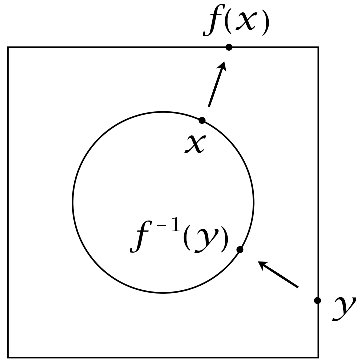
Notes on Introductory Point-Set Topology

Allen Hatcher

Chapter 1. Basic Point-Set Topology

One way to describe the subject of Topology is to say that it is qualitative geometry. The idea is that if one geometric object can be continuously transformed into another, then the two objects are to be viewed as being *topologically* the same. For example, a circle and a square are topologically equivalent. Physically, a rubber band can be stretched into the form of either a circle or a square, as well as many other shapes which are also viewed as being topologically equivalent. On the other hand, a figure eight curve formed by two circles touching at a point is to be regarded as topologically distinct from a circle or square. A qualitative property that distinguishes the circle from the figure eight is the number of connected pieces that remain when a single point is removed: When a point is removed from a circle what remains is still connected, a single arc, whereas for a figure eight if one removes the point of contact of its two circles, what remains is two separate arcs, two separate pieces.

The term used to describe two geometric objects that are topologically equivalent is *homeomorphic*. Thus a circle and a square are homeomorphic. Concretely, if we place a circle C inside a square S with the same center point, then projecting the circle radially outward to the square defines a function $f:C\to S$, and this function is continuous: small changes in x produce small changes in $f(x)$. The function f has an inverse $f^{-1}:S\to C$ obtained by projecting the square radially inward to the circle, and this is continuous as well. One says that f is a homeomorphism between C and S.



One of the basic problems of Topology is to determine when two given geometric objects are homeomorphic. This can be quite difficult in general.

Our first goal will be to define exactly what the ‘geometric objects’ are that one studies in Topology. These are called *topological spaces.* The definition turns out to be extremely general, so that many objects that are topological spaces are not very geometric at all, in fact.

Topological Spaces

Rather than jump directly into the definition of a topological space we will first spend a little time motivating the definition by discussing the notion of continuity of a function. One could say that topological spaces are the objects for which continuous functions can be defined.

For the sake of simplicity and concreteness let us talk about functions $f:\mathbb{R}\to\mathbb{R}$. There are two definitions of continuity for such a function that the reader may already be familiar with, the ε-δ definition and the definition in terms of limits. But it is a third definition, equivalent to these two, that is the one we want here. This definition is expressed in terms of the notion of an open set in $\mathbb{R}$, generalizing the familiar idea of an open interval (a,b).

Definition. A subset O of $\mathbb{R}$ is *open* if for each point $x \in O$ there exists an interval (a,b) that contains x and is contained in O.

With this definition an open interval certainly qualifies as an open set. Other examples are:

- $\mathbb{R}$ itself is an open set, as are semi-infinite intervals (a,∞) and $(-\infty,a)$.
- The complement of a finite set in $\mathbb{R}$ is open.
- If A is the union of the infinite sequence $x_n = 1/n$, $n = 1,2,\cdots$, together with its limit 0 then the complement $\mathbb{R} - A$ is open.
- Any union of open intervals is an open set. The preceding examples are special cases of this. The converse statement is also true: every open set O is a union of open intervals since for each $x \in O$ there is an open interval (a_x,b_x) with $x \in (a_x,b_x) \subset O$, and O is the union of all these intervals (a_x,b_x).
- The empty set $\varnothing$ is open, since the condition for openness is satisfied vacuously as there are no points x where the condition could fail to hold.

Here are some examples of sets which are not open:

- A closed interval $[a,b]$ is not an open set since there is no open interval about either a or b that is contained in $[a,b]$. Similarly, half-open intervals $[a,b)$ and $(a,b]$ are not open sets when $a < b$.
- A nonempty finite set is not open.

Now for the nice definition of a continuous function in terms of open sets:

Definition. A function $f:\mathbb{R}\to\mathbb{R}$ is *continuous* if for each open set O in $\mathbb{R}$ the inverse image $f^{-1}(O) = \{\, x \in \mathbb{R} \mid f(x) \in O \,\}$ is also an open set.

To see that this corresponds to the intuitive notion of continuity, consider what would happen if this condition failed to hold for a function f. There would then be an open set O for which $f^{-1}(O)$ was not open. This means there would be a point $x_0 \in f^{-1}(O)$ for which there was no interval (a,b) containing x_0 and contained in $f^{-1}(O)$. This is equivalent to saying there would be points x arbitrarily close to x_0 that are in the complement of $f^{-1}(O)$. For x to be in the complement of $f^{-1}(O)$

means that $f(x)$ is not in O. On the other hand, x_0 was in $f^{-1}(O)$ so $f(x_0)$ is in O. Since O was assumed to be open, there is an interval (c,d) about $f(x_0)$ that is contained in O. The points $f(x)$ that are not in O are therefore not in (c,d) so they remain at least a fixed positive distance from $f(x_0)$. To summarize: there are points x arbitrarily close to x_0 for which $f(x)$ remains at least a fixed positive distance away from $f(x_0)$. This certainly says that f is discontinuous at x_0.

This reasoning can be reversed. A reasonable interpretation of discontinuity of f at x_0 would be that there are points x arbitrarily close to x_0 for which $f(x)$ stays at least a fixed positive distance away from $f(x_0)$. Call this fixed positive distance ε. Let O be the open set $(f(x_0)-\varepsilon, f(x_0)+\varepsilon)$. Then $f^{-1}(O)$ contains x_0 but it does not contain any points x for which $f(x)$ is not in O, and we are assuming there are such points x arbitrarily close to x_0, so $f^{-1}(O)$ is not open since it does not contain all points in some interval (a,b) about x_0.

The definition we have given for continuity of functions $\mathbb{R}\to\mathbb{R}$ can be applied more generally to functions $\mathbb{R}^n\to\mathbb{R}^n$ and even $\mathbb{R}^m\to\mathbb{R}^n$ once one has a notion of what open sets in $\mathbb{R}^n$ are. The natural definition generalizing the case $n=1$ is to say that a set O in $\mathbb{R}^n$ is open if for each $x\in O$ there exists an open ball containing x and contained in O, where an open ball of radius r and center x_0 is defined to be the set of points x of distance less than r from x_0. Here the distance from x to x_0 is measured as in linear algebra, as the length of the vector $x-x_0$, the square root of the dot product of this vector with itself.

This definition of open sets in $\mathbb{R}^n$ does not depend as heavily on the notion of distance in $\mathbb{R}^n$ as might appear. For example in $\mathbb{R}^2$ where open “balls" are open disks, we could use open squares instead of open disks since if a point $x\in O$ is contained in an open disk contained in O then it is also contained in an open square contained in the disk and hence in O, and conversely, if x is contained in an open square contained in O then it is contained in an open disk contained in the open square and hence in O. In a similar way we could use many other shapes besides disks and squares, such as ellipses or polygons with any number of sides.

After these preliminary remarks we now give the definition of a topological space.

Definition. A *topological space* is a set X together with a collection $\mathcal{O}$ of subsets of X, called *open sets*, such that:

(1) The union of any collection of sets in $\mathcal{O}$ is in $\mathcal{O}$.

(2) The intersection of any finite collection of sets in $\mathcal{O}$ is in $\mathcal{O}$.

(3) Both $\varnothing$ and X are in $\mathcal{O}$.

The collection $\mathcal{O}$ of open sets is called a *topology* on X.

All three of these conditions hold for open sets in $\mathbb{R}$ as defined earlier. To check that (1) holds, suppose that we have a collection of open sets O_α where the index α ranges over some index set I, either finite or infinite. A point $x \in \bigcup_\alpha O_\alpha$ lies in some O_α, which is open so there is an interval (a,b) with $x \in (a,b) \subset O_\alpha$, hence $x \in (a,b) \subset \bigcup_\alpha O_\alpha$ so $\bigcup_\alpha O_\alpha$ is open. To check (2) it suffices by induction to check that the intersection of two open sets O_1 and O_2 is open. If $x \in O_1 \cap O_2$ then x lies in open intervals in O_1 and O_2, and there is a smaller open interval in the intersection of these two open intervals that contains x. This open interval lies in $O_1 \cap O_2$, so $O_1 \cap O_2$ is open. Finally, condition (3) obviously holds for open sets in $\mathbb{R}$.

In a similar fashion one can check that open sets in $\mathbb{R}^2$ or more generally $\mathbb{R}^n$ also satisfy (1)–(3).

Notice that the intersection of an infinite collection of open sets in $\mathbb{R}$ need not be open. For example, the intersection of all the open intervals $(-1/n, 1/n)$ for $n = 1, 2, \cdots$ is the single point $\{0\}$ which is not open. This explains why condition (2) is only for finite intersections.

It is always possible to construct at least two topologies on every set X by choosing the collection $\mathcal{O}$ of open sets to be as large as possible or as small as possible:

- The collection $\mathcal{O}$ of *all* subsets of X defines a topology on X called the *discrete topology*.
- If we let $\mathcal{O}$ consist of just X itself and $\varnothing$, this defines a topology, the *trivial topology*.

Thus we have three different topologies on $\mathbb{R}$, the usual topology, the discrete topology, and the trivial topology. Here are two more, the first with fewer open sets than the usual topology, the second with more open sets:

- Let $\mathcal{O}$ consist of the empty set together with all subsets of $\mathbb{R}$ whose complement is finite. The axioms (1)–(3) are easily verified, and we leave this for the reader to check. Every set in $\mathcal{O}$ is open in the usual topology, but not vice versa.
- Let $\mathcal{O}$ consist of all sets O such that for each $x \in O$ there is an interval $[a,b)$ with $x \in [a,b) \subset O$. Properties (1)–(3) can be checked by almost the same argument as for the usual topology on $\mathbb{R}$, and again we leave this for the reader to do. Intervals $[a,b)$ are certainly in $\mathcal{O}$ so this topology is different from the usual topology on $\mathbb{R}$. Every interval (a,b) is in $\mathcal{O}$ since it can be expressed as a union of a sequence of intervals $[a_n, b)$ in $\mathcal{O}$ where the numbers a_n are chosen to satisfy $a < a_n < b$

and to approach a from above. It follows that $\mathcal{O}$ contains all sets that are open in the usual topology since these can be expressed as unions of intervals (a,b).

These examples illustrate how one can have two topologies $\mathcal{O}$ and $\mathcal{O}'$ on a set X, with every set that is open in the $\mathcal{O}$ topology is also open in the $\mathcal{O}'$ topology, so $\mathcal{O} \subset \mathcal{O}'$. In this situation we say that the topology $\mathcal{O}'$ is *finer* than $\mathcal{O}$ and that $\mathcal{O}$ is *coarser* than $\mathcal{O}'$. Thus the discrete topology on X is finer than any other topology and the trivial topology is coarser than any other topology. In the case $X = \mathbb{R}$ we have interpolated three other topologies between these two extremes, with the finite-complement topology being coarser than the usual topology and the half-open-interval topology being finer than the usual topology. In general, given two topologies on a set X, it need not be true that either one is finer or coarser than the other.

Here is another piece of basic terminology:

Definition. A subset A of a topological space X is *closed* if its complement $X - A$ is open.

For example, in $\mathbb{R}$ with the usual topology a closed interval $[a,b]$ is a closed subset. Similarly, in $\mathbb{R}^2$ with its usual topology a closed disk, the union of an open disk with its boundary circle, is a closed subset.

Instead of defining a topology on a set X to be a collection of open sets satisfying the three axioms we gave earlier, one could equally well consider the collection of complementary closed sets, and define a topology on X to be a collection of subsets called closed sets, such that the intersection of any collection of closed sets is closed, the union of any finite collection of closed sets is closed, and both the empty set and the whole set X are closed. Notice that the role of intersections and unions is switched compared with the original definition. This is because of the general set theory fact that the complement of a union is the intersection of the complements, and the complement of an intersection is the union of the complements.

Interior, Closure, and Boundary

Consider an open disk D in the plane $\mathbb{R}^2$, consisting of all the points inside a circle C. We would like to assign precise meanings to certain intuitive statements like the following:

- C is the boundary of the open disk D, and also of the closed disk $D \cup C$.
- D is the interior of the closed disk $D \cup C$, and $D \cup C$ is the closure of the open disk D.

The key distinction between points in the boundary of the disk and points in its interior is that for points in the boundary, every open set containing such a point also contains points inside the disk and points outside the disk, while each point in the interior of the disk lies in some open set entirely contained inside the disk.

With this observation in mind let us consider what happens in general. Given a subset A of a topological space X, then for each point $x \in X$ exactly one of the following three possibilities holds:

(1) There exists an open set O in X with $x \in O \subset A$.
(2) There exists an open set O in X with $x \in O \subset X - A$.
(3) Every open set O with $x \in O$ meets both A and $X - A$.

Points x such that (1) holds form a subset of A called the *interior* of A, written $\operatorname{int}(A)$. The points where (2) holds then form $\operatorname{int}(X - A)$. Points x where (3) holds form a set called the *boundary* or *frontier* of A, written ∂A. The points x where either (1) or (3) hold are the points x such that every open set O containing x meets A. Such points are called *limit points* of A, and the set of these limit points is called the *closure* of A, written $\overline{A}$. Note that $A \subset \overline{A}$, so we have $\operatorname{int}(A) \subset A \subset \overline{A} = \operatorname{int}(A) \cup \partial A$, this last union being a disjoint union. We will use the symbol $\amalg$ to denote union of disjoint subsets when we want to emphasize the disjointness, so $\overline{A} = \operatorname{int}(A) \amalg \partial A$ and $X = \operatorname{int}(A) \amalg \partial A \amalg \operatorname{int}(X - A)$.

As an example, in $\mathbb{R}$ with the usual topology the intervals (a,b), $[a,b]$, $[a,b)$, and $(a,b]$ all have interior (a,b), closure $[a,b]$ and boundary $\{a,b\}$. Similarly, in $\mathbb{R}^2$ with the usual topology, if A is the union of an open disk D with any subset of its boundary circle C then $\operatorname{int}(A) = D$, $\overline{A} = D \cup C$, and $\partial A = C$. For a somewhat different type of example, let $A = \mathbb{Q}$ in $X = \mathbb{R}$ with the usual topology on $\mathbb{R}$. Then $\operatorname{int}(A) = \varnothing$ and $\overline{A} = \partial A = \mathbb{R}$.

Proposition 1.1. *For every subset $A \subset X$ the following statements hold:*

(a) $\operatorname{int}(A)$ *is open.*
(b) $\overline{A}$ *is closed.*
(c) A *is open if and only if* $A = \operatorname{int}(A)$.
(d) A *is closed if and only if* $A = \overline{A}$.

Proof. (a) If x is a point in $\operatorname{int}(A)$ then there is an open set O_x with $x \in O_x \subset A$. We have $O_x \subset \operatorname{int}(A)$ since for each $y \in O_x$, O_x is an open set with $y \in O_x \subset A$ so $y \in \operatorname{int}(A)$. It follows that $\operatorname{int}(A) = \bigcup_x O_x$, the union as x ranges over all points of $\operatorname{int}(A)$. This is a union of open sets and hence open.

(b) Since $X = \text{int}(A) \amalg \partial A \amalg \text{int}(X - A)$, we have $\overline{A}$ as the complement of $\text{int}(X - A)$, so $\overline{A}$ is closed, being the complement of an open set by part (a).

(c) If $A = \text{int}(A)$ then A is open by (a). Conversely, suppose A is open. Then every $x \in A$ is in $\text{int}(A)$ since we can take $O = A$ in condition (1). Thus $A \subset \text{int}(A)$. The opposite inclusion $\text{int}(A) \subset A$ always holds, so $A = \text{int}(A)$.

(d) If $A = \overline{A}$ then A is closed by (b). Conversely, if A is closed then $X - A$ is open, so each point of $X - A$ is contained in an open set disjoint from A, namely the set $X - A$ itself. This means that no point of $X - A$ is a limit point of A, or in other words we have $\overline{A} \subset A$. We always have $A \subset \overline{A}$, so $A = \overline{A}$. □

A small caution: Some authors use the term ‘limit point’ in a more restricted sense than we are using it here, requiring that every open set containing x contains points of A *other than* x *itself*. Other names for this more restricted concept that one sometimes finds are ‘point of accumulation’ and ‘cluster point’.

One might have expected the definition of a limit point to be expressed in terms of convergent sequences of points. In an arbitrary topological space X it is natural to define $\lim_{n\to\infty} x_n = x$ to mean that for every open set O containing x the points x_n lie in O for all but finitely many values of n, or in other words there exists an $N > 0$ such that $x_n \in O$ for all $n > N$. It is obvious that $\lim_{n\to\infty} x_n = x$ implies that x is a limit point of the subset of X formed by the sequence of points x_n. However there exist topological spaces in which a limit point of a subset need not be the limit of any convergent sequence of points in the subset. For subspaces X of $\mathbb{R}^n$ this strange behavior does not occur since if x is a limit point of a subset $A \subset X$ then for each $n > 0$ there is a point $x_n \in A$ in the open ball of radius $1/n$ centered at x, and for this sequence of points x_n we have $\lim_{n\to\infty} x_n = x$.

Basis for a Topology

Many arguments with open sets in $\mathbb{R}$ reduce to looking at what happens with open intervals since open sets are defined in terms of open intervals. A similar statement holds for $\mathbb{R}^2$ and $\mathbb{R}^n$ with open disks and balls in place of open intervals. In each case arbitrary open sets are unions of the special open sets given by open intervals, disks, or balls. This idea is expressed by the following terminology:

Definition. A collection $\mathcal{B}$ of open sets in a topological space X is called a *basis* for the topology if every open set in X is a union of sets in $\mathcal{B}$.

A topological space can have many different bases. For example, in $\mathbb{R}^2$ another basis besides the basis of open disks is the basis of open squares with edges parallel to the coordinate axes. Or we could take open squares with edges at 45 degree angles to the coordinate axes, or all open squares without restriction. Many other shapes besides squares could also be used. Another variation would be to fix a number $c > 0$ and take just the open disks of radius less than c. These disks also form a basis for the topology on $\mathbb{R}^2$.

If $\mathcal{B}$ is a basis for a topology on X, then $\mathcal{B}$ satisfies the following two properties:

(1) Every point $x \in X$ lies in some set $B \in \mathcal{B}$.

(2) For each pair of sets B_1, B_2 in $\mathcal{B}$ and each point $x \in B_1 \cap B_2$ there exists a set B_3 in $\mathcal{B}$ with $x \in B_3 \subset B_1 \cap B_2$.

The first statement holds since X is open and is therefore a union of sets in $\mathcal{B}$. The second statement holds since $B_1 \cap B_2$ is open and hence is a union of sets in $\mathcal{B}$.

Proposition 1.2. *If $\mathcal{B}$ is a collection of subsets of a set X satisfying* (1) *and* (2) *then $\mathcal{B}$ is a basis for a topology on X.*

The open sets in this topology have to be exactly the unions of sets in $\mathcal{B}$ since $\mathcal{B}$ is a basis for this topology.

Proof. Let $\mathcal{O}$ be the collection of subsets of X that are unions of sets in $\mathcal{B}$. Obviously the union of any collection of sets in $\mathcal{O}$ is in $\mathcal{O}$. To show the corresponding result for finite intersections it suffices by induction to show that $O_1 \cap O_2 \in \mathcal{O}$ if $O_1, O_2 \in \mathcal{O}$. For each $x \in O_1 \cap O_2$ we can choose sets $B_1, B_2 \in \mathcal{B}$ with $x \in B_1 \subset O_1$ and $x \in B_2 \subset O_2$. By (2) there exists a set $B_3 \in \mathcal{B}$ with $x \in B_3 \subset B_1 \cap B_2 \subset O_1 \cap O_2$. The union of all such sets B_3 as x ranges over $O_1 \cap O_2$ is $O_1 \cap O_2$, so $O_1 \cap O_2 \in \mathcal{O}$.

Finally, X is in $\mathcal{O}$ by (1), and $\varnothing \in \mathcal{O}$ since we can regard $\varnothing$ as the union of the empty collection of subsets of $\mathcal{B}$. □

Terminology: Neighborhoods

We have frequently had to deal with open sets O containing a given point x. Such an open set is called a 'neighborhood' of x. Actually, it is useful to use the following broader definition:

Definition. A *neighborhood* of a point x in a topological space X is any set $A \subset X$ that contains an open set O containing x.

The more restricted kind of neighborhood can then be described as an open neighborhood.

As an example of the usefulness of this terminology, we can rephrase the condition for a point x to be a limit point of a set A to say that every open neighborhood of x meets A. The word 'open' here can in fact be omitted, for if every neighborhood of x meets A then in particular every open neighborhood meets A, and conversely, if every open neighborhood meets A then so does every other neighborhood since every neighborhood contains an open neighborhood.

Similarly, a boundary point of A is a point x such that every neighborhood of x meets both A and $X - A$.

Metric Spaces

The topology on $\mathbb{R}^n$ is defined in terms of open balls, which in turn are defined in terms of distance between points. There are many other spaces whose topology can be defined in a similar way in terms of a suitable notion of distance between points in the space. Here is the ingredient needed to do this:

Definition. A *metric* on a set X is a function $d : X \times X \to \mathbb{R}$ such that

(1) $d(x,y) \geq 0$ for all $x,y \in X$, with $d(x,x) = 0$ and $d(x,y) > 0$ if $x \neq y$.

(2) $d(x,y) = d(y,x)$ for all $x,y \in X$.

(3) $d(x,y) \leq d(x,z) + d(z,y)$ for all $x,y,z \in X$.

This last condition is called the 'triangle inequality' because for the usual distance function in the plane it says that length of one side of a triangle is always less than or equal to the sum of the lengths of the other two sides.

Given a metric on X one defines the open ball of radius r centered at x to be the set $B_r(x) = \{\, y \in X \mid d(x,y) < r \,\}$.

Proposition. *The collection of all balls $B_r(x)$ for $r > 0$ and $x \in X$ forms a basis for a topology on X.*

A topological space together with a metric that defines the topology in this way is called a *metric space*.

Proof. First a preliminary observation: For a point $y \in B_r(x)$ the ball $B_s(y)$ is contained in $B_r(x)$ if $s \leq r - d(x,y)$, since for $z \in B_s(y)$ we have $d(z,y) < s$ and hence $d(z,x) \leq d(z,y) + d(y,x) < s + d(x,y) \leq r$.

Now to show the condition to have a basis is satisfied, suppose we are given a point $y \in B_{r_1}(x_1) \cap B_{r_2}(x_2)$. Then the observation in the preceding paragraph implies that $B_s(y) \subset B_{r_1}(x_1) \cap B_{r_2}(x_2)$ for any $s \le \min\{r_1 - d(x_1,y), r_2 - d(x_2,y)\}$. □

Different metrics on the same set X can give rise to different bases for the same topology. For example, in $\mathbb{R}^2$ the usual metric defined in terms of lengths of vectors by $d(x,y) = |x - y|$ has 'balls' which are disks, but another metric whose 'balls' are squares is $d(x,y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}$ for $x = (x_1, y_1)$ and $y = (y_1, y_2)$. We will leave it to the reader to verify that this satisfies the properties of a metric. Another metric is $d(x,y) = |x_1 - y_1| + |x_2 - y_2|$, which has balls that are also squares, but rotated 45 degrees from the squares in the previous metric.

Many important spaces that arise in the subject of Topology are metric spaces, but the point of view of Topology is to ignore any particular choice of metric as much as possible, and just focus on the open sets, on the topology itself.

Subspaces

We turn now to a topic which ought to be simple, and seems simple enough at first glance, but turns out to be a source of many headaches until one finally becomes comfortable with it.

Given a topology $\mathcal{O}$ on a space X and a subset $A \subset X$, we would like to use the topology on X to define a topology $\mathcal{O}_A$ on A. There is an easy way to do this: Just define a set $O \subset A$ to be in $\mathcal{O}_A$ if there exists an open set O' in $\mathcal{O}$ such that $O = A \cap O'$. Axiom (1) holds since if $O_\alpha = A \cap O'_\alpha$ for sets $O'_\alpha \in \mathcal{O}$ then $\bigcup_\alpha O_\alpha = \bigcup_\alpha (A \cap O'_\alpha) = A \cap (\bigcup_\alpha O'_\alpha)$, so $\bigcup_\alpha O_\alpha$ is in $\mathcal{O}_A$. Axiom (2) is similar since $\bigcap_\alpha O_\alpha = \bigcap_\alpha (A \cap O'_\alpha) = A \cap (\bigcap_\alpha O'_\alpha)$, which is in $\mathcal{O}_A$ if there are just finitely many indices α. Axiom (3) is obvious.

The topology $\mathcal{O}_A$ on A is called the *subspace topology*, and A with this topology is called a *subspace* of X. For example, if we take X to be $\mathbb{R}^2$ with its usual topology, then every subset of $\mathbb{R}^2$ becomes a topological space. In particular, geometric figures such as circles and polygons can now be viewed as topological spaces. Likewise, geometric figures in $\mathbb{R}^3$ such as spheres and polyhedra become topological spaces, with the subspace topology from the usual topology on $\mathbb{R}^3$.

If $\mathcal{B}$ is a basis for the topology on X and A is a subspace of X, then we can obtain a basis for the subspace topology on A by taking the collection $\mathcal{B}_A$ of all intersections $A \cap B$ as B ranges over all the sets in $\mathcal{B}$. This gives a basis for A because an arbitrary open set in the subspace topology on A has the form $A \cap (\bigcup_\alpha B_\alpha) = \bigcup_\alpha (A \cap B_\alpha)$ for some collection of basis sets $B_\alpha \in \mathcal{B}$. In particular this says that for any subspace X of $\mathbb{R}^n$, a basis for the topology on X is the collection of open sets $X \cap B$ as B ranges over all open balls in $\mathbb{R}^n$. For example, for a circle in $\mathbb{R}^2$ the open arcs in the circle form a basis for its topology.

If X is a metric space, any subset $A \subset X$ becomes a metric space by restricting the metric $X \times X \to \mathbb{R}$ to $A \times A$, since the three defining properties of a metric obviously still hold for the restricted distance function. The following Proposition gives some strong evidence that the subspace topology is a natural topology to use on subsets.

Proposition. *The metric topology on a subset A of a metric space X is the same as the subspace topology.*

Proof. Observe first that for a ball $B_r(x)$ in X, the intersection $A \cap B_r(x)$ consists of all points in A of distance less than r from x, so this is a ball in A regarded as a metric space in itself. For a collection of such balls $B_{r_\alpha}(x_\alpha)$ we have

$$A \cap (\bigcup_\alpha B_{r_\alpha}(x_\alpha)) = \bigcup_\alpha (A \cap B_{r_\alpha}(x_\alpha))$$

The left side of this equation is a typical open set in A with the subspace topology, and the right side is a typical open set in the metric topology, so the two topologies coincide. □

A subspace $A \subset X$ whose subspace topology is the discrete topology is called a *discrete subspace* of X. This is equivalent to saying that for each point $x \in A$ there is an open set in X whose intersection with A is just x. For example, $\mathbb{Z}$ is a discrete subspace of $\mathbb{R}$, but $\mathbb{Q}$ is not discrete. The sequence $1/2, 1/3, 1/4 \cdots$ without its limit 0 is a discrete subspace of $\mathbb{R}$, but with 0 it is not discrete.

For a subspace $A \subset X$, a subset of A which is open or closed in A need not be open or closed in X. However, there are times when this is true:

Lemma. *For an open set $A \subset X$, a subset $B \subset A$ is open in the subspace topology on A if and only if B is open in X. This is also true when 'open' is replaced by 'closed' throughout the statement.*

Proof. If $B \subset A$ is open in A, it has the form $A \cap O$ for some open set O in X. This intersection is open in X if A is open in X. Conversely, if $B \subset A$ is open in X then $A \cap B = B$ is open in A. (Note that the converse does not use the assumption that A is open in X.) The argument for closed sets is just the same. □

Closures behave nicely with respect to subspaces:

Lemma. *Given a space X, a subspace Y, and a subset $A \subset Y$, then the closure of A in the space Y is the intersection of the closure of A in X with Y.*

This amounts to saying that a point $y \in Y$ is a limit point of A in Y (i.e., using the subspace topology on Y) if and only if y is a limit point of A in X.

Proof. For a point $y \in Y$ to be a limit point of A in X means that every open set O in X that contains y meets A. Since $A \subset Y$, this is equivalent to $O \cap Y$ meeting A, or in other words, that every open set in Y containing y meets A. □

The analogous statement for interiors is not true. For example, if A is a line segment in the x-axis in $\mathbb{R}^2$, then the interior of A in the x-axis is an open interval, but the interior of A in $\mathbb{R}^2$ is empty.

Continuity and Homeomorphisms

Recall the definition: A function $f : X \to Y$ between topological spaces is continuous if $f^{-1}(O)$ is open in X for each open set O in Y. For brevity, continuous functions are sometimes called *maps* or *mappings*. (A map in the everyday sense of the word is in fact a function from the points on the map to the points in whatever region is being represented by the map.)

Lemma. *A function $f : X \to Y$ is continuous if and only if $f^{-1}(C)$ is closed in X for each closed set C in Y.*

Proof. An evident set-theory fact is that $f^{-1}(Y - A) = X - f^{-1}(A)$ for each subset A of Y. Suppose now that f is continuous. Then for any closed set $C \subset Y$, we have $Y - C$ open, hence the inverse image $f^{-1}(Y - C) = X - f^{-1}(C)$ is open in X, so its complement $f^{-1}(C)$ is closed. Conversely, if the inverse image of every closed set is closed, then for O open in Y the complement $Y - O$ is closed so $f^{-1}(Y - O) = X - f^{-1}(O)$ is closed and thus $f^{-1}(O)$ is open, so f is continuous. □

Here is another useful fact:

Lemma. *Given a function $f : X \to Y$ and a basis $\mathcal{B}$ for Y, then f is continuous if and only if $f^{-1}(B)$ is open in X for each $B \in \mathcal{B}$.*

Proof. One direction is obvious since the sets in $\mathcal{B}$ are open. In the other direction, suppose $f^{-1}(B)$ is open for each $B \in \mathcal{B}$. Then any open set O in Y is a union $\bigcup_\alpha B_\alpha$

of basis sets B_α, hence $f^{-1}(O) = f^{-1}(\bigcup_\alpha B_\alpha) = \bigcup_\alpha f^{-1}(B_\alpha)$ is open in X, being a union of the open sets $f^{-1}(B_\alpha)$. □

Lemma. *If $f:X\to Y$ and $g:Y\to Z$ are continuous, then their composition $gf:X\to Z$ is also continuous.*

Proof. This uses the easy set-theory fact that $(gf)^{-1}(A) = f^{-1}(g^{-1}(A))$ for any $A\subset Z$. Thus if f and g are continuous and A is open in Z then $g^{-1}(A)$ is open in Y so $f^{-1}(g^{-1}(A))$ is open in X. This means gf is continuous. □

Lemma. *If $f:X\to Y$ is continuous and A is a subspace of X, then the restriction $f|_A$ of f to A is continuous as a function $A\to Y$.*

Proof. For an open set $O\subset Y$ we have $(f|_A)^{-1}(O) = f^{-1}(O)\cap A$, which is an open set in A since $f^{-1}(O)$ is open in X. □

Definition. A continuous map $f:X\to Y$ is a *homeomorphism* if it is one-to-one and onto, and its inverse function $f^{-1}:Y\to X$ is also continuous.

[To be added: some examples of homeomorphisms, e.g., an open interval (a,b) is homeomorphic to $\mathbb{R}$, an open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$.]

Product Spaces

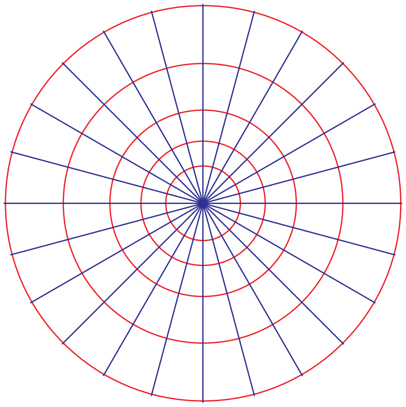
Given two sets X and Y, their product is the set $X\times Y = \{\,(x,y)|x\in X \text{ and } y\in Y\,\}$. For example $\mathbb{R}^2 = \mathbb{R}\times\mathbb{R}$, and more generally $\mathbb{R}^m\times\mathbb{R}^n = \mathbb{R}^{m+n}$. If X and Y are topological spaces, we can define a topology on $X\times Y$ by saying that a basis consists of the subsets $U\times V$ as U ranges over open sets in X and V ranges over open sets in Y. The criterion for a collection of subsets to be a basis for a topology is satisfied since $(U_1\times V_1)\cap(U_2\times V_2) = (U_1\cap U_2)\times(V_1\cap V_2)$. This is called the *product topology* on $X\times Y$. The same topology could also be produced by taking the smaller basis consisting of products $U\times V$ where U ranges over a basis for the topology on X and V ranges over a basis for the topology on Y. This is because $(\bigcup_\alpha U_\alpha)\times(\bigcup_\beta V_\beta) = \bigcup_{\alpha,\beta}(U_\alpha\times V_\beta)$.

For example, a basis for the product topology on $\mathbb{R}\times\mathbb{R}$ consists of the open rectangles $(a_1,b_1)\times(a_2,b_2)$. This is also a basis for the usual topology on $\mathbb{R}^2$, so the product topology coincides with the usual topology.

More generally one can define the product $X_1\times\cdots\times X_n$ to consist of all ordered n-tuples $(x_1,\cdots,x_n)$ with $x_i\in X_i$ for each i. A basis for the product topology on

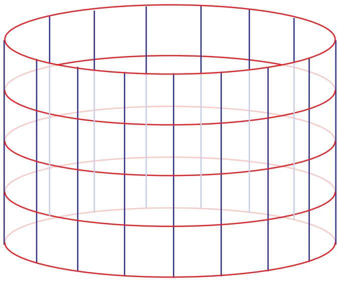
$X_1 \times \cdots \times X_n$ consists of all products $U_1 \times \cdots \times U_n$ as each U_i ranges over open sets in X_i, or just over a basis for the topology on X_i. Thus $\mathbb{R}^n$ with its usual topology is also describable as the product of n copies of $\mathbb{R}$, with basis the open 'boxes' $(a_1, b_1) \times \cdots \times (a_n, b_n)$.

Example. If we view points in the unit circle S^1 in $\mathbb{R}^2$ as angles θ, then polar coordinates give a homeomorphism $f : S^1 \times (0,\infty) \to \mathbb{R}^2 - \{0\}$ defined by $f(\theta, r) = (r\cos\theta, r\sin\theta)$. This is one-to-one and onto since each point in $\mathbb{R}^2$ other than the origin has unique polar coordinates (θ, r). To see that f is a homeomorphism, just observe that it takes a basis set $U \times V$, where U is an open interval (θ_0, θ_1) of θ values and V is an open interval (r_0, r_1) of r values, to an open 'polar rectangle' and such rectangles form a basis for the topology on $\mathbb{R}^2 - \{0\}$ as a subspace of $\mathbb{R}^2$. By restricting f to a product $S^1 \times [a, b]$ for $0 < a < b$ we obtain a homeomorphism from this product to a closed annulus in $\mathbb{R}^2$, the region between two concentric circles.

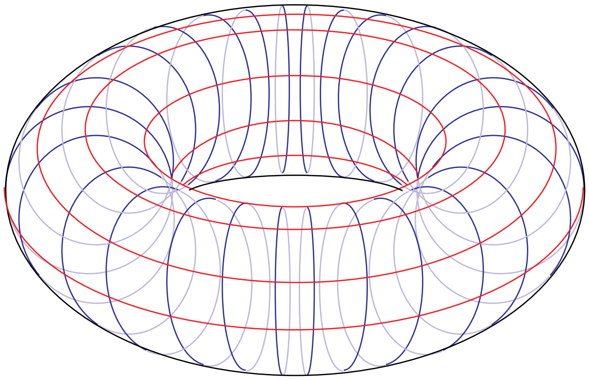


More generally, $\mathbb{R}^n - \{0\}$ is homeomorphic to $S^{n-1} \times (0,\infty)$ where S^{n-1} is the unit sphere in $\mathbb{R}^n$. Using vector notation, a homeomorphism $f : S^{n-1} \times (0,\infty) \to \mathbb{R}^n - \{0\}$ is given by $f(v, r) = rv$, with inverse $f^{-1}(v) = (v/|v|, |v|)$. The continuity of f and f^{-1} can be deduced from the explicit algebraic formulas for them, as we will see later in this section.

Example. A product $S^1 \times [a, b]$ is homeomorphic to a cylinder as well as to an annulus. If we use cylindrical coordinates (r, θ, z) in $\mathbb{R}^3$ then a cylinder is specified by taking r to be a constant r_0, letting θ range over the circle S^1, and restricting z to an interval $[a, b]$.



Example. The product $S^1 \times S^1$ is homeomorphic to a torus, say the torus T in $\mathbb{R}^3$ obtained by taking a circle C in the yz-plane disjoint from the z-axis and rotating this circle about the z-axis. We can parametrize points on T by a pair of angles (θ_1, θ_2) where θ_1 is the angle through which the yz-plane has been rotated and θ_2 is the angle between the horizontal radial vector of C pointing away from the z-axis and the radial vector to a given point of C. One can think of θ_1 and θ_2 as longitude and latitude on T. A basic open set $U \times V$ in $S^1 \times S^1$ is a product of two open arcs, and this corresponds to an open curvilinear rectangle on T. Such rectangles form a basis for the topology on T as a subspace of $\mathbb{R}^3$, so it follows that T is homeomorphic to $S^1 \times S^1$.



A product space $X \times Y$ has two projection maps $p_1 : X \times Y \to X$ and $p_2 : X \times Y \to Y$ defined by $p_1(x,y) = x$ and $p_2(x,y) = y$. These maps are continuous since if $U \subset X$ is open then so is $p_1^{-1}(U) = U \times Y$, and if $V \subset Y$ is open then so is $p_2^{-1}(V) = X \times V$.

For each $y \in Y$ there is an inclusion map $i_y : X \to X \times Y$ given by $i_y(x) = (x,y)$. This is continuous because $i_y^{-1}(U \times V)$ is either U if $y \in V$ or $\varnothing$ if $y \notin V$. The map i_y is a homeomorphism onto its image $X \times \{y\}$ since it has a continuous inverse, the restriction of the projection p_1 to $X \times \{y\}$. One can think of $X \times Y$ as the union of the family of subspaces $X \times \{y\}$, each homeomorphic to X, with one such subspace for each $y \in Y$. The situation is of course symmetric with respect to interchanging X and Y, so for each $x \in X$ there is a continuous inclusion map $i_x : Y \to X \times Y$ which is a homeomorphism onto its image $\{x\} \times Y$, and $X \times Y$ is the union of these copies of the space X, one for each point of Y.

A function $f : Z \to X \times Y$ has the form $f(z) = (f_1(z), f_2(z))$. A basic property of the product topology on $X \times Y$ is:

Proposition. *A function $f : Z \to X \times Y$ is continuous if and only if its component functions $f_1 : Z \to X$ and $f_2 : Z \to Y$ are both continuous.*

Proof. We have $f_1 = p_1 f$ and $f_2 = p_2 f$ so f_1 and f_2 are continuous if f is continuous. For the converse, note that $f^{-1}(U \times V) = f_1^{-1}(U) \cap f_2^{-1}(V)$, so this will be open if U and V are open and f_1 and f_2 are continuous. □

As an application, we can give a topological proof that a function $\mathbb{R}^n \to \mathbb{R}$ given by a polynomial in n variables is continuous. The first step is the following fact:

- If two functions $f, g : \mathbb{R}^n \to \mathbb{R}$ are continuous, then so also are the sum function $f+g$ and the product function $f \cdot g$. Namely, we can view $f+g$ as the composition $\mathbb{R}^n \to \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ where the first map is $x \mapsto (f(x), g(x))$ and the second map is $(x, y) \mapsto x + y$. We know the first map is continuous if f and g are continuous, and it is easy to check that the second map is continuous by seeing directly that the inverse image of an open interval is open. For $f \cdot g$ the argument is similar, replacing the second map by the product map $(x, y) \mapsto xy$.

A general polynomial in n variables is built up using repeated addition and multiplication from constant functions, which are certainly continuous, and the coordinate functions x_i, which are nothing but the projections of $\mathbb{R}^n$ onto its n $\mathbb{R}$ factors so are continuous as well.

Exercises

1. Show that every open set in $\mathbb{R}$ is the union of a collection of *disjoint* open intervals (a, b) where we allow $a = -\infty$ and $b = \infty$.

2. For a subset A of a topological space X show:

(a) Every open set contained in A is contained in $\text{int}(A)$. Thus $\text{int}(A)$ is the largest open set contained in A.

(b) Every closed set containing A contains $\overline{A}$. Thus $\overline{A}$ is the smallest closed set containing A.

3. Let $\mathcal{O}$ be the collection of all intervals $I_a = (a, \infty)$ in $\mathbb{R}$, including the cases $I_\infty = \varnothing$ and $I_{-\infty} = \mathbb{R}$. Show that $\mathcal{O}$ defines a topology on $\mathbb{R}$. In this topology, what is the closure of a set $A \subset \mathbb{R}$?

4. Show that if A is a subset of a topological space X then:

(a) $\overline{X - A} = X - \text{int}(A)$.

(b) $\text{int}(X - A) = X - \overline{A}$.

5. Verify each of the following for arbitrary subsets A, B of a topological space X:

(a) $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

(b) $\overline{A \cap B} \subset \overline{A} \cap \overline{B}$.

(c) $\text{int}(A \cap B) = \text{int}(A) \cap \text{int}(B)$.

(d) $\text{int}(A \cup B) \supset \text{int}(A) \cup \text{int}(B)$.

Give examples where equality fails to hold in (b) and (d).

6. Show that ∂A always contains $\partial(\text{int}(A))$. How does $\partial(A \cup B)$ relate to ∂A and ∂B?

7. If Y is a subspace of X and Z is a subspace of Y, show that Z is a subspace of X.

8. For a subspace A of $\mathbb{R}^2$, show that a set $O \subset A$ is open in the subspace topology if and only if for each $x \in O$ there exists an $\varepsilon > 0$ such that all points of A of distance less than ε from x lie in O.

9. Let Y be a subspace of X and let A be a subset of Y. Denote by $\text{int}_X(A)$ the interior of A regarded as a subset of X and by $\text{int}_Y(A)$ the interior of A regarded as a subset of Y. Show that $\text{int}_X(A) \subset \text{int}_Y(A)$ and give an example where equality does not hold.

10. For Y a subspace of X show that if a set $A \subset Y$ is open in Y (with the subspace topology on Y) and Y is open in X then A is open in X. Do the same with 'open' replaced by 'closed'.

11. For a function $f:X \to Y$ the image $f(X) = \{\, f(x) \mid x \in X \,\}$ is a subspace of Y. Show that $f:X \to Y$ is continuous if and only if $f:X \to f(X)$ is continuous.

12. A map $f:X \to Y$ is said to be open if $f(O)$ is open in Y whenever O is open in X. Similarly, $f:X \to Y$ is said to be closed if $f(C)$ is closed in Y whenever C is closed in X. (a) Give an example of a map that is open but not closed, and an example of a map that is closed but not open. (b) Determine whether the projection map $\mathbb{R}^2 \to \mathbb{R}$ sending (x,y) to x is open or closed. (c) Do the same for the map $f:\mathbb{R} \to S^1$, $f(x) = (\cos x, \sin x)$, where S^1 is the unit circle $x^2 + y^2 = 1$ in $\mathbb{R}^2$.

13. Show the two maps $\mathbb{R}^2 \to \mathbb{R}$ sending (x,y) to $x+y$ and xy are continuous, using only definitions and results from this class, not results from calculus for example.

14. Suppose a space X is the union of a collection of open subsets O_α. Show that a map $f:X \to Y$ is continuous if its restriction to each subspace O_α is continuous.

15. Let $\mathbb{R}_h$ denote $\mathbb{R}$ with the 'half-open interval topology' having as basis the intervals $[a,b)$. (a) For a subset $A \subset \mathbb{R}_h$ show that a point x lies in the closure of A if and only if there is a sequence $\{x_n\}$ in A such that $x_n \geq x$ and $|x_n - x| \to 0$. (b) Show that a function $f:\mathbb{R}_h \to \mathbb{R}$ (with the usual topology on $\mathbb{R}$) is continuous if and only if it is continuous from the right at each point x, that is, $\lim f(x+\varepsilon) = f(x)$ where the limit is as $\varepsilon \to 0$ with $\varepsilon > 0$.

Chapter 2. Connectedness

Some spaces are in a sense ‘disconnected’, being the union of two or more completely separate subspaces. For example the space $X \subset \mathbb{R}$ consisting of the two intervals $A = [0,1]$ and $B = [2,3]$ should certainly be disconnected, and so should a subspace X of $\mathbb{R}^2$ which is the union of two disjoint circles A and B. As these examples show, it is reasonable to interpret the idea of A and B being ‘completely separate’ as saying not only that they are disjoint, but no point of A is a limit point of B and no point of B is a limit point of A. Since we are assuming that X is the union of A and B, this is equivalent to saying that A and B each contain all their limit points. In other words, A and B are both closed subsets of X. Since each of A and B is the complement of the other, it would be equivalent to say that both A and B are open sets. Thus we have arrived at the following basic definition:

Definition. A space X is *connected* if it cannot be decomposed as the union of two disjoint nonempty open sets. This is equivalent to saying X cannot be decomposed as the union of two disjoint nonempty closed sets.

A third equivalent condition is that the only sets in X that are both open and closed are $\varnothing$ and X itself. For if A were any other set that was both open and closed, then X would be decomposed as the union of the disjoint nonempty open sets A and $X - A$. Conversely, if X were the disjoint union of the nonempty open sets A and B then A would be closed as well as open, being equal to the complement of B, and A would be neither $\varnothing$ nor X.

Example. The subspace of $\mathbb{R}$ consisting of the rational numbers $\mathbb{Q}$ is not connected, since we can decompose it as the union of the two open sets $\mathbb{Q} \cap (-\infty, \sqrt{2})$ and $\mathbb{Q} \cap (\sqrt{2}, \infty)$. More generally, any subspace $X \subset \mathbb{R}$ that is not an interval is not connected. For if X is not an interval, then there exist numbers $a < c < b$ with $a, b \in X$ but $c \notin X$, so X is the disjoint union of the open sets $X \cap (-\infty, c)$ and $X \cap (c, \infty)$. We will see below that intervals in $\mathbb{R}$ are in fact connected.

Example. If we give $\mathbb{R}$ the topology having as basis all the intervals $[a, b)$, then with this topology $\mathbb{R}$ is not connected, because the intervals $[a, b)$ are closed as well as open since their complements $(-\infty, a) \cup [b, \infty)$ are open, being the unions of the basis intervals $[a - n, a)$ and $[b, b + n)$ for $n = 1, 2, \cdots$.

Now we return to the usual topology on $\mathbb{R}$ to prove an extremely fundamental result:

Theorem. *An interval $[a, b]$ in $\mathbb{R}$ is connected.*

Proof. We may assume $a < b$ since it is obvious that $[a,a]$ is connected. Suppose $[a,b]$ is decomposed as the disjoint union of sets A and B that are open in $[a,b]$ with the subspace topology, so they are also closed in $[a,b]$. After possibly changing notation we may assume that a is in A. Since A is open in $[a,b]$ there is an interval $[a,a+\varepsilon)$ contained in A for some $\varepsilon > 0$, and hence there is an interval $[a,c] \subset A$, with $a < c$. The set $C = \{c \mid [a,c] \subset A\}$ is bounded above by b, so it has a least upper bound $L \le b$. (A fundamental property of $\mathbb{R}$ is that any set that is bounded above has a least upper bound.) We know that $L > a$ by the earlier observation that there is an interval $[a,c] \subset A$ with $c > a$. Since no number smaller than L is an upper bound for C, there exist intervals $[a,c] \subset A$ with $c \le L$ and c arbitrarily close to L. These numbers c are in A, hence L must also be in A since A is closed. Thus we have $[a,L] \subset A$. Now if we assume that $L < b$ we can derive a contradiction in the following way. Since A is open and contains $[a,L]$, it follows that A contains $[a,L+\varepsilon]$ for some $\varepsilon > 0$. But this means that C contains numbers bigger than L, contradicting the fact that L was an upper bound for C. Thus the assumption $L < b$ leads to a contradiction, so we must conclude that $L = b$ since we know $L \le b$. We already saw that $[a,L] \subset A$, so now we have $[a,b] \subset A$. This means that B must be empty, and we have shown that it is impossible to decompose $[a,b]$ into two disjoint nonempty open sets. Hence $[a,b]$ is connected. □

Path-connected Spaces

Here is another kind of connectedness that is often easier to deal with:

Definition. A space X is *path-connected* if for each pair of points $a,b \in X$ there exists a path in X from a to b, that is, a continuous map $f:[0,1] \to X$ with $f(0) = a$ and $f(1) = b$.

The choice of the interval $[0,1]$ rather than some other closed interval as the domain of f is not of much importance. All closed intervals are homeomorphic, so it is easy to change from one domain interval to another.

Often it is easy to see ‘by inspection’ that a given space is path-connected. This then tells us more:

Proposition. *If a space is path-connected, then it is connected.*

Proof. To argue by contradiction, suppose X is the disjoint union of nonempty open sets A and B. Choose a point $a \in A$ and a point $b \in B$. If X is path-connected there is a path $f:[0,1]\to X$ from a to b. Then $f^{-1}(A)$ and $f^{-1}(B)$ are nonempty disjoint open sets in $[0,1]$ whose union is all of $[0,1]$, contradicting the fact that $[0,1]$ is connected, proved in the previous theorem. □

This implies for example that all intervals, not just closed intervals, are connected since it is obvious that they are path-connected. Likewise circles are path-connected and hence connected. Another example is $\mathbb{R}^n$, where any two points a and b can be joined by the path obtained by parametrizing the straight line segment joining them, say by the function $f(t) = (1-t)a + tb$, in vector notation. The same construction shows that a subspace $X \subset \mathbb{R}^n$ is path-connected if it is *convex*, meaning that the line segment joining any two points in X also lies in X. For example, the union of an open ball in $\mathbb{R}^n$ with any subset of its boundary sphere is convex and hence path-connected.

We will give an example below of a connected space that is not path-connected, so the converse of the preceding proposition fails to hold in general.

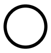
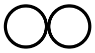
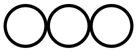
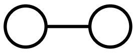
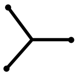
Proposition. *Suppose $f:X\to Y$ is continuous and onto. Then:*
(a) *If X is connected, so is Y.*
(b) *If X is path-connected, so is Y.*

Proof. (a) Suppose Y is the union of disjoint nonempty open sets A and B. Then X is the union of the disjoint open sets $f^{-1}(A)$ and $f^{-1}(B)$, which are both nonempty if f is onto.
(b) For any points $a, b \in Y$ there exist points $a', b' \in X$ with $f(a') = a$ and $f(b') = b$ since f is onto. If X is path-connected there is a path $g:[0,1]\to X$ from a' to b'. Then the composition $fg:[0,1]\to Y$ is a path in Y from a to b, so Y is path-connected. □

A consequence of this proposition is that if two spaces are homeomorphic and one is connected, then so is the other, and the same holds also with ‘path-connected’ in place of ‘connected’. This gives a way of showing that two spaces are not homeomorphic, if one is connected or path-connected and the other is not.

Cut Points

There is a more refined way to apply this fact that connectedness is preserved by homeomorphisms, using the following idea. In a connected space X, a point $x \in X$ is called a *cut point* if removing x from X produces a disconnected space $X - \{x\}$. Note that if $f:X\to Y$ is a homeomorphism, then a point $x \in X$ is a cut point of X if and only if $f(x)$ is a cut point of Y, since $X - \{x\}$ and $Y - \{f(x)\}$ are homeomorphic via the restriction of f. Thus by counting numbers of cut points and non-cut points we can sometimes show that two spaces are not homeomorphic. Here are some examples.

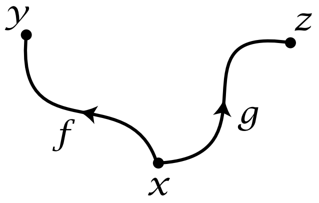
	♯ cut points	*♯ non-cut points*
closed interval	∞	2
open interval	∞	0
half-open interval	∞	1
circle 	0	∞
	1	∞
	2	∞
	∞	∞
	∞	3

Cut points can also be used to show that $\mathbb{R}$ is not homeomorphic to $\mathbb{R}^n$ for $n > 1$, since the latter spaces have no cut points whereas in $\mathbb{R}$ every point is a cut point. It is true more generally that $\mathbb{R}^m$ is not homeomorphic to $\mathbb{R}^n$ if $m \neq n$, but this is a much harder theorem. To distinguish $\mathbb{R}^2$ from $\mathbb{R}^3$ for example, one might try to use ‘cut curves’ instead of cut points, motivated by the fact that the complement of a line in $\mathbb{R}^2$ is disconnected whereas the complement of a line in $\mathbb{R}^3$ is connected. However, a hypothetical homeomorphism from $\mathbb{R}^2$ to $\mathbb{R}^3$ might take a line to a very complicated curve which might conceivably wander around so badly that its complement was not connected, unlike the complement of a straight line. To prove that this can’t happen takes quite a bit of work, and the best way to do it is to develop some heavy machinery first, the machinery in a branch of topology called algebraic topology.

Connected Components and Path Components

We will show that if a space X is not connected, then it decomposes as the union of a collection of disjoint connected subspaces that are maximal in the sense that none of them is contained in any larger connected subspace. These maximal connected subspaces are called the connected components of X. Similarly, if X is not path-connected, it decomposes as the union of disjoint maximal path-connected subspaces called path components.

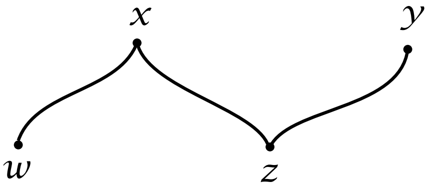
Let us begin by talking about path components. For a point $x \in X$ let $P(x)$ be the subspace of X consisting of points y such that there is a path in X from x to y. This path in fact lies in $P(x)$ since any point along the path is joined to x by an initial segment of the path. Notice also that any two points y and z in $P(x)$ can be joined by a path in $P(x)$ since if f is a path from x to y and g is a path from x to z then we obtain a path from y to z by first going backward along f from y to x and then forward along g from x to z.



Here is the key fact about the subspaces $P(x)$:

Lemma. *If* $P(x) \cap P(y) \neq \emptyset$ *then* $P(x) = P(y)$*.*

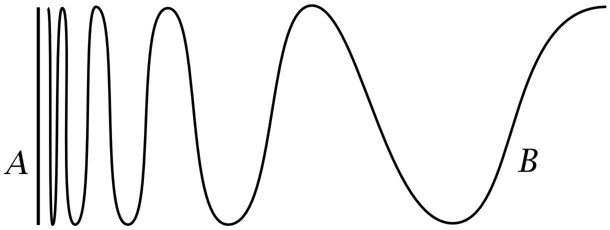
Proof. Suppose there is a point $z \in P(x) \cap P(y)$. This means there are paths from x and y to z. If w is any point in $P(x)$, then there is a path from x to w, and hence also a path from y to w by going first from y to z, then to x, then to w. Thus $P(x) \subset P(y)$. The same argument shows that $P(y) \subset P(x)$, so $P(x) = P(y)$. □



The lemma implies that the subspaces $P(x)$ that are distinct provide a decomposition of X into a collection of disjoint subspaces. Each $P(x)$ is path-connected, and if A is any path-connected subspace of X then A must be contained in some $P(x)$ since if $x \in A$ then all points in A can be connected to x by a path, so $A \subset P(x)$. This shows that the subspaces $P(x)$ are the maximal path-connected subspaces of X. They are called the *path components* of X.

Note that a continuous map $f : X \to Y$ takes each path component of X into some path component of Y. This is because a path component $P(x)$ is path-connected so its image $f(P(x))$ must also be path-connected, hence $f(P(x))$ must be contained in some path component of Y. In fact it will be contained in $P(f(x))$ since it contains the point $f(x)$.

Example. Let X be the subspace of $\mathbb{R}^2$ which is the closure of the graph of the function $f(x) = \sin(1/x)$ for $x > 0$. Thus X consists of this graph together with the line segment A in the y-axis



from $(0,-1)$ to $(0,1)$. Let B denote the graph of f itself, so X is the disjoint union of A and B. Obviously A and B are both path-connected. We will show that X itself is not path-connected, hence A and B are the two path components of X

Let $f:[0,1]\to X$ be a path starting at a point in A. We know that $f^{-1}(A)$ is closed since A is closed. We will show that $f^{-1}(A)$ is also open. This implies $f^{-1}(A)=[0,1]$ since $[0,1]$ is connected and $f^{-1}(A)$ is nonempty. The equation $f^{-1}(A)=[0,1]$ says that $f([0,1])\subset A$, and thus there is no path in X joining a point in A to a point in B. To show that $f^{-1}(A)$ is open let t_0 be any point in $f^{-1}(A)$. Choose a small open disk D in $\mathbb{R}^2$ centered at $f(t_0)$. Then $D\cap X$ has infinitely many path-components, one of which is $D\cap A$. Since f is continuous, $f^{-1}(D)$ is an open set in $[0,1]$ containing t_0, so there is an interval $I\subset[0,1]$ which is open in $[0,1]$, contains t_0, and is contained in $f^{-1}(D)$. This interval is path-connected, so its image $f(I)$ is also path-connected and therefore lies inside one of the path components of $D\cap X$. This path component contains the point $f(t_0)\in A$ so it has to be the path component $D\cap A$. This says that $f(I)\subset A$, or in other words $I\subset f^{-1}(A)$. Thus $f^{-1}(A)$ contains a neighborhood of t_0 in $[0,1]$. Since t_0 was an arbitrary point of $f^{-1}(A)$ we conclude that $f^{-1}(A)$ is open in $[0,1]$, finishing the argument.

To determine whether the space X in this example is connected we will use the following general fact:

Lemma. *If a subspace A of a space X is connected, then so is $\overline{A}$.*

Proof. Suppose $\overline{A}$ is the disjoint union of subsets B and C which are closed in $\overline{A}$ and hence also closed in X since $\overline{A}$ is closed in X. Then $A\cap B$ and $A\cap C$ are disjoint closed sets in A whose union is A, so one of these two sets must equal A, say $A\cap B=A$. This says $A\subset B$, so $\overline{A}\subset\overline{B}=B$. Since we originally had $\overline{A}=B\amalg C$ this implies $B=\overline{A}$ and $C=\varnothing$. □

From this lemma we can conclude that the space X in the preceding example is connected since it is the closure of the subspace B which is path-connected and hence connected. Thus X is an example of a space which is connected but not path-connected. Another example, very similar to this one, would be the closure of the graph of the function $r=\theta/(\theta+1)$ for $\theta\geq 0$, in polar coordinates. This space consists of a circle together with a curve that spirals out to it from inside.

Now we turn to the question of decomposing a space X into maximal connected subspaces.

Lemma. *If a subset A of a space X is both open and closed, then any connected subspace $C \subset X$ which meets A must be contained in A.*

Proof. If A is open and closed in X then $C \cap A$ is open and closed in C. If C is connected this implies that $C \cap A = C$, which says that C is contained in A. □

Proposition. *If $\{C_\alpha\}$ is a family of connected subspaces of a space X, any two of which have nonempty intersection, then $\bigcup_\alpha C_\alpha$ is connected.*

Proof. Let $Y = \bigcup_\alpha C_\alpha$ and let $A \subset Y$ be open and closed in Y. Then $A \cap C_\alpha$ is open and closed in C_α for each α, hence is either $\emptyset$ or C_α. If $A \neq \emptyset$ choose a point $x \in A$. Since $A \subset Y$ and Y is the union of all the C_α's we have $x \in C_\alpha$ for some α. Thus $A \cap C_\alpha \neq \emptyset$, so by the preceding lemma we must have $C_\alpha \subset A$. Any other C_β meets C_α by assumption, hence meets A and so is contained in A by the Lemma again. Thus Y, the union of the C_α's, is contained in A. We were assuming $A \subset Y$, so we have $A = Y$. Since A was any nonempty closed and open set in Y, we conclude that Y is connected. □

We can apply the proposition to the collection of all connected subsets C_α of X that contain a given point x, to conclude that the union of all these subsets is connected. Call this union $C(x)$. It is obviously the largest connected set in X containing x since it contains every connected set that contains x. Obviously $C(x)$ is also the largest connected set containing any other point $y \in C(x)$, since a larger connected set containing y would also be a larger connected set containing x. It follows that X is the disjoint union of all the different sets $C(x)$. These are the maximal connected sets in X, the *connected components* of X.

Note that the connected components of X are closed subsets by the earlier fact that if a subspace $A \subset X$ is connected then so is its closure. Connected components do not have to be open, however. For example, consider the rational numbers $\mathbb{Q}$, topologized as a subspace of $\mathbb{R}$. We know from earlier that the only connected subspaces of $\mathbb{R}$ are intervals, including the degenerate intervals $[a,a]$. Since $\mathbb{Q}$ contains no nondegenerate intervals, it follows that the only connected subspaces of $\mathbb{Q}$ are points. Thus the connected components of $\mathbb{Q}$ are just the individual points of $\mathbb{Q}$, since connected components are maximal connected subspaces. The connected components of $\mathbb{Q}$ are therefore not open subsets of $\mathbb{Q}$.

Path components do not have to be either open or closed, as one can see in the example of the closure of the graph of $\sin(1/x)$. This graph is a path component that is open but not closed, while the remaining line segment of the space is a path component that is closed but not open.

Each path component of a space is contained in a connected component since path-connected subspaces are connected, hence are contained in maximal connected subspaces, the connected components. One commonly encountered situation where path components are the same as connected components is given by the following result:

Proposition. *If each point in a space X has a neighborhood which is path-connected, then the path components of X are also the connected components.*

Proof. The hypothesis guarantees that path components are open subsets. This implies that they are also closed since the complement of a path component is a union of path components and hence is open. Any connected set must intersect some path component and hence be contained in it by the previous lemma, so the connected components are contained in the path components. □

This proposition applies for example to an open set X in $\mathbb{R}^n$, since each point in X then has a neighborhood which is a ball, which is path-connected. Notice that the proposition definitely does not apply to the closure of the $\sin(1/x)$ graph.

The Cantor Set

A space X whose connected components are points is said to be *totally disconnected*. This is equivalent to saying that the only connected subspaces of X are points, since connected components are maximal connected subspaces. As a trivial example, a space with the discrete topology is certainly totally disconnected since every subset is open and closed, so no subspace with more than one point can be connected.

For a subspace of $\mathbb{R}$ to be totally disconnected means that it contains no intervals, other than points, since the only connected subspaces of $\mathbb{R}$ are intervals. We mentioned the example of $\mathbb{Q}$ previously, but there are many more totally disconnected subspaces of $\mathbb{R}$. For a start, one can have subspaces such as the integers where every point is isolated, being both open and closed. This is just saying that the subspace topology is the discrete topology, so one has a discrete subspace.

Let us ask the following question: Is there a totally disconnected subspace of $\mathbb{R}$ which is closed in $\mathbb{R}$ but contains no isolated points. Such a subspace would thus combine features of both the integers (a closed subspace) and the rationals (a subspace with no isolated points). The answer to the question is yes, and there is a famous example which we will now describe, known as the *Cantor set*.

The Cantor set C is a subspace of $[0,1]$ constructed as an intersection of an infinite sequence of sets $C_0 \supset C_1 \supset C_2 \supset \cdots$. We start with $C_0 = [0,1]$, and we form C_1 be removing the open interval $(1/3, 2/3)$, so $C_1 = [0,1/3] \cup [2/3,1]$. Next we form C_2 by removing the open middle thirds $(1/9,2/9)$ and $(7/9,8/9)$ of the two intervals of C_1. This leaves the four closed intervals $[0,1/9]$, $[2/9,1/3]$, $[2/3,7/9]$, and $[8/9,1]$. Now we repeat the same process over and over again, at each stage removing the open middle thirds of each interval created at the previous stage. Thus C_n consists of 2^n closed intervals of length $1/3^n$. Finally we set $C = \bigcap_n C_n$. This is a closed subset of $\mathbb{R}$, being the intersection of the closed subsets C_n.

The set C is totally disconnected since it was constructed so as to contain no intervals other than points. Namely, if C contained an interval of positive length ε then this interval would be contained in each C_n, but C_n contains no interval of length greater than $1/3^n$ so if n is chosen to be large enough so that $1/3^n$ is less than ε, then there is no interval of length ε in C_n.

Now let us check that C contains no isolated points. Each point x in C is in the intersection of all the C_n's so it is in the intersection of the sequence of interval components I_n of C_n that contain x. If we let x_n be either endpoint of I_n then $x = \lim x_n$ since the lengths of the I_n's are approaching 0. We can assume $x \neq x_n$ for all n since we have two choices for each x_n so we can always choose one of them different from x if x happens to be an endpoint of some I_n. All the endpoints of the I_n's are contained in C, so this shows that each $x \in C$ is a limit of a sequence of points $x_n \in C$ different from x. This shows that C contains no isolated points.

There is a nice way of characterizing which points of $[0,1]$ lie in C in terms of base 3 decimals. The middle-third interval $(1/3,2/3)$ consists of base 3 decimals $.a_1a_2\cdots$ with $a_1 = 1$, excluding only the decimal $.1 = 1/3$. Thus C_1 consists of the decimals $.a_1a_2\cdots$ with $a_1 \neq 1$, including $1/3$ as $.0222\cdots$ and 1 as $.222\cdots$. In similar fashion C_2 consists of decimals $.a_1a_2\cdots$ with $a_1 \neq 1$ and $a_2 \neq 1$. More generally C_n consists of decimals $.a_1a_2\cdots$ with none of $a_1,\cdots,a_n$ equal to 1. Hence C consists exactly of all decimals $.a_1a_2\cdots$ with no a_i equal to 1. The points of C that are endpoints of an interval component of some C_n are the decimals ending either in all 0's or all 2's.

It is interesting to think of the process of constructing C in physical terms as taking a length of string and repeatedly cutting it into shorter pieces. If we think of the first piece as the interval $[0,1]$ and cut it at the point $1/2$, then it becomes two pieces of string each with two endpoints. In effect the single point $1/2$ has become two points, the two new endpoints of the half pieces, the intervals $[0,1/2]$ and $[1/2,1]$. At the next stage one cuts each of these two pieces in half, producing four separate pieces $[0,1/4]$, $[1/4,1/2]$, $[1/2,3/4]$, and $[3/4,1]$, and so on for later stages. In order to make all these pieces disjoint subsets of $\mathbb{R}$ one can image the string as being stretched so tightly that each time it is cut, it pulls apart at the cut and shrinks to two-thirds of its length, so after the first cut, $[0,1/2]$ shrinks to $[0,1/3]$ and $[1/2,1]$ shrinks to $[2/3,1]$. Then at the next stage we cut $[0,1/3]$ at its midpoint and the two pieces $[0,1/6]$ and $[1/6,1/3]$ shrink to $[0,1/9]$ and $[2/9,1/3]$, and similarly for the piece $[2/3,1]$, and so on.

All this cutting and shrinking can be described very neatly in terms of decimal expansions. Initially we cut $[0,1]$ at $1/2$ producing two copies of the point $1/2$. If we use base 2 decimals, then these two copies of $1/2$ can be distinguished by regarding one of them as $.0111\cdots$ and the other one as $.1000\cdots$. Normally these two decimals are regarded as the same number, but now we want to keep them distinct, viewing them as the two new endpoints of the cut string. Similarly at the next stage the point $1/4$ duplicates itself as the two decimals $.00111\cdots$ and $.01000\cdots$, and the point $3/4$ duplicates itself as $.10111\cdots$ and $.11000\cdots$, etc. The process of shrinking pieces to two-thirds of their lengths can then be achieved by simply changing the decimal base 2 to 3 and replacing all decimal digits 1 by 2.

Here is a question for the reader to ponder: What happens if we do not replace decimal digits 1 by 2 but simply consider all base 3 decimals consisting of 0's and 1's, with no 2's? How does the resulting set compare with the Cantor set?

We can determine the total length of the Cantor set by adding up the lengths of the complementary deleted intervals. First we deleted an interval of length $1/3$, then 2 intervals of length $1/9$, then 4 intervals of length $1/27$, and so on. Thus the total length deleted is the sum of the series $\frac{1}{3}+\frac{2}{9}+\frac{4}{27}+\cdots$, which is a geometric series with initial term $1/3$ and ratio $2/3$ so the sum is 1. Thus the total length of C is 0. Another way to see this is to observe that the length of C_n is $2/3$ of the length of C_{n-1}, so the length of C_n is $(2/3)^n$, which approaches 0 as n goes to infinity.

One can imagine many similar constructions of sets that look much like the Cantor set. For example, instead of removing middle thirds at each step one could remove middle halves or middle quarters, or any other fractions. The fractions could even vary from one step to the next. Or instead of removing a single middle interval from each interval, one could remove several interior open intervals, provided that the lengths of the remaining intervals approach zero. Interestingly enough, all these constructions produce 'Cantor sets' that are homeomorphic to the original Cantor set C. In fact, it is possible to list several properties that the Cantor set has, including being totally disconnected and having no isolated points, such that every space with these properties is homeomorphic to the Cantor set. We will study the other properties in the next couple chapters, and eventually (hopefully) prove the theorem that the list of properties characterizes the Cantor set up to homeomorphism.

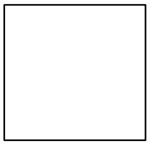
Exercises

1. Let D_1 and D_2 be two open disks in $\mathbb{R}^2$ whose closures $\overline{D}_1$ and $\overline{D}_2$ intersect in exactly one point, so the boundary circles of the two disks are tangent. Determine which of the following subspaces of $\mathbb{R}^2$ are connected: (a) $D_1 \cup D_2$. (b) $\overline{D}_1 \cup \overline{D}_2$. (c) $\overline{D}_1 \cup D_2$.

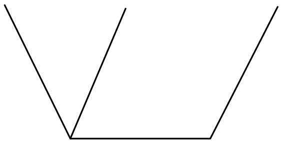
2. Show that if a subspace A of a space X is connected then so is its closure $\overline{A}$.

3. Show the subspace $X \subset \mathbb{R}^2$ consisting of points (x, y) such that at least one of x and y is rational is connected.

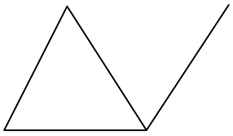
4. By counting cut points and non-cut points show that no two of the following four graphs in $\mathbb{R}^2$ are homeomorphic. (Each graph is a closed subspace of $\mathbb{R}^2$, the union of a finite number of closed line segments.)



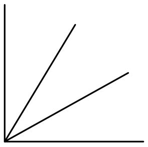
(a)



(b)



(c)



(d)

5. Show that if a space X has only a finite number of connected components, then these components are open subsets of X. (We already know they are closed.)

6. From the fact that an interval $[a, b]$ is connected, deduce the Intermediate Value Theorem: If $f : [a, b] \to \mathbb{R}$ is continuous and $f(a) < c < f(b)$ then there exists a number $x \in [a, b]$ with $f(x) = c$.

7. (a) Show that if $A \subset B \subset \overline{A}$ and A is connected then so is B.
(b) Let $X \subset \mathbb{R}^2$ be the closure of the graph of $\sin(1/x)$ for $x > 0$. Using part (a), determine all the connected subspaces of X.

8. For a space X let X' be the subspace of X obtained by deleting all points x which are isolated (i.e. $\{x\}$ is open and closed in X). Let B_n be the subspace of $[0,1]$ consisting of numbers having a base 2 decimal expansion $.a_1a_2\cdots$ in which at most n of the digits a_i are 1, and let $B = \bigcup_n B_n$. Draw a picture of B_1 and B_2 and determine B_n' and B'. Deduce that there exist spaces X for which the sequence $X \supset X' \supset X'' \supset \cdots$ becomes the empty set only after n stages, for any given number n.

9. (a) Show that the sets obtained by intersecting the Cantor set C with the interval components of the spaces C_n form a basis for the topology on C. (Here $C = \bigcap_n C_n$.)
(b) Describe the basis sets in part (a) in terms of the base 3 decimal expansions of elements of C with only 0's and 2's.
(c) For each subset S of $\{1,2,\cdots\}$ define a function $f_S : C \to C$ by changing the i-th digit of a base 3 decimal $.a_1a_2\cdots \in C$ from 0 to 2 or vice versa if $i \in S$ and leaving this digit unchanged if $i \notin S$. Using parts (a) and (b), show that f_S is a homeomorphism of C. (Note that $f_S^{-1} = f_S$.) Explain why this shows the somewhat surprising fact that for any two points $x, y \in C$ there is a homeomorphism $f : C \to C$ with $f(x) = y$. In particular, the points of C that are endpoints of intervals of C_n for some n are no different intrinsically from all the other points of C.

Chapter 3. Compactness

Compactness is a sort of finiteness property that some spaces have and others do not. The rough idea is that spaces which are 'infinitely large' such as $\mathbb{R}$ or $[0,\infty)$ are not compact. However, we want compactness to depend just on the topology on a space, so it will have to be defined purely in terms of open sets. This means that any space homeomorphic to a noncompact space will also be noncompact, so finite intervals (a,b) and $[a,b)$ will also be noncompact in spite of their 'finiteness'. On the other hand, closed intervals $[a,b]$ will be compact — they cannot be stretched to be 'infinitely large'.

How can this idea be expressed just in terms of open sets rather than in some numerical measure of size? This would seem to be difficult since open sets themselves can be large or small. But large open sets can be expressed as unions of small open sets, so perhaps we should think about counting how many small open sets are needed when a large open set in a space X, such as the whole space X itself, is expressed as a union of small open sets. The most basic question in this situation is whether the number of small open sets needed is finite or infinite. For example, if X is a metric space, then X is the union of all its balls $B_\varepsilon(x)$ of fixed radius $\varepsilon > 0$, so we could ask whether X is in fact the union of a finite collection of these balls $B_\varepsilon(x)$ of fixed radius. To generalize this idea to arbitrary spaces which need not have a metric, we replace balls by arbitrary open sets, and this leads to the following general definition:

Definition. A space X is *compact* if for each collection of open sets O_α in X whose union is X, there exist a finite number of these O_α's whose union is X.

More concisely, one says that every open cover of X has a finite subcover, where an *open cover* of X is a collection of open sets in X whose union is X, and a *finite subcover* is a finite subcollection whose union is still X.

For example, $\mathbb{R}$ is not compact because the cover by the open intervals $(-n,n)$ for $n = 1,2,\cdots$ has no finite subcover, since infinitely many of these intervals are needed to cover all of $\mathbb{R}$. Another open cover which has no finite subcover is the collection of intervals $(n-1,n+1)$ for $n \in \mathbb{Z}$.

In a similar vein, the interval $(0,1)$ fails to be compact since the cover by the open intervals $(1/n,1)$ for $n \geq 1$ has no finite subcover. Of course, there do exist open covers of $(0,1)$ which have finite subcovers, for example the cover by $(0,1)$ itself, or a little less trivially, the cover by all open subintervals of fixed length, say $1/4$, which has the finite subcover $(0,1/4)$, $(1/8,3/8)$, $(1/4,1/2)$, $(3/8,5/8)$, $(1/2,3/4)$, $(5/8,7/8)$, $(3/4,1)$. To be compact means that every possible open cover has a finite subcover. This could be difficult to check in individual cases, so we will develop general theorems to test for compactness.

Compact Sets in Euclidean Space

Spaces with only finitely many points are obviously compact, or more generally spaces whose topology has only finitely many open sets. However, such spaces are not very interesting. Our goal in this section will be to characterize exactly which subspaces of $\mathbb{R}^n$ are compact. We start with an important special case:

Theorem. *A closed interval* $[a,b]$ *is compact.*

Proof. This will be somewhat similar in flavor to the proof we gave that closed intervals are connected. The case $a = b$ is trivial, so we may assume $a < b$. Let a cover of $[a,b]$ by open sets O_α in $[a,b]$ be given. Since $a \in O_\alpha$ for some α, there exists $c > a$ such that the interval $[a,c]$ is contained in this O_α, and hence $[a,c]$ is contained in the union of finitely many O_α's. Let L be the least upper bound of the set of numbers $c \in [a,b]$ such that $[a,c]$ is contained in the union of finitely many O_α's. We know that $L > a$ by the preceding remarks, and by the definition of L we certainly have $L \le b$.

There is some O_α, call it O_β, that contains L. This O_β is open in $[a,b]$, so since $L > a$ there is an interval $[L-\varepsilon, L]$ contained in O_β for some $\varepsilon > 0$. By the definition of L there exist numbers $c < L$ arbitrarily close to L such that $[a,c]$ is contained in the union of finitely many O_α's. In particular, there are such numbers c in the interval $[L-\varepsilon, L]$. For such a c we can take a finite collection of O_α's whose union contains $[a,c]$ and add the set O_β containing $[L-\varepsilon, L]$ to this collection to obtain a finite collection of O_α's containing the interval $[a,L]$. If $L = b$ we would now be done, so it remains only to show that $L < b$ is not possible.

If $L < b$, the number ε could have been chosen so that not only is $[L-\varepsilon, L] \subset O_\beta$ but also $[L-\varepsilon, L+\varepsilon] \subset O_\beta$, since O_β is open in $[a,b]$. Then by adding O_β to the finite collection of O_α's whose union contains $[a,c]$, as in the preceding paragraph, we would have a finite collection of O_α's whose union contains $[a, L+\varepsilon]$. However, this means that L is not an upper bound for the set of c's such that $[a,c]$ is contained in a finite union of O_α's. This contradiction shows that $L < b$ is not possible, so we must have $L = b$. □

For a subspace A of a space X to be compact means of course that every open cover of A has a finite subcover. The open cover of A would consist of sets of the form $A \cap O_\alpha$ for O_α open in X. To say that $A = \bigcup_\alpha (A \cap O_\alpha)$ is equivalent to saying that $A \subset \bigcup_\alpha O_\alpha$. Thus for A to be compact means that for every collection of open sets in X whose union contains A, there is a finite subcollection whose union contains A. So it does no harm to interpret 'every open cover of A has a finite subcover' to mean precisely this.

Proposition. *A closed subset of a compact space is compact, in the subspace topology.*

Proof. Let $\{O_\alpha\}$ be a cover of A by open sets in X. We then obtain an open cover of X by adding the set $X - A$, which is open if A is closed. If X is compact this open cover of X has a finite subcover. The sets O_α in this finite subcover then give a finite cover of A since the set $X - A$ contributes nothing to covering A. □

As an example, the Cantor set is closed in $[0,1]$, so it is compact because $[0,1]$ is compact.

Here is another way to show that a space is compact:

Proposition. *If $f : X \to Y$ is continuous and onto, and if X is compact, then so is Y.*

Proof. Let a cover of Y by open sets O_α be given. Then the sets $f^{-1}(O_\alpha)$ form an open cover of X. If X is compact, this cover has a finite subcover. Call this finite subcover $f^{-1}(O_1), \cdots, f^{-1}(O_n)$. Assuming that f is onto, the corresponding sets $O_1, \cdots, O_n$ then cover Y since for each $y \in Y$ there exists $x \in X$ with $f(x) = y$, and this x will be in some set $f^{-1}(O_i)$ of the finite cover of X, so y will be in the corresponding set O_i. □

This implies for example that a circle is compact since it is the image of a continuous map $f : [0,1] \to \mathbb{R}^2$.

In order to expand our range of compact spaces we use the notion of product spaces, introduced in Chapter 1.

Theorem. *If X and Y are compact then so is their product $X \times Y$.*

By induction this implies that the product of any finite collection of compact spaces is compact.

Proof. Let a cover of $X \times Y$ by open sets O_α in $X \times Y$ be given. Each point $(x,y) \in X \times Y$ lies in some O_α, and this O_α is a union of basis sets $U \times V$, so there exists a basis set $U_{xy} \times V_{xy}$ containing (x,y) and contained in some O_α.

Suppose we choose a fixed x and let y vary. Then the sets $U_{xy} \times V_{xy}$ cover $\{x\} \times Y$, so the sets V_{xy} with fixed x and varying y form an open cover of Y. Since Y is compact, this cover has a finite subcover $V_{xy_1}, \cdots, V_{xy_n}$, where n may depend on x. The intersection $U_x = \bigcap_{j=1}^{n} U_{xy_j}$ is then an open set containing x with two key properties: The sets $U_x \times V_{xy_1}, \cdots, U_x \times V_{xy_n}$ cover $U_x \times Y$, and each $U_x \times V_{xy_j}$ is contained in some O_α.

Now we let x vary. The sets U_x form an open cover of X, so since X is compact there is a finite subcover $U_{x_1}, \cdots, U_{x_m}$. The products $U_{x_i} \times V_{x_i y_j}$ of the sets U_{x_i} with the corresponding sets $V_{x_i y_j}$ chosen earlier then form a finite cover of $X \times Y$. Each set in this finite cover is contained in some O_α, so by choosing an O_α containing each $U_{x_i} \times V_{x_i y_j}$ we obtain a finite cover of $X \times Y$. □

We can use this result to determine exactly which subspaces of $\mathbb{R}^n$ are compact. The result is usually called the Heine-Borel Theorem.

Theorem. *A subspace $X \subset \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

For a subset $X \subset \mathbb{R}^n$ to be bounded means that it lies inside some ball of finite radius centered at the origin.

Proof. First let us assemble previously-proved results to show the ‘if’ half of the theorem. If we assume X is bounded, then it lies in a ball of finite radius and hence in some closed cube $[-r,r] \times \cdots \times [-r,r]$. This cube is compact, being a product of closed intervals which are compact. Since X is a closed subset of a compact space, it is also compact.

Now for the converse, suppose X is compact. The collection of all open balls in $\mathbb{R}^n$ centered at the origin and of arbitrary radius forms an open cover of X, so there is a finite subcover, which means X is contained in a single ball of finite radius, the largest radius of the finitely many balls covering X. Hence X is bounded.

To show X is closed if it is compact, suppose x is a limit point of X that is not in X. Then every neighborhood of x contains points of X. In particular each open ball $B_r(x)$ of radius r centered at x contains points of X, so the same is true also for the closed balls $\overline{B}_r(x)$. The complements $\mathbb{R}^n - \overline{B}_r(x)$ form an open cover of X as r varies over $(0,\infty)$ since their union is $\mathbb{R}^n - \{x\}$ and $x \notin X$. This open cover of X has no finite subcover since each $\overline{B}_r$ contains points of X. Thus we have shown that if X is not closed, it is not compact. □

Hausdorff Spaces

We showed earlier that a compact subspace of $\mathbb{R}^n$ is closed. This is not something that remains true when $\mathbb{R}^n$ is replaced by an arbitrary space X. For example if X is a finite set with any topology then every subspace of X is compact since X has only finitely many subsets in total, but not every subset of X will be closed unless X has the discrete topology. Fortunately there is a fairly simple and natural condition to impose on a topology which will guarantee that compact subspaces are closed:

Definition. A topological space X is a *Hausdorff space* if for each pair of distinct points $x, y \in X$ there exist open neighborhoods U of x and V of y that are disjoint.

For example, every metric space X is a Hausdorff space since if x and y are distinct points in X then the open balls of radius ε centered at x and y will be disjoint if ε is less than half the distance between x and y. This follows from the triangle inequality, since if there were a point z in the intersection of these two balls, then the distances between x, y, and z would satisfy $d(x,y) \le d(x,z) + d(z,y) < \varepsilon + \varepsilon = 2\varepsilon < d(x,y)$, a contradiction.

The trivial topology on a space with more than one point is not Hausdorff. A more interesting example of a non-Hausdorff space is $\mathbb{R}$ with the topology having as its nonempty open sets the complements of finite sets, since in this topology any two nonempty open sets have a nonempty intersection. Many more examples of non-Hausdorff spaces can be constructed but they do not arise all that often ‘in nature’. Most non-Hausdorff spaces are in one way or another artificial. It is probably not a great exaggeration to say that most interesting spaces are Hausdorff.

Here are three nice properties of Hausdorff spaces:

Proposition. (a) *In a Hausdorff space, points are closed subsets.*
(b) *A subspace of a Hausdorff space is Hausdorff.*
(c) *A product of two Hausdorff spaces is Hausdorff.*

Proof. (a) If X is Hausdorff and x is a point in X, then for any other point y there is an open neighborhood V_y of y that is disjoint from a neighborhood of x and hence disjoint from x itself. The union of all these open sets V_y as y ranges over $X - \{x\}$ is $X - \{x\}$, which is therefore open, so $\{x\}$ is closed.

(b) Let A be a subspace of X and let $x, y \in A$. If X is Hausdorff these two points have disjoint open neighborhoods U and V in X, so $A \cap U$ and $A \cap V$ are disjoint open neighborhoods of x and y in A.

(c) Suppose X and Y are Hausdorff and we have two distinct points (x_1, y_1) and (x_2, y_2) in $X \times Y$, so either $x_1 \neq x_2$ or $y_1 \neq y_2$. In the former case, if U_1 and U_2 are disjoint neighborhoods of x_1 and x_2 in X then $U_1 \times Y$ and $U_2 \times Y$ are disjoint neighborhoods of (x_1, y_1) and (x_2, y_2) in $X \times Y$. The opposite case $y_1 \neq y_2$ is handled similarly. □

Proposition. *A compact subspace of a Hausdorff space is closed.*

Proof. Let X be Hausdorff with $A \subset X$ compact. The main work will be to show that for each $x \in X - A$ there exist disjoint open sets U and V in X such that $x \in U$ and $A \subset V$. This will imply that A is closed since no point x in the complement of A can be a limit point of A since it has a neighborhood U disjoint from A.

Now we construct U and V with the desired properties. The point $x \in X - A$ and an arbitrary point $a \in A$ have disjoint open neighborhoods U_x and V_a in X since X is Hausdorff. As a ranges over A the open sets V_a form an open cover of A. Since A is compact, this open cover has a finite subcover. Call this subcover $V_1, \cdots, V_n$, and let the sets U_x that correspond to these V_i's be labelled $U_1, \cdots, U_n$. The intersection $U = U_1 \cap \cdots \cap U_n$ is an open neighborhood of x that is disjoint from each V_i and hence also from the union $V = V_1 \cup \cdots \cup V_n$, which is an open set containing A. □

Corollary. *A map $f : X \to Y$ from a compact space X to a Hausdorff space Y is a homeomorphism if it is continuous, one-to-one, and onto.*

Proof. All that needs to be checked is that f^{-1} is continuous, and to do this it suffices to show that if $C \subset X$ is closed in X then $f(C)$ is closed in Y. But since C is a closed subset of a compact space, it is compact, hence $f(C)$ is compact and therefore also closed since Y is Hausdorff. □

For example, consider continuous maps $[0,1] \to [0,1] \times [0,1]$ defining paths in the square. Intuition would suggest that no such map can be onto, but in fact this intuition is incorrect, and there do exist continuous maps $[0,1] \to [0,1] \times [0,1]$ whose image is the whole square. However, no such map can be one-to-one, otherwise the Corollary would say it is a homeomorphism, but $[0,1]$ and $[0,1] \times [0,1]$ are not homeomorphic since $[0,1]$ has cutpoints but $[0,1] \times [0,1]$ does not.

The Corollary implies that the topology on a space X that is both compact and Hausdorff has a certain optimality property. Notice first that any topology that is finer than a Hausdorff topology is still Hausdorff, while a topology that is coarser than a compact topology is still compact.

- If X is compact and Hausdorff with respect to the topology $\mathcal{O}$, then any other topology that is strictly finer than $\mathcal{O}$ is Hausdorff but not compact, while any other topology that is strictly coarser than $\mathcal{O}$ is compact but not Hausdorff.

To see this, just apply the Corollary to the identity map $X \to X$, which is continuous if the topology in the domain is finer than the topology in the range. If the topology in the domain is compact Hausdorff and the topology in the range is strictly coarser, it would be compact but not Hausdorff, otherwise the Corollary would imply the identity map was a homeomorphism so the two topologies would be the same. Similarly, if the topology in the range is compact Hausdorff and the topology in the domain is strictly finer, it would be Hausdorff but not compact, otherwise the Corollary would say the identity map was a homeomorphism.

Normal Spaces

Often in a Hausdorff space X there is a stronger form of the Hausdorff condition that is satisfied: If A and B are disjoint closed sets in X, then there are disjoint open sets U and V in X with $A \subset U$ and $B \subset V$. In short, disjoint closed sets have disjoint open neighborhoods. A Hausdorff space with this property is called a *normal* space. We include the Hausdorff condition as part of the definition of a normal space to guarantee that points are closed, so that the condition of disjoint closed sets having disjoint open neighborhoods includes the condition that distinct points have disjoint open neighborhoods. The reason for restricting A and B to be closed is that one would not expect arbitrary disjoint subsets to have disjoint open neighborhoods, since for example in $\mathbb{R}$ the sets $\{0\}$ and $(0,1]$ do not have disjoint open neighborhoods.

In some cases the Hausdorff condition automatically implies normality:

Proposition. *A compact Hausdorff space is normal.*

Proof. Let A and B be disjoint closed sets in a compact Hausdorff space X. In particular this implies that A and B are compact since they are closed subsets of a compact space. By the argument in the proof of the preceding Proposition we know that for each $x \in A$ there exist disjoint open sets U_x and V_x with $x \in U_x$ and $B \subset V_x$. Letting x vary over A, we have an open cover of A by the sets U_x, so there is a finite subcover. Let U be the union of the sets U_x in this finite subcover and let V be the intersection of the corresponding sets V_x. Then U and V are disjoint open neighborhoods of A and B. □

It is natural to ask whether all Hausdoff spaces are normal, and here is an artificially constructed example to show that this need not be true:

Example. Let X be the subset of $\mathbb{R}^2$ consisting of all points (x,y) with $y \geq 0$. We construct a topology on X that is finer than the usual topology in the following way. Let $X' \subset X$ be the complement of the x-axis, the points (x,y) with $y > 0$. Then a basis for the new topology on X consists first of all of the usual open disks $D_r(x)$ of radius r and center x that are contained in X', and then in addition for each point x in the x-axis we take the sets $D'_r(x) = \{x\} \cup (X' \cap D_r(x))$ obtained by adding x to an open half-disk centered at x. It is easy to check that the condition for having a basis is satisfied, and we leave this for the reader to do. The Hausdorff condition is also easily verified since any two distinct points of X are contained in disjoint basis sets. Notice that any subset of the x-axis is closed in this topology since its complement is open, being a union of basis sets. However, if we let A be the rational numbers in the x-axis and B the irrational numbers, then A and B are disjoint closed sets in X that do not have disjoint open neighborhoods, since any open set U containing A would contain basis sets $D'_r(a)$ for $a \in A$ that intersect all the basis sets $D'_s(b)$ for $b \in B$ with $|a - b| < r$, hence U would intersect any open set V containing B. Thus X is not normal.

Notice that the subspace topology on the x-axis is the discrete topology since the intersection of the open set $D'_r(x)$ with the x-axis is $\{x\}$, making $\{x\}$ an open set in this subspace. On the other hand, the subspace topology on X', the complement of the x-axis, is the usual topology. So the topology on X somehow mingles these two rather different sorts of topologies.

Proposition. *Metric spaces are normal.*

Before proving this we need a preliminary fact. Let X be a metric space with metric d. Given a subset $A \subset X$ define the distance $d(x,A)$ from a point $x \in X$ to A to the the greatest lower bound of the set of distances $d(x,a)$ from x to points $a \in A$. Note that $d(x,A) \geq 0$, and $d(x,A) = 0$ if and only if x is in the closure of A since $d(x,A) = 0$ is equivalent to saying that every ball $B_r(x)$ contains points of A.

Lemma. *$d(x,A)$ is a continuous function of x.*

Proof. Given $\varepsilon > 0$, let x and y be points in X with $d(x,y) < \varepsilon$. For each $\delta > 0$ there exists $a \in A$ with $d(x,a) < d(x,A) + \delta$. Then we have:

$$d(y,A) \leq d(y,a) \leq d(y,x) + d(x,a) < \varepsilon + \delta + d(x,A)$$

Since this holds for each $\delta > 0$ it follows that $d(y,A) \le \varepsilon + d(x,A)$. Rewriting this as $d(y,A) - d(x,A) \le \varepsilon$, we can switch the roles of x and y to obtain also $d(x,A) - d(y,A) \le \varepsilon$ so $|d(x,A) - d(y,A)| \le \varepsilon$. This holds whenever $d(x,y) < \varepsilon$, from which we conclude that $d(x,A)$ is a continuous function of x. □

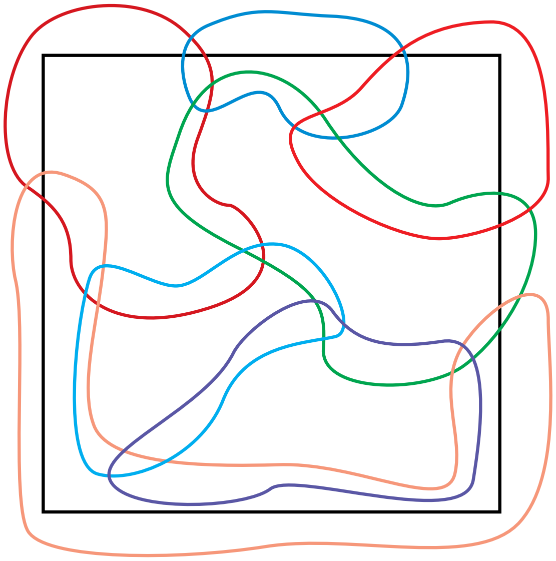
Proof of the Proposition. If A and B are disjoint closed sets in the metric space X, we claim that the sets $U = \{\, x \mid d(x,A) < d(x,B) \,\}$ and $V = \{\, x \mid d(x,B) < d(x,A) \,\}$ are disjoint open neighborhoods of A and B, respectively. Certainly they are disjoint, and they are open since for the difference function $d(x,A) - d(x,B)$, which is continuous by the Lemma, U and V are the inverse images of the open sets $(-\infty, 0)$ and $(0, \infty)$. To see that $A \subset U$ note that $d(x,A) = 0$ for points $x \in A$ and $d(x,A) > 0$ for points $x \notin A$ since A is closed. Similarly we have $B \subset V$ using the assumption that B is closed. □

Lebesgue Numbers

Given a cover of a metric space X by subsets A_α, a *Lebesgue number* for the cover is a number $\varepsilon > 0$ such that every ball $B_\varepsilon(x)$ in X is contained in at least one set A_α of the cover.

Theorem. *Every open cover of a compact metric space X has a Lebesgue number.*

For example, what would a Lebesgue number for the following open cover of a square be?



Proof. Since X is compact, each open cover has a finite subcover, and there is no loss of generality to find a Lebesgue number for this subcover. Thus we may assume the cover is by open sets $U_1, \cdots, U_n$.

For a ball $B_\varepsilon(x)$ to be contained in U_i is equivalent to the requirement that $d(x, X - U_i) \ge \varepsilon$. Thus we are led to consider the functions $d_i(x) = d(x, X - U_i)$, and to ask whether there exists an $\varepsilon > 0$ such that for each $x \in X$ at least one of the numbers $d_1(x), \cdots, d_n(x)$ is at least ε. Phrased differently, we are asking whether there is an $\varepsilon > 0$ such that the function $f(x) = \max\{d_1(x), \cdots, d_n(x)\}$ satisfies $f(x) \ge \varepsilon$ for all x. Each function $d_i(x)$ is continuous by the preceding Lemma, and we leave it as an exercise for the reader to verify that the maximum of a finite set of continuous functions is again continuous. (By induction it suffices to do the case of two functions.) So f is continuous. Since U_i is open, its complement $X - U_i$ is closed and hence $d_i(x) > 0$ for all $x \in U_i$. Thus $f(x) > 0$ for all $x \in X$ since the U_i's cover X. Since X is compact, the image $f(X)$ is compact in $(0, \infty)$, so there is a lower bound $\varepsilon > 0$ for the values of f on X. By the remarks at the beginning of this paragraph, this ε is a Lebesgue number for the cover. □

Corollary. *Let $f : X \to Y$ be a continuous map from a compact metric space X to a metric space Y. Then f is uniformly continuous: For each $\varepsilon > 0$ there exists a $\delta > 0$ such that $d(f(x), f(x')) < \varepsilon$ for all $x, x' \in X$ with $d(x, x') < \delta$.*

Proof. Given $\varepsilon > 0$, cover X by the open sets $f^{-1}(B_{\varepsilon/2}(y))$ as y ranges over Y. Let δ be a Lebesgue number for this cover of X. Then f takes each ball $B_\delta(x)$ to some ball $B_{\varepsilon/2}(y)$, so if $d(x, x') < \delta$ then $d(f(x), f(x')) \le d(f(x), y) + d(y, f(x')) < \varepsilon/2 + \varepsilon/2 = \varepsilon$. □

Infinite Products

There are times when one is interested in the product $X_1 \times X_2 \times \cdots$ of an infinite sequence of spaces $X_1, X_2, \cdots$. As a set, this product consists of all sequences $(x_1, x_2, \cdots)$ with $x_i \in X_i$ for each i. We will also use the notation $\prod_i X_i$ for this product. As an example, the Cantor set can be viewed as the product $\prod_i X_i$ where each X_i is the set $\{0, 2\}$ by identifying an infinite base 3 decimal $.a_1 a_2 \cdots$ with the sequence $(a_1, a_2, \cdots)$.

A first guess for how to define a topology on an infinite product $\prod_i X_i$ would be to do the same thing as for finite products, taking as a basis the products $U_1 \times U_2 \times \cdots$

of open sets $U_i \subset X_i$. The same argument as for finite products shows that this does in fact define a topology, called the *box topology* on $\prod_i X_i$. However, if one tries to work with this topology, one finds it has some undesirable features. For example, many maps $Z \to \prod_i X_i$ that one would expect to be continuous are not continuous if the box topology is used on the infinite product. To take a concrete instance, if Z and each X_i is $\mathbb{R}$ then the harmless-looking map $x \mapsto (x, x, \cdots)$ fails to be continuous, since the inverse image of the product $(-1/2, 1/2) \times (-1/3, 1/3) \times (-1/4, 1/4) \times \cdots$ is just $\{0\}$, which is not open.

Fortunately there is another topology on infinite products that avoids defects like this. Instead of taking a basis to consist of all products $U_1 \times U_2 \times \cdots$ of open sets $U_i \subset X_i$, one only takes products for which $U_i = X_i$ for all but finitely many values of i. This guarantees that a map $f : Z \to \prod_i X_i$, $f(z) = (f_1(z), f_2(z), \cdots)$, is continuous if and only if each f_i is continuous, since we have $f^{-1}(U_1 \times U_2 \times \cdots) = f_1^{-1}(U_1) \cap f_2^{-1}(U_2) \cap \cdots$, and this will be open if each f_i is continuous, as it is really only a finite intersection since $f_i^{-1}(U_i) = f_i^{-1}(X_i) = Z$ for all but finitely many values of i.

This topology on $\prod_i X_i$ is called the *product topology* since it turns out to be the one that is used for infinite products almost all the time. Notice that for finite products it gives the same thing as the product topology defined earlier.

Example. The Cantor set is homeomorphic to the product $\{0, 2\} \times \{0, 2\} \times \cdots$ where each $\{0, 2\}$ is given the discrete topology. [Proof: Use the basis for the topology on the Cantor set given by the sets $O(a_1, \cdots, a_n)$ consisting of all decimals whose first n digits are $a_1, \cdots, a_n$ and whose remaining digits are arbitrary.]

Another advantage of the product topology over the box topology is that it preserves compactness:

Theorem. *An infinite product $\prod_{i=1}^{\infty} X_i$ with the product topology is compact if each X_i is compact.*

This would not be true for the box topology. For example, if each X_i is a finite set having at least two elements, with the discrete topology, then the box topology on the infinite product is the discrete topology, which is never compact for an infinite set.

Proof. We will argue by contradiction. Suppose there is an open cover $\mathcal{O}$ of $\prod_i X_i$ that has no finite subcover. Assuming this, we claim first that there exists a point $x_1 \in X_1$ such that no basis set $U_1 \times X_2 \times X_3 \times \cdots$ with $x_1 \in U_1$ is covered by finitely many sets in $\mathcal{O}$. For otherwise if for every x_1 there was a basis set $U_1 \times X_2 \times X_3 \times \cdots$ covered by finitely many sets in $\mathcal{O}$, with $x_1 \in U_1$, then the collection of all these U_1's as x_1 varied over X_1 would be an open cover of X_1, so since X_1 is compact this cover would have a finite subcover, and it would follow that the cover $\mathcal{O}$ had a finite subcover, contrary to assumption.

Having chosen the point x_1, we claim next that there is a point $x_2 \in X_2$ such that no basis set $U_1 \times U_2 \times X_3 \times \cdots$ with $(x_1, x_2) \in U_1 \times U_2$ is covered by finitely many sets in $\mathcal{O}$. For otherwise if for every x_2 there was such a basis set $U_1 \times U_2 \times X_3 \times \cdots$, then the collection of these U_2's as x_2 varied over X_2 would be an open cover of X_2, and compactness of X_2 would give a finite subcover. Letting U_1' be the intersection of the corresponding finite set of U_1's, we would then have a basis set $U_1' \times X_2 \times X_3 \times \cdots$ covered by finitely many sets in $\mathcal{O}$, with $x_1 \in U_1'$, contrary to how x_1 was chosen in the preceding paragraph.

Repeating this same argument, we can choose an infinite sequence of points $x_i \in X_i$ for $i = 1, 2, \cdots$, such that for each n, no basis set $U_1 \times \cdots \times U_n \times X_{n+1} \times \cdots$ with $(x_1, \cdots, x_n) \in U_1 \times \cdots \times U_n$ is covered by finitely many sets in $\mathcal{O}$. But this is impossible because the point $(x_1, x_2, \cdots)$ has to lie in some set in $\mathcal{O}$ and since this set is open, it will contain a basis set $U_1 \times \cdots \times U_n \times X_n \times \cdots$ containing $(x_1, x_2, \cdots)$, so in particular this basis set will be covered by a single set in $\mathcal{O}$. □

We have considered the product of a countable collections of spaces X_n but it is also possible to define the product of an arbitrary infinite collection of spaces X_α where α ranges over any index set I. If I is uncountable we cannot write elements of the product $\prod_\alpha X_\alpha$ as sequences $(x_1, x_2, \cdots)$. Instead, elements of $\prod_\alpha X_\alpha$ are regarded as functions $\alpha \mapsto x_\alpha \in X_\alpha$, assigning an element of X_α to each index α, just as sequences $(x_1, x_2, \cdots)$ can be viewed as functions $i \mapsto x_i \in X_i$. One can write an element of $\prod_\alpha X_\alpha$ briefly as (x_α), it being understood that this is not just a single x_α but a family of them parametrized by $\alpha \in I$. As in the case of countable products, the product topology on $\prod_\alpha X_\alpha$ has as basis the products $\prod_\alpha U_\alpha$ of open sets $U_\alpha \subset X_\alpha$ such that $U_a = X_\alpha$ for all but finitely many values of α. It is again a theorem that $\prod_\alpha X_\alpha$ is compact if each X_α is compact, but the proof requires a fancier form of induction that depends on a preliminary discussion of set theory topics called the 'well-ordering principle' and the 'axiom of choice', so we will not go into this here.

Exercises

1. Give a detailed proof that if two spaces are homeomorphic and one of them is compact, then so is the other.

2. Show that a space X is compact if and only if it has the following property: For every collection of closed sets C_α in X such that the intersection of every finite subcollection of C_α's is nonempty, the intersection of all the C_α's is nonempty. [Hint: just use the fact the closed sets are complements of open sets.]

3. Let X be the real numbers with the topology in which the open sets are the complements of finite sets (and $\varnothing$). Show that X is compact.

4. Show that if X is compact and $f : X \to \mathbb{R}$ is continuous, then f is bounded and takes on a minimum and a maximum value.

5. Show that if $f : X \to Y$ is continuous and one-to-one and if Y is Hausdorff then so is X.

6. Show that if A and B are compact subspaces of a space X, then so is $A \cup B$. If in addition X is Hausdorff, show that $A \cap B$ is compact.

7. For subspaces $A \subset X$ and $B \subset Y$ show that $\overline{A \times B} = \overline{A} \times \overline{B}$ and $\text{int}(A \times B) = \text{int}(A) \times \text{int}(B)$.

8. Show that $[0,1) \times [0,1)$ is homeomorphic to $[0,1] \times [0,1)$ but not to $[0,1] \times [0,1]$.

9. Consider the orthogonal group $O(n)$ consisting of all $n \times n$ orthogonal matrices, i.e., $n \times n$ matrices whose columns form an orthonormal basis $v_1, \cdots, v_n$ for $\mathbb{R}^n$. We put a topology on $O(n)$ by regarding it as a subspace of $\mathbb{R}^{n^2}$, taking the n^2 entries of a matrix in $O(n)$ as the n^2 coordinates of a point in $\mathbb{R}^{n^2}$. (a) Show that $O(n)$ is a closed subset of $\mathbb{R}^{n^2}$ by considering the dot products $v_i \cdot v_j$ of the columns of matrices in $O(n)$ as functions $\mathbb{R}^{n^2} \to \mathbb{R}$. (b) Show that $O(n)$ is compact.

10. Consider the function $f : S^1 \to \S^1 \times S^1$ given by $f(\theta) = (2\theta, 3\theta)$ where we think of points in S^1 as angles θ. Show that f is a homeomorphism onto its image, and draw a picture of this image on the torus $S^1 \times S^1$.

11. Let X be the subset of $\mathbb{R}^2$ which is the union of the line segments L_n from $(0,0)$ to $(1, 1/n)$ for $n = 1, 2, \cdots$, together with the limiting segment L_∞ from $(0,0)$ to $(1,0)$. Define a topology on X by saying that a set $O \subset X$ is open if $O \cap L_n$ is open in L_n for all n, including $n = \infty$. (Here L_n is given the subspace topology from $\mathbb{R}^2$.) (a) Show the axioms for a topology are satisfied.

(b) Find a set that is open in this topology but not in the topology on X as a subspace of $\mathbb{R}^2$.
(c) Show that X is not compact in this new topology, in contrast with the subspace topology where it is compact since it is a closed bounded subset of $\mathbb{R}^2$.

12. For a set X define $d : X \times X \to \mathbb{R}$ by setting $d(x,x) = 0$ and $d(x,y) = 1$ if $x \neq y$. Show that d is a metric. What is the topology defined by this metric?

13. (a) Given a metric d on a set X, show that each of the functions $d'(x,y)$ listed below is again a metric, and that d and d' define the same topology on X. Note that in (b) and (c) the values of d' are bounded above by 1.
(a) $d'(x,y) = kd(x,y)$ for any constant $k > 0$.
(b) $d'(x,y) = \min\{d(x,y), 1\}$.
(c) $d'(x,y) = d(x,y)/(1 + d(x,y))$.

14. For a metric space X defined by a metric d show that the function $d : X \times X \to \mathbb{R}$ is continuous.

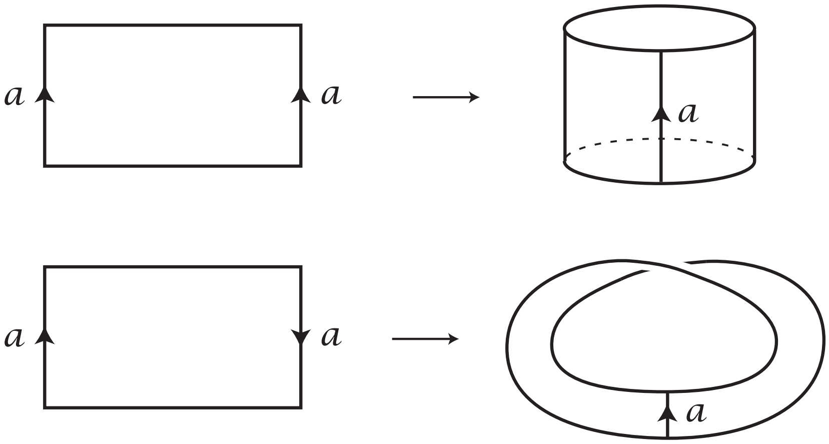
15. Let X be a metric space with metric d. For subsets $A, B \subset X$ define $d(A,B) = \inf\{\, d(a,b) \mid a \in A \text{ and } b \in B \,\}$. Show that if A and B are compact then there exist points $a \in A$ and $b \in B$ with $d(a,b) = d(A,B)$, so in particular $d(A,B) > 0$ if $A \cap B = \varnothing$. Show by an example that both these statements can be false if A and B are only assumed to be closed instead of compact.

16. Let X be the subset of $\mathbb{R}^2$ which is the union of the line segments L_n from $(0,0)$ to $(1, 1/n)$ for $n = 1, 2, \cdots$, together with the limiting segment L_∞ from $(0,0)$ to $(1,0)$. Define a topology $\mathcal{O}$ on X by saying that a set $O \subset X$ is in $\mathcal{O}$ if $O \cap L_n$ is open in L_n for each $n \leq \infty$, where L_n is given the subspace topology from $\mathbb{R}^2$. Show that this topology on X is normal but is not defined by any metric on X. Hint: Given a metric on X, find a sequence of points $x_n \in L_n$, $n = 1, 2, \cdots$, converging to $(0,0)$ in the topology defined by the metric but not in the topology $\mathcal{O}$.

Chapter 4. Quotient Spaces

Our next objective is to describe a general procedure for building complicated spaces out of simpler spaces. This is the topological analog of how all sorts of objects are made in the real world. Just think of all the different words there are in the English language for putting things together: gluing, pasting, taping, stapling, stitching, sewing, welding, riveting, soldering, brazing, bonding, nailing, bolting, clamping, ...

To take two simple examples, we can construct either a cylinder or a Möbius band by taking a rectangular strip and gluing two opposite edges together:



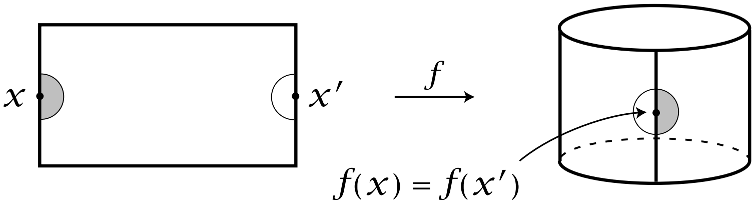
The gluing process can be described as a map $f:X\to Y$ where X is the rectangle and Y is either the cylinder or the Möbius band. The map f is onto, but not quite one-to-one because of the gluing: Two points on opposite edges of the rectangle that are glued together have the same image under f, but other than this, f is one-to-one.

Is it possible to describe the topology on Y just in terms of the topology on X and the map f? If so, then we can construct the space Y just by saying “start with the space X and glue this to that", since the gluing instructions are telling what the set Y and the map f are. In many cases there is an easy way to characterize open sets in Y in terms of open sets in X:

Proposition. *If $f:X\to Y$ is continuous and onto, with X compact and Y Hausdorff, then a set $U\subset Y$ is open in Y if and only if $f^{-1}(U)$ is open in X.*

Proof. Continuity of f says that $f^{-1}(U)$ is open if U is open, so we need to see that the converse is true if X is compact and Y is Hausdorff. If $f^{-1}(U)$ is open in X then $X-f^{-1}(U)$ is closed and hence compact since X is compact. The image $f(X-f^{-1}(U))$ is then compact in Y, and therefore also closed since Y is Hausdorff. But since f is onto we have $f(X-f^{-1}(U))=Y-U$. Thus $Y-U$ is closed, so U is open. □

This Proposition applies to the cylinder and Möbius band examples. One's first guess might have been that the open sets in Y are the images $f(O)$ of open sets O in X. However, this isn't right because an open neighborhood of a point x on the left or right edge of the rectangle is sent to a set in Y that is not an open neighborhood of $f(x)$, but only 'half' of an actual open neighborhood U of $f(x)$. To get the other half of U one would have to enlarge the open set O in X to include also an open neighborhood of the other point x' with $f(x') = f(x)$. This amounts to replacing O by $f^{-1}(U)$, as stated in the Proposition.



The previous Proposition gives a characterization of the topology on Y in terms of the topology on X. It is also possible to make this characterization into a *definition* of a topology on Y, as the following shows:

Proposition. *Given a space X, a set Y, and a function $f:X\to Y$, then the collection of all sets $U \subset Y$ such that $f^{-1}(U)$ is open in X is a topology on Y.*

Proof. The axioms for unions and intersections are satisfied since $f^{-1}(\bigcup_\alpha U_\alpha) = \bigcup_\alpha f^{-1}(U_\alpha)$ and $f^{-1}(\bigcap_\alpha U_\alpha) = \bigcap_\alpha f^{-1}(U_\alpha)$. For the other two axioms we have $f^{-1}(\varnothing) = \varnothing$ and $f^{-1}(Y) = X$. □

Notice that we did not require f to be onto in this Proposition. If f is not onto then points $y \in Y - f(X)$ form open sets $\{y\}$ since their inverse images $f^{-1}(y)$ are empty and hence open. So these points are just isolated points of Y having nothing to do with the topology on $f(X)$. Assuming f is onto is thus not a very serious restriction.

Definition. If the map $f:X\to Y$ is onto, the topology on Y given by this Proposition is called the *quotient topology* on Y, and the map f is called a *quotient map.* One also says that Y is a *quotient space* of X.

In these terms, the earlier Proposition then says that any continuous map from a compact space to a Hausdorff space that is onto is a quotient map.

In the examples of the torus and Möbius band as quotient spaces of a rectangle we already had existing models of the quotient space, but sometimes one wants to construct a quotient space without having a model of it available beforehand. One would just like to start with a space X and say “glue this to that" and know that these instructions will automatically produce a quotient space of X. To see that this is always possible, suppose first that we do have a quotient map $f:X\to Y$. The points of X that are ‘glued together’ to produce a point $y \in Y$ are exactly the points in $f^{-1}(y)$. Since quotient maps are onto, $f^{-1}(y)$ is always nonempty. Thus the points $y \in Y$ are in one-to-one correspondence with the subsets $f^{-1}(y)$ of X. These subsets are disjoint for different choices of y, and their union is all of X since each point of X is sent by f to some point of Y. Thus the sets $f^{-1}(y)$ form a *partition* of X into disjoint subsets. For example, for the cylinder and Möbius band these subsets of the rectangle each have either one or two points — either a pair of points on the left and right edges of the rectangle or a single point not on either of these edges.

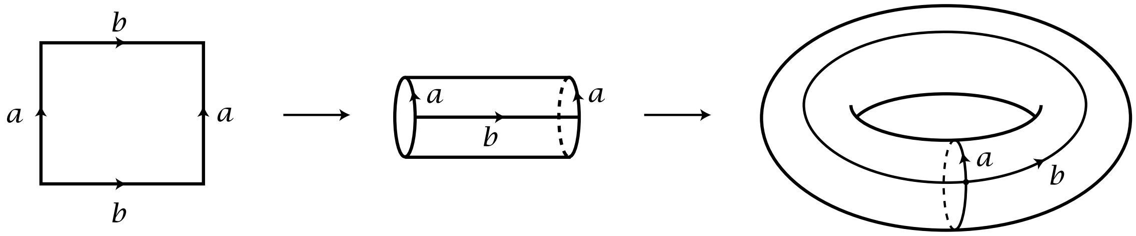
Conversely, suppose we are given a partition of X into disjoint nonempty subsets. For $x \in X$ we denote the set in the partition containing x by $[x]$. If we simply define Y to be the set of all these subsets $[x]$, then we have a natural function $f:X\to Y$ sending x to $[x]$, and this f is onto so we can use it to define a topology on Y by specifying that a set $U \subset Y$ is open exactly when $f^{-1}(U)$ is open in X. This makes f into a quotient map. We will use the notation $X/\sim$ for this quotient space Y whose points are the sets $[x]$ in the partition of X. Here $\sim$ is the relation on X given by $x \sim x'$ if $[x] = [x']$, i.e., x and x' lie in the same set in the partition of X. This is an *equivalence relation*, meaning that it satisfies the three properties

(i) $x \sim x$ for all $x \in X$.
(ii) If $x \sim y$ then $y \sim x$.
(iii) If $x \sim y$ and $y \sim z$ then $x \sim z$.

Conversely, every equivalence relation on X determines a partition of X into the equivalence classes $[x] = \{\, x' \in X \mid x' \sim x \,\}$.

In practice the partition of X is often specified implicitly by saying which points we wish to glue together, or in other words, which points of X we wish to become identified to single points in Y. For example, to obtain the cylinder from the rectangle $[0,1] \times [0,1]$ we make the identifications $(0,t) \sim (1,t)$ for $t \in [0,1]$, and to obtain the Möbius band we instead identify $(0,t) \sim (1,1-t)$. If identifications in a space X are specified by means of a relation $\sim$ then this determines a quotient space $X/\sim$ associated to the partition of X into the smallest subsets that contain all the specified pairs $x \sim x'$. In particular, if no identifications are specified for a point $x \in X$ then x becomes a set $[x] = \{x\}$ of the partition all by itself.

Example. The torus $S^1 \times S^1$ can be realized as a quotient space of a rectangle $[0,1] \times [0,1]$ by identifying both pairs of opposite edges via $(0,t) \sim (1,t)$ and $(s,0) \sim (s,1)$.



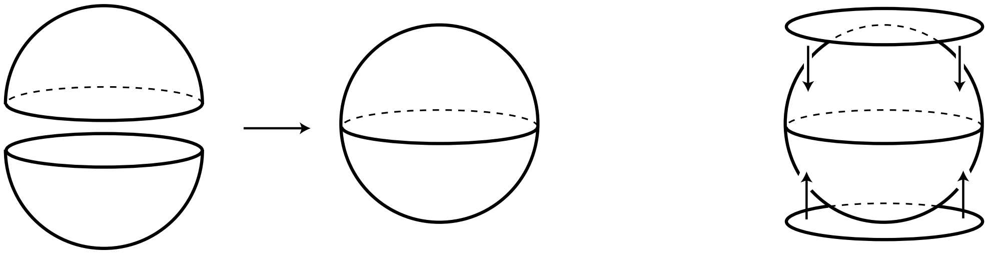
If we first identify the top and bottom edges of the rectangle we obtain a cylinder, and then identifying the two ends of the cylinder produces a torus. Notice that the identifications of opposite edges of the rectangle force all four corners to be identified to a single point. In particular, we are making the identification $(0,0) \sim (1,1)$ even though this is not part of either of the original identifications $(0,t) \sim (1,t)$ and $(s,0) \sim (s,1)$, but follows since $(0,0) \sim (0,1) \sim (1,1)$.

Given a concrete quotient map $f : X \to Y$ we can also form the abstract quotient space $X/\sim$ associated to the partition of X into the subsets $f^{-1}(y)$, and it would be nice to know that this is really the same as Y. The map f induces a well-defined function $\overline{f} : X/\sim \to Y$ sending $f^{-1}(y)$ to y, and this $\overline{f}$ is obviously one-to-one and onto.

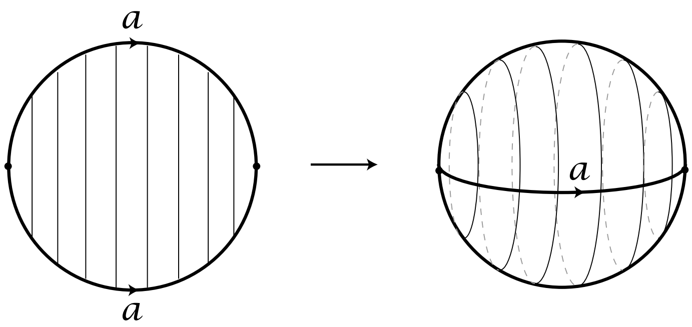
Proposition. *This function $\overline{f} : X/\sim \to Y$ is a homeomorphism.*

Proof. Let $p : X \to X/\sim$ be the quotient map $x \mapsto [x] = \{\, x' \in X \mid f(x') = f(x) \,\}$, so $f = \overline{f}p$. Then a set $U \subset Y$ is open if and only if $f^{-1}(U)$ is open (since f is a quotient map), but since $f^{-1}(U) = p^{-1}(\overline{f}^{-1}(U))$ and p is a quotient map, this is equivalent to $\overline{f}^{-1}(U)$ being open. □

Example. Let us describe three ways to obtain the sphere S^2 as a quotient space. First, as shown in the figure below, we can form S^2 from two closed disks by identifying their boundary circles together to form a single circle, the equator of S^2. The two disks then become the upper and lower hemispheres of S^2. More formally, we can take the two disks as the space $X = D^2 \times \{-1,1\} \subset \mathbb{R}^2 \times \mathbb{R} = \mathbb{R}^3$, where D^2 is the closed unit disk in $\mathbb{R}^2$, and then the identifications are $(x,1) \sim (x,-1)$ for $x \in \partial D^2$, with the map $X \to S^2$ given by vertical projection of each disk onto the hemisphere it is tangent to.

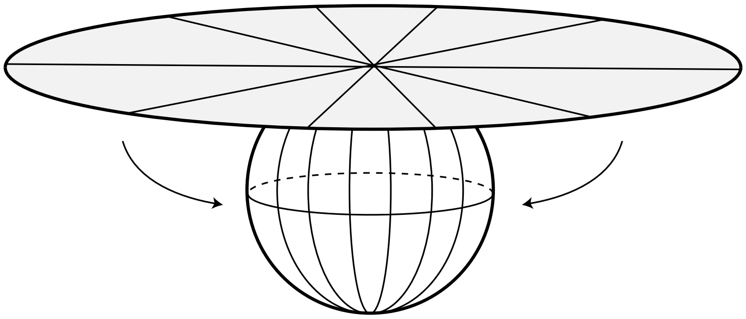


A second way to realize S^2 as a quotient space is to start with a single closed disk D^2 and glue the top half of its boundary circle to the bottom half, each point in the upper half being identified with the point directly below it in the lower half.



Parallel vertical lines in the disk then become parallel vertical circles in S^2. It would not be difficult to write down a formula for an explicit quotient map $D^2 \to S^2$ that does this, although there is really no need for explicit formulas.

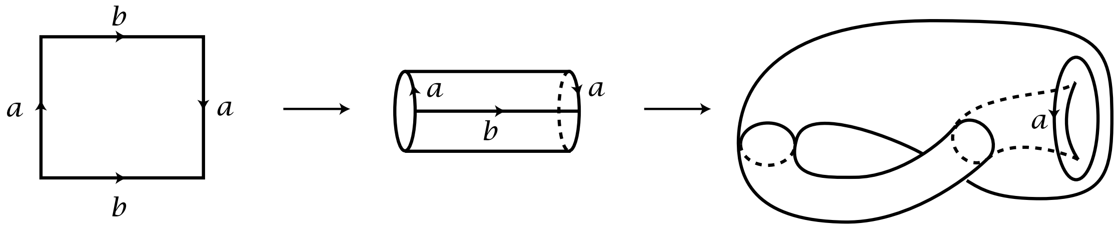
The third construction also realizes S^2 as a quotient of D^2, but this time all the points of ∂D^2 are identified to a single point of S^2, the south pole. This is achieved by a map $D^2 \to S^2$ that sends each radial line segment to a meridian of S^2 going from the north pole to the south pole. We can view this as taking a disk tangent to the sphere and wrapping it completely around the sphere.



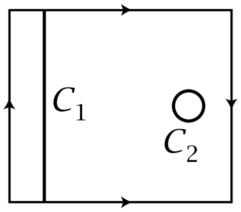
This example differs from all the previous ones in that the map $f : X \to Y$ is no longer finite-to-one.

In the next example we will not have a ready-made model of the quotient space, so we will be constructing a new space as a quotient of a known space.

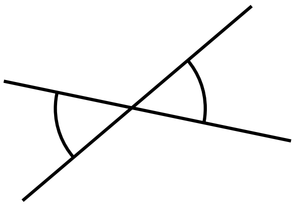
Example. Let us construct a surface, commonly called the *Klein bottle*, by modifying the earlier construction of the torus from a rectangle by reversing the orientation of one of the edges of the rectangle:



By definition, the Klein bottle is the quotient space of the rectangle obtained by identifying opposite edges according to the orientations shown. We will use the notation K^2 for this space, the superscript indicating that it is a 2-dimensional surface. Identifying the top and bottom edges of the rectangle gives a cylinder, but to identify the two ends of the cylinder by deforming the cylinder in $\mathbb{R}^3$ requires that the cylinder pass through itself in order to make the orientations match, as in the third figure. Thus we have a map from K^2 into $\mathbb{R}^3$ which is not one-to-one because two circles in K^2 have the same image circle C in $\mathbb{R}^3$. In fact, it is a theorem that there is no embedding of K^2 into $\mathbb{R}^3$, i.e., there is no subspace of $\mathbb{R}^3$ homeomorphic to K^2. However, there is an embedding into $\mathbb{R}^4$. The first three coordinates of this embedding are the map $K^2 \to \mathbb{R}^3$ indicated in the figure. For the fourth coordinate we just need a map $f : K^2 \to \mathbb{R}$ that takes different values on the two circles C_1 and C_2 of K^2 that map to C. In the rectangle these two circles are positioned as in the figure at the right. We can choose f to have a value $\varepsilon > 0$ on C_1 and be 0 everywhere outside a small neighborhood of C_1.



Example. Let P^2 be the set of all lines through the origin in $\mathbb{R}^3$. This is a metric space if we define the 'distance' between two lines through the origin to be the smaller of the two angles between them, so this angle lies in the interval $[0, \pi/2]$. The first two properties required of a metric are obviously satisfied, and we will leave it for the reader to check the triangle inequality. There is a natural map $p : S^2 \to P^2$ sending a point $x \in S^2$ to the line through the origin containing x. Each line through the origin intersects S^2 in two antipodal points x and

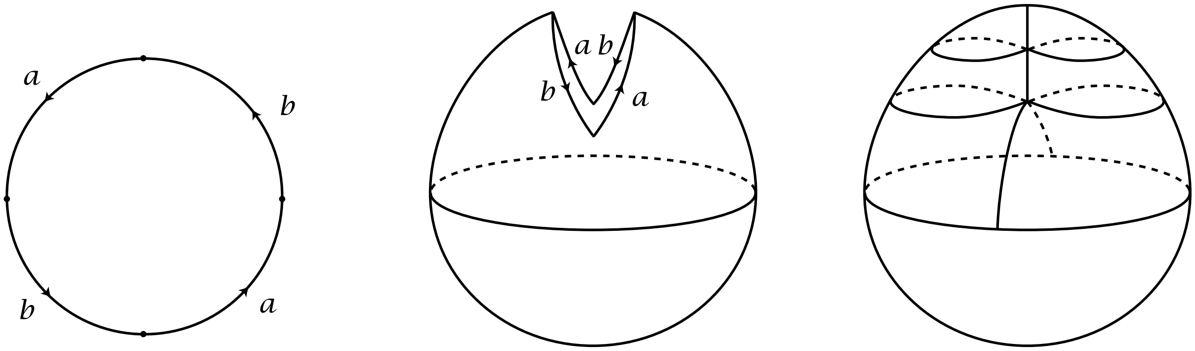


$-x$, so p is onto and is exactly two-to-one. We can see that p is continuous since p^{-1} of an ε-ball in P^2 consists of two antipodal ε-balls in S^2. It follows that p is a quotient map since it is a continuous map from a compact space onto a Hausdorff space (metric spaces are Hausdorff). This means that we can regard P^2 as the quotient space of S^2 in which each point x is identified with the antipodal point $-x$, so $P^2 = S^2/(x \sim -x)$.

The classical name for P^2 is the *projective plane*. This comes from thinking of how 2-dimensional pictures (drawings or paintings, for example) of 3-dimensional objects are obtained. Each point of a 3-dimensional object projects onto the plane of the picture along the line from the point to the viewer's eye. The points that we see are then really just all the different lines converging to our eye. Normally one's angle of vision is something less than 180 degrees, and a picture encompasses considerably less than 180 degrees. But imagine what might happen if we could see more than 180 degrees at a time. Two points in space that were exactly 180 degrees apart would presumably be recognized as distinct points. So perhaps if we were making a mathematical model of how we see things we should not identify antipodal points of S^2 but just consider S^2 itself. This leads to a kind of geometry called spherical geometry. The other alternative of identifying antipodal points leads to a slightly different geometry called projective geometry, which is in some ways closer to Euclidean geometry, and therefore preferable for some purposes.

Each line through the origin in $\mathbb{R}^3$ intersects the upper hemisphere of S^2 in either one or two points, the latter possibility holding only for lines in the xy-plane. This means that we can also describe P^2 as the quotient space of the upper hemisphere obtained by identifying antipodal points in its boundary, the equator. Since the upper hemisphere is homeomorphic to a disk via vertical projection onto the unit disk in the xy-plane, this means that P^2 is describable as the quotient space of the disk D^2 with the identifications $x \sim -x$ for $x \in \partial D^2$.

As with the Klein bottle, it is impossible to embed P^2 as a subspace of $\mathbb{R}^3$, so one looks instead for a map $P^2 \to \mathbb{R}^3$ that is almost an embedding. We will describe a map that is an embedding on the interior of D^2 and folds the circle $\partial D^2/(x \sim -x)$ onto a line segment. The image of this map is shown in the figure on the right below:



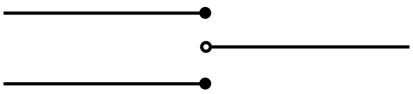
To describe the construction in words, we first deform the D^2 in the figure on the left so that it becomes a sphere with a slit. The boundary circle of the slit is divided into four arcs and we would like to identify these four arcs in pairs, a with a and b with b according to the orientations shown. Identifying the first pair presents no problems, but to identify the second pair then forces all four arcs to coincide in $\mathbb{R}^3$, so that one has a sphere pinched along a line segment as the image of the map $P^2 \to \mathbb{R}^3$. The preimage of the interior of this line segment consists of two distinct open arcs in P^2.

The upper half of the third figure is known as a *crosscap*. It is the image of a map of a Möbius band to $\mathbb{R}^3$, the Möbius band obtained by removing an open disk from P^2. Thinking of P^2 as D^2 with antipodal boundary points identified, the Möbius band is obtained by removing a smaller concentric disk from D^2, leaving an annulus, with antipodal points on one of its boundary circles identified. Thinking of P^2 as $S^2/(x \sim -x)$, one first removes symmetric disk neighborhoods of the two poles of S^2 leaving an equatorial band, and then one identifies antipodal points in this band to produce a Möbius band.

This example of the projective plane easily generalizes to higher dimensions. The n-dimensional projective space P^n is the space of lines through the origin in $\mathbb{R}^{n+1}$, the quotient space $S^n/(x \sim -x)$, or equivalently the quotient of D^n with antipodal points of ∂D^n identified. When $n = 1$ this is just identifying the endpoints of a semicircle, so P^1 is homeomorphic to S^1.

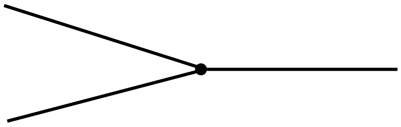
It can easily happen that a quotient space of a Hausdorff space is non-Hausdorff. Here is a simple example:

Example. Let $X = \mathbb{R} \times \{0,1\}$, two disjoint copies of $\mathbb{R}$, and make the identifications $(x,0) \sim (x,1)$ for all $x > 0$. Then the two points $x_0 = (0,0)$ and $x_1 = (0,1)$ are distinct in the quotient although any two neighborhoods U_0 of x_0 and U_1 of x_1 in the quotient intersect since they contain points $(x,0) \sim (x,1)$ for $x > 0$.



Quotient maps are not generally open maps, taking open sets to open sets, but this happens to be the case in this example.

If we make a small modification of this example by setting $(x,0) \sim (x,1)$ for $x \geq 0$ rather than $x > 0$ then the quotient becomes Hausdorff, three lines in $\mathbb{R}^2$ meeting at a point.



Here is a technical result that is used quite often when proving things about quotient spaces:

Lemma. *Given a quotient map $f:X\to Y$ and a space Z, then a function $g:Y\to Z$ is continuous if and only if the composition $gf:X\to Z$ is continuous.*

Proof. Let $O \subset Z$ be open. Then by the definition of a quotient map, the set $g^{-1}(O)$ is open in Y if and only if $f^{-1}(g^{-1}(O)) = (gf)^{-1}(O)$ is open in X. This is saying that g is continuous if and only if gf is continuous. □

Exercises

For the following problems recall that a map $f:X\to Y$ is a quotient map if it is onto and has the property that a set $O \subset Y$ is open if and only if $f^{-1}(O)$ is open in X.

1. Show that a continuous map $f:X\to Y$ is a quotient map if there exists a continuous map $g:Y\to X$ such that the composition $fg:Y\to Y$ is the identity map. Apply this to show that a projection map $X\times Y\to X$, $(x,y)\mapsto x$, is a quotient map.

2. Show that the composition of two quotient maps is a quotient map.

3. Show that there exist quotient maps $D^2\to S^1$, $S^2\to D^2$, $S^2\to S^1$, $S^2\to S^1\times S^1$, and $S^1\times S^1\to S^2$. [You do not need to give formulas for these maps, a precise geometric description will be sufficient. Careful proofs of continuity are not required, but the quotient maps do have to be continuous. Remember that $S^1\times S^1$ is a torus.]

4. (a) Show that there is a quotient map $(0,1)\to[0,1]$, but not $[0,1]\to(0,1)$.
(b) If C is the Cantor set, show there is a quotient map $C\to[0,1]$ but not $[0,1]\to C$.

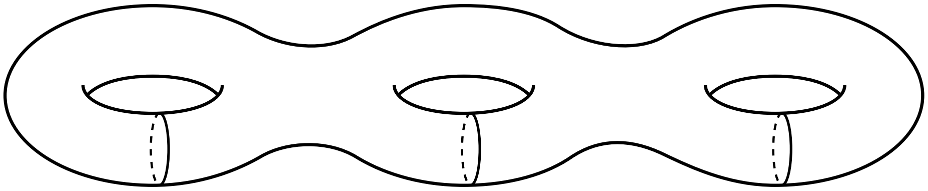
Surfaces

In this section we finally get to make a detailed study of a particular class of spaces that are important in several branches of mathematics, surfaces. Our point of view is purely topological, so some of the more refined structural questions about surfaces that are studied in these other branches of mathematics, such as analytic structure and metric properties like curvature, are ignored here. It is just the basic topology that we are interested in.

Definition. A *surface* is a Hausdorff space X such that each point $x \in X$ has an open neighborhood homeomorphic to $\mathbb{R}^2$, or equivalently, an open disk in $\mathbb{R}^2$.

Certainly $\mathbb{R}^2$ itself is a surface by this definition, or any open set in $\mathbb{R}^2$. A compact example is S^2, since every open hemisphere in S^2, the region on one side of a plane in $\mathbb{R}^3$ through the center of the sphere, is homeomorphic to an open disk. Another compact example is the torus $S^1 \times S^1$. The projective plane and the Klein bottle are two more examples. For the projective plane one can obtain open neighborhoods homeomorphic to a disk by taking the images of open hemispheres in S^2 under the natural quotient map $S^2 \to P^2$.

The torus is the first in an infinite sequence of compact surfaces, the n-holed tori. Here is a picture for $n = 3$:



More generally one can define an n*-dimensional manifold* to be a Hausdorff space in which each point has an open neighborhood homeomorphic to $\mathbb{R}^n$. The Hausdorff condition is included to exclude strange spaces that would otherwise have to be considered as manifolds too. For example, when $n = 1$ the quotient space of $\mathbb{R} \times \{0, 1\}$ obtained by identifying $(x, 0)$ with $(x, 1)$ for all $x > 0$ has the property that each point of the quotient has a neighborhood homeomorphic to $\mathbb{R}$.

In a manifold every point has a path-connected neighborhood, so by a Proposition in Chapter 2 we know that path-components are the same as connected components, and these components are open sets. So each component is itself a manifold that is connected. A compact manifold can have only finitely many components since the different components form an open cover of the manifold.

We will focus mainly on compact surfaces in this chapter, with the goal of obtaining a complete list of all of them. Before starting on this, however, let us digress to derive the analogous but simpler result for 1-dimensional manifolds:

Theorem. *A compact connected 1-dimensional manifold is homeomorphic to S^1.*

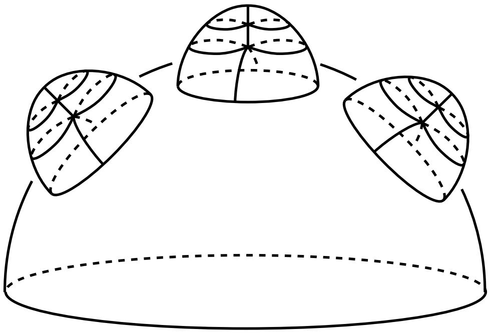
Proof. Let X be a compact connected 1-dimensional manifold. Compactness implies that there is a finite cover of X by open subsets $U_1, \cdots, U_n$ homeomorphic to open intervals. We must have $n \geq 2$ since a single open interval is not compact. We may assume there are no inclusions $U_i \subset U_j$ for $i \neq j$. Since X is connected there must exist two of these sets U_i and U_j that intersect. By examining how U_i and U_j can intersect we will show that either (a) $U_i \cup U_j$ is homeomorphic to S^1 or (b) $U_i \cup U_j$ is homeomorphic to an open interval. In the latter case we can replace U_i and U_j by this open interval to produce an open cover with fewer sets, so by induction this will leave only case (a) to consider. In case (a) the subset $U_i \cup U_j$ of X is open since it is a union of two open sets, and it is closed since it is a compact subspace of a Hausdorff space, being homeomorphic to S^1. So if X is connected we would have $X = U_i \cup U_j$ homeomorphic to S^1, finishing the proof.

It remains to show that if $U_i \cap U_j$ is nonempty then $U_i \cup U_j$ is homeomorphic to either S^1 or an open interval. We know that $U_i \cap U_j$ is open in X and hence in U_i which is homeomorphic to an open interval, so $U_i \cap U_j$ is a union of disjoint open intervals in U_i. Let A be one of these intervals. Then A is an open connected set in U_j, so since U_j is homeomorphic to an open interval, A is also an open interval in U_j. We claim that A must be an end interval of U_i, or in other words, if we choose a homeomorphism between U_i and an interval (a, b) then this homeomorphism takes A to an interval (a, c) or (c, b). For suppose A corresponds to an interval (c, d) with $a < c < d < b$. We know that A is not all of U_j since U_j is not contained in U_i, so this means that at one end of A there are distinct points $x_i \in U_i$ and $x_j \in U_j$. However, this violates the Hausdorff property of X since every neighborhood of x_i meets every neighborhood of x_j in points in A. This contradiction shows that A must be an end interval of U_i, and the same reasoning shows that it must also be an end interval of U_j. Moreover, at one end of A we must have a point of U_i and at the other end a point of U_j, otherwise the Hausdorff property would again be violated. Since U_i and U_j each have only two ends, this means that $U_j \cap U_j$ can consist of at most two intervals. If there are two intervals we have option (a) and if there is only one interval we have option (b). □

By pushing this reasoning a little farther it would not be hard to show that a noncompact connected 1-dimensional manifold that is the union of a countable collection of open sets homeomorphic to $\mathbb{R}$ must be homeomorphic to $\mathbb{R}$ itself. (The option (b) cannot occur, so it follows that the manifold can be expressed as the union of an increasing sequence of open intervals $U_1 \subset U_2 \subset \cdots$, which implies that the manifold must be homeomorphic to $\mathbb{R}$.) One may ask whether there are noncompact connected 1-dimensional manifolds that are not the union of a countable collection of open sets homeomorphic to $\mathbb{R}$. The answer is that there are, but these Frankensteinian monsters exist only in laboratories, not in nature.

Our aim is to provide an explicit list of all compact connected surfaces, at least for the surfaces that can be built from a finite collection of disjoint polygons by identifying pairs of edges. We call such surfaces *polygonal*. It is in fact true that all compact surfaces are homeomorphic to polygonal surfaces, although we will not prove this here.

The list of surfaces will contain the sphere S^2, the n-hole torus which will write as nT^2, and a sequence of surfaces nP^2 generalizing the projective plane P^2. To construct nP^2 start with the sphere S^2, then remove the interiors of n disjoint closed disks in S^2 and attach a cross-cap to the boundary circle of each of the resulting holes by forming a quotient space in which the boundary circle of each cross-cap is identified with a boundary circle of a hole:



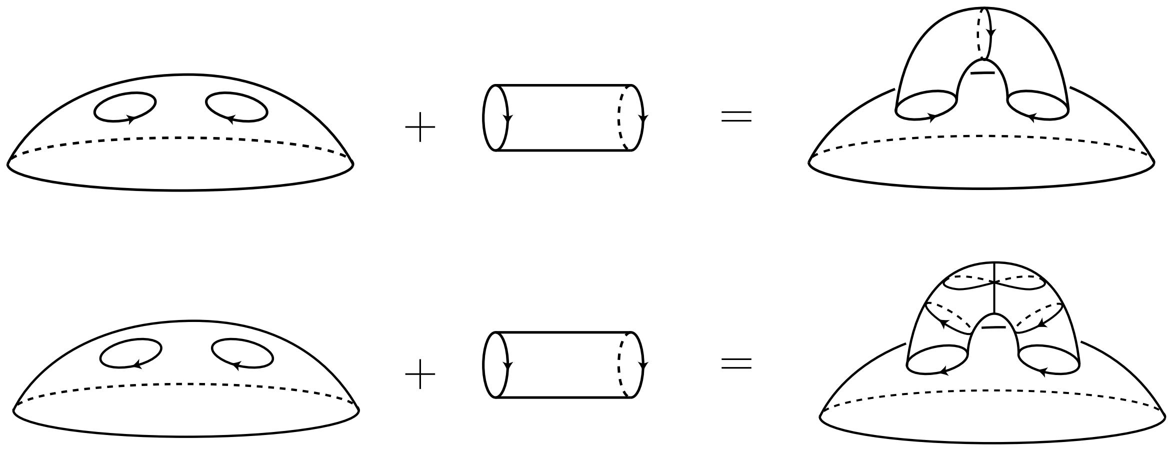
When $n = 1$ this gives P^2. Since P^2 can also be obtained from a disk D^2 by identifying antipodal pairs of points in its boundary, a cross-cap can be obtained from an annulus $S^1 \times [0,1]$ by identifying antipodal pairs of points in one of its boundary circles. It follows that nP^2 could also be obtained by taking a sphere, removing the interiors of n disjoint closed disks, then identifying pairs of antipodal points on each of the resulting boundary circles.

Theorem. *Every compact connected polygonal surface is homeomorphic to either the sphere S^2, an n-hole torus nT^2, or a sphere with n crosscaps nP^2.*

One might wonder where the Klein bottle K^2 is in this list. As we will see, K^2 is homeomorphic to $2P^2$.

Proof. Let S be a polygonal surface, obtained by identifying pairs of edges of some finite set of polygons. If there are two or more polygons and S is connected, there must be a pair of edges in two different polygons that are identified. If we identify these two edges, the two polygons become a single polygon, reducing the the total number of polygons needed to construct S. After repeating this reduction a finite number of times we may assume S is obtained by identifying the edges of a single polygon.

The plan of the proof is to identify one pair of edges at a time and to see inductively what the resulting surface becomes after each such identification. What we will show by induction is that at each stage the surface is homeomorphic to one of a family of *standard models* which we call 'spheres with handles, cross-handles, cross-caps, and holes'. To construct one of these standard models, we start with a sphere with the interiors of a finite set of disjoint closed disks removed. These will be taken to be round disks bounded by actual circles and to lie all in a row on the sphere. After the interiors of the disks are removed we have a sphere with holes. We allow cross-caps to be attached to some of the holes, and we allow cylinders to be attached to certain pairs of adjacent holes. There are two essentially different ways to attach such a cylinder, depending on choices of orientations. These are shown in the following figure:



In the first case we say we are *attaching a handle*, and in the second case we say we are *attaching a cross-handle*. Just as attaching a cross-cap is equivalent to identifying pairs of antipodal points in the boundary of a hole, attaching a handle or cross-handle is equivalent to identifying the boundary circles of two holes.

After we have shown that the given surface S is homeomorphic to one of the standard models, the rest of the proof will consist of showing:

(A) Attaching a cross-handle is equivalent to attaching two cross-caps, i.e., the two surfaces obtained by attaching a cross-handle or a pair of cross-caps are homemorphic.

(B) If there is at least one cross-cap already present, then attaching a handle is equivalent to attaching a cross-handle.

By applying (A) repeatedly we can replace all cross-handles by cross-caps, so we may assume S is a sphere with some number (perhaps zero) of handles and some number (perhaps zero) of cross-caps attached. Then if there is at least one cross-cap we can apply (B) to eliminate all handles, first replacing them by cross-handles and then by pairs of cross-caps using (A). The result will be that we have either a sphere, a sphere with handles, or a sphere with cross-caps, as stated in the Theorem.

Exercises

1. (a) Show that a map $f:X\to Y$ is continuous if there is a finite cover of X by closed sets A_i such that the restriction of f to each A_i is continuous. (b) Would this also be true for an infinite cover of X by closed sets?

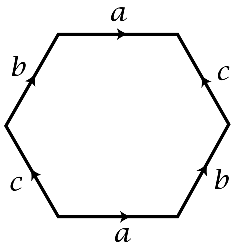
2. Suppose we are given numbers $0 < p_1 < \cdots < p_n < 1$ and $0 < q_1 < \cdots < q_n < 1$. Show that there is a homeomorphism $f:[0,1]\to[0,1]$ with $f(0)=0$, $f(1)=1$, and $f(p_i)=q_i$ for each i.

3. (a) If we are given points p and q in a closed disk $D^2 \subset \mathbb{R}^2$, show that there is a homeomorphism $f:D^2\to D^2$ with $f(p)=q$ and $f(x)=x$ for each $x\in\partial D^2$.

(b) Using part (a) show that for any two points p and q in a connected surface S there is a homeomorphism $f:S\to S$ with $f(p)=q$. [One approach: For fixed p, show that the set of q's for which this is true is both open and closed.]

4. Show that every homeomorphism $\partial D^2\to\partial D^2$ is the restriction of a homeomorphism $D^2\to D^2$. [One possible approach: Think of D^2 as the quotient space of $S^1\times[0,1]$ obtained by identifying $S^1\times\{0\}$ to a point.]

5. Show that the surface obtained by identifying opposite edges of a hexagon as shown in the figure is a torus. (A proof by pictures will be sufficient.)



6. Show that if we start with a finite set of disjoint polygons and identify all their edges in pairs in any way, preserving given orientations of these edges, then the resulting space is always a surface. (Don't forget the Hausdorff property.)