1
Basics

Topology, mainly algebraic topology, is the fundamental mathematical subject on which topological data analysis is based. In this chapter, we introduce some of the very basics of this subject that are used in this book. First, in Section 1.1, we give the definition of a topological space and other notions such as open and closed sets, covers, and subspace topology that are derived from it. These notions are quite abstract in the sense that they do not require any geometry. However, the intuition of topology becomes more concrete to nonmathematicians when we bring geometry into the mix. Section 1.2 is devoted to make the connection between topology and geometry through what is called metric spaces.

Maps such as homeomorphism and homotopy equivalence play a significant role to relate topological spaces. They are introduced in Section 1.3. At the heart of these definitions sits the important notion of continuous functions which generalizes the concept mainly known for Euclidean domains to topological spaces. Certain categories of topological spaces become important for their wide presence in applications. Manifolds are one such category which we introduce in Section 1.4. Functions on them satisfying certain conditions are presented in Section 1.5. They are well known as Morse functions. The critical points of such functions relate to the topology of the manifold they are defined on. We introduce these concepts in the smooth setting in this chapter, and later adapt them for the piecewise-linear domains that are amenable to finite computations.

1.1 Topological Space

The basic object in a topological space is a ground set whose elements are called points. A topology on these points specifies how they are *connected* by listing what points constitute a neighborhood – the so-called *open set*.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

The expression “rubber-sheet topology” commonly associated with the term “topology” exemplifies this idea of connectivity of neighborhoods. If we bend and stretch a sheet of rubber, it changes shape but always preserves the neighborhoods in terms of the points and how they are connected.

We first introduce basic notions from point set topology. These notions are prerequisites for more sophisticated topological ideas – manifolds, homeomorphism, isotopy, and other maps – used later to study algorithms for topological data analysis. Homeomorphisms, for example, offer a rigorous way to state that an operation preserves the topology of a domain, and isotopy offers a rigorous way to state that the domain can be deformed into a shape without ever colliding with itself.

Perhaps it is more intuitive to understand the concept of topology in the presence of a metric because then we can use the metric balls such as Euclidean balls in a Euclidean space to define neighborhoods – the open sets. Topological spaces provide a way to abstract out this idea without a metric or point coordinates, so they are more general than metric spaces. In place of a metric, we encode the connectivity of a point set by supplying a list of all of the open sets. This list is called a *system* of subsets of the point set. The point set and its system together describe a topological space.

Definition 1.1. (Topological space) A *topological space* is a point set $\mathbb{T}$ endowed with a *system of subsets* T, which is a set of subsets of $\mathbb{T}$ that satisfies the following conditions:

- $\emptyset, \mathbb{T} \in T$.
- For every $U \subseteq T$, the union of the subsets in U is in T.
- For every finite $U \subseteq T$, the common intersection of the subsets in U is in T.

The system T is called a *topology* on $\mathbb{T}$. The sets in T are called the *open sets* in $\mathbb{T}$. A *neighborhood* of a point $p \in \mathbb{T}$ is an open set containing p.

First, we give examples of topological spaces to illustrate the definition above. These examples have the set $\mathbb{T}$ finite.

Example 1.1. *Let* $\mathbb{T} = \{0, 1, 3, 5, 7\}$. *Then,* $T = \{\emptyset, \{0\}, \{1\}, \{5\}, \{1, 5\}, \{0, 1\}, \{0, 1, 5\}, \{0, 1, 3, 5, 7\}\}$ *is a topology because* $\emptyset$ *and* $\mathbb{T}$ *are in* T *as required by the first axiom, the union of any sets in* T *is in* T *as required by the second axiom, and the intersection of any two sets is also in* T *as required by the third axiom. However,* $T = \{\emptyset, \{0\}, \{1\}, \{1, 5\}, \{0, 1, 5\}, \{0, 1, 3, 5, 7\}\}$ *is not a topology because the set* $\{0, 1\} = \{0\} \cup \{1\}$ *is missing.*

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

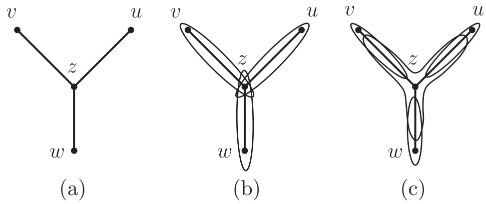


Figure 1.1 Example 1.3: (a) a graph as a topological space, stars of the vertices and edges as open sets; (b) a closed cover with three elements; and (c) an open cover with four elements.

Example 1.2. *Let $\mathbb{T} = \{u, v, w\}$. The power set $2^{\mathbb{T}} = \{\emptyset, \{u\}, \{v\}, \{w\}, \{u, v\}, \{u, w\}, \{v, w\}, \{u, v, w\}\}$ is a topology. For any ground set $\mathbb{T}$, the power set is always a topology on it which is called the* discrete topology.

One may take a subset of the power set as a ground set and define a topology, as the next example shows. We will recognize later that the ground set here corresponds to simplices in a simplicial complex and the "stars" of simplices generate all open sets of a topology.

Example 1.3. *Let $\mathbb{T} = \{u, v, w, z, (u, z), (v, z), (w, z)\}$; this can be viewed as a graph with four vertices and three edges as shown in Figure 1.1. Let*

- $T_1 = \{\{(u, z)\}, \{(v, z)\}, \{(w, z)\}\}$ *and*
- $T_2 = \{\{(u, z), u\}, \{(v, z), v\}, \{(w, z), w\}, \{(u, z), (v, z), (w, z), z\}\}$.

Then, $T = 2^{T_1 \cup T_2}$ is a topology because it satisfies all three axioms. All open sets of T are generated by the union of elements in $B = T_1 \cup T_2$ and there is no smaller set with this property. Such a set B is called a basis of T. We will see later in the next chapter (Section 2.1) that these are open stars *of all vertices and edges.*

We now present some more definitions that will be useful later.

Definition 1.2. (Closure; Closed sets) A set Q is *closed* if its complement $\mathbb{T} \setminus Q$ is open. The *closure* $\mathrm{Cl}\, Q$ of a set $Q \subseteq T$ is the smallest closed set containing Q.

In Example 1.1, the set $\{3, 5, 7\}$ is closed because its complement $\{0, 1\}$ in $\mathbb{T}$ is open. The closure of the open set $\{0\}$ is $\{0, 3, 7\}$ because it is the smallest closed set (complement of open set $\{1, 5\}$) containing 0. In Example 1.2, all sets are both open and closed. In Example 1.3, the set $\{u, z, (u, z)\}$ is closed, but the set $\{z, (u, z)\}$ is neither open nor closed. Interestingly, observe that

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

$\{z\}$ is closed. The closure of the open set $\{u, (u, z)\}$ is $\{u, z, (u, z)\}$. In all examples, the sets $\varnothing$ and $\mathbb{T}$ are both open and closed.

Definition 1.3. Given a topological space $(\mathbb{T}, T)$, the *interior* $\operatorname{Int} A$ of a subset $A \subseteq \mathbb{T}$ is the union of all open subsets of A. The *boundary* of A is $\operatorname{Bd} A = \operatorname{Cl} A \setminus \operatorname{Int} A$.

The interior of the set $\{3, 5, 7\}$ in Example 1.1 is $\{5\}$ and its boundary is $\{3, 7\}$.

Definition 1.4. (Subspace topology) For every point set $\mathbb{U} \subseteq \mathbb{T}$, the topology T induces a *subspace topology* on $\mathbb{U}$, namely the system of open subsets $U = \{P \cap \mathbb{U} \colon P \in T\}$. The point set $\mathbb{U}$ endowed with the system U is said to be a *topological subspace* of $\mathbb{T}$.

In Example 1.1, consider the subset $\mathbb{U} = \{1, 5, 7\}$. It has the subspace topology

$$U = \{\varnothing, \{1\}, \{5\}, \{1, 5\}, \{1, 5, 7\}\}.$$

In Example 1.3, the subset $\mathbb{U} = \{u, (u, z), (v, z)\}$ has the subspace topology

$$\{\varnothing, \{u, (u, z)\}, \{(u, z)\}, \{(v, z)\}, \{(u, z), (v, z)\}, \{u, (u, z), (v, z)\}\}.$$

Definition 1.5. (Connected) A topological space $(\mathbb{T}, T)$ is *disconnected* if there are two disjoint non-empty open sets $U, V \in T$ so that $\mathbb{T} = U \cup V$. A topological space is *connected* if it is not disconnected.

The topological space in Example 1.1 is connected. However, the topological subspace (Definition 1.4) induced by the subset $\{0, 1, 5\}$ is disconnected because it can be obtained as the union of two disjoint open sets $\{0, 1\}$ and $\{5\}$. The topological space in Example 1.3 is also connected, but the subspace induced by the subset $\{(u, z), (v, z), (w, z)\}$ is disconnected.

Definition 1.6. (Cover; Compact) An *open (closed) cover* of a topological space $(\mathbb{T}, T)$ is a collection C of open (closed) sets so that $\mathbb{T} = \bigcup_{c \in C} c$. The topological space $(\mathbb{T}, T)$ is called *compact* if every open cover C of it has a finite *subcover*, that is, there exists $C' \subseteq C$ such that $\mathbb{T} = \bigcup_{c \in C'} c$ and C' is finite.

In Figure 1.1(b), the cover consisting of $\{\{u, z, (u, z)\}, \{v, z, (v, z)\}, \{w, z, (w, z)\}\}$ is a closed cover whereas the cover consisting of $\{\{u, (u, z)\}, \{v, (v, z)\},$

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

$\{w, (w, z)\}$, $\{z, (u, z), (v, z), (w.z)\}\}$ in Figure 1.1(c) is an open cover. Any topological space with finite point set $\mathbb{T}$ is compact because all of its covers are finite. Thus, all topological spaces in the discussed examples are compact. We will see examples of noncompact topological spaces where the ground set is infinite.

In the above examples, the ground set $\mathbb{T}$ is finite. It can be infinite in general and a topology may have uncountably infinitely many open sets containing uncountably infinitely many points.

Next, we introduce the concept of *quotient topology*. Given a space $(\mathbb{T}, T)$ and an equivalence relation $\sim$ on elements in $\mathbb{T}$, one can define a topology induced by the original topology T on the quotient set $\mathbb{T}/\sim$ whose elements are equivalence classes $[x]$ for every point $x \in \mathbb{T}$.

Definition 1.7. (Quotient topology) Given a topological space $(\mathbb{T}, T)$ and an equivalence relation $\sim$ defined on the set $\mathbb{T}$, a quotient space $(\mathbb{S}, S)$ induced by $\sim$ is defined by the set $\mathbb{S} = \mathbb{T}/\sim$ and the *quotient topology* S where

$$S := \big\{U \subseteq \mathbb{S} \mid \{x\colon [x] \in U\} \in T\big\}.$$

We will see the use of quotient topology in Chapter 7 when we study Reeb graphs.

Infinite topological spaces may seem baffling from a computational point of view, because they may have uncountably infinitely many open sets containing uncountably infinitely many points. The easiest way to define such a topological space is to inherit the open sets from a metric space. A topology on a metric space excludes information that is not topologically essential. For instance, the act of stretching a rubber sheet changes the distances between points and thereby changes the metric, but it does not change the open sets or the topology of the rubber sheet. In the next section, we construct such a topology on a metric space and examine it from the concept of limit points.

1.2 Metric Space Topology

Metric spaces are a special type of topological space commonly encountered in practice. Such a space admits a *metric* that specifies the scalar *distance* between every pair of points satisfying certain axioms.

Definition 1.8. (Metric space) A *metric space* is a pair $(\mathbb{T}, \mathsf{d})$ where $\mathbb{T}$ is a set and d is a distance function $\mathsf{d}\colon \mathbb{T} \times \mathbb{T} \to \mathbb{R}$ satisfying the following properties:

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

- $\mathsf{d}(p,q) = 0$ if and only if $p = q$ for all $p \in \mathbb{T}$;
- $\mathsf{d}(p,q) = \mathsf{d}(q,p)$ for all $p, q \in \mathbb{T}$;
- $\mathsf{d}(p,q) \leq \mathsf{d}(p,r) + \mathsf{d}(r,q)$ for all $p, q, r \in \mathbb{T}$.

It can be shown that the three axioms above imply that $\mathsf{d}(p,q) \geq 0$ for every pair $p, q \in \mathbb{T}$. In a metric space $\mathbb{T}$, an open *metric ball* with center c and radius r is defined to be the point set $B_o(c,r) = \{p \in \mathbb{T} : \mathsf{d}(p,c) < r\}$. Metric balls define a topology on a metric space.

Definition 1.9. (Metric space topology) Given a metric space $\mathbb{T}$, all metric balls $\{B_o(c,r) \mid c \in \mathbb{T}$ and $0 < r \leq \infty\}$ and their union constituting the open sets define a topology on $\mathbb{T}$.

All definitions for general topological spaces apply to metric spaces with the above defined topology. However, we give alternative definitions using the concept of limit points which may be more intuitive.

As we have mentioned already, the heart of topology is the question of what it means for a set of points to be *connected*. After all, two distinct points cannot be adjacent to each other; they can only be connected to one another by passing through uncountably many intermediate points. The idea of *limit points* helps express this concept more concretely, specifically in the case of metric spaces.

We use the notation $\mathsf{d}(\cdot,\cdot)$ to express minimum distances between point sets $P, Q \subseteq \mathbb{T}$:

$$\mathsf{d}(p,Q) = \inf\{\mathsf{d}(p,q) \colon q \in Q\},$$
$$\mathsf{d}(P,Q) = \inf\{\mathsf{d}(p,q) \colon p \in P, q \in Q\}.$$

Definition 1.10. (Limit point) Let $Q \subseteq \mathbb{T}$ be a point set. A point $p \in \mathbb{T}$ is a *limit point* of Q, also known as an *accumulation point* of Q, if for every real number $\epsilon > 0$, however tiny, Q contains a point $q \neq p$ such that $d(p,q) < \epsilon$.

In other words, there is an infinite sequence of points in Q that gets successively closer and closer to p – without actually being p – and gets arbitrarily close. Stated succinctly, $\mathsf{d}(p, Q \setminus \{p\}) = 0$. Observe that it does not matter whether $p \in Q$ or not.

To see the parallel between the definitions given in this subsection and the definitions given before, it is instructive to define limit points also for general topological spaces. In particular, a point $p \in \mathbb{T}$ is a limit point of a set $Q \subseteq \mathbb{T}$ if every open set containing p intersects Q.

Definition 1.11. (Connected) A point set $Q \subseteq \mathbb{T}$ is called *disconnected* if Q can be partitioned into two disjoint non-empty sets U and V so that there is no

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

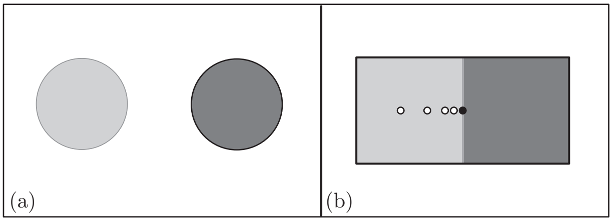


Figure 1.2 (a) The point set is disconnected; it can be partitioned into two connected subsets shaded differently. (b) The point set is connected; the black point at the center is a limit point of the points shaded lightly.

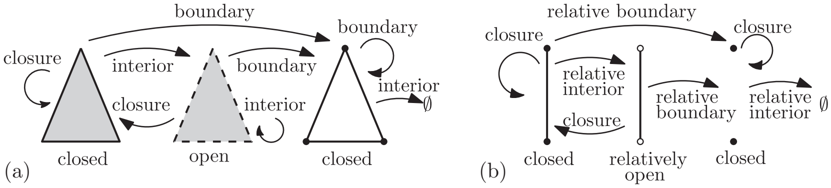


Figure 1.3 Closed, open, and relatively open point sets in the plane. Dashed edges and open circles indicate points missing from the point set.

point in U that is a limit point of V, and no point in V that is a limit point of U. (See Figure 1.2[a] for an example.) If no such partition exists, Q is *connected*, like the point set in Figure 1.2(b).

We can also distinguish between closed and open point sets using the concept of limit points. Informally, a triangle in the plane is *closed* if it contains all the points on its edges, and *open* if it excludes all the points on its edges, as illustrated in Figure 1.3. The idea can be formally extended to any point set.

Definition 1.12. (Closure; Closed; Open) The *closure* of a point set $Q \subseteq \mathbb{T}$, denoted $\text{Cl}\, Q$, is the set containing every point in Q and every limit point of Q. A point set Q is *closed* if $Q = \text{Cl}\, Q$, that is, Q contains all its limit points. The *complement* of a point set Q is $\mathbb{T} \setminus Q$. A point set Q is *open* if its complement is closed, that is, $\mathbb{T} \setminus Q = \text{Cl}\,(\mathbb{T} \setminus Q)$.

For example, consider the open interval $(0, 1) \subset \mathbb{R}$, which contains every $r \in \mathbb{R}$ so that $0 < r < 1$. Let $[0, 1]$ denote a *closed interval* $(0, 1) \cup \{0\} \cup \{1\}$. The numbers 0 and 1 are both limit points of the open interval, so $\text{Cl}\,(0, 1) = [0, 1] = \text{Cl}\,[0, 1]$. Therefore, $[0, 1]$ is closed and $(0, 1)$ is not. The numbers 0 and 1 are also limit points of the complement of the closed interval, $\mathbb{R} \setminus [0, 1]$, so $(0, 1)$ is open, but $[0, 1]$ is not.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

The definition of *open set* of course depends on the space being considered. A triangle τ that is missing the points on its edges is open in the two-dimensional affine Euclidean space supporting τ. However, it is not open in the Euclidean space $\mathbb{R}^3$. Indeed, every point in τ is a limit point of $\mathbb{R}^3 \setminus \tau$, because we can find sequences of points that approach τ from the side. In recognition of this caveat, a simplex $\sigma \subset \mathbb{R}^d$ is said to be *relatively open* if it is open relative to its affine hull. Figure 1.3 illustrates this fact where, in this example, the metric space is $\mathbb{R}^2$.

We can define the interior and boundary of a set using the notion of limit points also. Informally, the boundary of a point set Q is the set of points where Q meets its complement $\mathbb{T} \setminus Q$. The interior of Q contains all the other points of Q.

Definition 1.13. (Boundary; Interior) The *boundary* of a point set Q in a metric space $\mathbb{T}$, denoted $\mathrm{Bd}\,Q$, is the intersection of the closures of Q and its complement; that is, $\mathrm{Bd}\,Q = \mathrm{Cl}\,Q \cap \mathrm{Cl}\,(\mathbb{T} \setminus Q)$. The *interior* of Q, denoted $\mathrm{Int}\,Q$, is $Q \setminus \mathrm{Bd}\,Q = Q \setminus \mathrm{Cl}\,(\mathbb{T} \setminus Q)$.

For example, $\mathrm{Bd}\,[0, 1] = \{0, 1\} = \mathrm{Bd}\,(0, 1)$ and $\mathrm{Int}\,[0, 1] = (0, 1) = \mathrm{Int}\,(0, 1)$. The boundary of a triangle (closed or open) in the Euclidean plane is the union of the triangle's three edges, and its interior is an open triangle, illustrated in Figure 1.3. The terms *boundary* and *interior* have similar subtlety as open sets: the boundary of a triangle embedded in $\mathbb{R}^3$ is the whole triangle, and its interior is the empty set. However, relative to its affine hull, its interior and boundary are defined exactly as in the case of triangles embedded in the Euclidean plane. Interested readers can draw the analogy between this observation and the definition of interior and boundary of a manifold that appear later in Definition 1.23.

We have seen a definition of the compactness of a point set in a topological space (Definition 1.6). We define it differently here for a metric space. It can be shown that the two definitions are equivalent.

Definition 1.14. (Bounded; Compact) The *diameter* of a point set Q is $\sup_{p,q \in Q} \mathsf{d}(p, q)$. The set Q is *bounded* if its diameter is finite, and is *unbounded* otherwise. A point set Q in a metric space is *compact* if it is closed and bounded.

In the Euclidean space $\mathbb{R}^d$ we can use the standard Euclidean distance as the choice of metric. On the surface of a coffee mug, we could choose the Euclidean distance too; alternatively, we could choose the *geodesic distance*, namely the length of the shortest path from p to q on the mug's surface.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

Example 1.4. (Euclidean ball) *In $\mathbb{R}^d$, the Euclidean d-ball with center c and radius r, denoted $B(c,r)$, is the point set $B(c,r) = \{p \in \mathbb{R}^d : \mathsf{d}(p,c) \leq r\}$. A 1-ball is an edge, and a 2-ball is called a disk. A unit ball is a ball with radius 1. The boundary of the d-ball is called the Euclidean $(d-1)$-sphere and denoted $S(c,r) = \{p \in \mathbb{R}^d : \mathsf{d}(p,c) = r\}$. The name expresses the fact that we consider it a $(d-1)$-dimensional point set – to be precise, a $(d-1)$-dimensional manifold – even though it is embedded in d-dimensional space. For example, a circle is a 1-sphere, and a layman's "sphere" in $\mathbb{R}^3$ is a 2-sphere. If we remove the boundary from a ball, we have the open Euclidean d-ball $B_o(c,r) = \{p \in \mathbb{R}^d : \mathsf{d}(p,c) < r\}$.*

The topological spaces that are subspaces of a metric space such as $\mathbb{R}^d$ inherit their topology as a subspace topology. Examples of topological subspaces are the Euclidean d-ball $\mathbb{B}^d$, Euclidean d-sphere $\mathbb{S}^d$, open Euclidean d-ball $\mathbb{B}^d_o$, and Euclidean half-ball $\mathbb{H}^d$, where

$$\begin{aligned}
\mathbb{B}^d &= \{x \in \mathbb{R}^d : \ \|x\| \leq 1\}, \\
\mathbb{S}^d &= \{x \in \mathbb{R}^{d+1} : \ \|x\| = 1\}, \\
\mathbb{B}^d_o &= \{x \in \mathbb{R}^d : \ \|x\| < 1\}, \\
\mathbb{H}^d &= \{x \in \mathbb{R}^d : \ \|x\| < 1 \text{ and } x_d \geq 0\}.
\end{aligned}$$

1.3 Maps, Homeomorphisms, and Homotopies

The equivalence of two topological spaces is determined by how the points that comprise them are connected. For example, the surface of a cube can be deformed into a sphere without cutting or gluing it because they are connected the same way. They have the same topology. This notion of topological equivalence can be formalized via functions that send the points of one space to points of the other while preserving the connectivity.

This preservation of connectivity is achieved by preserving the open sets. A function from one space to another that preserves the open sets is called a *continuous function* or a *map*. Continuity is a vehicle to define topological equivalence, because a continuous function can send many points to a single point in the target space, or send no points to a given point in the target space. If the former does not happen, that is, when the function is injective, we call it an *embedding* of the domain into the target space. True equivalence is given by a *homeomorphism*, a bijective function from one space to another which has continuity as well as a continuous inverse. This ensures that open sets are preserved in both directions.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

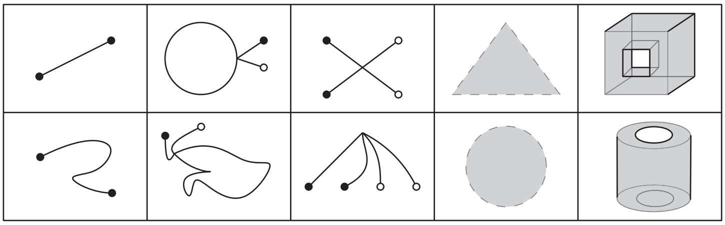


Figure 1.4 Each point set in this figure is homeomorphic to the point set above or below it, but not to any of the others. Open circles indicate points missing from the point set, as do the dashed edges in the point sets second from the right.

Definition 1.15. (Continuous function; Map) A function $f: \mathbb{T} \to \mathbb{U}$ from the topological space $\mathbb{T}$ to another topological space $\mathbb{U}$ is *continuous* if for every open set $Q \subseteq \mathbb{U}$, $f^{-1}(Q)$ is open. Continuous functions are also called *maps*.

Definition 1.16. (Embedding) A map $g: \mathbb{T} \to \mathbb{U}$ is an *embedding* of $\mathbb{T}$ into $\mathbb{U}$ if g is injective.

A topological space can be *embedded* into a Euclidean space by assigning coordinates to its points so that the assignment is continuous and injective. For example, drawing a triangle on paper is an embedding of $\mathbb{S}^1$ into $\mathbb{R}^2$. There are topological spaces that cannot be embedded into a Euclidean space, or even into a metric space – these spaces cannot be represented by any metric.

Next we define a homeomorphism that connects two spaces that have essentially the same topology.

Definition 1.17. (Homeomorphism) Let $\mathbb{T}$ and $\mathbb{U}$ be topological spaces. A *homeomorphism* is a bijective map $h: \mathbb{T} \to \mathbb{U}$ whose inverse is continuous too.

Two topological spaces are *homeomorphic* if there exists a homeomorphism between them.

Homeomorphism induces an equivalence relation among topological spaces, which is why two homeomorphic topological spaces are called *topologically equivalent*. Figure 1.4 shows pairs of homeomorphic topological spaces. A less obvious example is that the open d-ball $\mathbb{B}^d_o$ is homeomorphic to the Euclidean space $\mathbb{R}^d$, given by the homeomorphism $h(x) = x/(1 - ||x||)$. The same map also exhibits that the half-ball $\mathbb{H}^d$ is homeomorphic to the Euclidean half-space $\{x \in \mathbb{R}^d : x_d \geq 0\}$.

For maps between compact spaces, there is a weaker condition to be verified for homeomorphisms because of the following property.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

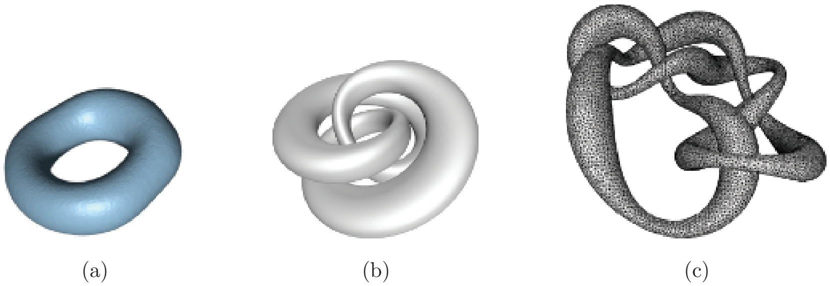


Figure 1.5 Two tori knotted differently, one triangulated (c) and the other not (b). Both are homeomorphic to the standard unknotted torus in (a), but not isotopic to it.

Proposition 1.1. *If $\mathbb{T}$ and $\mathbb{U}$ are compact metric spaces, every bijective map from $\mathbb{T}$ to $\mathbb{U}$ has a continuous inverse.*

One can take advantage of this fact to prove that certain functions are homeomorphisms by showing continuity only in the forward direction. When two topological spaces are subspaces of the same larger space, a notion of similarity called *isotopy* exists which is stronger than homeomorphism. If two subspaces are isotopic, one can be continuously deformed to the other while keeping the deforming subspace homeomorphic to its original form all the time. For example, a solid cube can be continuously deformed into a ball in this manner.

Homeomorphic subspaces are not necessarily isotopic. Consider a torus embedded in $\mathbb{R}^3$, illustrated in Figure 1.5(a). One can embed the torus in $\mathbb{R}^3$ so that it is knotted, as shown in Figure 1.5(b) and (c). The knotted torus is homeomorphic to the standard, unknotted one. However, it is not possible to continuously deform one to the other while keeping it embedded in $\mathbb{R}^3$ and homeomorphic to the original. Any attempt to do so forces the torus to be "self-intersecting" and thus not being a manifold. One way to look at this obstruction is by considering the topology of the space around the tori. Although the knotted and unknotted tori are homeomorphic, their complements in $\mathbb{R}^3$ are not. This motivates us to consider both the notion of an *isotopy*, in which a torus deforms continuously, and the notion of an *ambient isotopy*, in which not only the torus deforms, but the entire $\mathbb{R}^3$ deforms with it.

Definition 1.18. (Isotopy) An *isotopy* connecting two spaces $\mathbb{T} \subseteq \mathbb{R}^d$ and $\mathbb{U} \subseteq \mathbb{R}^d$ is a continuous map $\xi : \mathbb{T} \times [0, 1] \to \mathbb{R}^d$ where $\xi(\mathbb{T}, 0) = \mathbb{T}$, $\xi(\mathbb{T}, 1) = \mathbb{U}$, and for every $t \in [0, 1]$, $\xi(\cdot, t)$ is a homeomorphism between $\mathbb{T}$ and its image $\{\xi(x, t) : x \in \mathbb{T}\}$. An *ambient isotopy* connecting $\mathbb{T}$ and $\mathbb{U}$ is a map ξ :

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

$\mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ such that $\xi(\cdot, 0)$ is the identity function on $\mathbb{R}^d$, $\xi(\mathbb{T}, 1) = \mathbb{U}$, and for each $t \in [0,1]$, $\xi(\cdot, t)$ is a homeomorphism.

For an example, consider the map

$$\xi(x,t) = \frac{1-(1-t)\|x\|}{1-\|x\|}x$$

that sends the open d-ball $\mathbb{B}_o^d$ to itself if $t = 0$, and to the Euclidean space $\mathbb{R}^d$ if $t = 1$. The parameter t plays the role of time, that is, $\xi(\mathbb{B}_o^d, t)$ deforms continuously from a ball at time zero to $\mathbb{R}^d$ at time one. Thus, there is an isotopy between the open d-ball and $\mathbb{R}^d$.

Every ambient isotopy becomes an isotopy if its domain is restricted from $\mathbb{R}^d \times [0,1]$ to $\mathbb{T} \times [0,1]$. It is known that if there is an isotopy between two subspaces, then there exists an ambient isotopy between them. Hence, the two notions are equivalent.

There is another notion of similarity among topological spaces that is weaker than homeomorphism, called *homotopy equivalence*. It relates spaces that can be continuously deformed to one another but the transformation may not preserve homeomorphism. For example, a ball can shrink to a point, which is not homeomorphic to it because a bijective function from an infinite point set to a single point cannot exist. However, homotopy preserves some form of connectivity, such as the number of connected components, holes, and/or voids. This is why a coffee cup is homotopy equivalent to a circle, but not to a ball or a point.

To get to homotopy equivalence, we first need the concept of homotopies, which are isotopies without the homeomorphism.

Definition 1.19. (Homotopy) Let $g\colon \mathbb{X} \to \mathbb{U}$ and $h\colon \mathbb{X} \to \mathbb{U}$ be maps. A *homotopy* is a map $H\colon \mathbb{X} \times [0,1] \to \mathbb{U}$ such that $H(\cdot, 0) = g$ and $H(\cdot, 1) = h$. Two maps are *homotopic* if there is a homotopy connecting them.

For example, let $g\colon \mathbb{B}^3 \to \mathbb{R}^3$ be the identity map on a unit ball and $h\colon \mathbb{B}^3 \to \mathbb{R}^3$ be the map sending every point in the ball to the origin. The fact that g and h are homotopic is demonstrated by the homotopy $H(x,t) = (1-t) \cdot g(x)$. Observe that $H(\mathbb{B}^3, t)$ deforms continuously a ball at time zero to a point at time one. A key property of a homotopy is that, as H is continuous, at every time t the map $H(\cdot, t)$ remains continuous.

For developing more intuition, consider two maps that are not homotopic. Let $g\colon \mathbb{S}^1 \to \mathbb{S}^1$ be the identity map from the circle to itself, and let $h\colon \mathbb{S}^1 \to \mathbb{S}^1$ map every point on the circle to a single point $p \in \mathbb{S}^1$. Although apparently it seems that we can contract a circle to a point, that view is misleading because

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

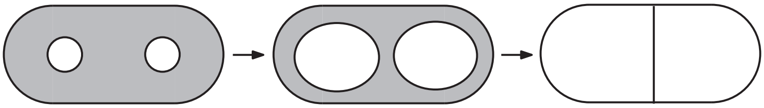


Figure 1.6 All three of the topological spaces are homotopy equivalent, because they are all deformation retracts of the leftmost space.

the map H is required to map every point on the circle at every time to a point on the circle. The contraction of the circle to a point is possible only if we break the continuity, say by cutting or gluing the circle somewhere.

Observe that a homeomorphism relates two topological spaces $\mathbb{T}$ and $\mathbb{U}$ whereas a homotopy or an isotopy (which is a special kind of homotopy) relates two maps, thereby indirectly establishing a relationship between two subspaces $g(\mathbb{X}) \subseteq \mathbb{U}$ and $h(\mathbb{X}) \subseteq \mathbb{U}$. That relationship is not necessarily an equivalent one, but the following is.

Definition 1.20. (Homotopy equivalent) Two topological spaces $\mathbb{T}$ and $\mathbb{U}$ are *homotopy equivalent* if there exist maps $g \colon \mathbb{T} \to \mathbb{U}$ and $h : \mathbb{U} \to \mathbb{T}$ such that $h \circ g$ is homotopic to the identity map $\iota_{\mathbb{T}} \colon \mathbb{T} \to \mathbb{T}$ and $g \circ h$ is homotopic to the identity map $\iota_{\mathbb{U}} \colon \mathbb{U} \to \mathbb{U}$.

Homotopy equivalence is indeed an equivalence relation, that is, if A, B and B, C are homotopy equivalent spaces, so are the pair A, C. Homeomorphic spaces necessarily have the same dimension though homotopy equivalent spaces may have different dimensions. To gain more intuition about homotopy equivalent spaces, we show why a 2-ball is homotopy equivalent to a single point p. Consider a map $h \colon \mathbb{B}^2 \to \{p\}$ and a map $g \colon \{p\} \to \mathbb{B}^2$ where $g(p)$ is any point q in $\mathbb{B}^2$. Observe that $h \circ g$ is the identity map on $\{p\}$, which is trivially homotopic to itself. In the other direction, $g \circ h \colon \mathbb{B}^2 \to \mathbb{B}^2$ sends every point in $\mathbb{B}^2$ to q. A homotopy between $g \circ h$ and the identity map $\mathrm{id}_{\mathbb{B}^2}$ is given by the map $H(x, t) = (1 - t)q + tx$.

An useful intuition for understanding the definition of homotopy equivalent spaces can be derived from the fact that two spaces $\mathbb{T}$ and $\mathbb{U}$ are homotopy equivalent if and only if there exists a third space $\mathbb{X}$ so that both $\mathbb{T}$ and $\mathbb{U}$ are *deformation retracts* of $\mathbb{X}$; see Figure 1.6.

Definition 1.21. (Deformation retract) Let $\mathbb{T}$ be a topological space, and let $\mathbb{U} \subset \mathbb{T}$ be a subspace. A *retraction* r of $\mathbb{T}$ to $\mathbb{U}$ is a map from $\mathbb{T}$ to $\mathbb{U}$ such that $r(x) = x$ for every $x \in \mathbb{U}$. The space $\mathbb{U}$ is a *deformation retract* of $\mathbb{T}$ if the identity map on $\mathbb{T}$ can be continuously deformed to a retraction with no motion of the points already in $\mathbb{U}$: specifically, there is a homotopy called *deformation*

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

retraction $R\colon \mathbb{T}\times[0,1]\to\mathbb{T}$ such that $R(\cdot,0)$ is the identity map on $\mathbb{T}$, $R(\cdot,1)$ is a retraction of $\mathbb{T}$ to $\mathbb{U}$, and $R(x,t)=x$ for every $x\in\mathbb{U}$ and every $t\in[0,1]$.

Fact 1.1. *If $\mathbb{U}$ is a deformation retract of $\mathbb{T}$, then $\mathbb{T}$ and $\mathbb{U}$ are homotopy equivalent.*

For example, any point on a line segment (open or closed) is a deformation retract of the line segment and is homotopy equivalent to it. The letter M is a deformation retract of the letter W, and also of a 1-ball. Moreover, as we said before, two spaces are homotopy equivalent if they are deformation retractions of a common space. The symbols $\varnothing$, ∞, and $\circ\!\!-\!\!\circ$ (viewed as one-dimensional point sets) are deformation retracts of a double doughnut – a doughnut with two holes. Therefore, they are homotopy equivalent to each other, though none of them is a deformation retract of any of the others because one is not a subspace of the other. They are not homotopy equivalent to A, X, O, $\oplus$, $\odot$, $\circledcirc$, a ball, nor a coffee cup.

1.4 Manifolds

A manifold is a topological space that is locally connected in a particular way. A 1-manifold has this local connectivity looking like a segment. A 2-manifold (with boundary) has the local connectivity looking like a complete or partial disk. In layman's terms, a 2-manifold has the structure of a piece of paper or rubber sheet, possibly with the boundaries glued together to form a closed surface – a category that includes disks, spheres, tori, and Möbius bands.

Definition 1.22. (Manifold) A topological space M is an *m-manifold*, or simply a *manifold*, if every point $x\in M$ has a neighborhood homeomorphic to $\mathbb{B}^m_o$ or $\mathbb{H}^m$. The *dimension* of M is m.

Every manifold can be partitioned into boundary and interior points. Observe that these words mean very different things for a manifold than they do for a metric space or topological space.

Definition 1.23. (Boundary; Interior) The *interior* $\operatorname{Int} M$ of an m-manifold M is the set of points in M that have a neighborhood homeomorphic to $\mathbb{B}^m_o$. The *boundary* $\operatorname{Bd} M$ of M is the set of points $M\setminus\operatorname{Int} M$. The boundary $\operatorname{Bd} M$, if not empty, consists of the points that have a neighborhood homeomorphic to $\mathbb{H}^m$. If $\operatorname{Bd} M$ is the empty set, we say that M is *without boundary*.

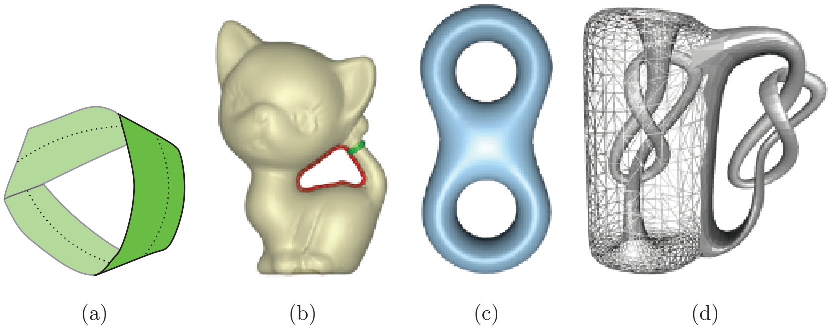


Figure 1.7 (a) A Möbius band. (b) Removal of the red and green loops opens up the torus into a topological disk. (c) A double torus: every surface without boundary in $\mathbb{R}^3$ resembles a sphere or a conjunction of one or more tori. (d) Double torus knotted.

A single point, a 0-ball, is a 0-manifold without boundary according to this definition. The closed disk $\mathbb{B}^2$ is a 2-manifold whose interior is the open disk $\mathbb{B}^2_o$ and whose boundary is the circle $\mathbb{S}^1$. The open disk $\mathbb{B}^2_o$ is a 2-manifold whose interior is $\mathbb{B}^2_o$ and whose boundary is the empty set. This highlights an important difference between Definitions 1.13 and 1.23 of "boundary": when $\mathbb{B}^2_o$ is viewed as a point set in the space $\mathbb{R}^2$, its boundary is $\mathbb{S}^1$ according to Definition 1.13; but viewed as a manifold, its boundary is empty according to Definition 1.23. The boundary of a manifold is *always* included in the manifold.

The open disk $\mathbb{B}^2_o$, the Euclidean space $\mathbb{R}^2$, the sphere $\mathbb{S}^2$, and the torus are all connected 2-manifolds without boundary. The first two are homeomorphic to each other, but the last two are not. The sphere and the torus in $\mathbb{R}^3$ are compact (bounded and closed with respect to $\mathbb{R}^3$) whereas $\mathbb{B}^2_o$ and $\mathbb{R}^2$ are not.

A d-manifold, $d \geq 2$, can have orientations whose formal definition we skip here. Informally, we say that a 2-manifold M is *non-orientable* if, starting from a point p, one can walk on one side of M and end up on the opposite side of M upon returning to p. Otherwise, M is *orientable*. Spheres and balls are orientable, whereas the *Möbius band* in Figure 1.7(a) is a non-orientable 2-manifold with boundary.

A *surface* is a 2-manifold that is a subspace of $\mathbb{R}^d$. Any compact surface without boundary in $\mathbb{R}^3$ is an orientable 2-manifold. To be non-orientable, a compact surface must have a non-empty boundary (like the Möbius band) or be embedded in a four- or higher-dimensional Euclidean space.

A surface can sometimes be disconnected by removing one or more *loops* (connected 1-manifolds without boundary) from it. The *genus* of an orientable and compact surface without boundary is g if $2g$ is the maximum number of

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

loops that can be removed from the surface without disconnecting it; here the loops are permitted to intersect each other. For example, the sphere has genus zero as every loop cuts it into two disks. The torus has genus one: a circular cut around its neck and a second circular cut around its circumference, illustrated in Figure 1.7(b), allow it to unfold into a topological disk. A third loop would cut it into two pieces. Figure 1.7(c) and (d) each shows a 2-manifold without boundary of genus two. Although a high-genus surface can have a very complex shape, all compact 2-manifolds in $\mathbb{R}^3$ that have the same genus and no boundary are homeomorphic to each other.

1.4.1 Smooth Manifolds

A purely topological manifold has no geometry. But if we embed it in a Euclidean space, it could appear smooth or wrinkled. We now introduce a "geometric" manifold by imposing a differential structure on it. For the rest of this chapter, we focus on only manifolds without boundary.

Consider a map $\phi\colon U \to W$ where U and W are open sets in $\mathbb{R}^k$ and $\mathbb{R}^d$, respectively. The map ϕ has d components, namely $\phi(x) = (\phi_1(x), \phi_2(x), \ldots, \phi_d(x))$, where $x = (x_1, x_2, \ldots, x_k)$ denotes a point in $\mathbb{R}^k$. The *Jacobian* of ϕ at x is the $d \times k$ matrix of the first-order partial derivatives

$$\begin{bmatrix} \frac{\partial \phi_1(x)}{\partial x_1} & \cdots & \frac{\partial \phi_1(x)}{\partial x_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_d(x)}{\partial x_1} & \cdots & \frac{\partial \phi_d(x)}{\partial x_k} \end{bmatrix}.$$

The map ϕ is *regular* if its Jacobian has rank k at every point in U. The map ϕ is C^i-continuous if the i-th-order partial derivatives of ϕ are continuous.

The reader may be familiar with *parametric surfaces*, for which U is a two-dimensional *parameter space* and its image $\phi(U)$ in d-dimensional space is a parametric surface. Unfortunately, a single parametric surface cannot easily represent a manifold with a complicated topology. However, for a manifold to be smooth, it suffices that each point on the manifold has a neighborhood that looks like a smooth parametric surface.

Definition 1.24. (Smooth embedded manifold) For any $i > 0$, an m-manifold M without boundary embedded in $\mathbb{R}^d$ is *C^i-smooth* if for every point $p \in M$, there exists an open set $U_p \subset \mathbb{R}^m$, a neighborhood $W_p \subset \mathbb{R}^d$ of p, and a map $\phi_p\colon U_p \to W_p \cap M$ such that (i) ϕ_p is C^i-continuous, (ii) ϕ_p is a homeomorphism, and (iii) ϕ_p is regular. If $m = 2$, we call M a *C^i-smooth surface*.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

The first condition says that each map is continuously differentiable at least i times. The second condition requires each map to be bijective, ruling out “wrinkles” where multiple points in U map to a single point in W. The third condition prohibits any map from having a directional derivative of zero at any point in any direction. The first and third conditions together enforce smoothness, and imply that there is a well-defined tangent m-flat at each point in M. The three conditions together imply that the maps ϕ_p defined in the neighborhood of each point $p \in M$ overlap smoothly. There are two extremes of smoothness. We say that M is C^∞-smooth if for every point $p \in M$, the partial derivatives of ϕ_p of all orders are continuous. On the other hand, M is *nonsmooth* if M is an m-manifold (therefore C^0-smooth) but not C^1-smooth.

1.5 Functions on Smooth Manifolds

In previous sections, we introduced topological spaces, including the special case of (smooth) manifolds. Very often, a space can be equipped with continuous functions defined on it. In this section, we focus on *real-valued* functions of the form $f\colon X \to \mathbb{R}$ defined on a topological space X, also called *scalar functions*; see Figure 1.8(a) for the graph of a function $f\colon \mathbb{R}^2 \to \mathbb{R}$. Scalar functions appear commonly in practice that describe space/data of interest (e.g., the elevation function defined on the surface of the Earth). We are interested in the topological structures behind scalar functions. In this section, we limit our discussion to nicely behaved scalar functions (called Morse functions) defined on smooth manifolds. Their topological structures are characterized by the so-called critical points which we will introduce below. Later in the book we will also discuss scalar functions on simplicial complex domains, as well as more complex maps defined on a space X, for example, a multivariate function $f\colon X \to \mathbb{R}^d$.

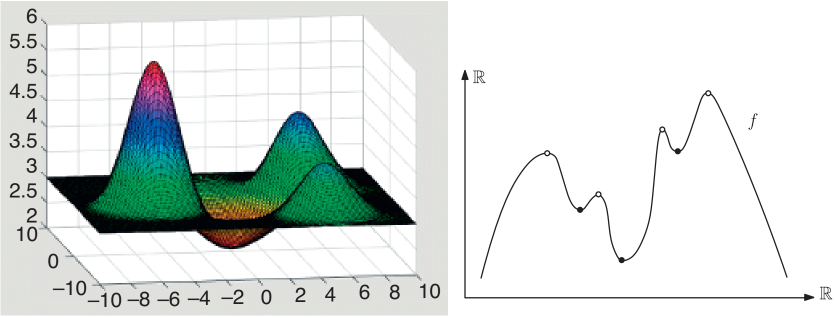
1.5.1 Gradients and Critical Points

In what follows, for simplicity of presentation, we assume that we consider smooth (C^∞-continuous) functions and smooth manifolds embedded in $\mathbb{R}^d$, even though often we only require the functions (resp. manifolds) to be C^2-continuous (resp. C^2-smooth).

To provide intuition, let us start with a smooth scalar function defined on the real line, $f\colon \mathbb{R} \to \mathbb{R}$; the graph of such a function is shown in Figure 1.8(b). Recall that the *derivative* of a function at a point $x \in \mathbb{R}$ is defined as

$$Df(x) = \frac{d}{dx} f(x) = \lim_{t \to 0} \frac{f(x+t) - f(x)}{t}. \tag{1.1}$$

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press



(a) (b)

Figure 1.8 (a) The graph of a function $f: \mathbb{R}^2 \to \mathbb{R}$. (b) The graph of a function $f: \mathbb{R} \to \mathbb{R}$ with critical points marked.

The value $Df(x)$ gives the rate of change of the value of f at x. This can be visualized as the slope of the tangent line of the graph of f at $(x, f(x))$. The *critical points* of f are the set of points x such that $Df(x) = 0$. For a function defined on the real line, there are two types of critical points in the generic case: maxima and minima, as marked in Figure 1.8(b).

Now suppose we have a smooth function $f: \mathbb{R}^d \to \mathbb{R}$ defined on $\mathbb{R}^d$. Fix an arbitrary point $x \in \mathbb{R}^d$. As we move a little around x within its local neighborhood, the rate of change of f differs depending on which direction we move. This gives rise to the *directional derivative* $D_v f(x)$ at x in direction (i.e., a unit vector) $v \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ is the unit $(d-1)$-sphere, defined as

$$D_v f(x) = \lim_{t \to 0} \frac{f(x + t \cdot v) - f(x)}{t}. \tag{1.2}$$

The gradient vector of f at $x \in \mathbb{R}^d$ intuitively captures the direction of steepest increase of the function f. More precisely, we have the following.

Definition 1.25. (Gradient for functions on $\mathbb{R}^d$) Given a smooth function $f: \mathbb{R}^d \to \mathbb{R}$, the *gradient vector field* $\nabla f: \mathbb{R}^d \to \mathbb{R}^d$ is defined as follows: for any $x \in \mathbb{R}^d$,

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}(x),\ \frac{\partial f}{\partial x_2}(x),\ \ldots,\ \frac{\partial f}{\partial x_d}(x) \right]^{\mathrm{T}}, \tag{1.3}$$

where $(x_1, x_2, \ldots, x_d)$ represents an orthonormal coordinate system for $\mathbb{R}^d$. The vector $\nabla f(x) \in \mathbb{R}^d$ is called the *gradient vector of f at x*. A point $x \in \mathbb{R}^d$ is a *critical point* if $\nabla f(x) = [0\ 0\ \ldots\ 0]^{\mathrm{T}}$; otherwise, x is *regular*.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

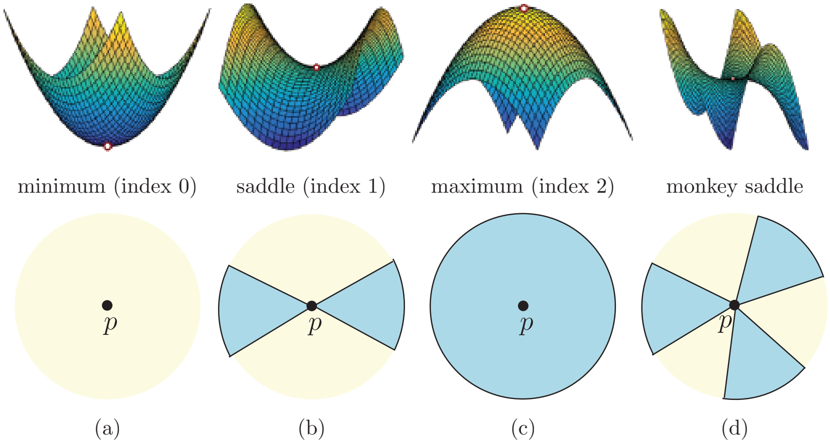


Figure 1.9 (top) The graph of the function around nondegenerate critical points for a smooth function on $\mathbb{R}^2$, and a degenerate critical point, called "monkey saddle." For example, for an index-0 critical point p, its local neighborhood can be written as $f(x) = f(p) + x_1^2 + x_2^2$, making p a local minimum. (bottom) The local (closed) neighborhood of the corresponding critical point in the domain $\mathbb{R}^2$, where the dark blue colored regions are the portion of the neighborhood of p whose function value is at most $f(p)$.

Observe that for any $v \in \mathbb{R}^d$, the directional derivative satisfies that $D_v f(x) = \langle \nabla f(x), v \rangle$. It then follows that $\nabla f(x) \in \mathbb{R}^d$ is along the unit vector v where $D_v f(x)$ is maximized among the directional derivatives in all unit directions around x; and its magnitude $\|\nabla f(x)\|$ equals the value of this maximum directional derivative. The critical points of f are those points where the directional derivative vanishes in all directions – locally, the rate of change of f is zero no matter which direction one deviates from x. See Figure 1.9 for the three types of critical points, minimum, saddle point, and maximum, for a generic smooth function $f : \mathbb{R}^2 \to \mathbb{R}$.

Finally, we can extend the above definitions of gradients and critical points to a smooth function $f : M \to \mathbb{R}$ defined on a smooth Riemannian m-manifold M. Here, a Riemannian manifold is a manifold equipped with a Riemannian metric, which is a smoothly varying inner product defined on the tangent spaces. This allows the measurements of length so as to define gradient. At a point $x \in M$, denote the tangent space of M at x by TM_x, which is the m-dimensional vector space consisting of all tangent vectors of M at x. For example, TM_x is an m-dimensional linear space $\mathbb{R}^m$ for an m-dimensional manifold M embedded in the Euclidean space $\mathbb{R}^d$, with Riemannian metric (inner product in the tangent space) induced from $\mathbb{R}^d$.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

The gradient ∇f is a vector field on M, that is, $\nabla f \colon M \to \mathrm{TM}$ maps every point $x \in M$ to a vector $\nabla f(x) \in \mathrm{TM}_x$ in the tangent space of M at x. Similar to the case for a function defined on $\mathbb{R}^d$, the gradient vector field ∇f satisfies that for any $x \in M$ and $v \in \mathrm{TM}_x$, $\langle \nabla f(x), v \rangle$ gives rise to the directional derivative $D_v f(x)$ of f in direction v, and $\nabla f(x)$ still specifies the direction of steepest increase of f along all directions in TM_x with its magnitude being the maximum rate of change. More formally, we have the following definition, analogous to Definition 1.25 for the case of a smooth function on $\mathbb{R}^d$.

Definition 1.26. (Gradient vector field; Critical points) Given a smooth function $f \colon M \to \mathbb{R}$ defined on a smooth m-dimensional Riemannian manifold M, the *gradient vector field* $\nabla f \colon M \to \mathrm{TM}$ is defined as follows: for any $x \in M$, let $(x_1, x_2, \ldots, x_m)$ be a local coordinate system in a neighborhood of x with orthonormal unit vectors x_i, then the gradient at x is

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

A point $x \in M$ is *critical* if $\nabla f(x)$ vanishes, in which case $f(x)$ is called a *critical value* for f. Otherwise, x is *regular*.

It follows from the chain rule that the criticality of a point x is independent of the local coordinate system being used.

1.5.2 Morse Functions and Morse Lemma

From the first-order derivatives of a function we can determine critical points. We can learn more about the "type" of the critical points by inspecting the second-order derivatives of f.

Definition 1.27. (Hessian matrix; Nondegenerate critical points) Given a smooth m-manifold M, the *Hessian matrix* of a twice differentiable function $f \colon M \to \mathbb{R}$ at x is the matrix of second-order partial derivatives,

$$\mathrm{Hessian}(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_m}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_2}2(x) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_m}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_m \partial x_1}(x) & \frac{\partial^2 f}{\partial x_m \partial x_2}2(x) & \cdots & \frac{\partial^2 f}{\partial x_m \partial x_m}(x) \end{bmatrix},$$

where $(x_1, x_2, \ldots, x_m)$ is a local coordinate system in a neighborhood of x.

A critical point x of f is *nondegenerate* if its Hessian matrix, Hessian(x), is nonsingular (has nonzero determinant); otherwise, it is a *degenerate critical point*.

For example, consider $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x^3 - 3xy^2$. The origin $(0, 0)$ is a degenerate critical point often referred to as a "monkey saddle:" see Figure 1.9(d), where the graph of the function around $(0, 0)$ goes up and down three times (instead of twice as for a nondegenerate saddle shown in Figure 1.9b). It turns out that, as a consequence of the Morse Lemma below, nondegenerate critical points are always isolated whereas the degenerate ones may not be so. A simple example is $f\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $f(x, y) = x^2$, where all points on the y-axis are degenerate critical points. The local neighborhood of nondegenerate critical points can be completely characterized by the following Morse Lemma.

Proposition 1.2. (Morse Lemma) *Given a smooth function $f\colon M \to \mathbb{R}$ defined on a smooth m-manifold M, let p be a nondegenerate critical point of f. Then there is a local coordinate system in a neighborhood $U(p)$ of p so that (i) the coordinate of p is* $(0, 0, \ldots, 0)$*, and (ii) locally for every point $x = (x_1, x_2, \ldots, x_m)$ in neighborhood $U(p)$,*

$$f(x) = f(p) - x_1^2 - \cdots - x_s^2 + x_{s+1}^2 \cdots + x_m^2, \qquad \textit{for some}\ \ s \in [0, m].$$

The number s of minus signs in the above quadratic representation of $f(x)$ is called the index of the critical point p.

For a smooth function $f\colon M \to \mathbb{R}$ defined on a 2-manifold M, an index-0, index-1, or index-2 (nondegenerate) critical point corresponds to a minimum, a saddle, or a maximum, respectively. For a function defined on an m-manifold, nondegenerate critical points include minima (index 0), maxima (index m), and $m - 1$ types of saddle points.

The behavior of degenerate critical points is more complicated to characterize. Instead, we now introduce a family of "nice" functions, called *Morse functions*, whose critical points cannot be degenerate.

Definition 1.28. (Morse function) A smooth function $f\colon M \to \mathbb{R}$ defined on a smooth manifold M is a *Morse function* if and only if: (i) none of f's critical points are degenerate; and (ii) the critical points have distinct function values.

Limiting our study only to well-behaved Morse functions is not too restrictive as the Morse functions form an open and dense subset of the space of all smooth functions $C^\infty(M)$ on M. So in this sense, a generic function is

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

a Morse function. On the other hand, it is much cleaner to characterize the topology induced by such a function, which we do now.

1.5.3 Connection to Topology

We now characterize how critical points influence the topology of M induced by the scalar function $f\colon M \to \mathbb{R}$.

Definition 1.29. (Interval, sublevel, and superlevel sets) Given $f\colon M \to \mathbb{R}$ and $I \subseteq \mathbb{R}$, the *interval levelset* of f with respect to I is defined as

$$M_I = f^{-1}(I) = \{x \in M \mid f(x) \in I\}.$$

The case for $I = (-\infty, a]$ is also referred to as the *sublevel set* $M_{\leq a} := f^{-1}((-\infty, a])$ of f, while $M_{\geq a} := f^{-1}([a, \infty))$ is called the *superlevel set*; and $f^{-1}(a)$ is called the *levelset* of f at $a \in \mathbb{R}$.

Given $f\colon M \to \mathbb{R}$, imagine sweeping M with increasing function values of f. It turns out that the topology of the sublevel sets can only change when we sweep through critical values of f. More precisely, we have the following classical result, where a diffeomorphism is a homeomorphism that is smooth in both directions.

Theorem 1.3. (Homotopy type of sublevel sets) *Let $f\colon M \to \mathbb{R}$ be a smooth function defined on a manifold M. Given $a < b$, suppose the interval levelset $M_{[a,b]} = f^{-1}([a, b])$ is compact and contains no critical points of f. Then $M_{\leq a}$ is diffeomorphic to $M_{\leq b}$.*

Furthermore, $M_{\leq a}$ is a deformation retract of $M_{\leq b}$, and the inclusion map $i\colon M_{\leq a} \hookrightarrow M_{\leq b}$ is a homotopy equivalence.

As an illustration, consider the example of height function $f\colon M \to \mathbb{R}$ defined on a vertical torus as shown in Figure 1.10(a). There are four critical points for the height function f, u (minimum), v, w (saddles), and z (maximum). We have that $M_{\leq a}$ is: (i) empty for $a < f(u)$; (ii) homeomorphic to a 2-disk for $f(u) < a < f(v)$; (iii) homeomorphic to a cylinder for $f(v) < a < f(w)$; (iv) homeomorphic to a compact genus-one surface with a circle as boundary for $f(w) < a < f(z)$; and (v) a full torus for $a > f(z)$.

Theorem 1.3 states that the homotopy type of the sublevel set remains the same until it passes a critical point. For Morse functions, we can also characterize the homotopy type of sublevel sets around critical points, captured by *attaching k-cells*.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

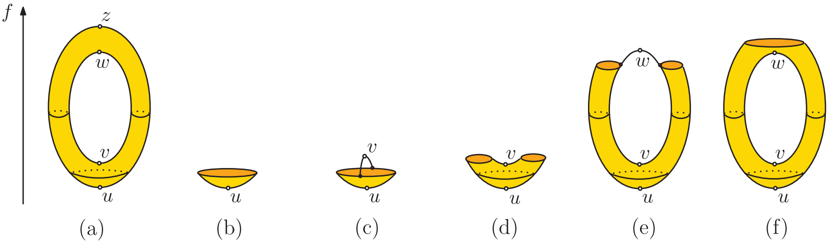


Figure 1.10 (a) The height function defined on a torus with critical points u, v, w, and z. (b)–(f) Passing through an index-k critical point is the same as attaching a k-cell from the homotopy point of view. For example, $M_{\leq a+\varepsilon}$ for $a = f(v)$ (as shown in (d)) is homotopy equivalent to attaching a 1-cell (shown in (c)) to $M_{\leq a-\varepsilon}$ (shown in (b)) for an infinitesimal positive ε.

Specifically, recall that $\mathbb{B}^k$ is the k-dimensional unit Euclidean ball, and its boundary is $\mathbb{S}^{k-1}$, the $(k-1)$-dimensional sphere. Let X be a topological space, and $g\colon \mathbb{S}^{k-1} \to X$ a continuous map. For $k > 0$, *attaching a k-cell to X (w.r.t. g)* is obtained by attaching the k-cell $\mathbb{B}^k$ to X along its boundary as follows: first, take the disjoint union of X and $\mathbb{B}^k$, and next, identify all points $x \in \mathbb{S}^{k-1}$ with $g(x) \in X$. For the special case of $k = 0$, attaching a 0-cell to X is obtained by simply taking the disjoint union of X and a single point.

The following theorem states that, from the homotopy point of view, sweeping past an index-k critical point is equivalent to attaching a k-cell to the sublevel set. See Figure 1.10 for illustrations.

Theorem 1.4. *Given a Morse function $f\colon M \to \mathbb{R}$ defined on a smooth manifold M, let p be an index-k critical point of f with $\alpha = f(p)$. Assume $f^{-1}([\alpha-\varepsilon, \alpha+\varepsilon])$ is compact for a sufficiently small $\varepsilon > 0$ such that there are no other critical points of f contained in this interval levelset other than p. Then the sublevel set $M_{\leq\alpha+\varepsilon}$ has the same homotopy type as $M_{\leq\alpha-\varepsilon}$ with a k-cell attached to its boundary* Bd $M_{\leq\alpha-\varepsilon}$.

Finally, we state the well-known Morse inequalities, connecting critical points with the so-called Betti numbers of the domain which we will define formally in Section 2.5. In particular, fixing a field coefficient, the i-th Betti number is the rank of the so-called i-th (singular) homology group of a topological space X.

Theorem 1.5. (Morse inequalities) *Let f be a Morse function on a smooth compact d-manifold M. For $0 \leq i \leq d$, let c_i denote the number of critical points of f with index i, and β_i be the i-th Betti number of M. We then have:*

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

(a) $c_i \geq \beta_i$ *for all* $i \geq 0$*; and* $\sum_{i=0}^{d}(-1)^i c_i = \sum_{i=0}^{d}(-1)^i \beta_i$ *(weak Morse inequality);*

(b) $c_i - c_{i-1} + c_{i-2} - \cdots \pm c_0 \geq \beta_i - \beta_{i-1} + \beta_{i-2} \cdots \pm \beta_0$ *for all* $i \geq 0$ *(strong Morse inequality).*

1.6 Notes and Exercises

A good source on point set topology is Munkres [243]. The concepts of various maps and manifolds are well described in Hatcher [186]. The books by Guillemin and Pollack [179] and Milnor [233, 234] are good sources for Morse theory on smooth manifolds and differential topology in general.

Exercises

1. A space is called Hausdorff if every two disjoint point sets have disjoint open sets containing them.
   (a) Give an example of a space that is not Hausdorff.
   (b) Give an example of a space that is Hausdorff.
   (c) Show the above examples on the same ground set $\mathbb{T}$.
2. In every space $\mathbb{T}$, the point sets $\varnothing$ and $\mathbb{T}$ are both closed and open.
   (a) Give an example of a space that has more than two sets that are both closed and open, and list all of those sets.
   (b) Explain the relationship between the idea of connectedness and the number of sets that are both closed and open.
3. A topological space $\mathbb{T}$ is called *path connected* if any two points $x, y \in \mathbb{T}$ can be joined by a path, that is, there exists a continuous map $f : [0, 1] \rightarrow \mathbb{T}$ of the segment $[0, 1] \subset \mathbb{R}$ onto $\mathbb{T}$ so that $f(0) = x$ and $f(1) = y$. Prove that a path connected space is also connected but the converse may not be true; however, if $\mathbb{T}$ is finite, then the two notions are equivalent.
4. Prove that for every subset X of a metric space, $\mathrm{Cl}\,\mathrm{Cl}\,X = \mathrm{Cl}\,X$. In other words, augmenting a set with its limit points does not give it more limit points.
5. Show that any metric on a finite set induces a discrete topology.
6. Prove that the metric is a continuous function on the Cartesian space $\mathbb{T} \times \mathbb{T}$ of a metric space $\mathbb{T}$.
7. Give an example of a bijective function that is continuous, but its inverse is not. In light of Proposition 1.1, the spaces need to be noncompact.
8. A space is called *normal* if it is Hausdorff and for any two disjoint closed sets X and Y, there are disjoint open sets $U_X \supset X$ and $U_Y \supset Y$. Show that any metric space is normal. Show the same for any compact space.

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press

9. Let $f : \mathbb{T} \to \mathbb{U}$ be a continuous function of a compact space $\mathbb{T}$ into another space $\mathbb{U}$. Prove that the image $f(\mathbb{T})$ is compact.
10. (a) Construct an explicit deformation retraction of $\mathbb{R}^k \setminus \{o\}$ onto $\mathbb{S}^{k-1}$ where o denotes the origin. Also, show $\mathbb{R}^k \cup \{\infty\}$ is homeomorphic to $\mathbb{S}^k$.

    (b) Show that any d-dimensional finite convex polytope is homeomorphic to the d-dimensional unit ball $\mathbb{B}^d$.
11. Deduce that homeomorphism is an equivalence relation. Show that the relation of homotopy among maps is an equivalence relation.
12. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined as $f(x_1, x_2, x_3) = 3x_1^2 + 3x_2^2 - 9x_3^2$. Show that the origin $(0, 0, 0)$ is a critical point of f. Give the index of this critical point. Let S denote the unit sphere centered at the origin. Show that $f^{(-\infty,0]} \cap S$ is homotopy equivalent to two points, whereas $f^{[0,\infty)} \cap S$ is homotopy equivalent to $\mathbb{S}^1$, the unit 1-sphere (i.e., circle).

https://doi.org/10.1017/9781009099950.002 Published online by Cambridge University Press