase. To define it we need to measure length, which we do by introducing a Riemannian metric, that is, a smoothly varying inner product defined on the tangent spaces. For example, if $\mathbb{M}$ is smoothly embedded in some Euclidean space then the tangent spaces are linear subspaces of the same Euclidean space and we can borrow the metric. Given a smooth manifold $\mathbb{M}$, a Riemannian metric on $\mathbb{M}$ and a smooth function $f : \mathbb{M} \to \mathbb{R}$, we define the *gradient* of f as the vector field $\nabla f : \mathbb{M} \to \mathrm{T}\mathbb{M}$ characterized by $\langle X(x), \nabla f(x) \rangle = X[f]$ for every vector field X. Assuming local coordinates with orthonormal unit vectors x_i, the gradient at the point x is

$$\nabla f(x) \quad = \quad \left[\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_d}(x) \right]^T .$$

We use the gradient to introduce a 1-*parameter group of diffeomorphisms* $\varphi : \mathbb{R} \times \mathbb{M} \to \mathbb{M}$. There are two characteristic properties of this group. First, the map $\varphi_t : \mathbb{M} \to \mathbb{M}$ defined by $\varphi_t(x) = \varphi(t, x)$ is a diffeomorphism of $\mathbb{M}$ to itself for each $t \in \mathbb{R}$, and second, $\varphi_{t+t_0} = \varphi_t \circ \varphi_{t_0}$ for all $t, t_0 \in \mathbb{R}$. Such a group defines a vector field by differentiation and we require that this vector field be the gradient vector field, modified by taking one over the original length:

$$\lim_{\varepsilon \to 0} \frac{f(\varphi_\varepsilon(x)) - f(x)}{\varepsilon} \quad = \quad \frac{\nabla f(x)}{\|\nabla f(x)\|^2}[f].$$

This group of diffeomorphisms follows the evolution of the sublevel set and can be used to prove that there are no topological changes that happen between contiguous critical values. Specifically, let $f : \mathbb{M} \to \mathbb{R}$ be smooth and $a < b$ such that $f^{-1}[a, b]$ is compact and contains no critical points of f. Then $\mathbb{M}_a$ is diffeomorphic to $\mathbb{M}_b$.

Attaching handles. The situation is different when we consider regular values $a < b$ such that $f^{-1}[a, b]$ is compact but contains one critical point of f. Let this critical point be u and its index be p. In this case, $\mathbb{M}_b$ has the homotopy type of $\mathbb{M}_a$ with a p-handle attached. To explain what this means we recall that $\mathbb{B}^p$ is the p-dimensional unit ball and $\mathbb{S}^{p-1}$ is its boundary. Let $g : \mathbb{S}^{p-1} \to \mathrm{bd}\, \mathbb{M}_a$ be a continuous map. To *attach* the handle to $\mathbb{M}_a$ we first take the topological sum (disjoint union) of $\mathbb{M}_a$ and $\mathbb{B}^p$ and then identify each point $x \in \mathbb{S}^{p-1}$ with its image $g(x) \in \mathrm{bd}\, \mathbb{M}_a$. The only case that is a bit different is $p = 0$. Then $\mathbb{S}^{-1}$ is empty and attaching the 0-handle just means adding a disjoint point.

We illustrate this construction for a 3-manifold $\mathbb{M}$. There are four types of critical points, namely minima with index 0, saddles with index 1 or 2, and maxima with index 3. The two types of saddles deserve some attention. To illustrate the local evolution of the sublevel set we draw spheres around them and shade the portion that belongs to the sublevel set, as in Figure V.4. The level set that passes through the saddle forms locally a double-cone with the apex at the saddle. This is the same for both types, the only difference being the side on which the sublevel set resides. For the index 1 saddle we imagine a two

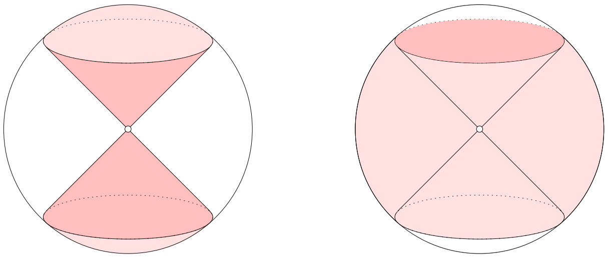

Figure V.4: The double-cone neighborhood of the index 1 saddle on the left and of the index 2 saddle on the right. The volume occupied by the sublevel set is shaded.

sheet hyperboloid approaching from two sides until the two sheets meet at the saddle. Thereafter the sublevel set thickens around the saddle as its boundary moves out as a one sheet hyperboloid (an hour glass). Homotopically, this evolution is the same as attaching a 1-handle (an interval) connecting the two sheets. For the index 2 saddle the sequence of events is reversed. Specifically, a one sheet hyperboloid approaches along a circle of directions until it reaches the saddle. Thereafter the sublevel set thickens around the saddle as its boundary moves out as two sheets of a hyperboloid. Homotopically, this evolution is the same as attaching a 2-handle (a disk) closing the tunnel formed by the one sheet hyperboloid.

Bibliographic notes. Morse theory developed first in infinite dimensions, as part of the calculus of variations, see Morse [4]. The classic source on the subject for finite-dimensional manifolds is the text by Milnor [3], but see also Matsumoto [2] and Banyaga and Hurtubis [1].

[1] A. Banyaga and D. Hurtubis. *Lectures on Morse Homology.* Kluwer, Dordrecht, the Netherlands, 2004.

[2] Y. Matsumoto. *An Introduction to Morse Theory.* Translated from Japanese by K. Hudson and M. Saito, Amer. Math. Soc., 2002.

[3] J. Milnor. *Morse Theory.* Princeton Univ. Press, New Jersey, 1963.

[4] M. Morse. *The Calculus of Variations in the Large.* Amer. Math. Soc., New York, 1934.