Lecture 1: August 23

Introduction. Topology grew out of certain questions in geometry and analysis about 100 years ago. As Wikipedia puts it, "the motivating insight behind topology is that some geometric problems depend not on the exact shape of the objects involved, but rather on the way they are put together. For example, the square and the circle have many properties in common: they are both one dimensional objects (from a topological point of view) and both separate the plane into two parts, the part inside and the part outside." In other words, topology is concerned with the qualitative rather than quantitative aspects of geometric objects.

The fundamental objects in topology are "topological spaces" and "continuous functions"; both were defined in more or less their current form by Felix Hausdorff in 1914. Like the concept of a group in algebra, topological spaces are very useful for unifying different parts of mathematics: they show up naturally in analysis, geometry, algebra, etc. Most mathematicians therefore end up using both ideas and results from topology in their research. On the other hand, topologist nowadays do not study all possible topological spaces – instead, they focus on specific classes such as 3-dimensional manifolds.

In the course, we will look at the most important definitions and results from basic point set topology and elementary algebraic topology. Our textbook will be the second edition of *Topology* by James Munkres, but I will not present things in exactly the same order. Most of the homework questions, however, will be from the textbook.

Metric spaces. The goal of today's class is to define topological spaces. Since it took people some time to find a good definition, let us try to retrace at least a small portion of this process. One concern of 19th century mathematics was to create rigorous foundations for analysis. This lead to the study of continuous and differentiable functions on subsets of the real line $\mathbb{R}$ and of Euclidean space $\mathbb{R}^n$. Here the notion of "distance" between points plays an important role: for example, a function $f\colon \mathbb{R} \to \mathbb{R}$ is continuous if, for every $x \in \mathbb{R}$ and every real number $\varepsilon > 0$, one can find another real number $\delta > 0$ such that

$$|f(y) - f(x)| < \varepsilon \text{ for every } y \in \mathbb{R} \text{ with } |y - x| < \delta.$$

By abstracting from the properties of distance in Euclidean space, people arrived at the idea of a "metric space".

Definition 1.1. Let X be a set. A *metric* on X is a function $d\colon X \times X \to \mathbb{R}$ with the following three properties:

(a) One has $d(x, y) \geq 0$ for every $x, y \in X$, with equality if and only if $x = y$.
(b) d is symmetric, meaning that $d(x, y) = d(y, x)$.
(c) The triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for every $x, y, z \in X$.

The pair (X, d) is then called a *metric space*.

The name of the triangle inequality comes from the fact that, in a triangle in Euclidean space, the length of each side is smaller than the sum of the lengths of the two other sides. Drawing pictures in the plane can be useful to visualize what is going on – but keep in mind that things like "straight line" or "triangle" do not make actually make sense in a general metric space. The only notions that make sense are those that can expressed in terms of distances between points. One such

notion, which is also useful in Euclidean space, is that of an "open ball": if $x_0 \in X$ is a point, and $r > 0$ a positive real number, the *open ball* of radius r and center x_0 is the set

$$B_r(x_0) = \big\{\, x \in X \;\big|\; d(x_0, x) < r \,\big\}.$$

We can get some idea of what a given metric space looks like by visualizing open balls of different radii. Here are some examples of metric spaces:

Example 1.2. Euclidean space $\mathbb{R}^n$ with the usual notion of distance. Denote the points of $\mathbb{R}^n$ in coordinates by $x = (x_1, x_2, \ldots, x_n)$, and define the length

$$\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}.$$

Then the distance between two points $x, y \in \mathbb{R}^n$ is given by

$$d(x, y) = \|x - y\| = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

It may seem pretty obvious that this satisfies the axioms for a metric, but let us make sure. In the first condition, $d(x, y) \geq 0$ is clear from the definition; moreover, $d(x, y) = 0$ if and only if $x_i - y_i = 0$ for every $i = 1, \ldots, n$ if and only if $x = y$. The second condition is also clear since $(x_i - y_i)^2 = (y_i - x_i)^2$.

Checking that the third condition holds requires a little bit more work. We first prove the following inequality for lengths:

$$\|x + y\| \leq \|x\| + \|y\|. \tag{1.3}$$

After taking the square and expanding, we get

$$\|x + y\|^2 = (x + y) \cdot (x + y) = \|x\|^2 + 2x \cdot y + \|y\|^2,$$

where $x \cdot y = x_1 y_1 + \cdots + x_n y_n$ is the dot product. From the Cauchy-Schwarz inequality in analysis, we obtain

$$x \cdot y \leq \|x\| \|y\|,$$

and therefore

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 = \big(\|x\| + \|y\|\big)^2,$$

which proves (1.3). Returning to the third condition, we now have

$$d(x, z) = \|x - z\| = \big\|(x - y) + (y - z)\big\| \leq \|x - y\| + \|y - z\| = d(x, y) + d(y, z),$$

and so the triangle inequality holds and d is a metric.

Example 1.4. Another metric on $\mathbb{R}^n$ is given by setting

$$d(x, y) = \max_{1 \leq i \leq n} |x_i - y_i|.$$

Unlike before, it takes almost no effort to verify all three axioms. The "open balls" in this metric are now actually open cubes.

Example 1.5. Let $X \subseteq \mathbb{R}^2$ be the union of all the vertical lines $x_1 = n$ and all the horizontal lines $x_2 = n$, for $n \in \mathbb{Z}$. The "taxicab metric" on X is defined by setting

$$d(x, y) = |x_1 - y_1| + |x_2 - y_2|.$$

Here is the proof of the triangle inequality:

$$\begin{aligned} d(x, z) &= |x_1 - z_1| + |x_2 - z_2| \\ &\leq |x_1 - y_1| + |y_1 - z_1| + |x_2 - y_2| + |y_2 - z_2| = d(x, y) + d(y, z). \end{aligned}$$

Example 1.6. Another interesting example is the "railroad metric" on $\mathbb{R}^2$. Choose a point $P \in \mathbb{R}^2$ in the plane, and define

$$d(x,y) = \begin{cases} \|x-y\| & \text{if the three points } x, y, P \text{ are collinear,} \\ \|x-P\| + \|y-P\| & \text{otherwise.} \end{cases}$$

To see the analogy with railroads, think of P as being the capital city of a country, in which all railroad lines go through the capital. I will leave it as an exercise to show that this defines a metric.

Example 1.7. On an arbitrary set X, one can define the trivial metric

$$d(x,y) = \begin{cases} 0 & \text{if } x = y, \\ 1 & \text{if } x \neq y. \end{cases}$$

In this case, open balls of radius $0 < r < 1$ consist of only one point.

The usual ε-δ-definition of continuity carries over to the setting of metric spaces. Suppose that (X, d_X) and (Y, d_Y) are two metric spaces.

Definition 1.8. A function $f\colon X \to Y$ is said to be *continuous* if, for every point $x \in X$ and every real number $\varepsilon > 0$, one can find a real number $\delta > 0$ such that

$$d_Y\big(f(x), f(x')\big) < \varepsilon \text{ for every } x' \in X \text{ with } d_X(x,x') < \delta.$$

More graphically, the condition says that f should map the entire open ball $B_\delta(x)$ into the open ball $B_\varepsilon\big(f(x)\big)$. Equivalently, we can look at the preimage

$$f^{-1}\Big(B_\varepsilon\big(f(x)\big)\Big) = \big\{\, x' \in X \;\big|\; f(x') \in B_\varepsilon\big(f(x)\big) \,\big\},$$

and the condition is that it should contain an open ball of some radius $\delta > 0$ around the point x. Sets that contain an open ball around any of their points are called "open"; this is the same use of the word open as in "open intervals". The precise definition is the following.

Definition 1.9. Let X be a metric space. A subset $U \subseteq X$ is called *open* if, for every point $x \in U$, there is some $r > 0$ such that $B_r(x) \subseteq U$.

Note that the empty set is considered to be open: the condition in the definition is vacuous in that case. It is also obvious that X itself is always open. The following lemma shows that open balls – as defined above – are indeed open.

Lemma 1.10. *In a metric space X, every open ball $B_r(x_0)$ is an open set.*

Proof. By picture. If $x \in B_r(x_0)$ is any point, then $d(x, x_0) < r$, and so the quantity $\delta = r - d(x, x_0)$ is positive. Intuitively, δ is the distance from the point x_0 to the boundary of the ball. Now $B_\delta(x) \subseteq B_r(x_0)$; indeed, if $y \in B_\delta(x)$, then we have

$$d(y, x_0) \leq d(y, x) + d(x, x_0) < \delta + d(x, x_0) = r$$

by virtue of the triangle inequality. □

It is clear from the definition that if $\{U_i\}_{i \in I}$ is any family of open subsets of X, indexed by some set I, then the union

$$\bigcup_{i \in I} U_i = \big\{\, x \in X \;\big|\; x \in U_i \text{ for some } i \in I \,\big\}$$

is again open. Similarly, given finitely many open subsets $U_1, \dots, U_n \subseteq X$, the intersection

$$U_1 \cap \dots \cap U_n = \big\{\, x \in X \;\big|\; x \in U_i \text{ for every } i = 1, \dots, n \,\big\}$$

is also open. In fact, if $x \in U_1 \cap \dots \cap U_n$, then $x \in U_i$; but U_i is open, and so $B_{r_i}(x) \subseteq U_i$ for some $r_i > 0$. Now if we set $r = \min(r_1, \dots, r_n)$, then

$$B_r(x) \subseteq U_1 \cap \dots \cap U_n,$$

proving that the intersection is again an open set.

Open sets can be used to give a criterion for continuity that does not depend on the actual values of the metric.

Proposition 1.11. *Let $f\colon X \to Y$ be a function between two metric spaces. Then f is continuous if and only if $f^{-1}(U)$ is open for every open subset $U \subseteq Y$.*

Proof. The proof is straightforward. Suppose first that f is continuous. Given an open set $U \subseteq Y$, we need to show that the preimage $f^{-1}(U)$ is again open. Take an arbitrary point $x \in f^{-1}(U)$. Since $f(x) \in U$, and U is open, we can find $\varepsilon > 0$ with $B_\varepsilon\big(f(x)\big) \subseteq U$. By definition of continuity, there exists $\delta > 0$ such that

$$f\big(B_\delta(x)\big) \subseteq B_\varepsilon\big(f(x)\big) \subseteq U;$$

but then $B_\delta(x) \subseteq f^{-1}(U)$, and so $f^{-1}(U)$ is an open set.

To prove the converse, suppose that f satisfies the condition in the statement. Given $x_0 \in X$ and $\varepsilon > 0$, the open ball $B_\varepsilon\big(f(x_0)\big)$ is an open subset of Y by Lemma 1.10; its preimage

$$f^{-1}\Big(B_\varepsilon\big(f(x_0)\big)\Big)$$

is therefore an open subset of X. Since it contains the point x_0, it has to contain an open ball around x_0; but this means exactly that

$$f\big(B_\delta(x_0)\big) \subseteq B_\varepsilon\big(f(x_0)\big)$$

for some $\delta > 0$. In other words, f is continuous. □

The proposition shows that we can decide whether or not a function is continuous without knowing the metric; all we have to know is which subsets of X and Y are open. This makes continuity a topological notion, in the sense we talked about at the beginning of class.

Topological spaces. We now come to the definition of topological spaces, which are the basic objects in topology. Rather than by a metric (which is something quantitative), a topological space is described by giving a collection of "open sets" (which is something qualitative). These "open sets" should behave in the same way as open sets in a metric space with respect to taking unions and intersections, and so we use those properties as axioms.

Definition 1.12. Let X be a set. A *topology* on X is a collection $\mathscr{T}$ of subsets of X with the following three properties:

(a) $\emptyset \in \mathscr{T}$ and $X \in \mathscr{T}$.

(b) If $\{U_i\}_{i \in I}$ is a family of subsets of X with $U_i \in \mathscr{T}$ for every $i \in I$, then

$$\bigcup_{i \in I} U_i \in \mathscr{T}.$$

(c) If $U \in \mathscr{T}$ and $V \in \mathscr{T}$, then $U \cap V \in \mathscr{T}$.

The sets in $\mathscr{T}$ are called *open*, and the pair $(X, \mathscr{T})$ is called a *topological space*.

By induction, the property in the third condition can easily be extended to all finite intersections: if $U_1, \ldots, U_n \in \mathscr{T}$, then also $U_1 \cap \cdots \cap U_n \in \mathscr{T}$.

Example 1.13. Every metric space (X, d) is naturally a topological space: the so-called *metric topology* consists of all subsets that are open in the sense of Definition 1.9. We have already checked that all three conditions in the definition are satisfied. Note that different metric spaces can give rise to the same topological space.

We will see many additional examples of topological spaces next time.

Lecture 2: August 30

Today, we are going to look at many additional examples of topological spaces, to become familiar with the definition from last time. Let me first point out that Hausdorff's original definition contained the following additional condition, known as the *Hausdorff axiom*.

Definition 2.1. A topological space $(X, \mathscr{T})$ is said to be *Hausdorff* if, for every pair of distinct points $x, y \in X$, there are open sets $U, V \in \mathscr{T}$ with $x \in U$, $y \in V$, and $U \cap V = \emptyset$.

One often says that the two points x and y can be "separated by open sets". The metric topology on a metric space (X, d) is always Hausdorff: if $x, y \in X$ are two distinct points, then $d(x, y) > 0$; now the open balls of radius $r = d(x, y)/2$ around x and y are disjoint open sets separating x and y. (Indeed, if there was a point $z \in B_r(x) \cap B_r(y)$, then we would have $d(x, z) < r$ and $d(y, z) < r$; by the triangle inequality, this would mean that $2r = d(x, y) \leq d(x, z) + d(z, y) < 2r$, which is absurd.) Since most of our intuition is derived from metric spaces such as $\mathbb{R}^n$, the Hausdorff axiom looks very natural. The reason for not making it part of the definition nowadays is that certain classes of topologies – most notably the ones used in algebra – do not satisfy the Hausdorff axiom.

Now on to some examples of topological spaces. We first observe that any set X can be made into a topological space in the following way.

Example 2.2. Let X be a set. The *trivial topology* on X is the topology $\{\emptyset, X\}$; in view of the conditions, it is the smallest possible topology. The *discrete topology* on X is the topology in which every subset of X is open; it is the largest possibly topology. Neither of these is very interesting.

Here is a small example of a topological space where the Hausdorff axiom does not hold.

Example 2.3. The *Sierpiński space* is the two-element set $\{0, 1\}$, with topology given by

$$\big\{\emptyset, \{1\}, \{0, 1\}\big\}.$$

It is not a Hausdorff space, because the two points 0 and 1 cannot be separated by open sets – in fact, the only open set containing the point 0 is $\{0, 1\}$.

Every subset of a topological space can itself be made into a topological space; this is similar to the fact that every subset of a metric space is again a metric space.

Example 2.4. Let $(X, \mathscr{T})$ be a topological space. Given a subset $Y \subseteq X$, we can put a topology on Y by intersecting the open sets in $\mathscr{T}$ with the subset Y. More precisely, the *subspace topology* on Y is defined to be

$$\mathscr{T}_Y = \big\{\, U \cap Y \;\big|\; U \in \mathscr{T} \,\big\}.$$

You will easily be able to verify that this really is a topology. Note that unless $Y \in \mathscr{T}$, the sets in $\mathscr{T}_Y$ are not usually open in X; to avoid confusion, people sometimes use the expression "open relative to Y" for the sets in $\mathscr{T}_Y$.

So for example, the sphere

$$\mathbb{S}^2 = \big\{\, (x, y, z) \in \mathbb{R}^3 \;\big|\; x^2 + y^2 + z^2 = 1 \,\big\}$$

is a topological space (in the subspace topology coming from $\mathbb{R}^3$); the Cantor set $C \subseteq \mathbb{R}$ is a topological space (in the subspace topology coming from $\mathbb{R}$).

The notion of basis. In the examples above, we described each topological space by saying which sets belonged to the topology. Since this can be a little cumbersome, people often use a more efficient way of presenting this information: a basis. Here is the definition.

Definition 2.5. Let $(X, \mathscr{T})$ be a topological space. A *basis* for the topology is a collection $\mathscr{B}$ of subsets of X with two properties: (1) Every set in $\mathscr{B}$ is open, hence $\mathscr{B} \subseteq \mathscr{T}$. (2) Every set in $\mathscr{T}$ can be written as the union of sets in $\mathscr{B}$.

This usage of the word "basis" is different from the one in linear algebra, because a given open set is allowed to be a union of sets in $\mathscr{B}$ in many different ways.

Example 2.6. Let (X, d) be a metric space. The collection of all open balls

$$\mathscr{B} = \big\{ B_r(x_0) \;\big|\; x_0 \in X \text{ and } r > 0 \big\}$$

is a basis for the metric topology: by definition, every open set can be written as a union of open balls; conversely, every open ball is an open set. Someone raised the question of whether the empty set is a union of open balls. The answer is yes: it is the union of zero many open balls. Another (and more efficient) choice of basis would be

$$\mathscr{B}' = \big\{ B_r(x_0) \;\big|\; x_0 \in X \text{ and } r \in \mathbb{Q} \cap (0, \infty) \big\},$$

using only those open balls whose radius is a rational number.

Example 2.7. Let X be a set. The collection of all one-point subsets of X is a basis for the discrete topology.

If we are given only the basis, we can recover the topology by taking all possible unions of sets in $\mathscr{B}$. Here the following notation is convenient: given a collection $\mathscr{C}$ of subsets of X, define

$$\bigcup \mathscr{C} = \bigcup_{U \in \mathscr{C}} U \subseteq X$$

to be the union of all the members of $\mathscr{C}$. If $\mathscr{C}$ is empty, let us agree that $\bigcup \mathscr{C}$ is the empty set. The following result allows us to specify a topology on a set X in terms of a basis.

Proposition 2.8. *Let $\mathscr{B}$ be a collection of subsets of X with the following properties: (1) $\bigcup \mathscr{B} = X$; (2) for every $U, V \in \mathscr{B}$, the intersection $U \cap V$ can be written as a union of elements of $\mathscr{B}$. Then*

$$\mathscr{T}(\mathscr{B}) = \left\{ \bigcup \mathscr{C} \;\middle|\; \mathscr{C} \subseteq \mathscr{B} \right\}$$

is a topology on X, and $\mathscr{B}$ is a basis for this topology.

Proof. Let us verify that $\mathscr{T}(\mathscr{B})$ is a topology on X. Since $X = \bigcup \mathscr{B}$ and $\emptyset = \bigcup \emptyset$, both the empty set and X itself are open. Moreover, any set in $\mathscr{B}$ is open, because $U = \bigcup\{U\}$. It is obvious from the definition that arbitrary unions of open sets are again open. To check the condition on intersections, observe first that the intersection of any two sets in $\mathscr{B}$ is open: by assumption, it can be written as a union of sets in $\mathscr{B}$. Now let

$$U_1 = \bigcup \mathscr{C}_1 = \bigcup_{V \in \mathscr{C}_1} V \quad \text{and} \quad U_2 = \bigcup \mathscr{C}_2 = \bigcup_{W \in \mathscr{C}_2} W$$

be two arbitrary open sets; then

$$U_1 \cap U_2 = \bigcup_{\substack{V \in \mathscr{C}_1 \\ W \in \mathscr{C}_2}} V \cap W$$

is a union of open sets, and therefore open. This shows that $\mathscr{T}(\mathscr{B})$ is a topology; that $\mathscr{B}$ is a basis is obvious, because every set in $\mathscr{B}$ is open, and every open set is a union of sets in $\mathscr{B}$. □

The topology $\mathscr{T}(\mathscr{B})$ is sometimes called the topology generated by the basis $\mathscr{B}$. From now on, we will usually describe topologies in terms of bases.

Ordered sets and the order topology. A nice class of examples comes from linearly ordered sets.

Definition 2.9. A relation $<$ on a set X is called a *linear order* if it has the following properties:

(a) For any pair of $x, y \in X$ with $x \neq y$, either $x < y$ or $y < x$.
(b) The relation $x < x$ is never satisfied for any $x \in X$.
(c) If $x < y$ and $y < z$, then $x < z$.

Given a linear order on X, we define $x \leq y$ to mean "$x < y$ or $x = y$".

The three conditions together imply that, for every pair of elements $x, y \in X$, exactly one of the three relations

$$x < y, \quad y < x, \quad x = y$$

holds. We can therefore visualize a linear order by thinking of the elements of X as being lined up in increasing order from left to right.

Example 2.10. The usual order relation $x < y$ on $\mathbb{R}$ is a linear order.

Example 2.11. The set $\{A, B, \ldots, Z\}$ of all uppercase letters is linearly ordered by the alphabetic order relation.

Example 2.12. The dictionary order on $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the following relation:

$$(x_1, x_2) < (y_1, y_2) \Leftrightarrow x_1 < y_1, \text{ or } x_1 = y_1 \text{ and } x_2 < y_2$$

You can easily check that this is a linear order.

Given a linear order $<$ on a set X, we can define *open intervals*

$$(a, b) = \{\, x \in X \mid a < x < b \,\}$$

and *open rays*

$$(a, \infty) = \{\, x \in X \mid a < x \,\}, \quad (-\infty, a) = \{\, x \in X \mid x < a \,\}$$

just as in the case of the real numbers. As the word "open" suggests, they form the basis for a topology on X, the so-called *order topology*.

Proposition 2.13. *Let $<$ be a linear order on a set X, and let $\mathscr{B}$ be the collection of all open intervals, all open rays, and X itself. Then $\mathscr{B}$ is a basis for a topology on X.*

Proof. We have to check that the assumptions of Proposition 2.8 are satisfied. Since $X \in \mathscr{B}$, we have $\bigcup \mathscr{B} = X$. Moreover, it is easy to see that the intersection of any two sets in $\mathscr{B}$ is again in $\mathscr{B}$. According to Proposition 2.8, $\mathscr{B}$ is a basis for a topology on X. After class, somebody asked me whether we really needs the set X in $\mathscr{B}$. The answer is yes, but only when X has exactly one element; as soon as X has at least two elements $a < b$, one has $X = (-\infty, b) \cup (a, \infty)$. □

Example 2.14. Since $\mathbb{R}$ is both a metric space and a linearly ordered set, it has two topologies: the metric topology and the order topology. In fact, both have the same open sets, and are therefore equal. Let us prove this. If a set U is open in the metric topology, then it is a union of open balls; but every open ball $B_r(x_0)$ is also an open interval $(x_0 - r, x_0 + r)$, and so U is open in the order topology. Conversely, all open intervals and open rays are clearly open sets in the metric topology, and so every open set in the order topology is also open in the metric topology.

Example 2.15. What do open sets look like in the dictionary topology on $\mathbb{R}^2$?

The product topology. Let X and Y be two topological spaces. Their cartesian product

$$X \times Y = \{\, (x, y) \mid x \in X \text{ and } y \in Y \,\}$$

can again be made into a topological space in a natural way.

Proposition 2.16. *The collection of all sets of the form $U \times V$, where U is an open subset of X and V is an open subset of Y, is a basis for a topology on $X \times Y$.*

Proof. Note that $X \times Y$ belongs to our collection of sets, and that we have

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2).$$

The assertion therefore follows from Proposition 2.8. □

The topology in the proposition is called the *product topology* on $X \times Y$.

Example 2.17. Consider the real line $\mathbb{R}$ with its standard topology. The product topology on $\mathbb{R} \times \mathbb{R}$ is the same as the metric topology; this can be proved in the same way as in Example 2.14

Lemma 2.18. *If $\mathscr{B}$ is a basis for the topology on X, and if $\mathscr{C}$ is a basis for the topology on Y, then*

$$\mathscr{B} \times \mathscr{C} = \{\, B \times C \mid B \in \mathscr{B} \text{ and } C \in \mathscr{C} \,\}$$

is a basis for the product topology on $X \times Y$.

Proof. By definition, every open set in $X \times Y$ is a union of open sets of the form $U \times V$, with U open in X and V open in Y. It is therefore enough to show that $U \times V$ can be written as a union of sets in $\mathscr{B} \times \mathscr{C}$. Since $\mathscr{B}$ is a basis for the topology on X, we have

$$U = \bigcup_{i \in I} U_i$$

for some collection $\{U_i\}_{i \in I}$ of sets in $\mathscr{B}$; for the same reason,

$$V = \bigcup_{j \in J} V_j$$

for some collection $\{V_j\}_{j\in J}$ of sets in $\mathscr{C}$. This means that

$$U \times V = \bigcup_{\substack{i\in I\\ j\in J}} U_i \times V_j$$

is a union of sets in $\mathscr{B} \times \mathscr{C}$, as required. □

Closed sets, interior, and closure. Here are a few additional definitions that are useful when talking about general topological spaces.

Definition 2.19. Let X be a topological space. A subset $A \subseteq X$ is called *closed* if its complement $X \setminus A$ is open.

The word "closed" is used in the same way as in "closed intervals", and is supposed to mean something like "closed under taking limits"; this usage comes from analysis, where a subset $A \subseteq \mathbb{R}$ is called closed if the limit of every convergent sequence in A also belongs to A. We shall come back to this point at the beginning of next class.

Obviously, $\emptyset = X \setminus X$ and $X = X \setminus \emptyset$ are closed sets – which makes those two sets both open and closed. The basic rules of set theory also imply that arbitrary intersections and finite unions of closed sets are again closed. The reason is that

$$X \setminus \bigcap_{i\in I} A_i = \bigcup_{i\in I} (X \setminus A_i),$$

and that a union of open sets is again open. In fact, one could define topological spaces entirely in terms of closed sets.

Definition 2.20. Let $Y \subseteq X$ be a subset. The *interior* of Y is defined to be

$$\operatorname{int} Y = \bigcup_{\substack{U\subseteq Y\\ \text{open}}} U;$$

it is the largest open subset contained in Y. The *closure* of Y is defined to be

$$\overline{A} = \bigcap_{\substack{A\supseteq Y\\ \text{closed}}} A;$$

it is the smallest closed set containing Y.

Intuitively, when we take the interior of a set, we are throwing away all points that lie at the edge of the set; when we take the closure, we add every possible point of this kind.

Lecture 3: September 1

Let me begin today's class by talking about closed sets again. Recall from last time that a subset $A \subseteq X$ of a topological space is called closed if its complement $X \setminus A$ is open. We also defined the closure of an arbitrary subset $Y \subseteq X$ to be the smallest closed subset containing Y; more precisely,

$$\overline{Y} = \bigcap_{\substack{A \supseteq Y \\ closed}} A$$

is the intersection of all closed sets containing Y. Intuitively, taking the closure means adding all those points of X that lie "at the edge of Y"; the goal is to understand this operation better.

Example 3.1. In a Hausdorff space X, every one-point set $\{x\}$ is closed. We have to convince ourselves that $X \setminus \{x\}$ is an open set. Let $y \in X \setminus \{x\}$ be an arbitrary point. Since X is Hausdorff, we can find two disjoint open sets U and V with $x \in U$ and $y \in V$. Clearly, $V \subseteq X \setminus \{x\}$; this shows that $X \setminus \{x\}$ is a union of open sets, and therefore open.

Example 3.2. Let Y be a subset of a topological space X. Then a set $A \subseteq Y$ is closed in the subspace topology on Y if and only if $A = Y \cap B$ for some closed set $B \subseteq X$. Note that this is not a definition, but a (very easy) theorem. Here is the proof: A is closed relative to Y if and only if $Y \setminus A$ is open relative to Y if and only if $Y \setminus A = Y \cap U$ for some open set $U \subseteq X$ if and only if $A = Y \cap (X \setminus U)$. (The last step is easiest to understand by drawing a picture.) But if U is open, $X \setminus U$ is closed, and so we get our result.

As I mentioned last time, the word "closed" comes from analysis, where it means something like "closed under taking limits". To help make the definition more concrete, let us now discuss the relationship between closed sets and limit points. If $x \in X$ is a point in a topological space, an open set containing x is also called a *neighborhood* of x. We usually think of a neighborhood as being a "small" open set containing x – but unless X is a metric space, this does not actually make sense, because we do not have a way to measure distances.

Definition 3.3. Let Y be a subset of a topological space X. A point $x \in X$ is called a *limit point* of Y if every neighborhood of x contains at least one point of Y other than x itself.

Note that a point $x \in Y$ may be a limit point of Y, provided that there are enough other points of Y nearby. An isolated point, however, is not considered to be a limit point. The following theorem describes the closure operation in terms of limit points.

Theorem 3.4. *Let Y be a subset of a topological space X. The closure of Y is the union of Y and all its limit points.*

Proof. Let us temporarily denote by Y' the union of Y and all its limit points. To prove that $Y' = \overline{Y}$, we have to show two things: $Y' \subseteq \overline{Y}$, and Y' is closed. Since Y' contains Y, this will be enough to give us $Y' = \overline{Y}$.

Let us first prove that $Y' \subseteq \overline{Y}$. Of course, we only have to argue that every limit point x of Y belongs to $\overline{Y}$. The proof is by contradiction: if $x \notin \overline{Y}$, then the open

set $X \setminus \overline{Y}$ is a neighborhood of x, and therefore has to contain some point of Y; but this is not possible because $Y \cap (X \setminus \overline{Y}) = \emptyset$.

Next, let us show that Y' is closed, or in other words, that $X \setminus Y'$ is open. Since a union of open sets is open, it will be enough to prove that for every $x \in X \setminus Y'$, some neighborhood of x is contained in $X \setminus Y'$. Now x is clearly not a limit point of Y, and because of how we defined limit points, this means that some neighborhood U of x does not contain any point of Y other than possibly x itself. Since we also know that $x \notin Y$, we deduce that $U \cap Y = \emptyset$. But then no point of U can be a limit point of Y (because U is open), and so $U \subseteq X \setminus Y'$. □

Sequences and limits. In metric spaces, the property of being closed can also be expressed in terms of convergent sequences. Let X be a metric space, and suppose that $x_1, x_2, \ldots$ is a sequence of points of X. We say that the sequence *converges* to a point $x \in X$, or that x is the *limit* of the sequence, if for every $\varepsilon > 0$, one can find an integer N such that

$$d(x_n, x) < \varepsilon \quad \text{for every } n \geq N.$$

Note that a sequence can have at most one limit: if x' is another potential limit of the sequence, the triangle inequality implies that

$$d(x, x') \leq d(x, x_n) + d(x_n, x');$$

as the right-hand side can be made arbitrarily small, $d(x, x') = 0$, which means that $x' = x$.

In view of how the metric topology is defined, we can rephrase the condition for convergence topologically: the sequence $x_1, x_2, \ldots$ converges to x if and only if every open set containing x contains all but finitely many of the x_n. This concept now makes sense in an arbitrary topological space.

Definition 3.5. Let $x_1, x_2, \ldots$ be a sequence of points in a topological space. We say that the sequence *converges* to a point $x \in X$ if, for every open set U containing x, there exists $N \in \mathbb{N}$ such that $x_n \in U$ for every $n \geq N$. In that case, x is called a *limit* of the sequence.

You should convince yourself that if x is a limit of a sequence $x_1, x_2, \ldots$, then it is also a limit point of the subset $\{x_1, x_2, \ldots\}$. (*Question*: What about the converse?) Unlike in metric spaces, limits are no longer necessarily unique.

Example 3.6. In the Sierpiński space, both 0 and 1 are limits of the constant sequence $1, 1, \ldots$, because $\{0, 1\}$ is the only open set containing the point 0.

In a Hausdorff space, on the other hand, limits are unique; the proof is left as an exercise.

Lemma 3.7. *In a Hausdorff space X, every sequence of points has at most one limit.*

The following result shows that in a metric space, "closed" really means "closed under taking limits".

Proposition 3.8. *Let X be a metric space. The following two conditions on a subset $Y \subseteq X$ are equivalent:*

(a) Y is closed in the metric topology.

(b) Y *is* sequentially closed*: whenever a sequence* $x_1, x_2, \ldots$ *of points in* Y *converges to a point* $x \in X$*, one has* $x \in Y$*.*

Proof. Suppose first that Y is closed. Let $x_1, x_2, \ldots$ be a sequence of points in Y that converges to a point $x \in X$; we have to prove that $x \in Y$. This is pretty obvious: because Y is closed, the complement $X \setminus Y$ is open, and if we had $x \in X \setminus Y$, then all but finitely many of the x_n would have to lie in $X \setminus Y$, which they don't.

Now suppose that Y is sequentially closed. To prove that Y is closed, we have to argue that $X \setminus Y$ is open. Suppose this was not the case. Because of how we defined the metric topology, this means that there is a point $x \in X \setminus Y$ such that no open ball $B_r(x)$ is entirely contained in $X \setminus Y$. So in each open ball $B_{1/n}(x)$, we can find at least one point $x_n \in Y$. Now I claim that the sequence $x_1, x_2, \ldots$ converges to x: indeed, we have $d(x_n, x) < 1/n$ by construction. Because $x \in X \setminus Y$, this contradicts the fact that Y is sequentially closed. □

This also gives us the following description of the closure: if $Y \subseteq X$ is a subset of a metric space, then the closure $\overline{Y}$ is the set of all limit points of convergent sequences in Y.

Unfortunately, Proposition 3.8 does not generalize to arbitrary topological spaces; you can find an example in this week's homework. Closed sets are always sequentially closed – the first half of the proof works in general – but the converse is not true. What made the second half of the proof work is that every point in a metric space has a *countable neighborhood basis*: for every point $x \in X$, there are countably many open sets $U_1(x), U_2(x), \ldots$, such that every open set containing x contains at least one of the $U_n(x)$. In a metric space, we can take for example $U_n(x) = B_{1/n}(x)$. Topological spaces with this property are said to satisfy the *first countability axiom*. So Proposition 3.8 is true (with the same proof) in every topological space where the first countability axiom holds. If this axiom does not hold in X, then there are simply "too many" open sets containing a point $x \in X$ to be able to describe closed sets in terms of sequences (which are by definition countable).

Note. If X is first countable, then the collection

$$\mathscr{B} = \left\{ U_n(x) \;\middle|\; x \in X \text{ and } n \geq 1 \right\}$$

is a basis for the topology on X. A stronger version of this condition is that X should have a basis consisting of countably many open sets; such spaces are said to satisfy the *second countability axiom*.

Continuous functions and homeomorphisms. As suggested in the first lecture, we define continuous functions by the condition that the preimage of every open set should be open.

Definition 3.9. Let $(X, \mathscr{T}_X)$ and $(Y, \mathscr{T}_Y)$ be two topological spaces. A function $f \colon X \to Y$ is called *continuous* if

$$f^{-1}(U) = \left\{ x \in X \;\middle|\; f(x) \in U \right\} \in \mathscr{T}_X$$

for every $U \in \mathscr{T}_Y$.

If the topology on Y is given in terms of a basis $\mathscr{B}$, then it suffices to check the condition for $U \in \mathscr{B}$; the reason is that

$$f^{-1}\left(\bigcup_{U\in\mathscr{C}} U\right) = \bigcup_{U\in\mathscr{C}} f^{-1}(U).$$

We could have just as well defined continuity used closed sets; the reason is that

$$f^{-1}(Y \setminus A) = \left\{ x \in X \;\middle|\; f(x) \notin A \right\} = X \setminus f^{-1}(A).$$

We have already seen that the topological definition is equivalent to the ε-δ one in the case of metric spaces; so we already know many examples of continuous functions from analysis. To convince ourselves that the topological definition is useful, let us prove some familiar facts about continuous functions in this setting. The first one is that the composition of continuous functions is again continuous.

Lemma 3.10. *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continous, then so is their composition $g \circ f$.*

Proof. Let $U \subseteq Z$ be an arbitrary open set. Since g is continuous, $g^{-1}(U)$ is open in Y; since f is continous,

$$(g \circ f)^{-1}(U) = f^{-1}\big(g^{-1}(U)\big)$$

is open in X. This proves that $g \circ f$ is continuous. □

In analysis, we often encounter functions that are defined differently on different intervals. Here is a general criterion for checking that such functions are continuous.

Proposition 3.11 (Pasting lemma)**.** *Let $X = A \cup B$, where both A and B are closed sets of X. Let $f\colon A \to Y$ and $g\colon B \to Y$ be two continuous functions. If $f(x) = g(x)$ for every $x \in A \cap B$, then the function*

$$h\colon X \to Y, \quad h(x) = \begin{cases} f(x) & \text{if } x \in A, \\ g(x) & \text{if } x \in B \end{cases}$$

is well-defined and continuous on X.

Proof. We can prove the continuity of h by showing that the preimage of every closed set in Y is closed. So let $C \subseteq Y$ be closed. We have

$$\begin{aligned} h^{-1}(C) &= \left\{ x \in X \;\middle|\; h(x) \in C \right\} \\ &= \left\{ x \in A \;\middle|\; f(x) \in C \right\} \cup \left\{ x \in B \;\middle|\; g(x) \in C \right\} = f^{-1}(C) \cup g^{-1}(C). \end{aligned}$$

Now f is continous, and so $f^{-1}(C)$ is closed in the subspace topology on A; but because A is itself closed in X, this means that $f^{-1}(C)$ is also closed in X. The same goes for $g^{-1}(C)$, and so $h^{-1}(C)$ is a closed set. □

Example 3.12. Let us consider the example of a product $X \times Y$ of two topological spaces (with the product topology). Denote by

$$p_1\colon X \times Y \to X \quad \text{and} \quad p_2\colon X \times Y \to Y$$

the projections to the two coordinates, defined by $p_1(x, y) = x$ and $p_2(x, y) = y$. Then both p_1 and p_2 are continuous. This is easy to see: for instance, if $U \subseteq X$ is an open subset, then $p^{-1}(U) = U \times Y$ is open by definition of the product topology.

Proposition 3.13. *A function $f\colon Z \to X \times Y$ is continuous if and only if the two coordinate functions*

$$f_1 = p_1 \circ f\colon Z \to X \quad \textit{and} \quad f_2 = p_2 \circ f\colon Z \to Y$$

are continuous.

Proof. One direction is easy: if f is continous, then f_1 and f_2 are compositions of continous functions, hence continuous. For the other direction, we use the definition. A basis for the product topology is given by sets of the form $U \times V$, with $U \subseteq X$ and $V \subseteq Y$ both open. Then

$$f^{-1}(U \times V) = \left\{ z \in Z \;\middle|\; f_1(z) \in U \text{ and } f_2(z) \in V \right\} = f_1^{-1}(U) \cap f_2^{-1}(V)$$

is the intersection of two open sets, hence open. □

Lecture 4: September 6

Homeomorphisms. As I mentioned in the first lecture, the purpose of topology is to look at qualitative properties of geometric objects that do not depend on the exact shape of an object, but more on how the object is put together. We formalized this idea by definiting topological spaces; but what does it mean to say that two different topological spaces (such as a circle and a square) are really "the same"?

Definition 4.1. Let $f\colon X \to Y$ be a bijective function between topological spaces. If both f and the inverse function $f^{-1}\colon Y \to X$ are continuous, then f is called a *homeomorphism*, and X and Y are said to be *homeomorphic*.

Intuitively, think of X as being made from some elastic material (like a balloon), and think of stretching, bending, or otherwise changing the shape of X without tearing the material. Any Y that you get in this way will be homeomorphic to the original X. Note that the actual definition is both more precise and more general, since we are allowing arbitrary functions.

Suppose that $f\colon X \to Y$ is a homeomorphism. For each open set $U \subseteq X$, we are assuming that its inverse image under $f^{-1}\colon Y \to X$ is open in X; but because f is bijective, this is the same as the image of U under f. In other words, a homeomorphism is a bijective function $f\colon X \to Y$ such that $f(U)$ is open if and only if U is open. We therefore get a bijective correspondence not only between the points of X and Y, but also between the open sets in both topologies. So any question about the topology of X or Y will have the same answer on both sides; we may therefore think of X and Y as being essentially the same topological space.

Example 4.2. The real numbers $\mathbb{R}$ are homeomorphic to the open interval $(0, 1)$. One possible choice of homeomorphism is the function

$$f\colon \mathbb{R} \to (0,1), \quad f(x) = \frac{e^x}{e^x + 1}$$

Both f and the inverse function $f^{-1}(y) = \log(y) - \log(1 - y)$ are continuous.

Example 4.3. Consider the function

$$f\colon [0,1) \to \mathbb{S}^1, \quad f(t) = (\cos t, \sin t)$$

that takes the interval (with the subspace topology from $\mathbb{R}$) to the unit circle (with the subspace topology from $\mathbb{R}^2$). It is bijective and continuous, but not a homeomorphism: $[0, 1/2)$ is open in $[0, 1)$, but its image is not open in $\mathbb{S}^1$.

Example 4.4. Let us classify the letters of the English alphabet

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

up to homeomorphism. Here we think of each letter as being made from line segments in $\mathbb{R}^2$; the topology is the subspace topology. By inspection, there are eight homeomorphism classes, depending on the number of loops and line segments in each letter:

B A R P Q D O C G I J L M N S U V W Z E F T Y H K X

For example, W can be bent to make I, and so the two are homeomorphic. On the other hand, there is no homeomorphism between T and I: if we remove the crossing point, we are left with three intervals in the case of T, but removing one point from I produces at most two intervals. (Think about how one can say this in terms of the topology on each letter.)

Topological manifolds. In the remainder of today's class, I want to introduce three additional examples of topological spaces. The first one is topological manifolds. A manifold is a space X that "locally" looks like Euclidean space: if you sit at any point of X, and only look at points nearby, you may think that you are in $\mathbb{R}^n$. Here is the precise definition.

Definition 4.5. An *n-dimensional topological manifold* is a Hausdorff topological space X with the following property: every point $x \in X$ has a neighborhood that is homeomorphic to an open subset in $\mathbb{R}^n$.

In geometry, people look at other classes of manifolds that are obtained by working with a smaller class of functions. For example, if a function and its inverse function are both differentiable, it is called a diffeomorphism; differentiable manifolds are defined by replacing "homeomorphic" by "diffeomorphic" in the above definition. In algebraic geometry, there is a similar definition with polynomials.

At this point, somebody asked why we need the Hausdorff condition; the answer is that we do not want to allow something like taking two copies of $\mathbb{R}$ and gluing them together along $\mathbb{R} \setminus \{0\}$. (More about this example later on, when we discuss quotient spaces.) Later in the semester, we will show that an open subset in $\mathbb{R}^n$ can never be homeomorphic to an open subset in $\mathbb{R}^m$ for $m \neq n$; this means that the dimension of a manifold really is a well-defined notion.

Example 4.6. The square and the circle are both one-dimensional manifolds; a homeomorphism between them is given by drawing the square inside the circle and projecting one to the other from their common center.

Example 4.7. The n-sphere

$$\mathbb{S}^n = \left\{ (x_0, x_1, \ldots, x_n) \in \mathbb{R}^{n+1} \;\middle|\; x_0^2 + x_1^2 + \cdots + x_n^2 = 1 \right\},$$

with the subspace topology coming from $\mathbb{R}^{n+1}$, is an n-dimensional manifold. Intuitively, this is clear; let me prove it for $n = 2$ by using stereographic projection. The plane $z = -1$ is tangent to the sphere at the south pole; given any point (x, y, z) not equal to the north pole $(0, 0, 1)$, we can see where the line connecting $(0, 0, 1)$ and (x, y, z) intersects the plane $z = -1$. In this way, we get a bijection

$$f \colon \mathbb{S}^2 \setminus \{(0, 0, 1)\} \to \mathbb{R}^2.$$

It is easy to work out the formulas to see that f and its inverse are continuous. The points on the line are parametrized by $(0, 0, 1) + t(x, y, z - 1)$, with $t \in \mathbb{R}$; the intersection point with the plane has

$$1 + t(z - 1) = -1 \quad \text{or} \quad t = \frac{2}{1 - z},$$

which means that

$$f(x, y, z) = \left(\frac{2x}{1 - z}, \frac{2y}{1 - z} \right).$$

One can show in a similar manner that f^{-1} is continuous. Since we can also do stereographic projection from the south pole, every point of $\mathbb{S}^2$ has a neighborhood that is homeomorphic to $\mathbb{R}^2$.

Example 4.8. The implicit function theorem from analysis gives us one way to define manifolds. Suppose that $f \colon \mathbb{R}^2 \to \mathbb{R}$ is a continuously differentiable function. The implicit function theorem says that if $f(x_0, y_0) = 0$, and if the partial derivative

$\partial f/\partial y$ does not vanish at the point (x_0, y_0), then all nearby solutions of the equation $f(x, y) = 0$ are of the form

$$y = \varphi(x)$$

for a continously differentiable function $\varphi\colon (x_0 - \varepsilon, x_0 + \varepsilon) \to \mathbb{R}$ with $\varphi(x_0) = y_0$. This function φ gives us a homeomorphism between a small neighborhood of the point (x_0, y_0) in the set $f^{-1}(0)$ and an open interval in $\mathbb{R}$. This shows that $f^{-1}(0)$ is a one-dimensional manifold, provided that at least one of the two partial derivatives $\partial f/\partial x$ or $\partial f/\partial y$ is nonzero at every point of $f^{-1}(0)$.

Example 4.9. If M_1 and M_2 are manifolds of dimension n_1 and n_2, respectively, then their product $M_1 \times M_2$ (with the product topology) is a manifold of dimension $n_1 + n_2$. The proof is left as an exercise. For instance, the product $\mathbb{S}^1 \times \mathbb{S}^1$ is a two-dimensional manifold called the *torus*.

An important general problem is to classify manifolds (or more general topological spaces) up to homeomorphism. In general, this is only possible if we impose sufficiently many other conditions (such as connectedness or compactness) to limit the class of topological spaces we are looking at. We will come back to this problem later in the semester.

Quotient spaces and the quotient topology. In geometry, it is common to describe spaces by "cut-and-paste" constructions like the following.

Example 4.10. If we start from the unit square and paste opposite edges (with the same orientation), we get the torus. If we start from the closed unit disk in $\mathbb{R}^2$ and collapse the entire boundary into a point, we obtain $\mathbb{S}^2$. To make a Möbius band, we take a strip of paper, twist one end by 180°, and then glue the two ends together. We can make a torus with two holes by taking two copies of the torus, removing a small disk from each, and then pasting them together along the two boundary circles.

In each of these cases, the result should again be a topological space. To formalize this type of construction, we start with a topological space X and an equivalence relation $\sim$ on it; intuitively, $\sim$ tells us which points of X should be glued together. (Recall that an equivalence relation is the same thing as a partition of X into disjoint subsets, namely the equivalence classes; two points are equivalent if and only if they are in the same equivalence class.) What we want to do is to build a new topological space in which each equivalence class becomes just one point. To do this, we let $X/\!\sim$ be the set of equivalence classes; there is an obvious function

$$p\colon X \to X/\!\sim,$$

which takes a point $x \in X$ to the equivalence class containing x. Now there is a natural way to make $X/\!\sim$ into a topological space.

Proposition 4.11. *The collection of sets*

$$\mathscr{T} = \left\{ U \subseteq X/\!\sim \;\middle|\; p^{-1}(U) \text{ is open in } X \right\}$$

defines a topology on $X/\!\sim$, called the quotient topology.

Proof. We have to check that the three conditions in the definition of topology are satisfied. First, $p^{-1}(\emptyset) = \emptyset$ and $p^{-1}(X/\!\sim) = X$, and so both $\emptyset$ and $X/\!\sim$ belong

to $\mathscr{T}$. The conditions about unions and intersections follow from the set-theoretic formulas

$$p^{-1}\left(\bigcup_{i\in I} U_i\right) = \bigcup_{i\in I} p^{-1}(U_i) \quad \text{and} \quad p^{-1}(U\cap V) = p^{-1}(U)\cap p^{-1}(V)$$

and the definition of $\mathscr{T}$. □

With this definition, p becomes a continuous function. In fact, the quotient topology is the largest topology with the property that p is continous. Even when X is Hausdorff, the quotient $X/\!\sim$ is not necessarily Hausdorff.

Example 4.12. Let us go back to the example of the line with two origins, made by gluing together two copies of $\mathbb{R}$ along $\mathbb{R}\setminus\{0\}$. Here we can take $X = \{0,1\}\times\mathbb{R}$, and define the equivalence relation so that $(0,t)\sim(1,t)$ for every $t\neq 0$. Most equivalence classes have two points, namely $\{(0,t),(1,t)\}$ with $t\neq 0$, except for $\{(0,0)\}$ and $\{(1,0)\}$. The quotient space $X/\!\sim$ is not Hausdorff (because the two equivalence classes $\{(0,0)\}$ and $\{(1,0)\}$ cannot be separated by open sets), but every point has a neighborhood homeomorphic to $\mathbb{R}$.

In fact, it is an interesting problem of finding conditions on X and $\sim$ that will guarantee that $X/\!\sim$ is Hausdorff. This does happen in real life: I work in algebraic geometry, but in one of my papers, I had to spend about a page on proving that a certain quotient space was again Hausdorff.

The most useful property of the quotient topology is the following.

Theorem 4.13. *Let $f\colon X\to Y$ be a continous function that is constant on equivalence classes: whenever $x_1\sim x_2$, one has $f(x_1)=f(x_2)$. Then f induces a function $\tilde{f}\colon X/\!\sim\to Y$, which is continuous for the quotient topology on $X/\!\sim$.*

Proof. The proof is left as an exercise. □

Product spaces and the product topology. We have already seen that the product of two topological spaces is again a topological space. Now we want to deal with the general case where we allow an arbitrary (and possibly infinite) number of factors. So let $(X_i,\mathscr{T}_i)$ be a collection of topological spaces, indexed by a (possibly infinite) set I. Consider the cartesian product

$$X = \prod_{i\in I} X_i = \big\{\,(x_i)_{i\in I} \;\big|\; x_i\in X_i \text{ for every } i\in I\,\big\},$$

whose elements are all (generally infinite) families of elements $x_i\in X_i$, one for each $i\in I$. It is not completely obvious that X has any elements at all – at least, this does not follow from the usual axioms of set theory. In addition to the Zermelo-Fraenkel axioms, one needs the so-called *axiom of choice*, which says that if $X_i\neq\emptyset$ for every $i\in I$, then $X\neq\emptyset$.

Note. The axiom of choice claims that one can simultaneously choose one element from each of a possibly infinite number of nonempty sets. The problem is that we cannot just "choose" the elements arbitrarily, because we do not have enough time to make infinitely many choices. This axiom may seem very natural, but it has a large number of strange consequences. For example, you may have heard of the Banach-Tarski paradox: the axiom of choice implies that one can divide the three-dimensional unit ball into finitely many pieces, and then put them back together

in a different way and end up with a ball of twice the radius. This kind of thing lead to many arguments about the validity of the axiom, until it was proved that the axiom of choice is logically independent from the other axioms of set theory. Nowadays, most people assume the axiom of choice since it makes it easier to prove interesting theorems.

We want to make X into a topological space. There are two different ways of generalizing the definition from the case of two factors.

Definition 4.14. The *box topology* on X is the topology generated by the basis

$$\left\{ \prod_{i \in I} U_i \;\middle|\; U_i \in \mathscr{T}_i \text{ for every } i \in I \right\}.$$

It is not hard to see that this is indeed a basis for a topology: it contains X, and since

$$\left(\prod_{i \in I} U_i \right) \cap \left(\prod_{i \in I} V_i \right) = \prod_{i \in I} U_i \cap V_i,$$

the intersection of any two basic open sets is again a basic open set. It it clear from the definition that the *coordinate functions*

$$p_j \colon X \to X_j, \quad p_j\left((x_i)_{i \in I}\right) = x_i$$

are continuous functions. The box topology is a perfectly good topology on X, but when I is infinite, it has a very large number of open sets, which leads to certain pathologies. (For example, it usually does not satisfy the first or second countability axiom.)

We can get a better topology by putting some finiteness into the definition.

Definition 4.15. The *product topology* on X is the topology generated by the basis

$$\left\{ \prod_{i \in I} U_i \;\middle|\; U_i \in \mathscr{T}_i \text{ for every } i \in I, \text{ and } U_i = X_i \text{ for all but finitely many } i \in I \right\}.$$

The difference with the box topology is that we are now allowed to specify only finitely many coordinates in each basic open set. The idea behind the product topology is that every set of the form $p_j^{-1}(U)$ should be open (since we want p_j to be continuous), and that finite intersections of open sets need to be open (since we want to have a topology). In fact, one can show that the product topology is the smallest topology on X that makes all the coordinate functions $p_j \colon X \to X_j$ continous.

Theorem 4.16. *If we give X the product topology, then a function $f \colon Y \to X$ is continuous if and only if $f_i = p_i \circ f \colon Y \to X_i$ is continous for every $i \in I$.*

Proof. The proof is the same as in the case of two factors. □

This nice result fails for the box topology. For that reason, we almost always use the product topology when talking about infinite products of topological spaces.

Example 4.17. Let X be the product of countably many copies of $\mathbb{R}$, indexed by the set $\{1, 2, \dots\}$. The function

$$f \colon \mathbb{R} \to X, \quad f(t) = (t, t, \dots)$$

is not continuous for the box topology, because

$$f^{-1}\left(\prod_{n=1}^{\infty}\left(-\frac{1}{n},\frac{1}{n}\right)\right)=\{0\}$$

is not open in $\mathbb{R}$.

Next time, we will talk about connectedness, §23 to §25 in the textbook.

Lecture 5: September 8

Now that we have seen several examples of topological spaces, it is time to begin our study of topology. The definition of a topological space is very broad, and there is not much that one can say in general. Instead, topologists focus on certain additional properties of topological spaces and try to prove interesting results about spaces that have those properties.

The first two important properties that we are going to consider are connectedness and compactness. In a sense, they are generalizations of two important results from calculus, namely the intermediate value theorem and the maximum value theorem. Both have to do with a continuous function $f\colon [a,b] \to \mathbb{R}$ defined on a closed interval $[a,b] \subseteq \mathbb{R}$. The *intermediate value theorem* says that f takes on every value that lies between the values at the two endpoints: for every r between $f(a)$ and $f(b)$, there is some $x \in [a,b]$ with $f(x) = r$. The *maximum value theorem* says that f has a maximum value: there is some $x_0 \in [a,b]$ such that $f(x_0) \geq f(x)$ for every $x \in [a,b]$. In calculus, these results are usually viewed as properties of continuous functions; but they also reflect two properties of closed intervals in $\mathbb{R}$, namely connectedness and compactness.

Connectedness. Let X be a topological space. A pair of open sets $U, V \subseteq X$ with $U \cup V = X$ and $U \cap V = \emptyset$ is called a *separation* of X, because it separates the points of X into two groups that have nothing to do with each other. Note that $U = X \setminus V$ is both open and closed; a separation of X is therefore the same thing as a subset $U \subseteq X$ that is open and closed and not equal to $\emptyset$ or X.

Definition 5.1. A topological space X is called *connected* if it has no separation.

Equivalently, X is connected if the only subsets that are both open and closed are $\emptyset$ and X itself. Connectedness depends only on the topology of X; if two topological spaces are homeomorphic, then they are either both connected or both not connected.

Example 5.2. The three-point space $\{a,b,c\}$ with the topology

$$\Big\{\emptyset, \{a\}, \{b\}, \{a,b\}, \{a,b,c\}\Big\}$$

is connected, because $\{a,b,c\}$ is the only open set containing the point c.

Example 5.3. The space $\mathbb{R} \setminus \{0\}$ is not connected, because $(-\infty, 0)$ and $(0, \infty)$ form a separation.

Example 5.4. The rational numbers $\mathbb{Q}$ (with the subspace topology coming from $\mathbb{R}$) are not connected. In fact, the only connected subspaces of $\mathbb{Q}$ are the points, and so $\mathbb{Q}$ is what is called *totally disconnected*. To see why, suppose that $X \subseteq \mathbb{Q}$ contains at least two points $a < b$. Let c be an irrational number with $a < c < b$; then $X \cap (-\infty, c)$ and $X \cap (c, \infty)$ form a separation of X.

Example 5.5. The Cantor set is also totally disconnected.

The real numbers $\mathbb{R}$ are an important example of a connected topological space. This is the content of the following theorem.

Theorem 5.6. *$\mathbb{R}$ is connected.*

Proof. The argument we are going to give depends on the following important property of real numbers: If a set of real numbers $A \subseteq \mathbb{R}$ is bounded from above, then there is a well-defined least upper bound $\sup A$. By definition, $\sup A$ is the smallest real number with the property that $x \leq \sup A$ for every $x \in A$.

Now let us assume that $\mathbb{R}$ is not connected and derive a contradiction. Let $\mathbb{R} = A \cup B$ be a separation, and choose a point $a \in A$ and $b \in B$; without loss of generality, we may suppose that $a < b$. Now we will find a point s in the interval $[a, b]$ where A and B touch each other, and get a contradiction by studying what happens at that point. Using the least upper bound property, define

$$s = \sup\big(A \cap [a, b]\big).$$

Since $\mathbb{R} = A \cup B$, the point s should lie either in A or in B, but we will see in a moment that neither $s \in A$ nor $s \in B$ can be true.

Let us first consider the case $s \in A$. Then $s < b$, because $b \in B$; now A is open, and so it has to contain an open interval of the form (a_1, a_2) with $a_1 < s < a_2 < b$. This contradicts the fact that s is an upper bound for the set $A \cap [a, b]$.

The other possibility is that $s \in B$. Here $a < s$, because $a \in A$; now B is open, and so it has to contain an open interval of the form (b_1, b_2) with $a < b_1 < s < b_2$. Now any $x \in A \cap [a, b]$ satisfies $x \leq s$, and therefore $s < b_1$. This shows that b_1 is also an upper bound for the set $A \cap [a, b]$, contradicting the fact that s is the least upper bound. □

The same proof shows that any interval (closed or open) and any half-interval (closed or open) is also connected. Note that the argument breaks down for $\mathbb{Q}$ precisely because the rational numbers do not have the least upper bound property.

General results about connectedness. We will now establish a few general results about connected spaces. They all agree with our intuition of what connectedness should mean. The following simple lemma will be a useful tool.

Lemma 5.7. *Let $X = C \cup D$ be a separation of a topological space. If $Y \subseteq X$ is a connected subspace, then $Y \subseteq C$ or $Y \subseteq D$.*

Proof. We have $Y = (Y \cap C) \cup (Y \cap D)$, and both sets are open in the subspace topology and disjoint. Since Y is connected, one of them must be empty; but then we either have $Y = Y \cap C$, which means that $Y \subseteq C$; or $Y = Y \cap D$, which means that $Y \subseteq D$. □

The first result is that if we join together any number of connected subspaces at a common point, the result is again connected.

Proposition 5.8. *Let $\{Y_i\}_{i \in I}$ be a collection of connected subspaces of a topological space X. If the intersection $\bigcap_{i \in I} Y_i$ is nonempty, then the union $Y = \bigcup_{i \in I} Y_i$ is again connected.*

Proof. We argue by contradiction. Suppose that $Y = C \cup D$ was a separation. Choose a point $x \in \bigcap_{i \in I} Y_i$; without loss of generality $x \in C$. Since each Y_i is connected, and since Y_i and C have the point x in common, Lemma 5.7 tells us that $Y_i \subseteq C$. This being true for every $i \in I$, we get $Y \subseteq C$, which contradicts the fact that D is nonempty. □

Similarly, the closure of a connected subspace is again connected.

Proposition 5.9. *Let Y be a connected subspace of a topological space X. If $Y \subseteq Z \subseteq \overline{Y}$, then Z is again connected.*

Proof. Suppose that there was a separation $Z = C \cup D$. By Lemma 5.7, the connected subspace Y has to lie entirely in one of the two sets; without loss of generality, we may assume that $Y \subseteq C$. Now $C = Z \cap A$ for some closed set $A \subseteq X$, because C is closed in the subspace topology of Z. Since $Y \subseteq A$, we also have $Z \subseteq \overline{Y} \subseteq A$, and therefore $Z = C$. This contradicts the fact that D is nonempty. □

The most interesting result is that the image of a connected space under a continuous function is again connected.

Theorem 5.10. *Let $f\colon X \to Y$ be a continous function. If X is connected, then the image $f(X)$ is also connected.*

Proof. Suppose that there was a separation $f(X) = C \cup D$. Then C is a nonempty open subset of $f(X)$, and because f is continuous, $f^{-1}(C)$ is a nonempty open subset of X; the same is true for $f^{-1}(D)$. But then

$$X = f^{-1}(C) \cup f^{-1}(D)$$

is a separation of X, contradicting the fact that X is connected. □

One can show quite easily (see Exercise 10 in §23) that an arbitrary product of connected spaces is connected in the product topology. The following example shows that this is not true for the box topology.

Example 5.11. Let $X = \mathbb{R}^{\mathbb{N}}$ be the set of all infinite sequences $x_0, x_1, \ldots$, in the box topology. Let B be the set of all bounded sequences, and U the set of all unbounded sequences. I claim that $X = B \cup U$ is a separation. B and U are clearly disjoint and nonempty; they are also open in the box topology. Indeed, if $x_0, x_1, \ldots \in B$ is a bounded sequence, then

$$(x_0 - 1, x_0 + 1) \times (x_1 - 1, x_1 + 1) \times \cdots$$

is an open set contained in B; if $x_0, x_1, \ldots \in U$ is unbounded, then the same set is contained in U.

Since we began our discussion of connectedness by talking about the intermediate value theorem in calculus, let us now prove a version for arbitrary connected topological spaces.

Theorem 5.12 (Intermediate value theorem)**.** *Let $f\colon X \to Y$ be a continuous function from a connected space X to a linearly ordered set $(Y, <)$ in the order topology. If $a, b \in X$, and if $r \in Y$ is any element lying between $f(a)$ and $f(b)$, then there is a point $x \in X$ with $f(x) = r$.*

Proof. We know from Theorem 5.10 that $f(X)$ is a connected subspace of Y. Now consider the two open sets $f(X) \cap (-\infty, r)$ and $f(X) \cap (r, \infty)$. They are obviously disjoint, and since r lies between $f(a)$ and $f(b)$, both are nonempty. If $r \notin f(X)$, then we would get a separation of $f(X)$, which is not possible because $f(X)$ is connected. The conclusion is that there is a point $x \in X$ with $f(x) = r$. □

The usual intermediate value theorem is of course the special case when X is a closed interval in $\mathbb{R}$, and $Y = \mathbb{R}$. By splitting up the proof into two parts – closed intervals in $\mathbb{R}$ are connected; connected spaces satisfy the intermediate value theorem – we get a better understanding of this calculus theorem, too.

Paths and path connectedness. We defined connectedness in a negative way, by saying that a space is connected if it cannot be separated into two disjoint chunks. There is another possible definition, namely that it should be possible to go from any point to any other point. This leads to another notion called "path connectedness". Let X be a topological space. A *path* from a point $x \in X$ to a point $y \in X$ is simply a continuous function $f\colon [a,b] \to X$ from a closed interval in $\mathbb{R}$ to the space X, such that $f(a) = x$ and $f(b) = y$.

Definition 5.13. We say that X is *path connected* if any two points in X can be joined by a path in X.

The advantage of this definition is that it is more intuitive. The disadvantage is that for many topological spaces, every continuous map from a closed interval in $\mathbb{R}$ is constant; this limits the usefulness of the definition in terms of paths. We will see later that the two definitions are equivalent when X is sufficiently nice; but in general, we only have the following implication.

Proposition 5.14. *If X is path connected, then it is connected.*

Proof. Suppose that there was a separation $X = C \cup D$. Since closed intervals in $\mathbb{R}$ are connected, the image of any path is a connected subspace of X by Theorem 5.10; according to Lemma 5.7, it has to be contained entirely inside C or entirely inside D. This means that there are no paths connecting C and D, contradicting the path connectedness of X. □

Example 5.15. Let $n \geq 2$, and consider $\mathbb{R}^n$ with the origin removed. This space is path connected: any two points on the unit sphere $\mathbb{S}^{n-1}$ can be joined by a path going along a great circle; and any point of $\mathbb{R}^n \setminus \{0\}$ can be joined to a point on $\mathbb{S}^{n-1}$ by a radial path. In particular, the space is connected.

Example 5.16. The rational numbers $\mathbb{Q}$ are not path connected. Since they are not even connected, this follows from the proposition; but one can also prove it directly by showing that any continuous function from a closed interval to $\mathbb{Q}$ must be constant.

Example 5.17. Here is an example of a topological space that is connected but not path connected. Let

$$S = \left\{ (x, \sin(1/x) \in \mathbb{R}^2 \;\middle|\; x > 0 \right\}$$

denote the graph of the function $\sin(1/x)$. Being the continuous image of the connected space $(0,\infty)$, it is connected; therefore its closure

$$\overline{S} = S \cup \left\{ (0,y) \in \mathbb{R}^2 \;\middle|\; -1 \leq y \leq 1 \right\}$$

is also connected. It is known as the *topologist's sine curve*. I claim that $\overline{S}$ is not path connected. To see why, suppose there was a path connecting the point $(0,0)$ to a point in S. Because S is open in $\overline{S}$, we can assume without loss of generality that the path is of the form

$$f\colon [0,1] \to \overline{S}$$

with $f(0) \in \overline{S} \setminus S$ and $f(t) \in S$ for $t > 0$. If we write $f(t) = \big(x(t), y(t)\big)$, then the coordinate functions x and y are continuous; we also have $x(t) > 0$ and $y(t) = \sin\big(1/x(t)\big)$ for $t > 0$, and $x(0) = 0$. But now the fact that $\sin(1/x)$ oscillates wildly leads to a contradiction: we can find a sequence of positive real numbers $t_n \to 0$ with $y(t_n) = (-1)^n$, whereas y was supposed to be continuous. To construct such a sequence, we first choose for every $n = 1, 2, \ldots$ a real number $0 < x_n < x(1/n)$ such that $\sin(1/x_n) = (-1)^n$. Because $x(0) = 0$, we can then use the intermediate value theorem to find some point $0 < t_n < 1/n$ with $x(t_n) = x_n$.

Lecture 6: September 13

Connected components. If a topological space X is not connected, it makes sense to investigate the maximal connected subspaces it contains. Given a point $x \in X$, we define

$$C(x) = \bigcup \{\, C \subseteq X \mid C \text{ is connected and } x \in C \,\}$$

as the union of all connected subspaces of X that contain the point x.

Proposition 6.1. *These sets have the following properties:*

(a) Each subspace $C(x)$ is connected and closed.
(b) If $x, y \in X$, then $C(x)$ and $C(y)$ are either equal or disjoint.
(c) Every nonempty connected subspace of X is contained in a unique $C(x)$.

Proof. $C(x)$ is a union of connected subspaces that all contain the point x, and therefore connected by Proposition 5.8. Now Proposition 5.9 shows that the closure $\overline{C(x)}$ is also connected; since it contains x, it must be contained in $C(x)$, and therefore equal to $C(x)$. This proves (a).

To prove (b), suppose that $C(x)$ and $C(y)$ are not disjoint. Take any $z \in C(x) \cap C(y)$. We shall argue that $C(x) = C(z)$; by symmetry, this will imply that $C(x) = C(z) = C(y)$. We have $z \in C(x)$, and since $C(x)$ is connected, we get $C(x) \subseteq C(z)$. Hence $x \in C(z)$, and for the same reason, $C(z) \subseteq C(x)$.

To prove (c), let $Y \subseteq X$ be nonempty connected subspace. Choose a point $x \in Y$; then $Y \subseteq C(x)$ by construction. Uniqueness follows from (b). □

The set $C(x)$ is called the *connected component* of x; the proposition shows that it is the maximal connected subspace of X containing the point x. Since any two connected components of X are either equal or disjoint, we get a partition of the space X into maximal connected subsets.

Example 6.2. The connected components of a space are always closed, but not necessarily open. In the case of $\mathbb{Q}$, the connected component of $x \in \mathbb{Q}$ is $\{x\}$, because no subspace with two or more points is connected. So the connected components of X do not give us a separation of X in general.

Example 6.3. If X has only finitely many connected components, then each connected component is open (being the complement of a finite union of closed sets).

There is a similar definition for path connectedness: given a point $x \in X$, we set

$$\begin{aligned} P(x) &= \bigcup \{\, P \subseteq X \mid P \text{ is path connected and } x \in P \,\} \\ &= \{\, y \in X \mid x \text{ and } y \text{ can be joined by a path in } X \,\}, \end{aligned}$$

and call it the *path component* of x. The following result can be proved in the same way as Proposition 6.1.

Proposition 6.4. *The path components of X have the following properties:*

(a) Each subspace $P(x)$ is path connected.
(b) If $x, y \in X$, then $P(x)$ and $P(y)$ are either equal or disjoint.
(c) Every nonempty path connected subspace of X is contained in a unique $P(x)$.

Since $P(x)$ is connected by Proposition 5.14, it is contained in $C(x)$; but the two need not be equal.

Example 6.5. The topologist's sine curve is connected, but it has two path components.

Path connectedness and connectedness. I mentioned last time that in sufficiently nice topological spaces (such as open subsets of $\mathbb{R}^n$) connectedness and path connectedness are equivalent. The point is that $\mathbb{R}^n$ contains lots and lots of paths, whereas a random topological space may not contain any nonconstant path at all. This suggests looking at spaces where at least any two nearby points can be joined by a path.

Definition 6.6. A topological space X is *locally path connected* if for every point $x \in X$ and every open set U containing x, there is a path connected open set V with $x \in V$ and $V \subseteq U$.

Open balls in $\mathbb{R}^n$ are obviously path connected. Consequently, every open set in $\mathbb{R}^n$ is locally path connected; more generally, every topological manifold is locally path connected. Note that in order to be locally path connected, every point needs to have *arbitrarily small* path connected neighborhoods. Note that a path connected space may not be locally path connected: we can get an example by taking the topologist's sine curve

$$\left\{ (0, y) \;\middle|\; -1 \leq y \leq 1 \right\} \cup \left\{ (x, \sin 1/x) \;\middle|\; 0 < x \leq 1 \right\} \subseteq \mathbb{R}^2$$

and joining the two points $(0, 1)$ and $(1, \sin 1)$ by a path that is disjoint from the rest of the set. The resulting space is path connected, but not locally path connected at any point of the form $(0, y)$ with $-1 \leq y \leq 1$.

Proposition 6.7. *Let X be a locally path connected topological space. Then $P(x)$ is open, and $P(x) = C(x)$ for every $x \in X$. In particular, the connected components of X are open.*

Proof. You will remember a similar result from one of last week's homework problems. We first prove that $P(x)$ is open for every $x \in X$. Since X is locally path connected, there is a path connected neighborhood U of x; we have $U \subseteq P(x)$, because $P(x)$ is the union of all path connected subspaces containing x. If $y \in P(x)$ is any point, then $P(y) = P(x)$, and so $P(x)$ also contains a neighborhood of y, proving that $P(x)$ is open.

Next, we show that $P(x) = C(x)$. The union of all the other path components of X is open, and so their complement $P(x)$ must be closed. Now $P(x) \subseteq C(x)$ is a nonempty subset that is both open and closed; because $C(x)$ is connected, it must be that $P(x) = C(x)$. □

The last step in the proof is a typical application of connectedness: to prove that a nonempty subset Y of a connected space X is equal to X, it is enough to show that Y is both open and closed.

Compactness. The second important property of closed intervals in $\mathbb{R}$ is compactness. Perhaps you know the definition that is used in analysis: a subset $A \subseteq \mathbb{R}^n$ is called compact if every sequence in A has a convergent subsequence. As with closed sets, definitions involving sequences are not suitable for general topological spaces. We therefore adopt the following definition in terms of open coverings.

Definition 6.8. An *open covering* of a topological space X is a collection $\mathscr{U}$ of open subsets with the property that

$$X = \bigcup_{U \in \mathscr{U}} U.$$

We say that X is *compact* if, for every open covering $\mathscr{U}$, there are finitely many open sets $U_1, \ldots, U_n \in \mathscr{U}$ such that $X = U_1 \cup \cdots \cup U_n$.

Compactness is a very important finiteness property of the topology on a space. In fact, every topological space with only finitely many points is trivially compact; in a sense, compact spaces are thus a generalization of finite topological spaces.

Note that compactness is a topological property: if X is compact, then any space homeomorphic to X is also compact. You should convince yourself that when the topology of X is given in terms of a basis, it suffices to check the condition for open coverings by basic open sets.

Example 6.9. $\mathbb{R}$ is not compact: in the open covering

$$\mathbb{R} = \bigcup_{n=1}^{\infty} (-n, n),$$

no finite number of subsets is enough to cover $\mathbb{R}$.

Example 6.10. $\mathbb{Q} \cap [0,1]$ is also not compact. This is obvious using the analyst's definition in terms of sequences; here is one way of proving it with our definition. We enumerate all the rational numbers between 0 and 1 in the form $\alpha_1, \alpha_2, \ldots$, and also choose an irrational number $\beta \in [0,1]$. For every k, let I_k be the open interval of length $\ell_k = \frac{1}{2}|\alpha_k - \beta|$ centered at the point α_k. Obviously, the sets

$$I_k \cap \mathbb{Q} \cap [0,1]$$

form an open covering of $\mathbb{Q} \cap [0,1]$; I claim that no finite number of sets can cover everything. Otherwise, there would be some integer $n \geq 1$ such that every rational number between 0 and 1 lies in $I_1 \cup \cdots \cup I_n$. But then the open interval of length

$$\min_{1 \leq k \leq n} \ell_k$$

centered at β would not contain any rational number, which is absurd.

These examples illustrate an important point: it is usually pretty easy to show that a space is not compact, because all we have to do is find one offending open covering. On the other hand, it can be quite hard to show that a space is compact, because we need to consider all possible open coverings.

Example 6.11. The closed unit interval $[0,1]$ is compact; we will prove this next time.

General properties of compactness. The following lemma is an easy consequence of the definitions (of compactness and of the subspace topology).

Lemma 6.12. *Let X be a topological space, and let $Y \subseteq X$ be a subspace. The following two conditions are equivalent:*

(a) Y is compact (in the subspace topology).

(b) Whenever $\mathscr{U}$ is a collection of open sets in X with

$$Y \subseteq \bigcup_{U \in \mathscr{U}} U,$$

there are finitely many sets $U_1, \ldots, U_n \in \mathscr{U}$ with $Y \subseteq U_1 \cup \cdots \cup U_n$.

We can use this lemma to prove some general properties of compact spaces.

Proposition 6.13. *Every closed subspace of a compact space is compact.*

Proof. Let X be compact and $Y \subseteq X$ closed. Given an open covering $\mathscr{U}$ of Y, we get an open covering of X by throwing in the open subset $U_0 = X \setminus Y$. Since X is compact, there are finitely many sets $U_1, \ldots, U_n \in \mathscr{U}$ such that $X = U_0 \cup U_1 \cup \cdots \cup U_n$. But then $Y \subseteq U_1 \cup \cdots \cup U_n$, proving that Y is compact. □

Proposition 6.14. *If X is compact and $f\colon X \to Y$ is continuous, then $f(X)$ is again compact.*

Proof. Let $\mathscr{U}$ be an arbitrary open covering of $f(X)$. Then

$$\left\{ f^{-1}(U) \;\middle|\; U \in \mathscr{U} \right\}$$

is a collection of open sets that cover X; because X is compact, there must be finitely many sets $U_1, \ldots, U_n \in \mathscr{U}$ with

$$X = f^{-1}(U_1) \cup \cdots \cup f^{-1}(U_n).$$

Since $U_1, \ldots, U_n \subseteq f(X)$, we now get $f(X) = U_1 \cup \cdots \cup U_n$, proving that $f(X)$ is also compact. □

Lecture 7: September 15

Closed sets and compact sets. Last time, we defined compactness in terms of open coverings: a topological space X is compact if, for every open covering $\mathscr{U}$, there are finitely many open sets $U_1, \dots, U_n \in \mathscr{U}$ such that $X = U_1 \cup \dots \cup U_n$. To prove that a given space is *not* compact is usually not difficult: we just have to find one open covering that violates the condition in the definition. To prove that a given space is compact can be quite difficult: we have to verify the condition for all possible open coverings. (This task becomes a little easier when the topology on X is given by a basis, because we only need to consider coverings by basic open subsets.) Since compactness is a strong requirement, compact spaces have a lot of wonderful properties. The following theorem about continuous functions on compact spaces shows this effect at work.

Theorem 7.1. *Let $f\colon X \to Y$ be a continous function. If X is compact and Y is Hausdorff, then the image $f(X)$ is closed.*

Recall from last time that the image of a compact space under a continuous function is again compact; this was an easy consequence of the definitions. Theorem 7.1 requires a little bit more work to prove. Let me first note the following surprising corollary.

Corollary 7.2. *Let $f\colon X \to Y$ be a bijective continous function. If X is compact and Y is Hausdorff, then f is a homeomorphism.*

Proof. Let us see how this follows from Theorem 7.1. We have to prove is that the inverse function $f^{-1}\colon Y \to X$ is continous; this is equivalent to saying that the preimage of every closed set $A \subseteq X$ is closed in Y. The preimage of A under f^{-1} is exactly $f(A)$, and so we need to argue that $f(A)$ is closed. As a closed subset of a compact space, A is compact (Proposition 6.13); therefore its image $f(A)$ is again compact by Proposition 6.14. Since Y is Hausdorff, Theorem 7.1 implies that $f(A)$ is closed. □

Now let us start proving the theorem. By Proposition 6.14, the image $f(X)$ is a compact subspace of the Hausdorff space Y. The following proposition explains why $f(X)$ must be closed.

Proposition 7.3. *Let X be a Hausdorff topological space. If $Y \subseteq X$ is a compact subspace, then Y is closed in X.*

Proof. We shall argue that $X \setminus Y$ is a union of open sets, and hence open. Fix a point $x \in X \setminus Y$. For any $y \in Y$, we can find disjoint open sets $U(y)$ and $V(y)$ with $x \in U$ and $y \in V$; this is because X is Hausdorff. Therefore

$$Y \subseteq \bigcup_{y \in C} V(y),$$

and because Y is compact, there are finitely many points $y_1, \dots, y_n \in Y$ with $Y \subseteq V(y_1) \cup \dots \cup V(y_n)$. But then the open set

$$U(y_1) \cap \dots \cap U(y_n)$$

is disjoint from $V(y_1) \cup \dots \cup V(y_n)$, hence contained in $X \setminus Y$; because it is a neighborhood of x, and because $x \in X \setminus Y$ was arbitrary, $X \setminus Y$ must be open. □

Note that this result is false when X is not Hausdorff; a simple example is given by the Sierpiński space. More generally, let X be any finite topological space; then all compact subsets of X are closed if and only if the topology on X is discrete if and only if X is Hausdorff.

Note. The proof of Proposition 7.3 shows that we can separate points and compact subspaces in a Hausdorff space by open sets. More precisely, given a compact subspace $Y \subseteq X$ and a point $x \notin Y$, there are disjoint open sets U and V with $x \in U$ and $Y \subseteq V$ – in the notation used during the proof,

$$U = U(y_1) \cap \cdots \cap U(y_n) \quad \text{and} \quad V = V(y_1) \cup \cdots \cup V(y_n).$$

If X is both compact and Hausdorff, so that every closed subspace is compact, we can even separate points and arbitrary closed subspaces by open sets. We shall investigate results of this type in more detail later.

Examples of compact spaces. Let us now look at a few topological spaces that are compact. The first example is closed intervals in $\mathbb{R}$. You probably already know that closed intervals are "compact" in the analysis sense – every sequence has a convergent subsequence – but we need to do some work to prove that they are also compact in the topology sense.

Theorem 7.4. *The closed interval $[0, 1]$ is compact.*

Proof. Let $\mathscr{U}$ be an open covering; we have to show that the interval can be covered by finitely many open sets in $\mathscr{U}$. The idea of the proof is very simple: 0 belongs to some open set $U_1 \in \mathscr{U}$, which contains some interval of the form $[0, a)$; then a belongs to some open set $U_2 \in \mathscr{U}$, which again contains a maximal interval $[a, b)$; and so on. Of course, the situation could be like in one of Zeno's paradoxes, with the points $a, b, \ldots$ converging to some limit inside the interval. To avoid this problem, we have to take a slightly less direct approach. Let us introduce the set

$$S = \{\, x \in [0, 1] \mid [0, x] \text{ can be covered by finitely many open sets in } \mathscr{U} \,\}.$$

We want to show that $1 \in S$, because this will mean that $[0, 1]$ can be covered by finitely many sets in $\mathscr{U}$. Using the least upper bound property of $\mathbb{R}$, define $s = \sup S$. Now the argument proceeds in three steps:

Step 1. Every point t with $0 \leq t < s$ has to belong to S. Suppose that we had $t \notin S$. Then every point $x \in S$ has to satisfy $x < t$: the reason is that $[0, x]$ can be covered by finitely many open sets in $\mathscr{U}$, and if $t \leq x$, then the same open sets would cover $[0, t]$. This means that t is an upper bound for S, contradicting our choice of s. Consequently, we have $[0, s) \subseteq S$.

Step 2. We show $s \in S$. Since $\mathscr{U}$ is an open covering, there is an open set $U \in \mathscr{U}$ with $s \in U$. If $s = 0$, this shows that $s \in S$. If $s > 0$, then U contains an interval of the form $(a, s]$ with $0 \leq a < s$. By Step 1, $a \in S$, and so $[0, a]$ is covered by finitely many open sets in $\mathscr{U}$; throwing in U, we get a finite open covering of $[0, s]$.

Step 3. We prove that $s = 1$. Suppose that $s < 1$. We already know that $[0, s]$ is covered by finitely many open sets in $\mathscr{U}$. Their union is a neighborhood of s, and therefore contains an interval of the form $[s, b)$ with $s < b \leq 1$. Because $\mathscr{U}$ is an open covering, $b \in U$ for some $U \in \mathscr{U}$; but now we have finitely many open sets covering $[0, b]$, contradicting our choice of s. □

In fact, every closed interval of the form $[a, b]$ is compact: this follows either by adapting the above proof, or by noting that any (nontrivial) closed interval in $\mathbb{R}$ is

homeomorphic to $[0,1]$. In the proof, we only used the ordering of $\mathbb{R}$ and the least upper bound property; for that reason, the theorem is true is any linearly ordered set with the least upper bound property.

The next example is $\mathbb{R}^n$ and its subsets. If you are taking analysis, you may already know that a subspace of $\mathbb{R}^n$ is compact if and only if it is closed and bounded (in the Euclidean metric). Here *bounded* means that it is contained in a ball $B_R(0)$ of some radius R. I want to explain why this result is also true with our definition of compactness. The following general result about products will be useful for the proof.

Proposition 7.5. *If X and Y are compact, then $X \times Y$ is compact.*

Proof. Recall that the product topology has a basis consisting of all open sets of the form $U \times V$, with U open in X and V open in Y. Let $\mathscr{U}$ be an open covering of $X \times Y$; as I explained earlier, it suffices to consider the case that $\mathscr{U}$ consists of basic open sets.

Step 1. Fix a point $x \in X$, and consider the vertical slice $p_1^{-1}(x) = \{x\} \times Y$. It is homeomorphic to Y, and therefore compact; one way of seeing this is to consider the continuous function $i\colon Y \to X \times Y$, $i(y) = (x, y)$, and to apply Proposition 6.14. Since it is covered by the union of all the open sets in $\mathscr{U}$, we can find finitely many open sets $U_1 \times V_1, \dots, U_n \times V_n \in \mathscr{U}$ such that $p_1^{-1}(x) \subseteq U_1 \times V_1 \cup \dots \cup U_n \times V_n$. Now define $U(x) = U_1 \cap \dots \cap U_n$; then the entire set $U(x) \times Y$ is covered by finitely many open sets in $\mathscr{U}$.

Step 2. The collection of open sets $U(x)$ from Step 1 is an open covering of X. By compactness, there are finitely many points $x_1, \dots, x_m \in X$ with

$$X = U(x_1) \cup \dots \cup U(x_m).$$

This means that

$$X \times Y = U(x_1) \times Y \cup \dots \cup U(x_m) \times Y,$$

and because each these m sets is covered by finitely many sets in $\mathscr{U}$, the same is true for the product $X \times Y$. □

More generally, any finite product of compact spaces is again compact; this can easily be proved by induction on the number of factors. We will see later in the semester that arbitrary products of compact spaces are compact (in the product topology); this is an extremely powerful result, but the proof is not that easy.

Theorem 7.6. *A subset $A \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. Let us first check that every compact subset $A \subseteq \mathbb{R}^n$ is both closed and bounded. As $\mathbb{R}^n$ is Hausdorff, Proposition 6.13 implies that A is closed. To prove boundedness, observe that

$$A \subseteq \bigcup_{n=1}^{\infty} B_n(0)$$

is an open covering; because A is compact, we must have $A \subseteq B_n(0)$ for some n.

The really interesting part is the converse. Suppose that $A \subseteq \mathbb{R}^n$ is closed and bounded. Choose some $R > 0$ with the property that $A \subseteq B_R(0)$; then also

$$A \subseteq [-R, R]^n.$$

We know from Theorem 7.4 that the closed interval $[-R, R]$ is compact; the space on the right-hand side is therefore compact by Proposition 7.5. Being a closed subset of a compact space, A is therefore compact as well. □

What about more general metric spaces? The proof above shows that a compact subset of a metric space must be closed and bounded; but the converse is not true in general. (The homework for next week will ask you to find an example of a subset in a metric space that is closed and bounded but not compact.) However, just as with closed sets, it is still true that compactness in metric spaces can be detected by sequences.

Theorem 7.7. *For a metric space X, the following two properties are equivalent:*

(a) X is a compact (in the metric topology).

(b) X is sequentially compact, meaning that every sequence in X has a convergent subsequence.

You can find the proof in §28 of the book by Munkres; I decided not to present this in class because it is more about choosing sequences and subsequences than about general topology.

Next time, we will look at some other results about compact spaces, and at one-point compactifications.

Lecture 8: September 20

Baire's theorem. Here is a cute problem. Suppose that $f\colon (0,\infty) \to \mathbb{R}$ is a continuous function with the property that the sequence of numbers $f(x), f(2x), f(3x), \dots$ converges to zero for every $x \in (0,\infty)$. Show that

$$\lim_{x\to\infty} f(x) = 0.$$

This problem, and several others of a similar nature, is basically unsolvable unless you know the following result.

Theorem 8.1 (Baire's theorem)**.** *Let X be a nonempty compact Hausdorff space.*

(a) If $X = A_1 \cup A_2 \cup \cdots$ can be written as a countable union of closed sets A_n, then at least one set A_n has nonempty interior.

(b) If $U_1, U_2, \dots \subseteq X$ is a countable collection of dense open sets, then the intersection $U_1 \cap U_2 \cap \cdots$ is dense in X.

Here a subset $Y \subseteq X$ in a topological space is called *dense* if its closure $\overline{Y}$ is equal to X, or equivalently, if Y intersects every nonempty open subset of X. In analysis, there is another version of Baire's theorem for complete metric spaces: the assumption is that X is a complete metric space, and the conclusion is the same as in Theorem 8.1.

Example 8.2. The problem from above uses the version for complete metric spaces. Pick some $\varepsilon > 0$. By assumption, for every $x > 0$, there is some integer $n \geq 1$ such that $|f(kx)| \leq \varepsilon$ for all $k \geq n$. This means that $(0,\infty)$ is the union of the countably many closed sets

$$A_n = \big\{\, x > 0 \;\big|\; |f(kx)| \leq \varepsilon \text{ for } k \geq n \,\big\}.$$

Baire's theorem ensures that at least one A_n contains an open interval; and from that, one can deduce that $|f(x)| \leq \varepsilon$ for $x \gg 0$.

In the first half of today's class, we are going to prove Theorem 8.1. The second portion of the theorem is actually stronger than the first one, so let me begin by explaining why (b) implies (a). Let X be a compact Hausdorff space, and suppose that we have countably many closed subsets $A_1, A_2, \dots$ with

$$X = \bigcup_{n=1}^{\infty} A_n.$$

If the interior $\operatorname{int} A_n = \emptyset$, then the open complement $U_n = X \setminus A_n$ must be dense in X: the reason is that $X \setminus \overline{U_n} \subseteq X \setminus U_n = A_n$ is an open subset of A_n, hence empty, which means that $\overline{U_n} = X$. Now if (a) was false, we would have a countable collection of dense open sets; since we are assuming that (b) holds, the intersection

$$\bigcap_{n=1}^{\infty} U_n$$

is dense in X, and therefore nonempty. But since

$$X \setminus \bigcap_{n=1}^{\infty} U_n = \bigcup_{n=1}^{\infty} X \setminus U_n = \bigcup_{n=1}^{\infty} A_n = X,$$

this contradicts our initial assumption that X is the union of the A_n. This argument also shows in which sense (b) is stronger than (a): it tells us not only that the

intersection of countably many dense open sets is nonempty, but that it is still dense in X.

The proof of Baire's theorem requires a little bit of preparation; along the way, we have to prove two other results that will also be useful for other things later on. As a first step, we restate the definition of compactness in terms of closed sets. To do that, we simply replace "open" by "closed" and "union" by "intersection"; we also make the following definition.

Definition 8.3. A collection $\mathscr{A}$ of subsets of X has the *finite intersection property* if $A_1 \cap \cdots \cap A_n \neq \emptyset$ for every $A_1, \ldots, A_n \in \mathscr{A}$.

The definition of compactness in terms of open coverings is great for deducing global results from local ones: if something is true in a neighborhood of every point in a compact space, then the fact that finitely many of those neighborhoods cover X will often imply that it is true on all of X. The following formulation emphasizes a different aspect of compactness: the ability to find points with certain properties.

Proposition 8.4. *A topological space X is compact if and only if, for every collection $\mathscr{A}$ of closed subsets with the finite intersection property, the intersection*

$$\bigcap_{A \in \mathscr{A}} A$$

is nonempty.

If we think of each closed set $A \in \mathscr{A}$ as being a certain condition on the points of X, and of the finite intersection property as saying that the conditions are consistent with each other, then the result can be interpreted as follows: provided X is compact, there is at least one point $x \in X$ that satisfies all the conditions at once.

Proof. Let us first show that the condition is necessary. Suppose that X is compact, and that $\mathscr{A}$ is a collection of closed sets with the finite intersection property. Consider the collection of open sets

$$\mathscr{U} = \big\{\, X \setminus A \;\big|\; A \in \mathscr{A} \,\big\}.$$

The finite intersection property of $\mathscr{A}$ means exactly that no finite number of sets in $\mathscr{U}$ can cover X: this is clear because

$$(X \setminus A_1) \cup \cdots \cup (X \setminus A_n) = X \setminus (A_1 \cap \cdots \cap A_n) \neq X$$

for every $A_1, \ldots, A_n \in \mathscr{A}$. Because X is compact, it follows that $\mathscr{U}$ cannot be an open covering of X; but then

$$X \setminus \bigcap_{A \in \mathscr{A}} A = \bigcup_{A \in \mathscr{A}} X \setminus A \neq X,$$

which means that the intersection of all the sets in $\mathscr{A}$ is nonempty. To see that the condition is also sufficient, we can use the same argument backwards. □

The next step in the proof of Baire's theorem is the following observation about compact Hausdorff spaces.

Proposition 8.5. *Let X be a compact Hausdorff space and let $x \in X$ be a point. Inside every neighborhood U of x, there is a smaller neighborhood V of x with $\overline{V} \subseteq U$.*

Proof. This follows easily from our ability to separate points and closed sets in a compact Hausdorff space (see the note after the proof of Proposition 7.3). The closed set $X \setminus U$ does not contain the point x; consequently, we can find disjoint open subsets $V, W \subseteq X$ with $x \in V$ and $X \setminus U \subseteq W$. Now $X \setminus W$ is closed, and so

$$\overline{V} \subseteq X \setminus W \subseteq U,$$

as asserted. □

The property in the proposition is called *local compactness*; we will study it in a little more detail during the second half of today's class. But first, let us complete the proof of Baire's theorem.

Proof of Theorem 8.1. Let $U_1, U_2, \ldots$ be countably many dense open subsets of X. To show that their intersection is again dense in X, we have to prove that

$$U \cap \bigcap_{n=1}^{\infty} U_n \neq \emptyset$$

for every nonempty open set $U \subseteq X$. It is not hard to show by induction that all finite intersections $U \cap U_1 \cap \cdots \cap U_n$ are nonempty: $U \cap U_1 \neq \emptyset$ because U_1 is dense; $U \cap U_1 \cap U_2 \neq \emptyset$ because U_2 is dense, etc. The problem is to find points that belong to all of these sets at once, and it is here that Proposition 8.4 comes into play.

First, consider the intersection $U \cap U_1$. It must be nonempty (because U_1 is dense in X), and so we can find a nonempty open set V_1 with $\overline{V_1} \subseteq U \cap U_1$ (by applying Proposition 8.5 to any point in the intersection). Next, consider the intersection $V_1 \cap U_2$. It must be nonempty (because U_2 is dense), and so we can find a nonempty open set V_2 with $\overline{V_2} \subseteq V_1 \cap U_2$ (by applying Proposition 8.5 to any point in the intersection). Observe that we have

$$\overline{V_2} \subseteq \overline{V_1} \quad \text{and} \quad \overline{V_2} \subseteq V_1 \cap U_2 \subseteq U \cap U_1 \cap U_2.$$

Continuing in this way, we obtain a sequence of nonempty open sets $V_1, V_2, \ldots$ with the property that $\overline{V_n} \subseteq V_{n-1} \cap U_n$; by construction, their closures satisfy

$$\overline{V_1} \supseteq \overline{V_2} \supseteq \overline{V_3} \supseteq \cdots$$

and so the collection of closed sets $\overline{V_n}$ has the finite intersection property. Since X is compact, Proposition 8.4 implies that

$$\bigcap_{n=1}^{\infty} \overline{V_n} \neq \emptyset.$$

But since $\overline{V_n} \subseteq U \cap U_1 \cap \cdots \cap U_n$ for every n, we also have

$$\bigcap_{n=1}^{\infty} \overline{V_n} \subseteq U \cap \bigcap_{n=1}^{\infty} U_n,$$

and so the intersection on the right-hand side is indeed nonempty. □

If you like a challenge, try to solve the following problem: Let $f \colon \mathbb{R} \to \mathbb{R}$ be an infinitely differentiable function, meaning that the n-th derivative $f^{(n)}$ exists and is continous for every $n \in \mathbb{N}$. Suppose that for every $x \in \mathbb{R}$, there is some $n \in \mathbb{N}$ with $f^{(n)}(x) = 0$. Prove that f must be a polynomial! (Warning: This problem is very difficult even if you know Baire's theorem.)

Local compactness and one-point compactification. In Proposition 8.5, we proved that every compact Hausdorff space has the following property.

Definition 8.6. A Hausdorff topological space X is called *locally compact* if for every $x \in X$ and every open set U containing x, there is an open set V containing x whose closure $\overline{V}$ is compact and contained in U.

Note that this definition is different from the one in the textbook: Munkres calls a topological space "locally compact" if every point $x \in X$ has some neighborhood whose closure is compact. In fact, several other definitions are in common use, too; in a sense, Munkres' definition is the least restrictive one, and the one I gave is the most restrictive one. When X is Hausdorff, however, Munkres' definition is equivalent to the one above; all interesting theorems involving local compactness are about locally compact Hausdorff spaces anyway. (I prefer this formulation because it is similar to the definition of "locally path connected".)

Of course, there are many examples of topological spaces that are locally compact but not compact.

Example 8.7. $\mathbb{R}^n$ is locally compact; in fact, we showed earlier that every closed and bounded subset of $\mathbb{R}^n$ is compact. It follows that every topological manifold is also locally compact.

The reason why $\mathbb{R}^n$ is not compact is because of what happens "at infinity". We can see this very clearly if we think of $\mathbb{R}^n$ as embedded into the n-sphere $\mathbb{S}^n$; recall that the n-sphere minus a point is homeomorphic to $\mathbb{R}^n$. By adding one point, which we may think of as a point at infinity, we obtain the compact space $\mathbb{S}^n$: the n-sphere is of course compact because it is a closed and bounded subset of $\mathbb{R}^{n+1}$. The following result shows that something similar is true for every locally compact Hausdorff space: we can always build a compact space by adding one point.

Theorem 8.8. *Every locally compact Hausdorff space X can be embedded into a compact Hausdorff space X^* such that $X^* \setminus X$ consists of exactly one point.*

Proof. To avoid confusion, we shall denote the topology on X by the letter $\mathscr{T}$. Define $X^* = X \cup \{\infty\}$, where ∞ is not already an element of X. Now we want to put a topology on X^* that induces the given topology on X and that makes X^* into a compact Hausdorff space. What should be the open sets containing the point at infinity? The complement of an open subset containing ∞ is a subset $K \subseteq X$; if we want X^* to be compact, K must also be compact, because closed subsets of compact spaces are compact. So we are lead to the following definition:

$$\mathscr{T}^* = \mathscr{T} \cup \left\{ X^* \setminus K \;\middle|\; K \subseteq X \text{ compact} \right\}$$

Keep in mind that since X is Hausdorff, any compact subset $K \subseteq X$ is closed (by Proposition 7.3), and so $X \setminus K$ is open in X. In particular, X itself is an open subset of X^*.

It is straightforward to check that $\mathscr{T}^*$ is a topology on X^*:

(1) Clearly, $\emptyset$ and $X^* = X^* \setminus \emptyset$ belong to $\mathscr{T}^*$.
(2) For unions of open sets, we consider three cases. First, an arbitrary union of sets in $\mathscr{T}$ again belongs to $\mathscr{T}$ (because $\mathscr{T}$ is a topology on X). Second,

$$\bigcup_{i \in I} X^* \setminus K_i = X^* \setminus \bigcap_{i \in I} K_i,$$

and the intersection of all the K_i is a closed subspace of a compact space, and therefore again compact (by Proposition 6.13). Third,

$$U \cup (X^* \setminus K) = X^* \setminus \big(K \cap (X \setminus U)\big)$$

and because $X \setminus U$ is closed, the intersection $K \cap (X \setminus U)$ is compact.

(3) For intersections of open sets, there are again three cases. First, a union of two sets in $\mathscr{T}$ again belongs to $\mathscr{T}$; second,

$$(X^* \setminus K_1) \cup (X^* \setminus K_2) = X^* \setminus (K_1 \cup K_2),$$

and $K_1 \cup K_2$ is obviously again compact; third,

$$U \cap (X^* \setminus K) = U \cap (X \setminus K)$$

is the intersection of two open sets in X, and therefore an element of $\mathscr{T}$.

It is also easy to see that the function $f\colon X \to X^*$ defined by $f(x) = x$ is an embedding: it gives a bijection between X and its image in X^*, and because of how we defined $\mathscr{T}^*$, the function f is both continuous and open.

Finally, we show that X^* is a compact Hausdorff space. Consider an arbitrary open covering of X^* by sets in $\mathscr{T}^*$. At least one of the open sets has to contain the point ∞; pick one such set $X^* \setminus K$. Then K has to lie inside the union of the remaining open sets in the covering; by compactness, finitely many of these sets will do the job, and together with the set $X^* \setminus K$, we obtain a finite subcovering of our original covering. This proves that X^* is compact. To prove that X^* is Hausdorff, consider two distinct points $x, y \in X^*$. If $x, y \in X$, they can be separated by open sets because X is Hausdorff. If, say, $y = \infty$, we use the local compactness of X to find a neighborhood U of x whose closure $\overline{U} \subseteq X$ is compact; then U and $X^* \setminus \overline{U}$ are disjoint neighborhoods of the points x and y, respectively. □

When X is not already compact, the space in the theorem is called the *one-point compactification* of X. In the homework for this week, you can find several examples of one-point compactifications.

Lecture 9: September 22

Our topic today is Tychonoff's theorem about the compactness of product spaces. We showed some time ago that the product of two compact spaces is again compact; by induction, it follows that the same holds for any finite product.

Theorem 9.1 (Tychonoff's theorem)**.** *Let $\{X_i\}_{i\in I}$ be a collection of topological spaces. If each X_i is compact, then the product*

$$X = \prod_{i\in I} X_i$$

is compact in the product topology.

For a product $X \times Y$ of two spaces, we were able to prove compactness by looking at the fibers of the projection to X, and then at X; obviously, this kind of argument will not work in the case of an infinite product. Instead, the proof of Tychonoff's theorem relies heavily of the axiom of choice. We noted earlier that it is needed to show that infinite products of nonempty sets are nonempty: in fact, one formulation of the axiom of choice is that

$$X_i \neq \emptyset \text{ for every } i \in I \quad \Longrightarrow \quad \prod_{i\in I} X_i \neq \emptyset.$$

But since the empty set is still compact, this is not the only reason why we need the axiom of choice: as explained in one of this week's homework problems, Tychonoff's theorem is actually equivalent to the axiom of choice.

Zorn's lemma. Constructions with infinite sets often involve making an infinite number of choices that are not independent from each other. To deal with such problems, set theorists have come up with several other results that, although logically equivalent to the axiom of choice, are more convenient in practice. Probably the most useful among these is Zorn's lemma about partially ordered sets.

Definition 9.2. A *partial ordering* on a set E is a relation $\leq$ with the following properties:

(1) Symmetry: $x \leq x$ for every $x \in E$.
(2) Reflexivity: $x \leq y$ and $y \leq x$ imply $x = y$.
(3) Transitivity: $x \leq y$ and $y \leq z$ implies $x \leq z$.

Note that we do not assume that every two elements of E are comparable; recall from earlier in the semester that a *linear ordering* is a partial ordering in which $x \leq y$ or $y \leq x$ is true for every $x, y \in E$.

Example 9.3. The power set of a set S is the collection of all subsets of S; it is partially ordered by inclusion. More generally, we can consider any collection of subsets of S; except in very special cases, this is not a linear ordering.

Zorn's lemma gives a condition under which a partial ordering $(E, \leq)$ has maximal elements. As some of the elements of E may not be comparable, it does not make sense to look for a maximum (= an element $m \in E$ with $x \leq m$ for every $x \in E$); instead, we settle for the following.

Definition 9.4. An element $m \in E$ is called *maximal* if $m \leq x$ for some $x \in E$ implies $m = x$.

Note that there may be more than one maximal element; it is easy to find examples in the case of subsets of a set. At this point, someone objected that not every partially ordered set has maximal elements: for example, there are no maximal elements in $\mathbb{R}$, because for every $m \in \mathbb{R}$, the number $m+1 \in \mathbb{R}$ is strictly greater; equivalently, the sequence $1, 2, 3, \ldots$ is unbounded. Interestingly, it turns out that the existence of increasing families of elements without upper bound is the only obstacle.

Theorem 9.5 (Zorn's lemma)**.** *Let $(E, \leq)$ be a partially ordered set. If every well-ordered subset of E has an upper bound in E, then E has at least one maximal element.*

Here a subset $W \subseteq E$ is said to be *well-ordered* if W is linearly ordered by $\leq$, in such a way that every nonempty subset of W contains a minimum; an *upper bound* for W is an element $b \in E$ such that $x \leq b$ for every $x \in W$. In the textbooks, Zorn's lemma is often stated with the condition that every linearly ordered subset of E should have an upper bound in E; our version is due to Kneser (1950).

Let me first explain very roughly how the axiom of choice comes into play. An obvious strategy for proving Zorn's lemma is the following. Pick an arbitrary element $x_0 \in E$. If x_0 happens to be maximal, we are done; otherwise, the set of strictly larger elements is nonempty. Pick one such element $x_1 \neq x_0$ with $x_0 \leq x_1$; if x_1 is maximal, we are done; etc. If we continue this process until we can no longer add new elements, we ought to end up with a (possibly uncountable) increasing family of elements of E. (If the family is countable, we can index it by the natural numbers; if not, we need to use a larger ordinal number than ω.) The family is well-ordered, and therefore has an upper bound; by construction, any upper bound must be a maximal element in E. Of course, this is not rigorous: one problem is that each choice in the construction depends on all the preceding ones, whereas the axiom of choice is about making independent choices.

In case you want to know exactly how the axiom of choice implies Zorn's lemma, let me now describe Kneser's proof (which is written in German). To shorten the notation, let us agree that $x < y$ stands for $x \leq y$ and $x \neq y$. For each $x \in E$ that is not maximal, the set of $y \in E$ with $x < y$ is nonempty; using the axiom of choice, we obtain the existence of a function $f\colon E \to E$ with the property that

$$\begin{cases} x < f(x) & \text{if } x \in E \text{ is not maximal,} \\ x = f(x) & \text{if } x \in E \text{ is maximal.} \end{cases}$$

By another application of the axiom of choice, we can choose for every well-ordered subset $W \subseteq E$ an upper bound $b(W)$; here b is a function from the set of well-ordered subsets of E to the set E. Having made these two choices, Zorn's lemma is now evidently a consequence of the following fixed point theorem.

Theorem 9.6. *Let $(E, \leq)$ be a partially ordered set in which every well-ordered subset W has an upper bound $b(W)$. If $f\colon E \to E$ is a function with the property that $x \leq f(x)$ for all $x \in E$, then there is some element $m \in E$ with $m = f(m)$.*

Kneser's observation is that Theorem 9.6 can be proved without further appeals to the axiom of choice. Before we can get into the proof, we have to introduce one item of notation and one definition. Given a well-ordered subset $W \subseteq E$ and an element $x \in W$, the set

$$W_x = \{\, y \in W \mid y < x \,\}$$

is called a *section* of W; it is again a well-ordered subset of E.

From now on, we consider a partially ordered set $(E, \leq)$ that satisfies the assumptions in Theorem 9.6. The following definition formalizes the construction of increasing families of elements that I sketched above.

Definition 9.7. A well-ordered subset $C \subseteq E$ is called a *chain* if every $x \in C$ satisfies $x = f\big(b(C_x)\big)$.

We observe that in any pair of chains, one of the two contains the other.

Lemma 9.8. *If C and D are two chains in E, then either $C \subseteq D$ or $D = C_c$ for some $c \in C$.*

Proof. If $C \not\subseteq D$, the set $C \setminus D$ is nonempty, and because C is well-ordered, there is a unique smallest element $c \in C \setminus D$. Of course, this means that

$$C_c = \big\{\, x \in C \;\big|\; x < c \,\big\} \subseteq D.$$

I claim that $D = C_c$. Suppose that this was not the case; then $D \setminus C_c$ is nonempty, and because D is well-ordered, there is a unique smallest element $d \in D \setminus C_c$; as before, the minimality of d means that

$$D_d \subseteq C_c \subseteq D.$$

Now we cannot have $D_d = C_c$, because then the chain property would give

$$d = f\big(b(D_d)\big) = f\big(b(C_c)\big) = c,$$

contradicting our choice of c. So there must be some element of C_c that does not belong to D_d; let $x \in C_c \setminus D_d$ be the unique smallest one. Since x is in D but not in D_d, it must satisfy $d \leq x$; in particular, we have $D_d \subseteq C_x$. Since x is minimal, it is clear that $D_d = C_x$; but now the chain property gives

$$d = f\big(b(D_d)\big) = f\big(b(C_x)\big) = x,$$

and since $x \in C_c$, this contradicts our choice of d. The conclusion is that $D = C_c$, and hence that $D \subseteq C$. □

It is now a relatively easy matter to show that there is a unique maximal chain.

Lemma 9.9. *Let K be the union of all the chains in E. Then K is again a chain.*

Proof. We first argue that K is again well-ordered. By Lemma 9.8, any two elements $x, y \in K$ are contained in a single chain C; because C is linearly ordered, x and y are comparable, and so K is also linearly ordered. Now let $S \subseteq K$ be an arbitrary nonempty subset. Take any element $s \in S$; it is contained in a chain C, and since C is well-ordered, the intersection $S \cap C$ has a smallest element s_0. I claim that s_0 is the minimum of S. To see why, take any $x \in S$. If $x \in C$, it is clear that $s_0 \leq x$; if not, x belongs to another chain D. Now D is not entirely contained in C, and so Lemma 9.8 shows that $C = D_d$ for some $d \in D$. Because x is not an element of C, it must be that $d \leq x$; but then

$$s_0 < d \leq x,$$

proving that s_0 is indeed the minimum of S.

To verify that K is a chain, we have to show that $x = f\big(b(K_x)\big)$ for every $x \in K$. Let C be a chain containing x. A similar argument as above shows that $C_x = K_x$, and hence that

$$x = f\big(b(C_x)\big) = f\big(b(K_x)\big),$$

proving that K is also a chain. □

We can now finish the proof of Theorem 9.6 by showing that $m = f\big(b(K)\big)$ is the desired fixed point. The properties of f guarantee that

$$b(K) \leq f\big(b(K)\big) = m,$$

and so m is an upper bound for K. If we had $m \notin K$, then the set $K \cup \{m\}$ would be a chain in E; this is not possible because of how we defined K. The conclusion is that $m \in K$, and therefore

$$m \leq b(K) \leq f\big(b(K)\big) = m.$$

These inequalities show that $m = b(K)$, and hence that $m = f(m)$.

Proof of Tychonoff's theorem. Now we come back to Tychonoff's theorem about the product

$$X = \prod_{i \in I} X_i,$$

of a family of compact topological spaces X_i. To show that X is compact, we have to argue that every open covering of X has a finite subcovering. Recall that the product topology has a basis

$$\mathscr{B} = \left\{ \prod_{i \in I} U_i \;\middle|\; U_i \subseteq X_i \text{ open and } U_i = X_i \text{ for all but finitely many } i \in I \right\};$$

as we observed earlier, it is enough to prove that every open covering $\mathscr{U} \subseteq \mathscr{B}$ has a finite subcovering. If we denote by $p_i \colon X \to X_i$ the projection to the i-th factor, and consider the smaller collection of open sets

$$\mathscr{C} = \left\{ p_i^{-1}(U_i) \;\middle|\; i \in I \text{ and } U_i \subseteq X_i \text{ open} \right\},$$

then every open set in $\mathscr{B}$ is the intersection of finitely many open sets in $\mathscr{C}$. This makes $\mathscr{C}$ a "subbasis" for the product topology, in the following sense.

Definition 9.10. A family of open sets $\mathscr{C}$ in a topological space is called a *subbasis* if the collection of all finite intersections of sets in $\mathscr{C}$ is a basis for the given topology.

As a first step towards proving the compactness of X, let us show that every open covering by sets in $\mathscr{C}$ has a finite subcovering.

Lemma 9.11. *Let $\mathscr{U} \subseteq \mathscr{C}$ be an open covering of X by subbasic open sets. Then there are finitely many open sets $U_1, \ldots, U_n \in \mathscr{C}$ such that $X = U_1 \cup \cdots \cup U_n$.*

Proof. For each $i \in I$, we consider the collection of open sets

$$\mathscr{U}_i = \big\{\, U \subseteq X_i \;\big|\; p_i^{-1}(U) \in \mathscr{U} \,\big\}.$$

If $\mathscr{U}_i$ does not cover X_i for any $i \in I$, then the axiom of choice would allow us to choose, for every index $i \in I$, an element $x_i \in X_i \setminus \bigcup \mathscr{U}_i$. Now consider the point $x = (x_i)_{i \in I}$. By construction, it does not lie in any open set of the form $p_i^{-1}(U)$ with $U \in \mathscr{U}_i$; but this is absurd because $\mathscr{U}$ was supposed to be an open covering of X by sets in $\mathscr{C}$. The conclusion is that there must be some index $i \in I$ with

$$X_i = \bigcup_{U \in \mathscr{U}_i} U.$$

Because X_i is compact, finitely many open sets in $\mathscr{U}_i$ suffice to cover X_i; now their preimages under p_i give us the desired finite subcovering of $\mathscr{U}$. □

Now we have the following general result, which says that in order to check whether a space X is compact, it suffices to look at coverings by open sets in a subbasis.

Theorem 9.12 (Alexander's subbasis theorem)**.** *Let X be a topological space and let $\mathscr{C}$ be a subbasis for the topology. If every open covering $\mathscr{U} \subseteq \mathscr{C}$ has a finite subcovering, then X is compact.*

Proof. We shall prove the contrapositive, namely that if X is not compact, then there must be an open covering $\mathscr{V} \subseteq \mathscr{C}$ without finite subcovering. Let $\mathscr{B}$ be the basis generated by the subbasis $\mathscr{C}$; every element of $\mathscr{B}$ is an intersection of finitely many open sets in $\mathscr{C}$.

Since X is not compact, there is at least one open covering $\mathscr{U} \subseteq \mathscr{B}$ without finite subcovering. Now the idea is to look for a maximal open covering with this property, in the hope that it will contain enough sets from the subbasis $\mathscr{C}$. This obviously requires Zorn's lemma. Consider the collection E of all open coverings $\mathscr{U} \subseteq \mathscr{B}$ without finite subcovering. It is partially ordered by inclusion: $\mathscr{U}_1 \subseteq \mathscr{U}_2$ means that the covering $\mathscr{U}_2$ contains more open sets than the covering $\mathscr{U}_1$. If $W \subseteq E$ is a well-ordered subset, then the open covering

$$\big\{\, U \in \mathscr{B} \;\big|\; U \in \mathscr{U} \text{ for some } \mathscr{U} \in W \,\big\}$$

again belongs to E, and is therefore an upper bound for W: indeed, any finite number of open sets in this covering will be in one $\mathscr{U} \in W$ (because W is linearly ordered), and therefore cannot cover X (because $\mathscr{U}$ does not have any finite subcovering). Now Zorn's lemma guarantees the existence of a maximal open covering $\mathscr{V} \subseteq \mathscr{B}$ without finite subcovering. Maximality translates into the following remarkable property: for every basic open set $V \in \mathscr{B}$, either $V \in \mathscr{V}$ or $\mathscr{V} \cup \{V\}$ has a finite subcovering.

Now let U be an arbitrary open set in $\mathscr{V}$; since $U \in \mathscr{B}$, it must be of the form $U = U_1 \cap \cdots \cap U_n$ for certain $U_1, \ldots, U_n \in \mathscr{C}$. I claim that for some $i = 1, \ldots, n$, we must have $U_i \in \mathscr{V}$. If this was not the case, then $U_i \notin \mathscr{V}$, and by maximality, $\mathscr{V} \cup \{U_i\}$ has a finite subcovering for every $i = 1, \ldots, n$. This means that finitely many open sets in $\mathscr{V}$ cover the complement $X \setminus U_i$; but since

$$(X \setminus U_1) \cup \cdots \cup (X \setminus U_n) = X \setminus (U_1 \cap \cdots \cap U_n),$$

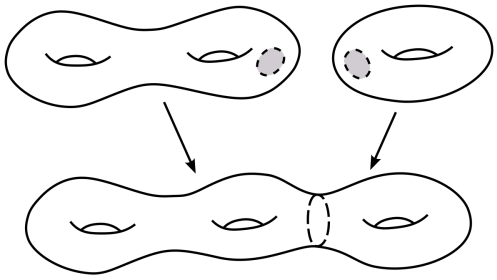
this would produce a finite covering of X by open sets in $\mathscr{V}$, which is not allowed. So every set $U \in \mathscr{V}$ is contained in a larger open set $U_i \in \mathscr{V} \cap \mathscr{C}$. The conclusion is that $\mathscr{V} \cap \mathscr{C}$ is still an open covering of X; since it cannot contain any finite subcovering either, we get the desired result. □

Taken together, Lemma 9.11 and Alexander's subbasis theorem prove Tychonoff's theorem. Note that the axiom of choice was used twice: once during the proof of Lemma 9.11 and once in the form of Zorn's lemma.

Connected sums. We can join two manifolds of the same dimension together to make another manifold with a more complicated topology. This construction gives us other interesting examples of compact connected surfaces.

Suppose that X and Y are two connected n-dimensional manifolds. In each of the two manifolds, we remove a small open set homeomorphic to a ball in $\mathbb{R}^n$, and then we paste the two pieces together along the boundaries of the two balls, which

are homeomorphic to the $(n-1)$-sphere $\mathbb{S}^{n-1}$. The resulting manifold is called the *connected sum* of X and Y and is denoted by the symbol $X\#Y$.



Here is a more careful description of the construction. Choose two points $x \in X$ and $y \in Y$. Since X is a manifold, some open neighborhood of x is homeomorphic to an open subset of $\mathbb{R}^n$. If we let $B = \left\{ x \in \mathbb{R}^n \;\middle|\; |x| < 1 \right\}$ be the open unit ball in $\mathbb{R}^n$, and $\bar{B} = \left\{ x \in \mathbb{R}^n \;\middle|\; |x| \leq 1 \right\}$ its closure, we can therefore choose a function

$$f\colon \bar{B} \to X$$

such that $f(0) = x$ and such that f is a homeomorphism between $\bar{B}$ and its image. The complement $X \setminus f(B)$ is now a manifold with boundary, the boundary being homeomorphic to $\mathbb{S}^{n-1}$ via f. Similarly choose $g\colon \bar{B} \to Y$. On the disjoint union

$$\big(X \setminus f(B)\big) \sqcup \big(Y \setminus g(B)\big),$$

we get an equivalence relation by declaring that $f(x) \sim g(x)$ for every $x \in \mathbb{S}^{n-1}$. The connected sum $X\#Y$ is then defined as the quotient by this equivalence relation, together with the quotient topology.

Note. The construction of the connected sum depends on the choice of $x \in X$ and $y \in Y$ (and of the functions f and g), but one can show that different choices lead to spaces that are homeomorphic. The key point is the following fact: If $x_1, x_2 \in X$ are two points on a connected manifold, then there is a homeomorphism $\phi\colon X \to X$ with the property that $\phi(x_1) = x_2$. (In other words, the group of homeomorphisms acts transitively on X.) Here is a sketch of the proof. First, one shows that for any two points $x_1, x_2 \in \bar{B}$ in the closed unit ball, there is a homeomorphism $f\colon \bar{B} \to \bar{B}$ such that $f(x_1) = x_2$ and such that f acts as the identity on the unit sphere $\mathbb{S}^{n-1}$. (One can actually write down such a homeomorphism explicitly.) By the pasting lemma, this means that if the two points $x_1, x_2 \in X$ are contained in a subset homeomorphic to $\bar{B}$, then there is a homeomorphism of X taking x_1 to x_2. To deal with the general case, we use the fact that X is path-connected to find a path joining x_1 and x_2; by compactness, this path can be covered by finitely many open sets whose closures are homeomorphic to $\bar{B}$. We can then construct the desired homeomorphism in steps, by moving x_1 from the first ball into the second ball, then into the third ball, and so on.

You should convince yourself that $X\#Y$ is again an n-dimensional manifold. It is also easy to see that if X and Y are connected (or compact), then $X\#Y$ is again connected (or compact).

Example 9.13. Recall that the torus is defined as $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$. The connected sum $\mathbb{T}\#\mathbb{T}$ is called the 2-holed torus; similarly, the connected sum of n copies of $\mathbb{T}$ is called the n-holed torus. It is a compact and connected 2-dimensional manifold.

Example 9.14. The *projective plane* is the space of lines in $\mathbb{R}^3$. As a topological space, we can define it for example as the quotient space of the closed unit disk

$$\left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 \leq 1 \right\}$$

by the equivalence relation $(x, y) \sim (-x, -y)$ for $(x, y) \in \mathbb{S}^1$. It is another example of a compact and connected 2-dimensional manifold; like the Klein bottle, it is not orientable. We denote the projective plane by the symbol $\mathbb{P}$.

Here is a nice exercise: Describe the connected sum $\mathbb{P}\#\mathbb{P}$. (Hint: What do you get if you remove an open disk from $\mathbb{P}$?)

Lecture 10: September 27

Countability axioms. Today, I am going to introduce several additional conditions – called "axioms" for historical reasons – that a topological space may or may not satisfy. All of these conditions are saying that the topology on a given space X is "nice" in some way; for example, we will see that $\mathbb{R}^n$ satisfies all of them. By assuming one or several of these conditions about a topological space X, we can usually prove more interesting theorems about X.

The first set of conditions is known as the *countability axioms*.

Definition 10.1. Let X be a topological space.

(a) X is said to be *first countable* if every point $x \in X$ has a countable neighborhood basis: there are countably many open sets $B_1(x), B_2(x), \ldots$ such that every open set U containing the point x contains at least one $B_n(x)$.
(b) X is said to be *second countable* if X has a countable basis: there are countably many open sets $B_1, B_2, \ldots$ that form a basis for the topology.

The second countability axiom is saying that the topology on X can be described by countably many conditions; the first countability axiom is saying that this is true locally at each point of X. Note that a second countable space is automatically first countable, since all those B_n with $x \in B_n$ will be countable neighborhood basis at the point x. Both conditions tell us that there are not too many open sets in X: for instance, if X is second countable, then every open set must be a union of certain B_n, and so the cardinality of the topology is at most that of the power set of $\mathbb{N}$.

Example 10.2. $\mathbb{R}^n$ is second countable: the collection of all open balls $B_r(x_0)$ with $x_0 \in \mathbb{Q}^n$ and $r \in \mathbb{Q} \cap (0, \infty)$ is a countable basis for the topology.

Example 10.3. Every metric space is first countable: at every point x, the collection of open balls $B_{1/n}(x)$ is a countable neighborhood basis. On the other hand, a metric space does not have to be second countable: we have seen before that the discrete topology on a set X always comes from a metric; when X is uncountable, the discrete topology is obviously not second countable.

Example 10.4. The infinite product $\mathbb{R}^{\mathbb{N}}$ in the box topology is not first countable; this was the content of a homework problem earlier in the semester.

Example 10.5. Here is another interesting example of a space that is not second countable. Let $\mathbb{R}_\ell$ denote the set of real numbers with the *lower limit topology*: a basis for the topology consists of all half-open intervals $[a, b)$ with $a < b$. Since

$$(a, b) = \bigcup_{c > a} [c, b),$$

this topology is finer than the usual topology on $\mathbb{R}$. It is not second countable, for the following reason: Suppose that $\mathscr{B}$ is any basis for the topology on $\mathbb{R}_\ell$. For each $x \in \mathbb{R}$, the set $[x, x + 1)$ is open, and so there would have to be a basic open set $B_x \in \mathscr{B}$ with

$$x \in B_x \subseteq [x, x + 1).$$

But now we have $B_x \neq B_y$ for $x \neq y$, because we have $x = \inf B_x$. This means that the cardinality of $\mathscr{B}$ is at least that of $\mathbb{R}$, and so there is no countable basis.

Example 10.6. Any subspace of a first/second countable space is again first/second countable.

Second countable spaces are nice because many properties such as closedness or compactness can be checked using sequences. We have already proved that in a first countable space, a subset is closed if and only if it is sequentially closed (see the remark just after Proposition 3.8). In a second countable space, the same is true for compactness.

Proposition 10.7. *Let X be a topological space that is second countable. Then X is compact if and only if it is* sequentially compact*, that is if every sequence in X has a convergent subsequence.*

For the proof, we need the following lemma about the existence of convergent subsequences. Note that we are not assuming that X is Hausdorff, and so a sequence may have more than one limit.

Lemma 10.8. *Let $x_1, x_2, \dots$ be a sequence of points in a first countable space X. Let $x \in X$ be a point such that for every neighborhood U of x, one has $x_n \in U$ for infinitely many n. Then a subsequence of $x_1, x_2, \dots$ converges to x.*

Proof. X has a countable neighborhood basis $B_1(x), B_2(x), \dots$ at the point x; by successively intersecting these open sets, we can arrange that

$$B_1(x) \supseteq B_2(x) \supseteq \cdots$$

The set of n with $x_n \in B_1(x)$ is infinite; let n_1 be the smallest element. The set of n with $x_n \in B_2(x)$ is also infinite; let n_2 be the smallest element with $n_2 > n_1$. Continuing in this way, we obtain an increasing sequence of integers $n_1 < n_2 < \cdots$, such that $x_{n_k} \in B_k(x)$ for every k. In particular, every $B_k(x)$ contains all but finitely many points of the subsequence $x_{n_1}, x_{n_2}, \dots$; since the $B_k(x)$ form a neighborhood basis, this implies that x is a limit of the subsequence. □

Proof of Proposition 10.7. We first show that compactness implies sequential compactness; this part of the proof only uses the fact that X is first countable. Let $x_1, x_2, \dots \in X$ be an arbitrary sequence. Suppose that it does not have convergent subsequence; then Lemma 10.8 tells us that every point $x \in X$ must have some neighborhood U_x that only contains x_n for finitely many values of n. Now the open covering

$$X = \bigcup_{x \in X} U_x$$

has a finite subcovering (because X is compact), and so at least one of the open sets in this finite subcovering should contain infinitely many terms of our sequence. This is a contradiction, and so there must be a convergent subsequence after all.

Next, we show that sequential compactness implies compactness. Let $\mathscr{B}$ be a basis for the topology on X. Recall that compactness of X is equivalent to saying that every open covering by basic open sets has a finite subcovering. So let $\mathscr{U} \subseteq \mathscr{B}$ be an open covering by basic open sets; since there are countably many open sets in the covering, we can enumerate them as $U_1, U_2, \dots$. To show that X is compact, we have to show that

$$U_1 \cup \cdots \cup U_n = X$$

for some n. Suppose this was false; then for each n, we can choose a point

$$x_n \in X \setminus U_1 \cup \cdots \cup U_n.$$

By assumption, a subsequence of this sequence has a limit $x \in X$; because $\mathscr{U}$ is an open covering of X, there must be some m with $x \in U_m$, and so U_m has to contain

x_n for infinitely many values of n. But by construction, $x_n \notin U_m$ for $n \geq m$, and so we have arrived at a contradiction. □

Another useful property of second countable spaces is the following; in analysis, spaces with this property are called "separable".

Proposition 10.9. *If X is second countable, then it contains a countable dense subset.*

Proof. Let $B_1, B_2, \dots$ be a countable basis for the topology; without loss of generality, each B_n is nonempty. For every n, choose a point $x_n \in B_n$; then $\{x_1, x_2, \dots\}$ is countable and dense, because by construction, every nonempty open subset of X contains at least one x_n. □

Separation axioms. The second set of conditions are the so-called *separation axioms*; they are generalizations of the Hausdorff axiom. Each of these conditions is saying that there are sufficiently many open sets in X to "separate" certain kinds of subsets of X; we will see later how they can be used to construct interesting nontrivial continous functions on X.

Definition 10.10. Let X be a topological space in which every one-point set is closed.

(a) X is called *Hausdorff* if any two distinct points can be separated by open sets: given two points $x \neq y$, there are disjoint open sets $U, V \subseteq X$ with $x \in U$ and $y \in V$.
(b) X is called *regular* if any point and any closed set not containing the point can be separated by open sets: given a point $x \in X$ and a closed set $A \subseteq X$ with $x \notin A$, there are disjoint open sets $U, V \subseteq X$ with $x \in U$ and $A \subseteq V$.
(c) X is called *normal* if any two disjoint closed sets can be separated by open sets: given two closed sets $A, B \subseteq X$ with $A \cap B = \emptyset$, there are disjoint open sets $U, V \subseteq x$ with $A \subseteq U$ and $B \subseteq V$.

The condition that sets of the form $\{x\}$ are closed is also called T_1; the Hausdorff axiom is T_2, regularity is T_3, and normality is T_4. This terminology comes from the German word "Trennungsaxiom", which translates as "separation axiom". Note that every Hausdorff space is automatically T_1; the other two conditions no longer imply that one-point sets are closed, and so we include this in the definition. (There is also a condition called T_0; if I remember correctly, it says that for two distinct points, there is an open set containing one but not the other. In addition, there are various intermediate notions, but those are mostly just used by specialists.)

Example 10.11. Here is an example of a non-Hausdorff space where all one-point sets are closed: take any infinite set X with the finite complement topology. Here any two nonempty open sets intersect each other, and so we obviously cannot separate anything.

As suggested by the words "regular" and "normal", regular or normal spaces are closer to our intuition: for example, regularity means that if A is a closed set and $x \notin A$, then x is a certain distance away from A. We will also see next week that normality in particular has very interesting consequences. For now, let us look at a few examples from earlier in the semester.

Example 10.12. We already know that every compact Hausdorff space X is regular (see the comments after Proposition 7.3). We can use essentially the same proof to show that X is also normal. Let A and B be disjoint closed subsets of X. For every $x \in A$, we have $x \notin B$, and since X is regular, we can find disjoint open sets U_x and V_x with $x \in U_x$ and $B \subseteq V_x$. Now A is closed, and therefore compact; since we have

$$A \subseteq \bigcup_{x \in A} U_x,$$

there are finitely many points $x_1, \ldots, x_n \in A$ with $A \subseteq U_{x_1} \cup \cdots \cup U_{x_n}$. If we now define

$$U = U_{x_1} \cup \cdots \cup U_{x_n} \quad \text{and} \quad V = V_{x_1} \cap \cdots \cap V_{x_n},$$

then U and V are open, disjoint, $A \subseteq U$, and $B \subseteq V$.

Example 10.13. Metric spaces are another interesting class of examples: every metric space is normal. Since we already know that one-point sets are closed, it suffices to show that any two disjoint closed sets A and B can be separated by open sets. The main point is that $X \setminus B$ is open, and so for every $a \in A$, we can find a small positive number $r(a) > 0$ such that the open ball $B_{r(a)}(a) \subseteq X \setminus B$. Likewise, for every $b \in B$, we can find some $r(b) > 0$ such that $B_{r(b)}(b) \subseteq X \setminus A$. Now consider the two open sets

$$U = \bigcup_{a \in A} B_{\frac{1}{2}r(a)}(a) \quad \text{and} \quad V = \bigcup_{b \in B} B_{\frac{1}{2}r(b)}(b).$$

Evidently, $A \subseteq U$ and $B \subseteq V$; but in fact, U and V are also disjoint. For suppose that there was a point $x \in U \cap V$. Then we would have

$$d(x, a) < \frac{1}{2}r(a) \quad \text{and} \quad d(x, b) < \frac{1}{2}r(b)$$

for some $a \in A$ and some $b \in B$; by the triangle inequality,

$$d(a, b) \leq d(a, x) + d(x, b) < \frac{1}{2}r(a) + \frac{1}{2}r(b).$$

Now if $r(a) \leq r(b)$, we get $d(a, b) < r(b)$, which means that $a \in B_{r(b)}(b)$; but this contradicts that fact that $B_{r(b)}(b) \subseteq X \setminus A$. Similarly, $r(b) \leq r(a)$ would imply that $b \in B_{r(a)}(a) \subseteq X \setminus B$. The conclusion is that U and V must be disjoint after all.

The following result, which we did not discuss in class, shows how the countability axioms and separation axioms work together.

Theorem 10.14. *Every regular space with a countable basis is normal.*

Proof. Let X be a regular topological space with a countable basis $\mathscr{B}$. We have to prove that every pair of disjoint closed sets $A, B \subseteq X$ can be separated by open sets. We will first use regularity to construct countable open coverings for A and B, and then explain how to produce disjoint open sets containing A and B.

For any $x \in A$, the open set $X \setminus B$ contains x; because X is regular, there is a neighborhood of x whose closure does not intersect B. Now $\mathscr{B}$ is a basis, and so inside this neighborhood, we can certainly find a basic open set containing x whose closure does not intersect B. Since $\mathscr{B}$ is countable, the collection of all basic

open sets obtained in this way must also be countable; let us enumerate them as $U_1, U_2, \ldots$. We then have

$$A \subseteq \bigcup_{n=1}^{\infty} U_n,$$

and the closure of every U_n is disjoint from B. By the same argument, we can find basic open sets $V_1, V_2, \ldots$ with

$$B \subseteq \bigcup_{n=1}^{\infty} V_n$$

and $\overline{V_n} \cap A = \emptyset$. In this way, we obtain two open sets containing A and B. There is no reason why they should be disjoint, but we can use the following trick to make them so. Consider the collection of open sets

$$U_n' = U_n \setminus \bigcup_{k=1}^{n} \overline{V_k} \quad \text{and} \quad V_n' = V_n \setminus \bigcup_{k=1}^{n} \overline{U_k}.$$

Since every $\overline{V_k}$ is disjoint from A, we have $A \cap U_n' = A \cap U_n$, and therefore

$$A \subseteq \bigcup_{n=1}^{\infty} U_n' = U;$$

for the same reason, B is contained in V, which is the union of the V_n'. I claim that U and V are disjoint. Indeed, any point $x \in U \cap V$ would have to belong to some U_m' and some V_n'; if $m \leq n$, then

$$x \in U_m' \subseteq U_m \quad \text{and} \quad x \in V_n' = V_n \setminus \bigcup_{k=1}^{n} U_k,$$

which is clearly a contradiction. Since $n \leq m$ also leads to a contradiction, we conclude that X must be normal. □

Before we discuss some additional examples, let us first investigate which of the separation axioms are preserved under taking subspaces and products. Since many constructions in topology involve looking at subspaces or products of known spaces, this is very useful in practice.

Proposition 10.15. *If a topological space X is Hausdorff/regular, then every subspace of X is also Hausdorff/regular. If $(X_i)_{i \in I}$ is a collection of topological spaces that are all Hausdorff/regular, then their product*

$$\prod_{i \in I} X_i$$

is also Hausdorff/regular.

Proof. We already know that every subspace of a Hausdorff space is again Hausdorff, and that an arbitrary product of Hausdorff spaces is again Hausdorff. We will therefore concentrate on proving the same results for regular spaces.

First, let us consider a subspace Y of a regular space X. Obviously, every point $x \in Y$ is closed relative to Y, since it is the intersection of Y and the closed set $\{x\}$. Now suppose we are also given a subset $B \subseteq Y$ that is closed relative to Y, such that $x \notin B$. By definition of the subspace topology, there is a closed set $A \subseteq X$ such that $B = Y \cap A$. Because $x \in Y$, we have $x \notin A$, and so we separate x and

A by disjoint open sets $U, V \subseteq X$. But then $Y \cap U$ and $Y \cap V$ are disjoint, open relative to Y, and separate x and B.

Second, let us consider an arbitrary product of regular spaces X_i, indexed by a set I. We shall use the other formulation of regularity in Proposition 11.1, because it is easier to check. Given a point $x = (x_i)_{i \in I}$ and an open set U containing x, we can certainly find inside U a basic open set of the form

$$\prod_{i \in I} U_i$$

containing x; recall that $U_i = X_i$ for all but finitely many $i \in I$. If $U_i = X_i$, we define $V_i = X_i$; otherwise, we use the regularity of X to find an open subset V_i with $x \in V_i$ and $\overline{V_i} \subseteq U_i$. Now

$$\prod_{i \in I} V_i$$

is a neighborhood of x whose closure

$$\prod_{i \in I} \overline{V_i}$$

is contained inside the original open set U. □

This result is no longer true for normality: subspaces of normal spaces need not be normal. The proof breaks down in the case of two disjoint closed subsets of Y, because their extensions to closed sets in X may no longer be disjoint.

Counterexamples involving normal spaces. Unlike in the case of regular spaces, products of normal spaces need not be normal. The simplest example of this involves the lower limit topology.

Example 10.16. The lower limit topology $\mathbb{R}_\ell$ is normal. Since the topology is finer than the usual topology on $\mathbb{R}$, it is clear that every one-point set is closed. To prove normality, let A and B be disjoint closed subsets of $\mathbb{R}_\ell$. Every $a \in A$ belongs to the open set $\mathbb{R}_\ell \setminus B$, and so there must exist a basic open set of the form $[a, x_a)$ that is disjoint from B. Similarly, there must exist a basic open set of the form $[b, x_b)$ that is disjoint from A. If we now define

$$U = \bigcup_{a \in A} [a, x_a) \quad \text{and} \quad V = \bigcup_{b \in B} [b, x_b),$$

it is easy to see that U and V are disjoint open sets with $A \subseteq U$ and $B \subseteq V$.

Example 10.17. The product space $\mathbb{R}_\ell \times \mathbb{R}_\ell$ is not normal. This example shows that a product of two normal spaces need not be normal; since $\mathbb{R}_\ell$ is regular, and since the product of two regular spaces is again regular, it is also provides us with an example of a regular space that is not normal.

The proof that $\mathbb{R}_\ell \times \mathbb{R}_\ell$ is not normal is surprisingly tricky. Since the basic open sets in $\mathbb{R}_\ell$ are half-open intervals, the product space has a basis consisting of half-open squares of the form

$$S_r(x, y) = [x, x + r) \times [y, y + r).$$

The idea is to consider the line

$$L = \{\, (x, -x) \mid x \in \mathbb{R} \,\} \subseteq \mathbb{R}_\ell \times \mathbb{R}_\ell.$$

Observe that L is closed (because its complement is open), and that the subspace topology on L is discrete (because $L \cap S_1(x,-x) = \{(x,-x)\}$ is open relative to L). This means that *every* subset of L is actually closed in $\mathbb{R}_\ell \times \mathbb{R}_\ell$. In particular, the two sets

$$A = \left\{ (x,-x) \;\middle|\; x \in \mathbb{Q} \right\} \quad \text{and} \quad B = \left\{ (x,-x) \;\middle|\; x \notin \mathbb{Q} \right\}$$

are disjoint closed subsets; we shall argue that whenever we have $A \subseteq U$ and $B \subseteq V$, the two open sets U and V must intersect.

What does it mean to have $A \subseteq U$ and $B \subseteq V$? Simply that for every $x \in \mathbb{Q}$, some basic open set $S_r(x,-x)$ with $r > 0$ is contained in U, and that for every $x \notin \mathbb{Q}$, some basic open set $S_r(x,-x)$ with $r > 0$ is contained in V. Since A and B are extremely interwoven, it looks very likely that some of the basic open sets around the points of A would have to intersect some of the basic open sets around the points of B, and hence that $U \cap V \neq \emptyset$; the only problem is that we do not have much control over how big each basic open set is.

In a situation like this, where something is true for every real number, it is often useful to think of Baire's theorem. In fact, we can write $\mathbb{R}$ as a countable union

$$\mathbb{R} = \mathbb{Q} \cup \bigcup_{n=1}^{\infty} E_n,$$

where $E_n = \left\{ x \notin \mathbb{Q} \;\middle|\; S_{1/n}(x,-x) \subseteq V \right\}$. The reason is that every $x \notin \mathbb{Q}$ has to belong to E_n for sufficiently large n, simply because $B \subseteq V$. Of course, the E_n will not be closed; but we also have

$$\mathbb{R} = \mathbb{Q} \cup \bigcup_{n=1}^{\infty} \overline{E_n},$$

and now Baire's theorem gurantees that some $\overline{E_n}$ contains an open interval I. What this means is that the points of $E_n \cap I$ are dense in I.

We now have a uniform bound on the size of the basic open sets, and so we can easily show that $U \cap V \neq \emptyset$. Let $y \in I$ be any rational number; then $(y,-y) \in A$, and so there exists some $r > 0$ with $S_r(y,-y) \subseteq U$. On the other hand, I contains a dense subset of points $x \in I \cap E_n$ with $S_{1/n}(x,-x) \subseteq V$. If we choose one such point with $|x-y| < \min(r, 1/n)$, then

$$\emptyset \neq S_r(y,-y) \cap S_{1/n}(x,-x) \subseteq U \cap V.$$

This shows that $\mathbb{R}_\ell \times \mathbb{R}_\ell$ is not normal.

The second example is a subspace of a normal space that fails to be normal. It involves the minimal uncountable well-ordered set S_Ω, which you may remember from the homework. Let me first recall the definition of well-ordered sets.

Definition 10.18. A linear ordering $\leq$ on a set X is called a *well-ordering* if every nonempty subset of X contains a least element.

A typical example of a well-ordered set is $(\mathbb{N}, \leq)$. In fact, any well-ordered set X looks like the natural numbers "at the beginning": being well-ordered, X has a unique smallest element, a unique second-smallest element, etc. It is not easy to imagine well-ordered sets of larger cardinality, but in fact, the axiom of choice implies that *every* set can be well-ordered in some way.

Theorem 10.19 (Well-Ordering Theorem)**.** *Every nonempty set has at least one well-ordering.*

In particular, there are well-ordered sets of every possible cardinality. When Zermelo announced this result, nobody really believed it could be true. A closer examination of his proof showed that it involved choosing elements from infinitely many nonempty sets, and thus the axiom of choice was introduced into set theory. In fact, the Well-Ordering Theorem is another result that is logically equivalent to the axiom of choice.

Example 10.20. Here is how one can use Zermelo's theorem to prove the existence of an uncountable well-ordered set in which every section is countable. Let S be any uncountable set (such as $\mathbb{R}$), and let $\leq$ be a well-ordering on S; as usual, we shall use the notation $x < y$ to mean $x \leq y$ and $x \neq y$. For every $x \in S$, we consider the so-called *section*

$$S_x = \{\, y \in S \mid y < x \,\}.$$

If it happens that S_x is countable for every $x \in S$, then $(S, \leq)$ is the desired minimal uncountable well-ordered set. Otherwise, the subset

$$\{\, x \in S \mid S_x \text{ is uncountable} \,\}$$

is nonempty, and because $(S, \leq)$ is well-ordered, it contains a smallest element m. But then every section of S_m must be countable, and so $(S_m, \leq)$ has the properties we want.

We denote the minimal uncountable well-ordered set by the symbol S_Ω; we also let $\overline{S}_\Omega = S_\Omega \cup \{\Omega\}$, with the ordering in which Ω is the largest element. The well-ordering on S_Ω has the following curious property.

Lemma 10.21. *Every countable subset of S_Ω has a least upper bound in S_Ω.*

Proof. Let $C \subseteq S_\Omega$ be a countable subset. Since every section of S_Ω is countable, the set

$$\{\, x \in S_\Omega \mid x < y \text{ for some } y \in C \,\} = \bigcup_{y \in C} \{\, x \in S_\Omega \mid x < y \,\}$$

is countable, and therefore a proper subset of the uncountable set S_Ω. Every element of S_Ω not in this subset is an upper bound for C; because S_Ω is well-ordered, there is a unique smallest element, which is a least upper bound for C. □

Example 10.22. The order topology on $\overline{S}_\Omega$ is compact Hausdorff, and therefore normal. Recall that $\overline{S}_\Omega$ has both a largest element Ω and a smallest element; being well-ordered, it also has the least upper bound property. The proof of Theorem 7.4 shows that $\overline{S}_\Omega$ is compact in the order topology; on the other hand, every order topology is Hausdorff.

Example 10.23. The product $S_\Omega \times \overline{S}_\Omega$ is not normal. Since it is a subspace of $\overline{S}_\Omega \times \overline{S}_\Omega$, which is again compact Hausdorff and therefore normal, this example shows that a subspace of a normal space does not have to be normal.

The proof is again not easy. Recall that because S_Ω is Hausdorff, the diagonal in $S_\Omega \times S_\Omega$ is closed; consequently,

$$A = \{\, (x, x) \mid x \in S_\Omega \,\}$$

is a closed subset of $S_\Omega \times \overline{S}_\Omega$. Likewise, the set

$$B = \{\, (x, \Omega \mid x \in S_\Omega \,\}$$

is also closed, and clearly disjoint from A. We shall prove that it is not possible to separate A and B by disjoint open sets in $S_\Omega \times \overline{S}_\Omega$.

Suppose to the contrary that we had $A \subseteq U$ and $B \subseteq V$ with $U \cap V = \emptyset$. For every point $x \in S_\Omega$, the vertical line

$$\{\, (x, y) \mid y \in \overline{S}_\Omega \,\}$$

contains the point $(x, x) \in A$. Since U is a neighborhood of (x, x), and V is a neighborhood of (x, Ω), it is clear that there exists an element $y \in \overline{S}_\Omega$ with $x < y < \Omega$ and $(x, y) \notin U$. Since $\overline{S}_\Omega$ is well-ordered, we can define $f(x)$ to be the smallest element for which this holds; thus we obtain a function

$$f\colon S_\Omega \to S_\Omega$$

with the property that $x < f(x)$ and $\big(x, f(x)\big) \notin U$. Now choose any $x_1 \in S_\Omega$ and consider the increasing sequence

$$x_1 < x_2 < x_3 < \cdots$$

defined by setting $x_{n+1} = f(x_n)$ for $n = 1, 2, \ldots$. According to Lemma 10.21, the sequence has a least upper bound $b \in S_\Omega$; being increasing, it is forced to converge to b. We now obtain a contradiction by looking at the sequence of points

$$\big(x_n, f(x_n)\big) = \big(x_n, x_{n+1}\big).$$

On the one hand, it converges to the point $(b, b) \in U$; on the other hand, none of the points $\big(x_n, f(x_n)\big)$ belongs to U. Since U is open, this is absurd, and so $S_\Omega \times \overline{S}_\Omega$ cannot be normal.

Lecture 11: September 29

We can formulate the definition of regularity and normality differently, in a way that emphasizes separating subsets inside an open set from the "boundary" of the open set.

Proposition 11.1. *Let X be a topological space in which one-point sets are closed.*

(a) X is regular if and only if for every $x \in X$ and every open set U containing x, there is a smaller open set V with $x \in V$ and $\overline{V} \subseteq U$.

(b) X is normal if and only if for every closed subset $A \subseteq X$ and every open set U containing A, there is a smaller open set V with $A \subseteq V$ and $\overline{V} \subseteq U$.

Proof. We will only prove (a), since (b) is very similar. Let us first show that regularity implies the condition above. Any point $x \in U$ does not belong to the closed set $X \setminus U$; since X is regular, we can find disjoint open sets V and W with $x \in V$ and $X \setminus U \subseteq W$. Now V is contained in the closed set $X \setminus W$, and so

$$\overline{V} \subseteq X \setminus W \subseteq U,$$

as claimed. To prove the converse, we can use the same argument backwards. □

Urysohn's lemma. Our topic today is the following useful theorem about normal spaces – which, for historical reasons, is known as Urysohn's lemma.

Theorem 11.2 (Urysohn's lemma)**.** *Let X be a normal topological space, and let $A, B \subseteq X$ be disjoint closed subsets. Then there is a continuous function*

$$f \colon X \to [0,1]$$

with the property that $f(x) = 0$ for every $x \in A$, and $f(x) = 1$ for every $x \in B$.

Urysohn's lemma tells us that when X is normal, there are many continuous functions from X to $\mathbb{R}$. This is very useful, because one can then try to use such functions to embed X into a product of copies of $\mathbb{R}$; as we will see later, applications of this idea include a sufficient condition for a topological space to be metrizable, and an embedding theorem for abstract compact manifolds.

Example 11.3. Before we try to prove Urysohn's lemma in general, let us first consider the example of metric spaces; as we know, every metric space is normal. When X is a metric space, one can actually write down a function f explicitly in terms of distances. First, given any closed set $A \subseteq X$, we can easily create a continuous function that vanishes on A and is positive everywhere else: simply take

$$X \to [0, \infty), \quad x \mapsto d(x, A).$$

Here the distance from the point x to the set A is by definition

$$d(x, A) = \inf\big\{\, d(x, y) \;\big|\; y \in A \,\big\};$$

note that the infimum may not be achieved (for example when $X = \mathbb{Q}$ and A is the set of all rational numbers greater than $\sqrt{2}$). One can show that this function is continuous on X (see the homework for this week). Note that $d(x, A) = 0$ if $x \in A$. If $x \notin A$, then we have $B_r(x) \subseteq X \setminus A$ for some $r > 0$, and so $d(x, A) \geq r > 0$.

Now suppose that we are given two disjoint closed sets A and B. Consider the function

$$f \colon X \to [0,1], \quad f(x) = \frac{d(x, A)}{d(x, A) + d(x, B)},$$

which is well-defined and continuous because, due to the fact that $A \cap B = \emptyset$, the denominator is always positive. By inspection, $f(x) = 0$ for $x \in A$, and $f(x) = 1$ for $x \in B$.

Proof of Urysohn's lemma. Now we return to the case where A, B are disjoint closed sets in an arbitrary normal space X. Since we do not know any $\mathbb{R}$-valued continuous functions on X, constructing the desired function f will be more involved. Note that continuity of f is the main point: otherwise, we could just define f to be 0 on A, to be 1 on B, and to be something else on $X \setminus (A \cup B)$. In a nutshell, the idea for the construction is the following: using the normality of X, we will construct very many open sets in X, and then we use the position of $x \in X$ with respect to these open sets to decide what the value of $f(x)$ should be.

One way to approach the problem is the following. Suppose we already had a continuous function $f\colon X \to [0,1]$ with the desired properties. For each $t \in [0,1]$, the *sublevel set*

$$U_t = f^{-1}[0,t) \subseteq X$$

would then be open. Moreover, for $s < t$, we would have not only $U_s \subseteq U_t$, but actually

$$\overline{U_s} \subseteq U_t,$$

due to the fact that $f^{-1}[0,s]$ is a closed set containing U_s and contained in U_t. These containments look very similar to the alternative definition of normality in Proposition 11.1: whenever we have $A \subseteq U$ with A closed and U open, we can find V open with $A \subseteq V \subseteq \overline{V} \subseteq U$. So what we will do is to construct a collection of open sets that have the same property as the sublevel sets of our hypothetical function f; and then we will use these open sets to actually define the function f.

Now we begin the actual proof. The first step is to choose a countable dense subset of $\mathbb{R}$; for the sake of convenience, we shall use the set of *dyadic rationals*

$$D = \left\{ \frac{k}{2^n} \;\middle|\; k \in \mathbb{Z} \text{ and } n \geq 1 \right\}.$$

We order the set of dyadic rationals in $[0,1]$ as follows:

$$D \cap [0,1] = \left\{0, 1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{8}, \tfrac{3}{8}, \dots\right\}$$

For every $\alpha \in D \cap [0,1]$, we now select an open set $U_\alpha \subseteq X$ by the following recursive procedure. To start, define $U_1 = X \setminus B$ and note that A is a closed subset of the open set U_1; using normality, we can find an open set U_0 with

$$A \subseteq U_0 \subseteq \overline{U_0} \subseteq U_1.$$

The next dyadic rational on the list is $\frac{1}{2}$, which lies between 0 and 1; we use normality to find an open set $U_{\frac{1}{2}}$ with

$$\overline{U_0} \subseteq U_{\frac{1}{2}} \subseteq \overline{U_{\frac{1}{2}}} \subseteq U_1.$$

Next comes $\frac{1}{4}$, which lies between 0 and $\frac{1}{2}$; accordingly, we choose

$$\overline{U_0} \subseteq U_{\frac{1}{4}} \subseteq \overline{U_{\frac{1}{4}}} \subseteq U_{\frac{1}{2}};$$

and for $\frac{3}{4}$, we choose

$$\overline{U_{\frac{1}{2}}} \subseteq U_{\frac{3}{4}} \subseteq \overline{U_{\frac{3}{4}}} \subseteq U_1.$$

Continuing in this way, we obtain a family of open subsets U_α, indexed by the set $D \cap [0,1]$. Since it will simplify the proof later, we extend the definition to all dyadic

rationals by putting $U_\alpha = \emptyset$ for $\alpha < 0$ and $U_\beta = X$ for $\beta > 1$. The resulting family of open sets has the following properties:

(1) $A \subseteq U_0$ and $U_1 = X \setminus B$.
(2) $U_\alpha = \emptyset$ for $\alpha < 0$ and $U_\beta = X$ for $\beta > 1$.
(3) If $\alpha < \beta$, then $\overline{U_\alpha} \subseteq U_\beta$.

Although this is not going to be quite exactly true, think of the sets U_α as the sublevel sets of the function f that we are trying to construct.

The second step of the proof is to define a suitable function $f\colon X \to [0,1]$. For any $x \in X$, we consider the set

$$D(x) = \left\{ \alpha \in D \;\middle|\; x \in U_\alpha \right\}$$

that keeps track of which U_α the given point belongs to. If our open sets were actually sublevel sets of a function f, the condition $x \in U_\alpha$ would mean that $f(x) < \alpha$; this observation suggests the following definition:

$$f\colon X \to [0,1], \quad f(x) = \inf D(x)$$

Why does this make sense? Since $U_\beta = X$ for every $\beta > 1$, the set $D(x)$ contains all dyadic rationals greater than 1; on the other hand, $U_\alpha = \emptyset$ for $\alpha < 0$, and so $D(x)$ contains no dyadic rationals less than 0. In particular, $D(x)$ is nonempty and bounded from below; using the properties of $\mathbb{R}$, it has a well-defined greatest lower bound in $[0,1]$. It is not obvious that f is continuous, but we can easily show that it has the correct values on A and B. Indeed, if $x \in A$, we have $x \in U_0$, hence $0 \in D(x)$, which shows that $f(x) = \inf D(x) = 0$. On the other hand, $x \in B$ implies that $x \notin U_\alpha$ for any $\alpha \leq 1$; but then $f(x) = \inf D(x) = 1$.

The following lemma will help us prove that f is continuous.

Lemma 11.4. *With notation as above, the following is true:*

(a) If $x \in \overline{U_\alpha}$, then $f(x) \leq \alpha$; equivalently, if $f(x) > \alpha$, then $x \notin \overline{U_\alpha}$.
(b) If $X \notin U_\beta$, then $f(x) \geq \beta$; equivalently, if $f(x) < \beta$, then $x \in U_\beta$.

Proof. For (a), suppose that $x \in \overline{U_\alpha}$. Then $x \in U_\beta$ for every $\beta > \alpha$, which means that $D(x)$ contains every dyadic rational $\beta > \alpha$. Because D is dense in the unit interval, this can only happen if $f(x) = \inf D(x) \leq \alpha$: otherwise, we could find some $\beta \in D(x)$ with $\alpha < \beta < f(x)$, contradicting the definition of $f(x)$.

For (b), suppose that $x \notin U_\beta$. Then $x \notin U_\alpha$ for any $\alpha \leq \beta$, which means that $D(x)$ does not contain any $\alpha \leq \beta$; this clearly makes β a lower bound for $D(x)$ and so $f(x) = \inf D(x) \geq \beta$. □

Now we argue that f is continuous. Since $[0,1]$ has the subspace topology, it will be enough to show that f is continuous as a function from X to $\mathbb{R}$. Given any point $x_0 \in X$, and any open interval (c,d) containing the point $f(x_0)$, we have to show that there is a neighborhood U of x_0 with $f(U) \subseteq (c,d)$. Using the fact that D is dense, we choose $\alpha, \beta \in D$ with

$$c < \alpha < f(x_0) < \beta < d.$$

I claim that the open set

$$U = U_\beta \setminus \overline{U_\alpha}$$

contains the point x_0 and satisfies $f(U) \subseteq I$. To see why, observe that we have $f(x_0) < \beta$, and therefore $x_0 \in U_\beta$ by the lemma; likewise, $f(x_0) > \alpha$, and therefore

$x_0 \notin \overline{U_\alpha}$. To show that $f(U) \subseteq (c, d)$, take an arbitrary point $x \in U$. Then $x \in \overline{U_\beta}$, and so $f(x) \leq \beta$ by the lemma; likewise, $x \notin U_\alpha$, and so $f(x) \geq \alpha$. This yields

$$f(U) \subseteq [\alpha, \beta] \subseteq (c, d)$$

and shows that f is continuous.

Urysohn's lemma versus normality. Urysohn's lemma shows that any two disjoint closed subsets in a normal space can be separated by a continuous function, in the following sense.

Definition 11.5. Let X be a topological space and let $A, B \subseteq X$ be disjoint closed sets. We say that A and B can be *separated by a continuous function* if there is a continuous function $f\colon X \to [0, 1]$ with $f(x) = 0$ for every $x \in A$ and $f(x) = 1$ for every $x \in B$.

This condition looks a lot more useful than being normal, because it is more useful to have a lot of continuous functions that to have a lot of open sets. But in fact, it is precisely equivalent to normality.

Corollary 11.6. *Let X be a topological space in which all one-point sets are closed. Then X is normal if and only if every pair of disjoint closed sets can be separated by a continuous function.*

Proof. Urysohn's lemma shows that every normal space satisfies this condition. Conversely, suppose that every pair of disjoint closed sets $A, B \subseteq X$ can be separated by a continuous function $f\colon X \to [0, 1]$. Then

$$U = f^{-1}\left[0, \tfrac{1}{2}\right) \quad \text{and} \quad V = f^{-1}\left(\tfrac{1}{2}, 1\right]$$

are disjoint open sets containing A and B, respectively. □

Lecture 12: October 4

Midterm. The midterm will be held in class on Tuesday, October 18. It can include all the material from point set topology up until the end of this week.

Urysohn's metrization theorem. Today, I want to explain some applications of Urysohn's lemma. The first one has to do with the problem of characterizing metric spaces among all topological spaces. As we know, every metric space is also a topological space: the collection of open balls $B_r(x)$ is a basis for the metric topology. A natural question is exactly which topological spaces arise in this way.

Definition 12.1. A topological space X is called *metrizable* if it is homeomorphic to a metric space (with the metric topology).

In fact, the answer is known: the *Nagata-Smirnov metrization theorem* gives a necessary and sufficient condition for metrizability. If you are interested, please see Chapter 6 in Munkres' book; to leave enough time for other topics, we will not discuss the general metrization theorem in class. We will focus instead on the special case of second countable spaces, which is all that one needs in practice.

Recall from last week that every metric space is normal. It turns out that if we restrict our attention to spaces with a countable basis, then normality is equivalent to metrizability; this is the content of the following theorem by Urysohn. In fact, it is enough to assume only regularity: by Theorem 10.14, every regular space with a countable basis is normal.

Theorem 12.2 (Urysohn's metrization theorem)**.** *Every regular space with a countable basis is metrizable.*

Example 12.3. For example, every compact Hausdorff space with a countable basis is metrizable: the reason is that every compact Hausdorff space is normal.

There are two ways to show that a given space X is metrizable: one is to construct a metric that defines the topology on X; the other is to find an embedding of X into a metric space, because every subspace of a metric space is again a metric space. To prove the metrization theorem, we first show that the product space $[0,1]^\omega$ is metrizable (by constructing a metric), and then we show that every regular space with a countable basis can be embedded into $[0,1]^\omega$ (by using Urysohn's lemma).

Proposition 12.4. *The product space $[0,1]^\omega$ is metrizable.*

Proof. We write the points of $[0,1]^\omega$ in the form $x = (x_1, x_2, \dots)$, where $0 \le x_k \le 1$. Recall from our discussion of the product topology that the collection of open sets

$$\left\{ y \in [0,1]^\omega \;\middle|\; |x_k - y_k| < r_k \text{ for } k = 1, \dots, n \right\}$$

is a basis for the topology $\mathscr{T}$; here $x \in [0,1]^\omega$ is any point, $n \ge 1$ is an integer, and $r_1, \dots, r_n$ are positive real numbers. Each of the basic open sets only imposes conditions on finitely many coordinates; our task is to find a metric in which the open balls have the same property.

We define a candidate metric on the product space by the formula

$$d(x,y) = \sup_k \frac{1}{k}|x_k - y_k| = \max_k \frac{1}{k}|x_k - y_k|.$$

Because $|x_k - y_k| \le 1$, the numbers $\frac{1}{k}|x_k - y_k|$ approach 0 as k gets large; the supremum is therefore achieved for some k. It is an easy exercise to check that d is

indeed a metric. To prove that the metric topology $\mathscr{T}_d$ is the same as the product topology $\mathscr{T}$, we have to compare the basic open sets in both topologies.

Let us first show that every open ball $B_r(x)$ is open in the product topology. By definition, we have

$$\begin{aligned} B_r(x) &= \big\{\, y \in [0,1]^\omega \;\big|\; d(x,y) < r \,\big\} \\ &= \big\{\, y \in [0,1]^\omega \;\big|\; |x_k - y_k| < kr \text{ for every } k \,\big\} \\ &= \big\{\, y \in [0,1]^\omega \;\big|\; |x_k - y_k| < kr \text{ for every } k \le r^{-1} \,\big\}; \end{aligned}$$

the last equality holds because $|x_k - y_k| < kr$ is automatically satisfied once $kr > 1$, due to the fact that $x_k, y_k \in [0,1]$. So the open ball $B_r(x)$ is actually a basic open set in the product topology, which means that $\mathscr{T}_d \subseteq \mathscr{T}$.

To prove that the two topologies coincide, let $U \in \mathscr{T}$ be an arbitrary open set in the product topology. For any point $x \in U$, some basic open set

$$\big\{\, y \in [0,1]^\omega \;\big|\; |x_k - y_k| < r_k \text{ for } k = 1, \dots, n \,\big\}$$

must be contained in U. If we now define

$$r = \min_{1 \le k \le n} \frac{r_k}{k},$$

then we have $kr < r_k$ for every $k = 1, \dots, n$, and so the open ball

$$B_r(x) = \big\{\, y \in [0,1]^\omega \;\big|\; |x_k - y_k| < kr \text{ for every } k \,\big\}$$

is contained in U. This is enough to conclude that $U \in \mathscr{T}_d$, and hence that $\mathscr{T}_d = \mathscr{T}$; thus the product space $[0,1]^\omega$ is indeed metrizable. □

One consequence is that $\mathbb{R}^\omega$ is also metrizable: it is homeomorphic to $(0,1)^\omega$, which is a subspace of the metrizable space $[0,1]^\omega$. You can see from the proof that the index set really had to be countable in order to define the metric; in fact, one can show that the product of uncountably many copies of $\mathbb{R}$ is no longer metrizable.

Proof of Urysohn's metrization theorem. We are now ready to prove Theorem 12.2. Let X be a regular space with a countable basis; let me remind you that X is automatically normal (by Theorem 10.14). The idea of the proof is to construct an embedding

$$F \colon X \to [0,1]^\omega, \quad F(x) = \big(f_1(x), f_2(x), \dots\big).$$

Recall that an *embedding* is an injective function $f \colon X \to Y$ that induces a homeomorphism between X and the subspace $f(X)$ of Y. Since a function into a product space is continuous if and only if all the coordinate functions $f_n = p_n \circ F$ are continuous, we have to look for countably many continuous functions $f_n \colon X \to [0,1]$. There are two other requirements: (1) F should be injective, meaning that whenever $x \neq y$, there should exist some n with $f_n(x) \neq f_n(y)$. (2) F should induce a homeomorphism between X and $F(X)$, meaning that whenever $U \subseteq X$ is open, $F(U)$ should be open in the subspace topology on $F(X)$.

Lemma 12.5. *There are countably many continuous functions $f_n \colon X \to [0,1]$ with the following property: for every open set $U \subseteq X$ and for every point $x_0 \in U$, there is some index n such that $f_n(x_0) = 1$, but $f_n(x) = 0$ for all $x \notin U$.*

Proof. Since X is normal, we can easily find such a function for every U and every x_0: we only have to apply Urysohn's lemma to the two closed sets $\{x_0\}$ and $X \setminus U$. The resulting collection of functions is not going to be countable in general, but we can use the existence of a countable basis to make it so.

Let $B_1, B_2, \dots$ be a countable basis for the topology on X. For every pair of indices m, n such that $\overline{B_m} \subseteq B_n$, the two closed sets $\overline{B_m}$ and $X \setminus B_n$ are disjoint; Urysohn's lemma gives us a continuous function $g_{m,n} \colon X \to [0,1]$ with

$$g_{m,n}(x) = \begin{cases} 1 & \text{if } x \in \overline{B_m}, \\ 0 & \text{if } x \notin B_n. \end{cases}$$

In this way, we obtain countably many continuous functions; I claim that they have the desired property. In fact, suppose that we have an open set $U \subseteq X$ and a point $x_0 \in U$. Because the B_n form a basis, there is some index n with $x \in B_n \subseteq U$. Now X is regular, and so we can find a smaller open set V with

$$x \in V \subseteq \overline{V} \subseteq B_n;$$

we can also choose another index m such that $x \in B_m \subseteq V$. Then $\overline{B_m} \subseteq \overline{V} \subseteq B_n$, and the function $g_{m,n}$ from above satisfies $g_{m,n}(x_0) = 1$ (because $x_0 \in B_m$), and $g_{m,n}(x) = 0$ for $x \in X \setminus U$ (because $X \setminus U \subseteq X \setminus B_n$). This shows that the countably many functions $g_{m,n}$ do what we want. □

Using the functions from the lemma, we can now define a function

$$F \colon X \to [0,1]^\omega, \quad F(x) = \big(f_1(x), f_2(x), \dots\big);$$

because all the individual functions f_n are continuous, the function F is also continuous (by Theorem 4.16). Our goal is to show that F is an embedding. Injectivity is straightforward: Suppose we are given two points $x, y \in X$ with $x \neq y$. Then x belongs to the open set $X \setminus \{y\}$, and so Lemma 12.5 guarantees that there is some index n for which $f_n(x) = 1$ and $f_n(y) = 0$. But then $F(x) \neq F(y)$, because their n-th coordinates are different.

This already proves that F is a continuous bijection between X and $F(X)$; the following lemma shows that it is even a homeomorphism.

Lemma 12.6. *F is a homeomorphism between X and the subspace $F(X)$ of $[0,1]^\omega$.*

Proof. We have to show that for any nonempty open set $U \subseteq X$, the image $F(U)$ is open in the subspace topology on $F(X)$. Consider an arbitrary point $t_0 \in F(U)$; note that $t_0 = F(x_0)$ for a unique $x_0 \in U$. According to Lemma 12.5, there is some index n with $f_n(x_0) = 1$ and $f_n(x) = 0$ for every $x \notin U$. Now consider the set

$$W = F(X) \cap \big\{\, t \in [0,1]^\omega \;\big|\; p_n(t) > 0 \,\big\}.$$

It is clearly open in $F(X)$, being the intersection of $F(X)$ with a basic open set in $[0,1]^\omega$. I claim that $t_0 \in W$ and $W \subseteq F(U)$. In fact, we have

$$p_n(t_0) = p_n\big(F(x_0)\big) = f_n(x_0) > 0,$$

which shows that $t_0 \in W$. On the other hand, every point $t \in W$ is of the form $t = F(x)$ for a unique $x \in X$, and

$$f_n(x) = p_n\big(F(x)\big) = p_n(t) > 0.$$

Since f_n vanishes outside of U, this means that $x \in U$, and so $t = F(x) \in F(U)$. It follows that $F(U)$ is a union of open sets, and therefore open. □

The conclusion is that X is homeomorphic to a subspace of $[0,1]^\omega$; because the latter space is metrizable, X itself must also be metrizable. We have proved Theorem 12.2.

Embeddings of manifolds. The strategy that we used to prove Theorem 12.2 – namely to embed a given space X into a nice ambient space, using the existence of sufficiently many continuous functions on X – has many other applications in topology and geometry. One such application is to the study of abstract manifolds. Recall the following definition.

Definition 12.7. An m-dimensional *topological manifold* is a Hausdorff space with a countable basis in which every point has a neighborhood homeomorphic to an open set in $\mathbb{R}^m$.

Earlier in the semester, I left out the condition that manifolds should have a countable basis. The integer m is called the *dimension* of the manifold; it is uniquely determined by the manifold, because one can show that a nontrivial open subset in $\mathbb{R}^m$ can only be homeomorphic to an open subset in $\mathbb{R}^n$ when $m = n$. One-dimensional manifolds are called *curves*, two-dimensional manifolds are called *surfaces*; the study of special classes of manifolds is arguably the most important object of topology.

Of course, people were already studying manifolds long before the advent of topology; back then, the word "manifold" did not mean an abstract topological space with certain properties, but rather a submanifold of some Euclidean space. So the question naturally arises whether every abstractly defined manifold can actually be realized as a submanifold of some $\mathbb{R}^N$. The answer is yes; we shall prove a special case of this result, namely that every compact manifold can be embedded into $\mathbb{R}^N$ for some large N.

Remark. One can ask the same question for manifolds with additional structure, such as smooth manifolds, Riemannian manifolds, complex manifolds, etc. This makes the embedding problem more difficult: for instance, John Nash (who was portrayed in the movie *A Beautiful Mind*) became famous for proving an embedding theorem for Riemannian manifolds.

Of course, every m-dimensional manifold can "locally" be embedded into $\mathbb{R}^m$; the problem is how to patch these locally defined embeddings together to get a "global" embedding. This can be done with the help of the following tool.

Definition 12.8. Let $X = U_1 \cup \cdots \cup U_n$ be an open covering of a topological space X. A *partition of unity* (for the given covering) is a collection of continuous functions $\phi_1, \ldots, \phi_n \colon X \to [0,1]$ with the following two properties:

(1) The support of ϕ_k is contained in the open set U_k.
(2) We have $\sum_{k=1}^n \phi_k(x) = 1$ for every $x \in X$.

Here the *support* $\operatorname{Supp}\phi$ of a continuous function $\phi \colon X \to \mathbb{R}$ means the closure of the set $\phi^{-1}(\mathbb{R} \setminus \{0\})$; with this definition, $x \notin \operatorname{Supp}\phi$ if and only if ϕ is identically zero on some neighborhood of x.

Lecture 13: October 6

Embeddings of manifolds (continued). Recall from last time the definition of a *partition of unity*. Given a finite open covering $X = U_1 \cup \cdots \cup U_n$ of a topological space, a partition of unity is a collection of continuous functions $\phi_1, \ldots, \phi_n \colon X \to [0,1]$ with the following two properties:

(1) The support of ϕ_k is contained in the open set U_k.
(2) We have $\sum_{k=1}^n \phi_k(x) = 1$ for every $x \in X$.

Example 13.1. Partitions of unity are useful for extending locally defined functions continuously to the entire space. For example of how this works, suppose that $X = U_1 \cup U_2$, and that we have a continuous function $f_1 \colon U_1 \to \mathbb{R}$. Assuming that there is a partition of unity $\phi_1 + \phi_2 = 1$, we can consider the function

$$f \colon X \to \mathbb{R}, \quad f(x) = \begin{cases} f_1(x)\phi_1(x) & \text{if } x \in U_1, \\ 0 & \text{if } x \in X \setminus \operatorname{Supp} \phi_1. \end{cases}$$

Since the two definitions are compatible on the intersection of the two open sets U_1 and $X \setminus \operatorname{Supp} \phi_1$, the function f is continuous. Now every $x \in X \setminus \operatorname{Supp} \phi_2$ belongs to $\operatorname{Supp} \phi_1 \subseteq U_1$, because of the relation $\phi_1(x) = \phi_1(x) + \phi_2(x) = 1$, and so

$$f(x) = f_1(x)\phi_1(x) = f_1(x).$$

As you can see, we have obtained a continuous function on X that agrees with the original function f_1 on the smaller open set $X \setminus \operatorname{Supp} \phi_2$.

We can use Urysohn's lemma to construct a partition of unity for every finite open covering of a normal space.

Lemma 13.2. *If X is normal, then every finite open covering $X = U_1 \cup \cdots \cup U_n$ admits a partition of unity.*

Proof. Let me first explain the general idea. This is easier to follow if you draw a picture. Consider one of the open sets in the covering, say U_1. Since we want $\operatorname{Supp} \phi_1$ to be contained inside U_1, the function ϕ_1 should presumably be equal to 1 somewhere inside of U_1, and then continuously go down to 0 near the boundary of U_1. One way to make sure that this happens is to choose two open sets W_1 and V_1 with the property that

$$\overline{W_1} \subseteq V_1 \subseteq \overline{V_1} \subseteq U_1,$$

and then arrange that ϕ_1 is equal to 1 on $\overline{W_1}$, and equal to 0 on $X \setminus V_1$. To construct ϕ_1, we can of course use Urysohn's lemma. Here are the details:

Step 1. We can find an open covering $X = V_1 \cup \cdots \cup V_n$ such that $\overline{V_k} \subseteq U_k$ for every $k = 1, \ldots, n$. To get started, consider the set

$$A = X \setminus \big(U_2 \cup \cdots \cup U_n\big).$$

It is closed and contained in the open set U_1, because $X = U_1 \cup \cdots \cup U_n$. Since X is normal, we can find an open set V_1 with $A \subseteq V_1 \subseteq \overline{V_1} \subseteq U_1$; by construction, $V_1, U_2, \ldots, U_n$ is still an open covering of X.

To get the remaining open sets, we proceed by induction. Suppose that we already have $V_1, \ldots, V_{k-1}$ with $X = V_1 \cup \cdots \cup V_{k-1} \cup U_k \cup \cdots \cup U_n$. Then

$$B = X \setminus \big(V_1 \cup \cdots \cup V_{k-1} \cup U_{k+1} \cup \cdots \cup U_n\big)$$

is closed and contained in the open set U_k. Since X is normal, we can find an open set V_k with $B \subseteq V_k \subseteq \overline{V_k} \subseteq U_k$; now $V_1, \dots, V_k, U_{k+1}, \dots, U_n$ still cover X, and so we can continue the process until we reach $k = n$.

Step 2. We construct the desired partition of unity $\phi_1 + \dots + \phi_n = 1$. Repeating the argument in Step 1 for the open covering $X = V_1 \cup \dots \cup V_n$, we can find an open covering $X = W_1 \cup \dots \cup W_n$ by even smaller open sets with

$$\overline{W_k} \subseteq V_k \subseteq \overline{V_k} \subseteq U_k.$$

Now the closed sets $\overline{W_k}$ and $X \setminus V_k$ are disjoint, and so Urysohn's lemma tells us that there is a continuous function $\psi_k \colon X \to [0,1]$ with

$$\psi_k(x) = \begin{cases} 1 & \text{if } x \in \overline{W_k}, \\ 0 & \text{if } x \in X \setminus V_k. \end{cases}$$

Because ψ_k is identically zero on $X \setminus V_k$, the support of ψ_k is contained in $\overline{V_k} \subseteq U_k$. To turn these functions into a partition of unity, we now consider their sum

$$\psi_1 + \dots + \psi_n.$$

At every $x \in X$, the value of the sum is at least 1: indeed, we have $\psi_k(x) = 1$ for $x \in W_k$, and the open sets $W_1, \dots, W_n$ cover X. Therefore

$$\phi_k = \frac{\psi_k}{\psi_1 + \dots + \psi_n}$$

is a continuous function from $\mathbb{R}$ into $[0,1]$, with the property that $\operatorname{Supp} \phi_k \subseteq U_k$. Since $\phi_1 + \dots + \phi_n = 1$ is clear, we have found the desired partition of unity. □

We can use the existence of partitions of unity to prove the following embedding theorem for compact manifolds.

Theorem 13.3. *Every compact manifold can be embedded into Euclidean space.*

Proof. Let X be a compact manifold of dimension m; we will construct an embedding $F \colon X \to \mathbb{R}^N$ for some large integer N. By definition, every point of X has a neighborhood homeomorphic to an open set in $\mathbb{R}^m$; since X is compact, finitely many of these neighborhoods will cover X. We thus get an open covering $X = U_1 \cup \dots \cup U_n$ as well as embeddings

$$g_k \colon U_k \to \mathbb{R}^m.$$

Let $\phi_1 + \dots + \phi_n = 1$ be a partition of unity; it exists by Lemma 13.2 because X is compact Hausdorff, hence normal. For every $k = 1, \dots, n$, consider the function

$$f_k \colon X \to \mathbb{R}^m, \quad f_k(x) = \begin{cases} \phi_k(x) g_k(x) & \text{if } x \in U_k, \\ 0 & \text{if } x \in X \setminus \operatorname{Supp} \phi_k. \end{cases}$$

The two definitions are compatible on the intersections of the open sets U_k and $X \setminus \operatorname{Supp} \phi_k$, and so f_k is continuous.

Now it is an easy matter to obtain the desired embedding. Set $N = n + mn$, and consider the function

$$F \colon X \to \mathbb{R}^N = \mathbb{R}^n \times (\mathbb{R}^m)^n, \quad F(x) = \big(\phi_1(x), \dots, \phi_n(x), f_1(x), \dots, f_n(x)\big).$$

Clearly, F is continuous; we want to show that it defines an embedding of X into $\mathbb{R}^N$. Because X is compact and $\mathbb{R}^N$ is Hausdorff, all we have to do is prove that F

is injective: a continuous bijection between a compact space and a Hausdorff space is automatically a homeomorphism!

So suppose that we have two points $x, y \in X$ with $F(x) = F(y)$. This means that $\phi_k(x) = \phi_k(y)$ and $f_k(x) = f_k(y)$ for every $k = 1, \dots, n$. Since $\phi_1 + \dots + \phi_n = 1$, we can find some index k for which $\phi_k(x) = \phi_k(y) > 0$. This forces $x, y \in U_k$ (because $\operatorname{Supp} \phi_k \subseteq U_k$); but then

$$\phi_k(x) g_k(x) = f_k(x) = f_k(y) = \phi_k(y) g_k(y).$$

After dividing by $\phi_k(x) = \phi_k(y)$, we see that $g_k(x) = g_k(y)$; but $g_k \colon U_k \to \mathbb{R}^m$ was injective, and so $x = y$. □

The choice of N is far from optimal: with some additional tricks, one can show that every compact manifold of dimension m can actually be embedded into $\mathbb{R}^{2m}$.

The Tietze extension theorem. Another important application of Urysohn's lemma is the following extension theorem for continuous real-valued functions. This result is very useful in analysis.

Theorem 13.4 (Tietze extension theorem)**.** *Let X be a normal topological space, and $A \subseteq X$ a closed subset.*

(a) Let $I \subseteq \mathbb{R}$ be a closed interval. Any continuous function $f \colon A \to I$ can be extended to a continuous function $g \colon X \to I$.

(b) Similarly, any continuous function $f \colon A \to \mathbb{R}$ can be extended to a continuous function $g \colon X \to \mathbb{R}$.

Saying that g extends f means that we have $g(a) = f(a)$ for every point $a \in A$. The assumption that A be closed is very important: for example, the function $f \colon (0, \infty) \to \mathbb{R}$ with $f(x) = 1/x$ cannot be extended continuously to all of $\mathbb{R}$.

Proof of Tietze's theorem. We did not discuss the proof of Theorem 13.4 in class; I am including it in the notes for people who want to see how it works. Roughly speaking, it goes like this. Using Urysohn's lemma, we construct a sequence of continuous functions $s_n \colon X \to I$ that approximates f more and more closely as n gets large. The desired function g will be the limit of this sequence. Since we want g to be continuous, we first have to understand under what conditions the limit of a sequence of continuous functions is again continuous. The keyword here is "uniform convergence".

Consider a sequence of functions $f_n \colon X \to \mathbb{R}$ from a topological space X to the real numbers (or, more generally, to a metric space). We say that the sequence converges (pointwise) to a function $f \colon X \to \mathbb{R}$ if, for every $x \in X$, the sequence of real numbers $f_n(x)$ converges to the real number $f(x)$. More precisely, this means that for every $x \in X$ and every $\varepsilon > 0$, there exists N with $|f_n(x) - f(x)| < \varepsilon$ for all $n \geq N$. Of course, N is allowed to depend on x; we get a more restrictive notion of convergence if we assume that the same N works for all $x \in X$ at the same time.

Definition 13.5. A sequence of functions $f_n \colon X \to \mathbb{R}$ *converges uniformly* to a function $f \colon X \to \mathbb{R}$ if, for every $\varepsilon > 0$, there is some N such that $|f(x) - f_n(x)| < \varepsilon$ for all $n \geq N$ and all $x \in X$.

The usefulness of uniform convergence is that it preserves continuity.

Lemma 13.6. *The limit of a uniformly convergent sequence of continuous functions is continuous.*

Proof. Suppose that the sequence of continuous functions $f_n\colon X \to \mathbb{R}$ converges uniformly to a function $f\colon X \to \mathbb{R}$. We have to show that f is continuous. Let $V \subseteq \mathbb{R}$ be an arbitrary open set; to show that $f^{-1}(V)$ is open, it suffices to produce for every point $x_0 \in f^{-1}(V)$ a neighborhood U with $f(U) \subseteq V$. This is straightforward. One, $f(x_0) \in V$, and so there is some $r > 0$ with

$$B_r\big(f(x_0)\big) \subseteq V.$$

Two, the sequence converges uniformly, and so we can find an index n such that $|f_n(x) - f(x)| < r/3$ for every $x \in X$. Three, f_n is continuous, and so there is an open set U containing x_0 with

$$f_n(U) \subseteq B_{r/3}\big(f_n(x_0)\big).$$

Now we can show that $f(U) \subseteq V$. Let $x \in U$ be any point; then

$$|f(x)-f(x_0)| \le |f(x)-f_n(x)|+|f_n(x)-f_n(x_0)|+|f_n(x_0)-f(x_0)| < \frac{r}{3}+\frac{r}{3}+\frac{r}{3} = r,$$

and so $f(x) \in B_r\big(f(x_0)\big) \subseteq V$. □

We will do the proof of Theorem 13.4 in three steps. Throughout, X denotes a normal topological space, and $A \subseteq X$ a closed subset.

The *first step* is to solve the following simpler problem. Given a continuous function $f\colon A \to [-r, r]$, we are going to construct a continuous function $h\colon X \to \mathbb{R}$ that is somewhat close to f on the set A, without ever getting unreasonably large. More precisely, we want the following two conditions:

$$|h(x)| \le \frac{1}{3}r \quad \text{for every } x \in X \tag{13.7}$$

$$|f(a) - h(a)| \le \frac{2}{3}r \quad \text{for every } a \in A \tag{13.8}$$

To do this, we divide $[-r, r]$ into three subintervals of length $\frac{2}{3}r$, namely

$$I_1 = \left[-r, \frac{1}{3}r\right], \quad I_2 = \left[-\frac{1}{3}r, \frac{1}{3}r\right], \quad I_3 = \left[\frac{1}{3}r, r\right].$$

Now consider the two sets $B = f^{-1}(I_1)$ and $C = f^{-1}(I_3)$. They are closed subsets of A (because f is continuous), and therefore of X (because A is closed); they are also clearly disjoint. Because X is normal, Urysohn's lemma produces for us a continuous function $h\colon X \to [-\frac{1}{3}r, \frac{1}{3}r]$ with

$$h(x) = \begin{cases} -\frac{1}{3}r & \text{for } x \in B, \\ \frac{1}{3}r & \text{for } x \in C. \end{cases}$$

Since $|f(x)| \le \frac{1}{3}r$, it is clear that (13.7) holds. To show that (13.8) is also satisfied, take any point $a \in A$. There are three cases. If $a \in B$, then $f(a)$ and $h(a)$ both belong to I_1; if $a \in C$, then $f(a)$ and $h(a)$ both belong to I_3; if $a \notin B \cup C$, then $f(a)$ and $h(a)$ both belong to I_2. In each case, the distance between $f(a)$ and $h(a)$ can be at most $\frac{2}{3}r$, which proves (13.8).

The *second step* is to use the construction from above to prove assertion (a) in Theorem 13.4. If I consists of a single point, the result is clear. On the other hand, any closed interval of positive length is homeomorphic to $[-1, 1]$; without loss of generality, we may therefore assume that we are dealing with a continuous function $f\colon A \to [-1, 1]$. As I said above, our strategy is to build a uniformly convergent

sequence of continuous functions $s_n\colon X \to [-1,1]$ that approximates f more and more closely on A.

To begin with, we can apply the construction in the first step to the function $f\colon A \to [-1,1]$; the result is a continuous function $h_1\colon X \to \mathbb{R}$ with

$$|h_1(x)| \le \frac{1}{3}r \quad \text{and} \quad |f(a) - h_1(a)| \le \frac{2}{3}r.$$

Now consider the difference $f - h_1$, which is a continuous function from A into the closed interval $[-\frac{2}{3}r, \frac{2}{3}r]$. By applying the construction from the first step again (with $r = \frac{2}{3}$), we obtain a second continuous function $h_2\colon X \to \mathbb{R}$ with

$$|h_2(x)| \le \frac{1}{3}\cdot\frac{2}{3} \quad \text{and} \quad |f(a) - h_1(a) - h_2(a)| \le \left(\frac{2}{3}\right)^2.$$

Notice how $h_1 + h_2$ is a better approximation for f than the initial function h_1. We can obviously continue this process indefinitely. After n steps, we have n continuous functions $h_1, \ldots, h_n\colon X \to \mathbb{R}$ with

$$|f(a) - h_1(a) - \cdots - h_n(a)| \le \left(\frac{2}{3}\right)^n.$$

By applying the construction to the function $f - h_1 - \cdots - h_n$ and the value $r = (2/3)^n$, we obtain a new continuous function $h_{n+1}\colon X \to \mathbb{R}$ with

$$|h_{n+1}(x)| \le \frac{1}{3}\cdot\left(\frac{2}{3}\right)^n \quad \text{and} \quad |f(a) - h_1(a) - \cdots - h_n(a) - h_{n+1}(a)| \le \left(\frac{2}{3}\right)^{n+1}.$$

Now I claim that the function

$$g(x) = \sum_{n=1}^{\infty} h_n(x)$$

is the desired continuous extension of f. To prove this claim, we have to show that the series converges for every $x \in X$; that the limit function $g\colon X \to [-1,1]$ is continuous; and that $g(a) = f(a)$ for every $a \in A$.

To prove the convergence, let us denote by $s_n(x) = h_1(x) + \cdots + h_n(x)$ the n-th partial sum of the series; clearly, $s_n\colon X \to \mathbb{R}$ is continuous. If $m > n$, then

$$|s_m(x) - s_n(x)| \le \sum_{k=n+1}^{m} |h_k(x)| \le \frac{1}{3}\sum_{k=n+1}^{m}\left(\frac{2}{3}\right)^{k-1} \le \left(\frac{2}{3}\right)^n.$$

This proves that the sequence of real numbers $s_n(x)$ is Cauchy; if we define $g(x)$ as the limit, we obtain a function $g\colon X \to \mathbb{R}$. Now we can let m go to infinity in the inequality above to obtain

$$|g(x) - s_n(x)| \le \left(\frac{2}{3}\right)^n$$

for every $x \in X$. This means that the sequence s_n converges uniformly to g, and so by Lemma 13.6, g is still continuous. It is also not hard to see that g takes values in $[-1,1]$: for every $x \in X$, we have

$$|g(x)| \le \sum_{n=1}^{\infty}|h_n(x)| \le \frac{1}{3}\sum_{n=1}^{\infty}\left(\frac{2}{3}\right)^{n-1} = 1,$$

by evaluating the geometric series. It remains to show that $g(a) = f(a)$ whenever $a \in A$. By construction, we have

$$|f(a) - s_n(a)| = |f(a) - h_1(a) - \cdots - h_n(a)| \leq \left(\frac{2}{3}\right)^n;$$

letting $n \to \infty$, it follows that $|f(a) - g(a)| = 0$, which is what we wanted to show.

The *third step* is to prove assertion (b) in Theorem 13.4, where we are given a continuous function $f\colon A \to \mathbb{R}$. Evidently, $\mathbb{R}$ is homeomorphic to the open interval $(-1, 1)$; the result of the second step therefore allows us to extend f to a continuous function $g\colon X \to [-1, 1]$. The remaining problem is how we can make sure that $g(X) \subseteq (-1, 1)$. Here we use the following trick. Given g, we consider the subset

$$D = g^{-1}\{-1, 1\} \subseteq X.$$

Because g is continuous, this set is closed; it is also disjoint from the closed set A, because $g(A) \subseteq (-1, 1)$. By Urysohn's lemma, there is a continuous function $\varphi\colon X \to [0, 1]$ with $\varphi(D) = \{0\}$ and $\varphi(A) = \{1\}$. Now consider the continuous function $\varphi \cdot g$. It is still an extension of f, because we have

$$\varphi(a) \cdot g(a) = g(a) = f(a)$$

for $a \in A$. The advantage is that $\varphi \cdot g$ maps X into the open interval $(-1, 1)$: if $x \in D$, then $\varphi(x) \cdot g(x) = 0$, while if $x \notin D$, then $|\varphi(x) \cdot g(x)| \leq |g(x)| < 1$. This completes the proof of Tietze's extension theorem.

Lecture 14: October 13

Complete metric spaces. Our topic this week is the space $C(X, Y)$ of continuous functions between two topological spaces X and Y. Understanding the space of all continuous functions can be very useful, for example to find interesting examples of continuous functions. This topic actually belongs more to analysis than to topology, but since it involves several results from general topology as well, it seems like a good way to finish our discussion of general topology. Before we get to the space of continuous functions, however, we have to return for a moment to the subject of metric spaces and introduce the important concept of *completeness*.

Let (X, d) be a metric space. We saw earlier that the metric topology on X is first countable – the open balls with radius in $\mathbb{Q}$ form a neighborhood basis at every point – and that many properties of subsets can therefore be detected by looking at sequences: a subset $A \subseteq X$ is closed iff it is sequentially closed; a subset $A \subseteq X$ is compact iff it is sequentially compact; etc.

In analysis, an important problem is to decide whether a given sequence in a metric space converges. Typically, one does not know ahead of time what the limit is – more often than not, the whole point of trying to show that the sequence converges is so that one can be sure that there is a limit. The notion of a "Cauchy sequence", which you have probably already seen in your analysis course, was introduced to deal with this problem: proving that a sequence converges without knowing what the limit is.

Definition 14.1. A sequence of points $x_n \in X$ is called a *Cauchy sequence* if, for every $\varepsilon > 0$, one can find an integer N such that

$$d(x_n, x_m) < \varepsilon \quad \text{for all } m, n \geq N.$$

Intuitively, a Cauchy sequence is one where the points x_n huddle closer and closer together as n gets large. Of course, not every Cauchy sequence actually converges: in the metric space $\mathbb{Q}$, any sequence that converges to an irrational number in $\mathbb{R}$ is a Cauchy sequence without limit in $\mathbb{Q}$.

Definition 14.2. A metric space (X, d) is called *complete* if every Cauchy sequence in X has a limit in X.

So $\mathbb{R}$ is complete, but $\mathbb{Q}$ is not; this is the reason why analysts prefer to work with real numbers instead of rational numbers. There is a general construction for "completing" a metric space, similar to the way is which $\mathbb{R}$ can be obtained from $\mathbb{Q}$. The result is that, given an arbitrary metric space (X, d), there is a complete metric space $(\hat{X}, \hat{d})$ that contains X has a dense subset. Roughly speaking, the idea is to define an equivalence relation on the set of Cauchy sequences in X: two Cauchy sequences x_n and y_n are equivalent if $d(x_n, y_n) \to 0$ as $n \to \infty$. The points of $\hat{X}$ are the equivalence classes of Cauchy sequences; there is a natural metric $\hat{d}$ on $\hat{X}$, and after some checking, one finds that $(\hat{X}, \hat{d})$ is complete. If you have not seen this construction in a course on analysis, have a look at Theorem 43.7 or at Exercise 43.9 in Munkres' book.

Example 14.3. Completeness of a metric space is a property of the metric, not of the topology: $\mathbb{R}$ and $(0, 1)$ are homeomorphic as topological spaces, but whereas $\mathbb{R}$ is complete, $(0, 1)$ is not.

Example 14.4. Every compact metric space is complete. To see why, suppose that $x_n \in X$ is a Cauchy sequence. Being a metric space, X is also sequentially compact, and so the sequence has a convergent subsequence $x_{n(k)}$ with

$$x = \lim_{k\to\infty} x_{n(k)}.$$

Now that we have a potential limit, we can easily show that the entire sequence x_n converges to x. Fix $\varepsilon > 0$; then since x_n is a Cauchy sequence, we can find N such that $d(x_n, x_m) < \frac{\varepsilon}{2}$ for every $n, m \geq N$. Using the triangle inequality, we get

$$d(x_n, x) \leq d(x_n, x_{n(k)}) + d(x_{n(k)}, x) < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

for $n \geq N$, by choosing k large enough so that $n(k) \geq N$ and $d(x_{n(k)}, x) < \frac{\varepsilon}{2}$.

The converse of this last result is not true: in the Euclidean metric, $\mathbb{R}$ is complete but not compact. Let us try to figure out what extra condition (besides completeness) is needed to ensure that a metric space is compact. From the definition of compactness in terms of open coverings, we know that every compact metric space has to be bounded; in fact, the following stronger form of boundedness holds.

Definition 14.5. A metric space (X, d) is called *totally bounded* if, for every $r > 0$, it is possible to cover X by finitely many open balls of radius r.

Example 14.6. Let X be an infinite set, with the metric $d(x, y) = 1$ if $x \neq y$, and $d(x, y) = 0$ if $x = y$. Then X is bounded, but not totally bounded.

Theorem 14.7. *A metric space is compact if and only if it is complete and totally bounded.*

Proof. We have already seen that every compact metric space is complete and totally bounded. Let us prove that the converse holds. Suppose that (X, d) is a complete metric space that is totally bounded. In a metric space, compactness is equivalent to sequential compactness, and so it suffices to show that every sequence in X has a convergent subsequence. Since X is complete, it suffices moreover to find a subsequence that is Cauchy.

Let $x_n \in X$ be a sequence of points. By assumption, we can cover X by finitely many open balls of radius 1; evidently, at least one of these balls must contain x_n for infinitely many $n \in \mathbb{N}$. Call this ball U_1, and let

$$J_1 = \big\{\, n \in \mathbb{N} \;\big|\; x_n \in U_1 \,\big\},$$

which is an infinite set. Similarly, we can cover X by finitely many open balls of radius $\frac{1}{2}$; since the set J_1 is infinite, at least one of these balls must contain x_n for infinitely many $n \in J_1$. Call this ball U_2, and let

$$J_2 = \big\{\, n \in J_1 \;\big|\; x_n \in U_2 \,\big\},$$

which is again infinite. Continuing in this way, we obtain a nested sequence $J_1 \supseteq J_2 \supseteq J_3 \supseteq \cdots$ of infinite subsets of $\mathbb{N}$, such that $x_n \in U_1 \cap \cdots \cap U_k$ whenever $n \in J_k$.

Now we can choose a suitable subsequence $x_{n(k)}$ of our original sequence. Pick any element $n(1) \in J_1$; then pick an element $n(2) \in J_2$ with $n(2) > n(1)$, which exists because J_2 is infinite; then pick an element $n(3) \in J_3$ with $n(3) > n(2)$; and so on. In this way, we obtain a subsequence $x_{n(k)}$ of the original sequence, with the property that

$$x_{n(k)} \in U_1 \cap \cdots \cap U_k$$

for every $k = 1, 2, \dots$. This sequence is obviously Cauchy: whenever $k \leq \ell$, both $x_{n(k)}$ and $x_{n(\ell)}$ belong to the open ball U_k of radius $\frac{1}{k}$, and so

$$d(x_{n(k)}, x_{n(\ell)}) \leq \frac{2}{k}.$$

This subsequence converges because X is complete; the conclusion is that X is sequentially compact, and therefore compact. □

Function spaces. Given two nonempty sets X and Y, we can consider the space $\mathrm{Fun}(X, Y)$ of all functions $f\colon X \to Y$. Since a function is uniquely determined by its values $f(x)$ for $x \in X$, it is clear that

$$\mathrm{Fun}(X, Y) = Y^X = \prod_{x \in X} Y$$

is simply the Cartesian product, indexed by X, of several copies of Y: a function $f\colon X \to Y$ corresponds to the element $\big(f(x)\big)_{x \in X}$ of the product.

Now suppose that X and Y are topological spaces; we are interested in the subset

$$C(X, Y) = \big\{\, f\colon X \to Y \;\big|\; f \text{ is continuous} \,\big\} \subseteq \mathrm{Fun}(X, Y)$$

of all continuous functions. There are several ways of making $\mathrm{Fun}(X, Y)$ and $C(X, Y)$ into topological spaces; each is useful in certain situations. The simplest way is to use the product topology on Y^X. The product topology is given by a basis; let us see what the usual basic open sets look like in terms of functions. Take finitely many points $x_1, \dots, x_n \in X$, and finitely many open sets $U_1, \dots, U_n \subseteq Y$; the corresponding basic open set is

$$B(x_1, \dots, x_n, U_1, \dots, U_n) = \big\{\, f\colon X \to Y \;\big|\; f(x_i) \subseteq U_i \text{ for every } i = 1, \dots, n \,\big\}.$$

Example 14.8. When does a sequence of functions $f_n\colon X \to Y$ converge to another function $f\colon X \to Y$ in this topology? For every $x \in X$ and every neighborhood U of $f(x)$, the basic open set $B(x, U)$ is a neighborhood of f, and therefore has to contain all but finitely many of the f_n. In other words, $f_n(x) \in U$ for all but finitely many n, which means that the sequence $f_n(x) \in Y$ converges to $f(x) \in Y$. So convergence in this topology is the same as pointwise convergence.

Definition 14.9. The product topology on $\mathrm{Fun}(X, Y) = Y^X$ is called the *topology of pointwise convergence*.

The disadvantage of this topology is that the subspace $C(X, Y)$ is not closed, because the pointwise limit of continuous functions may fail to be continuous.

Example 14.10. Let $X = [0, 1]$ and $Y = \mathbb{R}$, and consider the sequence of functions $f_n\colon [0, 1] \to \mathbb{R}$, $f_n(x) = x^n$. Since $x^n \to 0$ for $x < 1$, the sequence converges pointwise to the function

$$f\colon [0, 1] \to \mathbb{R}, \quad f(x) = \begin{cases} 0 & \text{if } x < 1, \\ 1 & \text{if } x = 1, \end{cases}$$

which is no longer continuous.

Another choice of topology on $\mathrm{Fun}(X, Y)$ is related to the notion of uniform convergence. Here we assume that (Y, d) is a metric space. Given two functions $f, g \in \mathrm{Fun}(X, Y)$, we define their uniform distance to be

$$\rho(f, g) = \sup_{x \in X} d\big(f(x), g(x)\big)$$

if the right-hand side is ≤ 1, and $\rho(f,g) = 1$ otherwise. (This slightly odd definition is to make sure that ρ takes values in $\mathbb{R}$.) Intuitively, $\rho(f,g)$ is a measurement of the "distance" between the graphs of f and g. You should convince yourself that ρ is in fact a metric on the space $\mathrm{Fun}(X,Y)$; for obvious reasons, it is called the *uniform metric*.

Definition 14.11. The metric topology on $\mathrm{Fun}(X,Y)$ defined by ρ is called the *uniform topology*.

The basic open sets in this metric are balls of some radius: if $0 < \varepsilon < 1$, then

$$B_\varepsilon(f_0) = \big\{\, f\colon X \to Y \;\big|\; d\big(f(x), f_0(x)\big) < \varepsilon \text{ for all } x \in X \,\big\}.$$

So a neighborhood of a given function f_0 consists of all functions whose graphs lie in a strip of width 2ε around the graph of f_0.

Example 14.12. When does a sequence of functions $f_n\colon X \to Y$ converge to another function $f\colon X \to Y$ in the uniform topology? For every $0 < \varepsilon < 1$, the open ball $B_\varepsilon(f)$ is a neighborhood of f, and therefore has to contain all but finitely many of the f_n. In other words, there should exist some N such that

$$d\big(f_n(x), f(x)\big) < \varepsilon$$

for every $n \geq N$ and every $x \in X$; but this is saying exactly that $f_n \to f$ uniformly.

Lemma 14.13. *In the uniform topology, $C(X,Y)$ is a closed subset of* $\mathrm{Fun}(X,Y)$.

Proof. The uniform topology is a metric topology, and so it is enough to show that $C(X,Y)$ is sequentially closed. Suppose a sequence of continuous functions $f_n \in C(X,Y)$ converges to a function $f \in \mathrm{Fun}(X,Y)$. Then $f_n \to f$ uniformly, and we know from Lemma 13.6 that f is also continuous. Thus $f \in C(X,Y)$, and so $C(X,Y)$ is closed. □

In analysis, people are most interested in the case when (Y,d) is a complete metric space (such as $\mathbb{R}^n$). Let me end today's class by showing that the space of (continuous) functions inherits this completeness property.

Proposition 14.14. *If (Y,d) is complete, then both* $\mathrm{Fun}(X,Y)$ *and $C(X,Y)$ are complete with respect to the uniform metric.*

Proof. We first show that $\mathrm{Fun}(X,Y)$ is complete with respect to ρ. Let $f_n \in \mathrm{Fun}(X,Y)$ be an arbitrary Cauchy sequence. For every $\varepsilon > 0$, we can therefore find some N such that

$$d\big(f_m(x), f_n(x)\big) < \varepsilon \tag{14.15}$$

for every $x \in X$ and every $m, n \geq N$. If we fix a point $x \in X$, this implies that the sequence $f_n(x) \in Y$ is a Cauchy sequence; but Y is complete, and so it converges to some limit that we shall denote by $f(x) \in Y$. In this way, we obtain a function $f\colon X \to Y$. For given $\varepsilon > 0$, we can now let m go to infinity in (14.15) to obtain

$$d\big(f(x), f_n(x)\big) \leq \varepsilon$$

for every $x \in X$ and every $n \geq N$. Therefore $\rho(f, f_n) \leq \varepsilon$ whenever $n \geq N$, which means that f is the limit of the Cauchy sequence f_n.

Completeness of $C(X,Y)$ follows from the fact that it is closed: any Cauchy sequence in $C(X,Y)$ has a limit in $\mathrm{Fun}(X,Y)$, because $\mathrm{Fun}(X,Y)$ is complete; the limit actually belongs to $C(X,Y)$, because $C(X,Y)$ is sequentially closed. □

Lecture 15: October 20

Let X be a topological space, and (Y, d) a metric space. Last time, we defined two topologies on the space of functions from X to Y: the topology of pointwise convergence, which is just the product topology on Y^X; and the uniform topology, which is the metric topology induced by the uniform metric. In both cases, we also get a topology on the subspace $C(X, Y)$ of all continuous functions. In the uniform topology, $C(X, Y)$ is a closed subspace, because uniform limits of continuous functions are again continuous; in the topology of pointwise convergence, this is generally not the case.

Today, we are going to study compact subspaces of $C(X, Y)$ in the uniform topology. This question is of some importance in analysis: because $C(X, Y)$ is a metric space, compactness means that every sequence has a convergent subsequence. We shall only look at the special case where X is compact. In that case, the uniform metric that we defined last time is equivalent to the sup metric

$$\rho(f, g) = \sup_{x \in X} d\big(f(x), g(x)\big);$$

note that the supremum is achieved at some point, because $x \mapsto d\big(f(x), g(x)\big)$ is a continuous function on the compact space X.

Equicontinuity. Consider a subset $\mathcal{F} \subseteq C(X, Y)$, that is to say, a collection of continuous functions from X to Y. Under what conditions is $\mathcal{F}$ compact? From Theorem 14.7, we know that a metric space is compact if and only if it is complete and totally bounded. We should therefore try to understand what total boundedness of $\mathcal{F}$ says about the functions in $\mathcal{F}$.

Definition 15.1. We say that $\mathcal{F} \subseteq C(X, Y)$ is *equicontinuous* if, for every $x_0 \in X$ and every $\varepsilon > 0$, there is a neighborhood U of the point x_0 such that

$$d\big(f(x), f(x_0)\big) < \varepsilon$$

for every $x \in U$ and every $f \in \mathcal{F}$.

For a single function f, this condition is just saying that f is continuous; the point of the definition is that the same open set U should work for all the functions in $\mathcal{F}$ at the same time.

Example 15.2. The collection of functions $f_n \colon [0, 1] \to \mathbb{R}$, $f_n(x) = x^n$, is not equicontinuous at the point $x_0 = 1$.

Roughly speaking, total boundedness is equivalent to being equicontinuous. This is the content of the next two lemmas.

Lemma 15.3. *If $\mathcal{F} \subseteq C(X, Y)$ is totally bounded, then it is equicontinuous.*

Proof. Let $\varepsilon > 0$ be an arbitrary positive number. Since $\mathcal{F}$ is totally bounded, it can be covered by finitely open balls of radius $\frac{\varepsilon}{3}$:

$$\mathcal{F} \subseteq B_{\frac{\varepsilon}{3}}(f_1) \cup \cdots \cup B_{\frac{\varepsilon}{3}}(f_n),$$

for certain $f_1, \ldots, f_n \in C(X, Y)$. Since

$$B_r(f_i) = \big\{\, f \colon X \to Y \;\big|\; d\big(f(x), f_i(x)\big) < r \text{ for every } x \in X \,\big\},$$

this means concretely that the graph of every $f \in \mathcal{F}$ stays within $\frac{\varepsilon}{3}$ of the graph of one of the f_i.

Now let $x_0 \in X$ be any point. Each f_i is continuous, and so there is an open set U_i containing x_0, such that $d\big(f_i(x), f_i(x_0)\big) < \frac{\varepsilon}{3}$ for all $x \in U_i$. Define $U = U_1 \cap \cdots \cap U_n$; then U is still a neighborhood of the point x_0, and we have

$$d\big(f_i(x), f_i(x_0)\big) < \frac{\varepsilon}{3}$$

for every $x \in U$ and every $i = 1, \ldots, n$. Now I claim that

$$d\big(f(x), f(x_0)\big) < \varepsilon$$

for every $x \in U$ and every $f \in \mathcal{F}$, which is enough to conclude that $\mathcal{F}$ is equicontinuous. Indeed, for any $f \in \mathcal{F}$, there is some i such that $f \in B_{\frac{\varepsilon}{3}}(f_i)$; but then

$$d\big(f(x), f(x_0)\big) \leq d\big(f(x), f_i(x)\big) + d\big(f_i(x), f_i(x_0)\big) + d\big(f_i(x_0), f(x_0)\big) < \varepsilon,$$

which is what we wanted. □

During the proof, we did not use the fact that X is compact. If we add this assumption, we can get the following better version of equicontinuity.

Lemma 15.4. *Suppose that X is compact and that $\mathcal{F} \subseteq C(X, Y)$ is equicontinuous. Then for every $\varepsilon > 0$, there is a finite open covering $X = U_1 \cup \cdots \cup U_n$ such that*

$$d\big(f(x), f(x')\big) < \varepsilon$$

for every $x, x' \in U_i$ and every $f \in \mathcal{F}$.

Proof. Since $\mathcal{F}$ is equicontinuous, every point $x_0 \in X$ has a neighborhood U such that $d\big(f(x), f(x_0)\big) < \frac{\varepsilon}{2}$ for every $x \in U$ and every $f \in \mathcal{F}$. Using the compactness of X, finitely many of these open sets cover X; if we denote the points by $x_1, \ldots, x_n$ and the open sets by $U_1, \ldots U_n$, we get $X = U_1 \cup \cdots \cup U_n$ and

$$d\big(f(x), f(x')\big) \leq d\big(f(x), f(x_i)\big) + d\big(f(x_i), f(x')\big) < \varepsilon$$

whenever $x, x' \in U_i$. □

Example 15.5. If $\mathcal{F}$ is equicontinuous, then every subset of $\mathcal{F}$ is also equicontinuous.

Example 15.6. One way of proving equicontinuity is by using the derivative. Suppose we are looking at a family $\mathcal{F}$ of continuous functions from $X = [0, 1]$ to $\mathbb{R}$. If each $f \in \mathcal{F}$ is differentiable, then by the mean value theorem,

$$|f(x) - f(y)| \leq |f'(\xi)| \cdot |x - y|$$

for some ξ in the interval between x and y. So if we happen to know that the derivatives of all the functions in $\mathcal{F}$ are uniformly bounded, then we can say that $\mathcal{F}$ must be equicontinuous.

Ascoli's theorem. The main result about compact subsets of $C(X, Y)$ is Ascoli's theorem. This is really a theorem in analysis, but the proof relies on many of the abstract concepts in topology that we have studied this semester.

Theorem 15.7. *Let X be a compact space, let (Y, d) be a metric space, and let $\mathcal{F} \subseteq C(X, Y)$ be a family of continuous functions. Then $\mathcal{F}$ is contained in a compact subspace of $C(X, Y)$ if and only if $\mathcal{F}$ is equicontinuous and, for every $x \in X$, the closure of the set*

$$\mathcal{F}_x = \big\{\, f(x) \;\big|\; x \in X \,\big\} \subseteq Y$$

is compact in Y.

So in particular, $\mathcal{F}$ is compact iff it is closed in the uniform topology on $C(X,Y)$ and satisfies both conditions in the theorem.

Let us begin by proving the easy half of the theorem. We first consider the case where $\mathcal{F}$ is compact. Then $\mathcal{F}$ is a compact metric space, hence totally bounded (by Theorem 14.7), hence equicontinuous (by Lemma 15.3). To prove that $\mathcal{F}_x$ is compact in Y, consider the function

$$\mathrm{ev}_x \colon \mathcal{F} \to Y, \quad \mathrm{ev}_x(f) = f(x).$$

Since $d\big(f(x), g(x)\big) \le \rho(f,g)$ for every $f, g \in \mathcal{F}$, this function is clearly continuous. As $\mathcal{F}$ is compact and Y is Hausdorff, the image

$$\mathcal{F}_x = \mathrm{ev}_x(\mathcal{F}) \subseteq Y$$

is compact. Now suppose that $\mathcal{F}$ is contained in a compact subspace $\mathcal{G}$ of $C(X,Y)$. By the above, $\mathcal{G}$ is equicontinuous, and so $\mathcal{F}$ is also equicontinuous; moreover, we have $\mathcal{F}_x \subseteq \mathcal{G}_x$, and because $\mathcal{G}_x$ is compact, the closure of $\mathcal{F}_x$ is also compact. This proves the easy half of Theorem 15.7.

The other half of the proof requires several steps, so let me first give an outline. We denote by Y^X the space of all functions with the product topology, and by $C(X,Y)$ the space of all continuous functions with the uniform topology. We will prove compactness with the help of Tychonoff's theorem; this is why we also need the product topology. Keep in mind that the topology on $\mathcal{F}$ is induced by the uniform topology on $C(X,Y)$. Here is the outline of the proof:

(1) Let $\mathcal{G}$ be the closure of the set $\mathcal{F}$ in Y^X. We use Tychonoff's theorem to prove that $\mathcal{G}$ is compact (in the product topology).
(2) We show that $\mathcal{G}$ is also equicontinuous; in particular, all the functions in $\mathcal{G}$ are continuous.
(3) We show that the uniform topology coincides with the topology on $\mathcal{G}$.
(4) We conclude the proof by noting that $\mathcal{G}$ is a compact subspace of $C(X,Y)$ containing $\mathcal{F}$.

Step 1. Let $\mathcal{G}$ be the closure of $\mathcal{F}$, considered as a subset of Y^X. More precisely, we have an injective (but not continuous) function

$$\mathcal{F} \to Y^X, \quad f \mapsto \big(f(x)\big)_{x \in X},$$

and we let $\mathcal{G}$ denote the closure of the image. Since the product topology is the topology of pointwise convergence, each function $g \in \mathcal{G}$ is a pointwise limit of a sequence of functions in $\mathcal{F}$; a priori, the functions in $\mathcal{G}$ could therefore be pretty crazy. But in return, it is easy to show that $\mathcal{G}$ is compact. Indeed, for each $x \in X$, we have $f(x) \in \mathcal{F}_x$, and so $\mathcal{G}$ is contained in the subspace

$$\prod_{x \in X} \overline{\mathcal{F}_x}.$$

Since the closure of each $\mathcal{F}_x$ is compact, Tychonoff's theorem tells us that the product is compact; being a closed subspace, $\mathcal{G}$ is therefore also compact.

Step 2. Now we prove that $\mathcal{G}$ is equicontinuous, which means in particular that all the functions in $\mathcal{G}$ are continuous; this is not at all obvious, because they are only pointwise limits of continuous functions. Let $x_0 \in X$ be an arbitrary point. Since $\mathcal{F}$ is equicontinuous, we can find a neighborhood U of x_0 such that

$$d\big(f(x), f(x_0)\big) < \frac{\varepsilon}{3}$$

for all $x \in U$ and all $f \in \mathcal{F}$. I claim that with this choice of U, we also have

$$d\big(g(x), g(x_0)\big) < \varepsilon$$

for all $x \in U$ and all $g \in \mathcal{G}$. To see why, let $x \in U$ and $g \in \mathcal{G}$ be arbitrary. Consider the basic open set

$$\big\{\, f\colon X \to Y \;\big|\; d\big(f(x), g(x)\big) < \tfrac{\varepsilon}{3} \text{ and } d\big(f(x_0), g(x_0)\big) < \tfrac{\varepsilon}{3} \,\big\}$$

in Y^X. It is a neighborhood of the function g, and therefore has to contain some $f \in \mathcal{F}$ (because $\mathcal{G}$ is the closure of $\mathcal{F}$). Then we have

$$d\big(g(x), g(x_0)\big) \le d\big(g(x), f(x)\big) + d\big(f(x), f(x_0)\big) + d\big(f(x_0), g(x_0)\big) < \varepsilon$$

from the triangle inequality, and so $\mathcal{G}$ is equicontinuous.

Step 3. Since all the functions in $\mathcal{G}$ are continuous, $\mathcal{G}$ is a subset of $C(X, Y)$. Our next task is to show that the topology on $\mathcal{G}$ is the same as the subspace topology coming from the uniform topology on $C(X, Y)$. Since every open set in the product topology is also open in the uniform topology, it is enough to prove that the intersection of an arbitrary open set in $C(X, Y)$ with $\mathcal{G}$ is open in $\mathcal{G}$. So let $\mathcal{U} \subseteq C(X, Y)$ be an arbitrary open set; for any $g_0 \in \mathcal{U} \cap \mathcal{G}$, we shall find a neighborhood of g_0 in the product topology whose intersection with $\mathcal{G}$ is contained in $\mathcal{U} \cap \mathcal{G}$. To get started, observe that $\mathcal{U}$ is open, and therefore contains a ball of some radius $r > 0$ around g_0; hence

$$B_r(g_0) \cap \mathcal{G} = \big\{\, g \in \mathcal{G} \;\big|\; d\big(g(x), g_0(x)\big) < r \text{ for all } x \in X \,\big\} \subseteq \mathcal{U} \cap \mathcal{G}.$$

Now X is compact and $\mathcal{G}$ is equicontinuous, and so we can apply Lemma 15.4. This gives us a finite open covering $X = U_1 \cup \cdots \cup U_n$ and points $x_i \in U_i$ with the property that

$$d\big(g(x), g(x_i)\big) < \frac{r}{3}$$

whenever $x \in U_i$ and $g \in \mathcal{G}$. Now consider the set

$$\big\{\, g \in \mathcal{G} \;\big|\; d\big(g(x_i), g_0(x_i)\big) < \tfrac{r}{3} \text{ for every } i = 1, \ldots, n \,\big\}. \tag{15.8}$$

It contains g_0; it is also open in $\mathcal{G}$, being the intersection of $\mathcal{G}$ with a basic open set in the product topology on Y^X. I claim that this entire open set is contained in the intersection $B_r(g_0) \cap \mathcal{G}$, and therefore in $\mathcal{U} \cap \mathcal{G}$. To show this, let $g \in \mathcal{G}$ be any function in (15.8). Every $x \in X$ belongs to some U_i, and so

$$d\big(g(x), g_0(x)\big) \le d\big(g(x), g(x_i)\big) + d\big(g(x_i), g_0(x_i)\big) + d\big(g_0(x_i), g_0(x)\big) < r.$$

Here the first and third term are less than $\frac{r}{3}$ because $g, g_0 \in \mathcal{G}$; the second term is less than $\frac{r}{3}$ because g belongs to the open set in (15.8). We conclude that $g \in B_r(g_0)$, and hence that $\mathcal{U} \cap \mathcal{G}$ is open in $\mathcal{G}$.

Step 4. Now we can easily conclude the proof of Ascoli's theorem. From Step 1, we know that $\mathcal{G}$ is compact in the product topology; from Step 3, we know that it is also compact in the uniform topology. Therefore $\mathcal{F}$ is contained in the compact subspace $\mathcal{G}$, as claimed. □

Note. Let me briefly explain how Ascoli's theorem is used in analysis. Consider a sequence of continuous real-valued functions $f_n\colon X \to \mathbb{R}$ on a compact space X; in the notation of Theorem 15.7, we are looking at the countable subset

$$\mathcal{F} = \{f_1, f_2, \ldots\} \subseteq C(X, \mathbb{R}).$$

Ascoli's theorem tells us that if $\mathcal{F}$ is equicontinuous and pointwise bounded, then the closure of $\mathcal{F}$ is compact; in particular, there is a subsequence that converges uniformly to a continuous real-valued function on X. (Note that a subset of $\mathbb{R}$ is compact if and only if it is closed and bounded.)

Lecture 16: October 25

Algebraic topology. In the remainder of the semester, we shall be talking about algebraic topology. A basic problem in topology is to decide whether two given topological spaces X and Y are homeomorphic or not. To show that X and Y are homeomorphic, we have to find a continuous bijection whose inverse is also continuous; this comes down to being able to construct continuous functions. On the other hand, to show that X and Y are not homeomorphic, we have to prove that there does not exist *any* homeomorphism between them. This can be difficult or even impossible: it requires a lot of work to show that $\mathbb{R}^n$ and $\mathbb{R}^m$ are not homeomorphic for $n \neq m$. In practice, one tries to do this with the help of certain "invariants": for example, topological properties such as connectedness or compactness or countability that are the same for homeomorphic spaces.

Example 16.1. In this way, we can distinguish $\mathbb{R}^2$ from the two-sphere (because one is compact and one is not), or $\mathbb{R}$ from $\mathbb{R}^2$ (because removing a point disconnects one but not the other). But none of the topological properties we have introduced can tell $\mathbb{R}^n$ from $\mathbb{R}^m$ for $n, m \geq 2$.

Another useful property of this type is that of being *simply connected*; roughly speaking, X is simply connected if every closed loop in X can be contracted continuously to a point. You may have seen this notion in complex analysis when discussing line integrals.

Example 16.2. $\mathbb{R}^2$ is simply connected, but $\mathbb{R}^2 \setminus \{(0,0)\}$ is not: for example, any circle around the origin cannot be contracted to a point. For the time being, this statement is only based on our geometric intuition; we shall prove it rigorously later.

More generally, we shall associate to every topological space X (and to a choice of point $x_0 \in X$) a certain group $\pi_1(X, x_0)$, called its *fundamental group*. Roughly speaking, the size of this group is related to how many essentially different closed loops there are in X; in particular, X is simply connected if and only its fundamental group is the trivial group with one element. If X and Y are homeomorphic, then their fundamental groups are isomorphic as groups; this means that if two spaces have different fundamental groups, they cannot be homeomorphic. In this way, the fundamental group can serve as an "algebraic invariant" of a topological space. There is also a whole set of higher homotopy groups $\pi_n(X, x_0)$; closed loops, which are the same as continuous functions from $\mathbb{S}^1$ into X, get replaced by continuous functions from $\mathbb{S}^n$ into X.

This then is the general idea behind algebraic topology: to a topological space X, one associates various algebraic objects such as groups or vector spaces that can serve as invariants. Examples are the fundamental group (first introduced by Poincaré) and various homology and cohomology groups. Algebraic topology also provides various tools for computing the groups of vector spaces in question. Very often, one can obtain a huge amount of information about a space just by knowing its homotopy groups; for that and other reasons, algebraic topology has become one of the most important branches in topology.

Paths and homotopies. Recall that a path in a topological space X is simply a continuous mapping $f\colon I \to X$ from the closed unit interval $I = [0,1]$ to X. The point $f(0)$ is called the starting point of the path, and the point $f(1)$ its

endpoint. Before we can define the fundamental group correctly, we first have to make sense of things like "deforming one path into another" or "contracting a path to a point". The correct notion is that of a homotopy, which roughly means a continuous deformation of one continuous function into another.

Definition 16.3. Let $f\colon X \to Y$ and $f'\colon X \to Y$ be two continuous functions. A *homotopy* between f and f' is a continuous function

$$F\colon X \times I \to Y$$

such that $F(x,0) = f(x)$ and $F(x,1) = f'(x)$ for every $x \in X$. If such a homotopy exists, we say that f and f' are *homotopic*; we often abbreviate this as $f \sim f'$.

A homotopy gives us a family of continuous functions $f_t(x) = F(x,t)$ from X to Y, depending continuously on $t \in I$, that interpolates between $f = f_0$ and $f' = f_1$. In the case of paths, we usually impose that conditions that all the paths in the family should have the same starting point and endpoint.

Definition 16.4. Let $f\colon I \to X$ and $f'\colon I \to X$ be two paths with $f(0) = f'(0) = x_0$ and $f(1) = f'(1) = x_1$. A *path homotopy* between f and f' is a continuous function

$$F\colon I \times I \to X$$

such that $F(s,0) = f(s)$ and $F(s,1) = f'(s)$ for every $s \in I$, and such that $F(0,t) = x_0$ and $F(1,t) = x_1$ for every $t \in I$. If such a path homotopy exists, we say that f and f' are *path homotopic*; we often abbreviate this as $f \sim_p f'$.

It is not hard to see that being (path) homotopic is an equivalence relation.

Lemma 16.5. *Both $\sim$ and $\sim_p$ are equivalence relations.*

Proof. We shall prove this in the case of $\sim$; you should convince yourself that the same argument works for $\sim_p$, too. There are three things to be checked:

(1) For every $f\colon X \to Y$, we have $f \sim f$: for the homotopy $F\colon X \times I \to Y$, one can take $F(x,t) = f(x)$.
(2) If $f \sim g$, then also $g \sim f$. Indeed, if $F\colon X \times I \to Y$ is a homotopy between f and g, then $G(x,t) = F(x,1-t)$ is a homotopy between g and f.
(3) If $f \sim g$ and $g \sim h$, then also $f \sim h$. Indeed, suppose that F is a homotopy between f and g, and that G is a homotopy between g and h. We can then define a homotopy between f and h by setting

$$H\colon X \times I \to Y, \quad H(x,t) = \begin{cases} F(x,2t) & \text{if } t \in [0, \frac{1}{2}], \\ G(x,2t-1) & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Since $F(x,1) = g(x) = G(x,1)$, the pasting lemma shows that H is well-defined and continuous. □

We usually denote the (path) homotopy class of f by the symbol $[f]$.

Example 16.6. Any two functions $f, g\colon X \to \mathbb{R}^n$ are homotopic. For the homotopy, we just move the point $f(x)$ to the point $g(x)$ along the straight line joining them; more formally,

$$F\colon X \times I \to \mathbb{R}^n, \quad F(x,t) = (1-t)f(x) + tg(x)$$

is a homotopy between f and g. More generally, we could replace $\mathbb{R}^n$ by any convex subset $Y \subseteq \mathbb{R}^n$, since the entire line segment between $f(x)$ and $g(x)$ will

still be contained in Y. When f and g are paths with the same starting points and endpoints, F is clearly a path homotopy.

Example 16.7. The two paths

$$f(s) = \big(\cos \pi s, \sin \pi s\big) \quad \text{and} \quad g(s) = \big(\cos \pi s, -\sin \pi s\big)$$

in $\mathbb{R}^2 \setminus \{(0,0)\}$ are not path homotopic; this should be pretty obvious, although we cannot prove it at the moment.

Composition of paths. To make it clear which continuous functions are paths, I will start using lowercase Greek letters like α, β, γ, ... for paths. If α is a path from a point x_0 to a point x_1, and β is a path from x_1 to x_2, we can obviously joint the two paths together into a path from x_0 to x_2. This operation is where we start to see some algebra.

Definition 16.8. Given two paths $\alpha\colon I \to X$ and $\beta\colon I \to X$ with $\alpha(1) = \beta(0)$, we define their *product*

$$\alpha * \beta\colon I \to X, \quad t \mapsto \begin{cases} \alpha(2t) & \text{if } t \in [0, \frac{1}{2}], \\ \beta(2t-1) & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Note that this is well-defined and continuous by the pasting lemma.

The notation should be read from left to right: $\alpha * \beta$ is the path obtained by first going along α and then along β. In fact, the product only depends on the path homotopy classes of α and β.

Lemma 16.9. *If $\alpha \sim_p \alpha'$ and $\beta \sim_p \beta'$, then $\alpha * \beta \sim_p \alpha' * \beta'$.*

Proof. If A is a path homotopy between α and α', and B a path homotopy between β and β', then

$$C\colon I \times I \to X, \quad C(s,t) = \begin{cases} A(2s,t) & \text{if } s \in [0, \frac{1}{2}], \\ B(2s-1,t) & \text{if } s \in [\frac{1}{2}, 1], \end{cases}$$

is a path homotopy between $\alpha * \beta$ and $\alpha' * \beta'$. □

It therefore makes sense to define the product for path homotopy classes by the formula $[\alpha] * [\beta] = [\alpha * \beta]$; the lemma shows that even if we use different representatives in each class, the result is the same. Keep in mind that the product is only defined when $\alpha(1) = \beta(0)$.

Proposition 16.10. *The operation $*$ has the following algebraic properties:*

(a) It is associative:

$$[\alpha] * \big([\beta] * [\gamma]\big) = \big([\alpha] * [\beta]\big) * [\gamma]$$

whenever $\alpha(1) = \beta(0)$ and $\beta(1) = \gamma(0)$.

(b) The class of the constant path $e_x(s) = x$ acts as an identity element:

$$[\alpha] * [e_{\alpha(1)}] = [\alpha] \quad \text{and} \quad [e_{\alpha(0)}] * [\alpha] = [\alpha].$$

(c) The class of the opposite path $\bar{\alpha}(s) = \alpha(1-s)$ acts as an inverse element:

$$[\alpha] * [\bar{\alpha}] = [e_{\alpha(0)}] \quad \text{and} \quad [\bar{\alpha}] * [\alpha] = [e_{\alpha(1)}].$$

Proof. One can prove everything by writing down suitable path homotopies: for example, a path homotopy between $\alpha * e_{\alpha(1)}$ and α is given by

$$H\colon I \times I \to X, \quad H(s,t) = \begin{cases} \alpha(1) & \text{if } t \leq 2s-1, \\ \alpha\left(\frac{2s}{1+t}\right) & \text{if } t \geq 2s-1. \end{cases}$$

Writing down all of these formulas is too much work, though, so let me present a different argument. Observe that in each identity, the two paths go through the same set of points of X, but at different speeds. So we only have to find homotopies that adjust the speed of each path. More precisely, we will make use of the following two simple observations:

(1) If $\alpha \sim_p \alpha'$, and if $f\colon X \to Y$ is continuous, then
$$f \circ \alpha \sim_p f \circ \alpha'.$$
This is obvious, because if F is a path homotopy between α and α', then $f \circ F$ is a path homotopy between $f \circ \alpha$ and $f \circ \alpha'$.

(2) If $\alpha(1) = \beta(0)$, and if $f\colon X \to Y$ is continuous, then
$$f \circ (\alpha * \beta) = (f \circ \alpha) * (f \circ \beta).$$
Again, this is easy to see from the definition of $*$.

Now let us get going with the proof. For (b), we have to show that $\alpha * e_{\alpha(1)} \sim_p \alpha$. Observe that

$$\alpha * e_{\alpha(1)} = \alpha \circ (i * e_1) \quad \text{and} \quad \alpha = \alpha \circ i,$$

where $i\colon I \to I$ is the identity path $i(s) = s$, and $e_1\colon I \to I$ is the constant path $e_1(s) = 1$. Now $i * e_1$ and i are two paths in I that start and end at the same points; since I is convex, they must be path homotopic. But then $i * e_1 \sim_p i$, and by the observation above,

$$\alpha * e_{\alpha(1)} = \alpha \circ (i * e_1) \sim_p \alpha \circ i = \alpha.$$

The same argument proves that $e_{\alpha(0)} * \alpha \sim_p \alpha$, and hence (b). To get (c), note that

$$\bar{\alpha} = \alpha \circ \bar{\imath},$$

where $\bar{\imath}\colon I \to I$ is the path $\bar{\imath}(s) = 1 - s$. For the same reason as above, $i * \bar{\imath} \sim_p e_0$ and $\bar{\imath} * i \sim_p e_1$, and then (c) follows by composing with α.

It remains to prove (a). Both paths in question are running through α, β, and γ is the same order, but at different speeds. Let us define

$$\pi\colon I \to X, \quad \pi(s) = \begin{cases} \alpha(3s) & \text{if } s \in [0, \frac{1}{3}], \\ \beta(3s-1) & \text{if } s \in [\frac{1}{3}, \frac{2}{3}], \\ \gamma(3s-2) & \text{if } s \in [\frac{2}{3}, 1] \end{cases}$$

Then you can check that $\alpha * (\beta * \gamma) = \pi \circ \ell$, where ℓ is the continuous function

$$\ell\colon I \to I, \quad \ell(s) = \begin{cases} \frac{2}{3}s & \text{if } s \in [0, \frac{1}{2}], \\ \frac{4}{3}s - \frac{1}{3} & \text{if } s \in [\frac{1}{2}, 1]. \end{cases}$$

Likewise, $(\alpha * \beta) * \gamma = \pi \circ r$, where r is the continuous function

$$r\colon I \to I, \quad r(s) = \begin{cases} \frac{4}{3}s & \text{if } s \in [0, \frac{1}{2}], \\ \frac{2}{3}s + \frac{1}{3} & \text{if } s \in [\frac{1}{2}, 1]. \end{cases}$$

Now r and ℓ are paths in I that both start at the point 0 and end at the point 1, and so $\ell \sim_p r$; as before, we conclude that

$$\alpha * (\beta * \gamma) = \pi \circ \ell \sim_p \pi \circ r = (\alpha * \beta) * \gamma,$$

which is what we needed to show. □

The fundamental group. We can now define the fundamental group. Let X be a topological space, and let $x_0 \in X$ be a point. A path $\alpha\colon I \to X$ is called a *loop based at* x_0 if it starts and ends at the point x_0, in the sense that $\alpha(0) = \alpha(1) = x_0$. We define

$$\pi_1(X, x_0) = \left\{ [\alpha] \;\middle|\; \alpha\colon I \to X \text{ is a path with } \alpha(0) = \alpha(1) = x_0 \right\}$$

by considering all loops based at x_0, up to path homotopy. In fact, $\pi_1(X, x_0)$ is not just a set, but a group; the group operation is given by composition by paths. More precisely, any two loops based at x_0 can be composed, and so we have a well-defined operation

$$*\colon \pi_1(X, x_0) \times \pi_1(X, x_0) \to \pi_1(X, x_0), \quad [\alpha] * [\beta] = [\alpha * \beta].$$

The properties in Proposition 16.10 are exactly the axioms for being a group. Recall the following definition from algebra.

Definition 16.11. Let G be a set with a binary operation $\circ\colon G \times G \to G$. Then $(G, \circ)$ is called a *group* if it has the following three properties:

(a) The operation is associative: for every $g_1, g_2, g_3 \in G$, one has $g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3$.
(b) There is an element $e \in G$ such that $g \circ e = e \circ g = g$ for every $g \in G$.
(c) For every $g \in G$, there is an element $g^{-1} \in G$ with $g \circ g^{-1} = g^{-1} \circ g = e$.

The element $e \in G$ is called the *unit*, and for given $g \in G$, the element $g^{-1} \in G$ is called the *inverse* of g; one can show that both are uniquely determined. In practice, the group operation is denoted by juxtaposition: gh means $g \circ h$.

In the case of $\pi_1(X, x_0)$, all three group axioms for the operation $*$ are proved in Proposition 16.10.

Definition 16.12. Let X be a topological space and $x_0 \in X$. Then $\pi_1(X, x_0)$, together with the operation $*$, is called the *fundamental group* of X (for the given base point x_0).

The fundamental group does depend on the base point; we shall investigate this dependence next time.

Definition 16.13. Let X be a path connected topological space. We say that X is *simply connected* if, for every point $x_0 \in X$, the fundamental group $\pi_1(X, x_0)$ is the trivial group (with one element).

This makes our earlier definition precise: a path connected space is simply connected if every loop in X can be contracted to a point.

Example 16.14. Our earlier example of the straight line homotopy shows that $\mathbb{R}^n$, or any convex subset of $\mathbb{R}^n$, is simply connected.

Lecture 17: October 27

Last time, we defined the fundamental group $\pi_1(X, x_0)$ of a topological space. Its elements are paths in X that start and end at the base point x_0, taken up to path homotopy equivalence. Thus a typical element looks like $[\alpha]$, where $\alpha\colon I \to X$ is a continuous function with $\alpha(0) = \alpha(1) = x_0$. The group structure is defined as $[\alpha] * [\beta] = [\alpha * \beta]$, where $\alpha * \beta$ is the path obtained by transversing α followed by β.

Dependence on the base point. Let us first investigate how the fundamental group depends on the choice of base point. Since any path in X has to stay inside the path component of x_0, it makes sense to assume that X is path connected. Let $x_0, x_1 \in X$ be two candidates for the base point. Since X is path connected, we can choose a path $\varphi\colon I \to X$ with $\varphi(0) = x_0$ and $\varphi(1) = x_1$. We obtain a function

$$\hat{\varphi}\colon \pi_1(X, x_0) \to \pi_1(X, x_1), \quad \hat{\varphi}[\alpha] = [\bar{\varphi}] * [\alpha] * [\varphi];$$

in other words, given a path α based at x_0, we build a new path based at x_1 by moving from x_1 to x_0 along the path $\bar{\varphi}$, then transversing α, and then moving back from x_0 to x_1 along φ.

Lemma 17.1. *The function $\hat{\varphi}\colon \pi_1(X, x_0) \to \pi_1(X, x_1)$ is an isomorphism of groups.*

Proof. Recall that a function $\phi\colon G \to H$ between two groups is called a *homomorphism* if $\phi(gh) = \phi(g)\phi(h)$ for every $g, h \in H$; from this condition one can deduce, with the help of the group axioms, that $\phi(e) = e$ and that $\phi(g^{-1}) = \phi(g)^{-1}$ for every $g \in G$. A bijective homomorphism is called an *isomorphism*; if ϕ is an isomorphism, then the inverse function $\phi^{-1}\colon H \to G$ is automatically a homomorphism as well. (Can you prove this?)

Now let us prove the statement. It is easy to see that $\hat{\varphi}$ is a homomorphism. Indeed, given two elements $[\alpha]$ and $[\beta]$ in $\pi_1(X, x_0)$, we have

$$\begin{aligned}\hat{\varphi}\big([\alpha] * [\beta]\big) &= [\bar{\varphi}] * [\alpha] * [\beta] * [\varphi] \\ &= [\bar{\varphi}] * [\alpha] * [\varphi] * [\bar{\varphi}] * [\beta] * [\varphi] = \big(\hat{\varphi}[\alpha]\big) * \big(\hat{\varphi}[\beta]\big),\end{aligned}$$

due to the identity $[\varphi] * [\bar{\varphi}] = [e_{x_0}]$. If we denote by $\psi = \bar{\varphi}$ the reverse path from x_1 to x_0, we get another homomorphism

$$\hat{\psi}\colon \pi_1(X, x_1) \to \pi_1(X, x_0),$$

and it is clear from the definitions that $\hat{\psi}$ is the inverse function of $\hat{\varphi}$, which is therefore an isomorphism. □

Functorial properties. The fundamental group was supposed to be an invariant of a topological space X, and so we should also convince ourselves that homeomorphic spaces have isomorphic fundamental groups. Suppose we have a continuous function $f\colon X \to Y$. If α is a path in X starting and ending at the point x_0, then $f \circ \alpha$ is a path in Y starting and ending at the point $y_0 = f(x_0)$; moreover, if $\alpha \sim_p \alpha'$, then $f \circ \alpha \sim_p f \circ \alpha'$. In this way, we obtain a well-defined function

$$f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0), \quad f_*[\alpha] = [f \circ \alpha].$$

This function is actually a group homomorphism.

Lemma 17.2. *If $f\colon (X, x_0) \to (Y, y_0)$ is continuous, $f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$ is a homomorphism of groups. Given a second continuous function $g\colon (Y, y_0) \to (Z, z_0)$, one has $(g \circ f)_* = g_* \circ f_*$.*

Here the notation $f\colon (X, x_0) \to (Y, y_0)$ means that $f\colon X \to Y$ is a function with the property that $f(x_0) = y_0$.

Proof. The first assertion holds because

$$f_*\big([\alpha] * [\beta]\big) = f_*[\alpha * \beta] = [f \circ (\alpha * \beta)] = [(f \circ \alpha) * (f \circ \beta)] = \big(f_*[\alpha]\big) * \big(f_*[\beta]\big).$$

The second assertion holds because

$$(g \circ f)_*[\alpha] = [(g \circ f) \circ \alpha] = [g \circ (f \circ \alpha)] = g_*[f \circ \alpha] = g_*\big(f_*[\alpha]\big). \qquad \square$$

In the language of category theory, the above construction is an example of a *functor* from the category of topological spaces to the category of groups. The most basic example of a *category* is the category of sets: a typical object is a set X, and a morphism between two sets X and Y is simply a function $f\colon X \to Y$. One can add extra structure and consider for example the category of topological spaces or the category of groups; in each case, the right notion of morphism is one that preserves the extra structure: a morphism between topological spaces is a continuous function, and a morphism between groups is a group homomorphism.

Example 17.3. One can also consider the category of topological spaces with base point: its objects are pairs (X, x_0), and its morphisms are continuous functions $f\colon (X, x_0) \to (Y, y_0)$.

In that sense, sending a continuous function $f\colon (X, x_0) \to (Y, y_0)$ to the group homomorphism $f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$ is a functor: it takes morphisms in the category of topological spaces with base point to morphisms in the category of groups, in a way that is compatible with composition.

Corollary 17.4. *If $f\colon X \to Y$ is a homeomorphism with $f(x_0) = y_0$, then*

$$f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$$

is an isomorphism of groups.

Proof. Let $g\colon Y \to X$ denote the inverse function; since $g(y_0) = x_0$, it induces a homomorphism

$$g_*\colon \pi_1(Y, y_0) \to \pi_1(X, x_0)$$

in the opposite direction. From the lemma, we get $f_* \circ g_* = (f \circ g)_* = \mathrm{id}_* = \mathrm{id}$ and $g_* \circ f_* = (g \circ f)_* = \mathrm{id}_* = \mathrm{id}$, and so f_* is an isomorphism. $\square$

Note. To make the notation less cumbersome, let us agree on the following conventions. If α is a path in X that starts and ends at the point x_0, we denote the corresponding element of $\pi_1(X, x_0)$ also by the letter α, instead of the more correct $[\alpha]$. For the product, we write $\alpha\beta$ instead of the more correct $[\alpha] * [\beta]$; for the unit element, we simply write e instead of the more correct $[e_{x_0}]$.

An example. Now let us compute our first nontrivial example, namely the fundamental group of the circle $\mathbb{S}^1$. For convenience, we use the point $b_0 = (1, 0) \in \mathbb{S}^1$ as the base point; if we think of $\mathbb{S}^1$ as the set of all complex numbers of modulus 1, then b_0 is nothing but $1 \in \mathbb{C}$. Suppose we have a path α in the circle that starts and ends at the point b_0. At least intuitively, the only thing that matters is how many times the path winds around the circle – with a plus or minus sign, depending on the orientation – and so the answer should be the following.

Theorem 17.5. *$\pi_1(\mathbb{S}^1, b_0)$ is isomorphic to $(\mathbb{Z}, +)$.*

The goal is to prove that this is indeed the case. We could try to define some sort of "winding number" for paths in $\mathbb{S}^1$, but the following approach is easier. Consider the continuous function

$$p\colon \mathbb{R} \to \mathbb{S}^1, \quad p(x) = \big(\cos 2\pi x, \sin 2\pi x\big).$$

Clearly, p is locally a homeomorphism: every sufficiently small neighborhood of $x \in \mathbb{R}$ is mapped homeomorphically to its image in $\mathbb{S}^1$. The preimage of our base point b_0 is exactly the set of all integers $\mathbb{Z}$. We can picture p by imagining an infinite spiral in $\mathbb{R}^3$, winding around the cylinder $\mathbb{S}^1 \times \mathbb{R}$ and projecting down to the circle in $\mathbb{R}^2$.

Here is the idea for the proof of Theorem 17.5. We first show that every path $\alpha\colon I \to \mathbb{S}^1$ with $\alpha(0) = b_0$ can be uniquely lifted to a path $\tilde{\alpha}\colon I \to \mathbb{R}$ with $\tilde{\alpha}(0) = 0$ and $p \circ \tilde{\alpha} = \alpha$; we also show that homotopic paths have homotopic liftings. The endpoint $\tilde{\alpha}(1)$ of the lifted path belongs to the set $\mathbb{Z} \in p^{-1}(0)$, and it turns out that the function

$$\pi_1(\mathbb{S}^1, b_0) \to \mathbb{Z}, \quad [\alpha] \mapsto \tilde{\alpha}(1),$$

is an isomorphism of groups.

Covering spaces. The function $p\colon \mathbb{R} \to \mathbb{S}^1$ is an example of a covering space, and for later use, we will work in this more general setting.

Definition 17.6. Let $p\colon E \to B$ be a surjective and continuous function. An open subset $U \subseteq B$ is said to be *evenly covered by* p if its preimage $p^{-1}(U)$ is a disjoint union of open subsets of E, each of which is homeomorphic to U (via p). We say that $p\colon E \to B$ is a *covering space* if every point of B has a neighborhood that is evenly covered by p.

In symbols, the condition on p is that

$$p^{-1}(U) = \bigsqcup_{i \in I} V_i$$

should be the disjoint union of open sets $V_i \subseteq E$, in such a way that

$$p\big|_{V_i}\colon V_i \to U$$

is a homeomorphism for each $i \in I$. The sets V_i are usually called the *sheets* of the covering space. In particular, p is a *local homeomorphism*: every point in E has a neighborhood that is mapped homeomorphically (under p) to its image in B.

Example 17.7. The function $p\colon \mathbb{R} \to \mathbb{S}^1$ is a covering space: the preimage of the open set $U = \mathbb{S}^1 \setminus \{b_0\}$ is

$$p^{-1}(U) = \mathbb{R} \setminus \mathbb{Z} = \bigcup_{m \in \mathbb{Z}} (m, m+1),$$

and the restriction of p to each open interval is a homeomorphism with U. Likewise, the preimage of $V = \mathbb{S}^1 \setminus \{-b_0\}$ is

$$p^{-1}(V) = \mathbb{R} \setminus \big(\tfrac{1}{2} + \mathbb{Z}\big) = \bigcup_{m \in \mathbb{Z}} \big(m - \tfrac{1}{2}, m + \tfrac{1}{2}\big),$$

and the restriction of p to each open interval is a homeomorphism with V. In this example, every point of $\mathbb{S}^1$ has infinitely many preimages, and quite naturally, one says that p is an *infinite-sheeted covering space*.

Example 17.8. To get an example with finitely many sheets, think of $\mathbb{S}^1$ as being a subset of $\mathbb{C}$, and consider the function $z \mapsto z^m$ from $\mathbb{S}^1$ to itself. Each point has exactly m preimages, and one can show that this is a covering space.

Note that in a covering space, the fiber $p^{-1}(b)$ over any point $b \in B$ is a discrete subspace of E; this is clear because some neighborhood of b is evenly covered by p.

Lifting paths and homotopies. Now let $p\colon E \to B$ be a covering space. Fix a base point $b_0 \in B$; since p is surjective, we can choose a base point $e_0 \in E$ such that $p(e_0) = b_0$. An important property of covering spaces is that paths (and homotopies) in B can be lifted uniquely to E, once we decide where the starting point should go.

Theorem 17.9. *Let $\alpha\colon I \to B$ be a path with $\alpha(0) = b_0$. Then there is a unique path $\tilde{\alpha}\colon I \to E$ with $\tilde{\alpha}(0) = e_0$ and $p \circ \tilde{\alpha} = \alpha$.*

For every $s \in I$, the point $\tilde{\alpha}(s) \in E$ lies over the corresponding point $\alpha(s) \in B$, and so one says that $\tilde{\alpha}$ is a *lifting* of the path α. The assertion in the theorem is known as the *path lifting property* of covering spaces.

Proof. To get the idea of the proof, suppose first that $\alpha(I)$ lies entirely inside an open set $U \subseteq B$ that is evenly covered by p. Then $p^{-1}(U)$ is a disjoint union of open sets; because $p(e_0) = b_0 = \alpha(0) \in U$, exactly one of these open sets contains the point e_0. If we denote this open set by $V \subseteq E$, then

$$p\big|_V\colon V \to U$$

is a homeomorphism. It therefore makes sense to define

$$\tilde{\alpha} = \big(p\big|_V\big)^{-1} \circ \alpha\colon I \to E.$$

This is a path in E with $p \circ \tilde{\alpha} = \alpha$; because e_0 is the unique point of V that lies over b_0, we also get $\tilde{\alpha}(0) = e_0$. Uniqueness is clear: any path lifting α must be contained in $p^{-1}(U)$, and because we are assuming that the lifting starts at the point e_0, actually in V; but $p\big|_V$ is a homeomorphism, and so $\tilde{\alpha}$ is uniquely determined.

Now let us deal with the general case. The idea is to subdivide I into smaller intervals $I_1, \dots, I_N$ that are each mapped into an evenly covered open set, and then construct the lifting $\tilde{\alpha}$ in N steps. By assumption, every point of B has a neighborhood that is evenly covered by p. Since $\alpha(I)$ is compact, it is contained in the union of finitely many such open sets; thus we can find evenly covered open subsets $U_1, \dots, U_n \subseteq B$ such that

$$I = \bigcup_{k=1}^{n} \alpha^{-1}(U_k).$$

Lemma 17.10 below shows that if we choose a sufficiently large integer N and divide I into subintervals $I_1, \dots, I_N$ of length $\frac{1}{N}$, then each subinterval is contained entirely in one open set $\alpha^{-1}(U_k)$. By applying the argument from above to the restriction of α to the first subinterval I_1, we obtain a unique lifting

$$\tilde{\alpha}\colon I_1 \to E$$

with $\tilde{\alpha}(0) = e_0$; it ends at some point $\tilde{\alpha}(\frac{1}{N}) \in E$, which lies over the point $\alpha(\frac{1}{N}) \in B$. One the second subinterval I_2, there is again a unique lifting of $\alpha\colon I_2 \to B$ that

maps the left endpoint $\frac{1}{N} \in I_2$ to the point $\tilde{\alpha}(\frac{1}{N})$; by the pasting lemma, we obtain a continuous function

$$\tilde{\alpha} \colon I_1 \cup I_2 \to E$$

with $p \circ \tilde{\alpha} = \alpha$ on $I_1 \cup I_2$. Continuing in this manner, we obtain the desired lifting of α (and its uniqueness) after N steps. □

During the proof, we used the following lemma; note that the same result is true (with the same proof) for finite open coverings of arbitrary compact sets in $\mathbb{R}^n$.

Lemma 17.10. *Let $I = U_1 \cup \cdots \cup U_n$ be a finite open covering of the closed unit interval $I \subset \mathbb{R}$. Then there exists $\varepsilon > 0$ such that for every $x \in I$, the open ball $B_\varepsilon(x)$ is contained entirely in some U_k.*

Proof. For each $k = 1, \ldots, n$, consider the function

$$f_k \colon I \to [0, \infty), \quad f_k(x) = \sup\big\{\, r \geq 0 \;\big|\; B_r(x) \subseteq U_k \,\big\}.$$

We have $f_k(x) = 0$ if $x \notin U_k$, and $f_k(x) > 0$ if $x \in U_k$; it is also not hard to see that f_k is continuous. If we define

$$f \colon I \to [0, \infty), \quad f(x) = \max_{k=1,\ldots,n} f_k(x),$$

then the pasting lemma shows that f is continuous; we have $f(x) > 0$ for every $x \in I$ because the open sets $U_1, \ldots, U_n$ cover I. Because I is compact, f achieves a minimum on I, and so we can find some $\varepsilon > 0$ such that $f(x) \geq \varepsilon$ for every $x \in I$. This gives us what we want. □

A very similar construction also allows us to lift homotopies; the following result is known as the *homotopy lifting property* of covering spaces.

Theorem 17.11. *Let $H \colon I \times I \to B$ be a homotopy with $H(0,0) = b_0$. Then there is a unique homotopy $\tilde{H} \colon I \times I \to E$ with $\tilde{H}(0,0) = e_0$ and $p \circ \tilde{H} = H$.*

Proof. When $H(I \times I)$ is contained entirely in an evenly covered open subset $U \subseteq B$, this can be proved by exactly the same argument as before. In the general case, we subdivide $I \times I$ into N^2 little squares of side length $\frac{1}{N}$; in the notation from above, these little squares are of the form $I_j \times I_k$ for $1 \leq j, k \leq N$. If we choose N sufficiently large, then each $\alpha(I_j \times I_k)$ is contained in an evenly covered subset of B; we can now construct the lifting $\tilde{H}$ (and prove its uniqueness) in N^2 steps, by going through the set of pairs (j, k) in lexicographic order. □

Lecture 18: November 1

We are in the middle of computing the fundamental group of $\mathbb{S}^1$ with the help of the covering space $\mathbb{R} \to \mathbb{S}^1$. With that goal in mind, we showed last time covering spaces have the *path lifting property* (and the *homotopy lifting property*). Let $p\colon E \to B$ be an arbitrary covering space, and choose base points $b_0 \in B$ and $e_0 \in E$ with the property that $p(e_0) = b_0$. Last time, we showed that any path $\alpha\colon I \to B$ with $\alpha(0) = b_0$ can be uniquely lifted to a path $\tilde{\alpha}\colon I \to E$ with $\tilde{\alpha}(0) = e_0$; here the word "lifting" means that $p \circ \tilde{\alpha} = \alpha$. If α is a loop based at the point b_0, then $\tilde{\alpha}(1)$ must belong to the fiber $p^{-1}(b_0)$, because $p\big(\tilde{\alpha}(1)\big) = \alpha(1) = b_0$.

The homotopy lifting property implies that the endpoint of a lifted path only depends on the path homotopy class of the path. Let $\alpha, \beta\colon I \to B$ be two paths with $\alpha(0) = \beta(0) = b_0$. Denote by $\tilde{\alpha}, \tilde{\beta}\colon I \to E$ the liftings constructed in Theorem 17.9, subject to the condition that $\tilde{\alpha}(0) = \tilde{\beta}(0) = e_0$.

Corollary 18.1. *Suppose that $\alpha(1) = \beta(1)$ and that $\alpha \sim_p \beta$. Then we also have $\tilde{\alpha}(1) = \tilde{\beta}(1)$ and $\tilde{\alpha} \sim_p \tilde{\beta}$.*

Proof. Let $H\colon I \times I \to B$ be a path homotopy between α and β; if we define $b_1 = \alpha(1) = \beta(1)$, then

$$H(s,0) = \alpha(s), \quad H(s,1) = \beta(s), \quad H(0,t) = b_0, \quad H(1,t) = b_1$$

for every $s, t \in I$. Theorem 17.11 shows that there is a unique lifting $\tilde{H}\colon I \times I \to E$ with $\tilde{H}(0,0) = e_0$. I claim that $\tilde{H}$ is a path homotopy between $\tilde{\alpha}$ and $\tilde{\beta}$.

To prove this claim, we have to analyze what $\tilde{H}$ does on the four edges of $I \times I$:

(1) To begin with, $\tilde{H}(-,0)\colon I \to E$ is a lifting of the path α with $\tilde{H}(0,0) = e_0$; the uniqueness statement in Theorem 17.9 shows that $\tilde{H}(s,0) = \tilde{\alpha}(s)$.
(2) Likewise, $\tilde{H}(0,-)\colon I \to E$ is a lifting of the constant path b_0; by uniqueness, we must have $\tilde{H}(0,t) = e_0$.
(3) Since $\tilde{H}(0,-)\colon I \to E$ is a lifting of the constant path b_1, we also have $\tilde{H}(1,t) = e_1$ for some $e_1 \in E$.
(4) Lastly, $\tilde{H}(-,1)\colon I \to E$ is a lifting of the path β with $\tilde{H}(0,1) = e_0$; for the same reason as before, $\tilde{H}(s,1) = \tilde{\beta}(s)$.

The conclusion is that $\tilde{\alpha}(1) = \tilde{\beta}(1) = e_1$, and that $\tilde{H}$ is a path homotopy between $\tilde{\alpha}$ and $\tilde{\beta}$. □

The corollary shows that the following *lifting correspondence* is well-defined:

$$\ell\colon \pi_1(B, b_0) \to p^{-1}(b_0), \quad \ell(\alpha) = \tilde{\alpha}(1)$$

In certain cases, one can use it to compute the fundamental group.

Theorem 18.2. *If E is path connected, the lifting correspondence ℓ is surjective. If E is simply connected, ℓ is bijective.*

Proof. Let $e \in p^{-1}(b_0)$ be an arbitrary point of the fiber. Since E is path connected, there is a path $\tilde{\alpha}\colon I \to E$ with $\tilde{\alpha}(0) = e_0$ and $\tilde{\alpha}(1) = e$. Obviously, $\tilde{\alpha}$ is a lifting of the path $\alpha = p \circ \tilde{\alpha}\colon I \to B$; moreover, $\alpha(0) = \alpha(1) = b_0$. Now the definition of the lifting correspondence shows that $\ell(\alpha) = \tilde{\alpha}(1) = e$, and so ℓ must be surjective.

If E is simply connected, we can prove moreover that ℓ is injective (and therefore bijective). Suppose we have two elements $\alpha, \beta \in \pi_1(B, b_0)$ with $\ell(\alpha) = \ell(\beta)$. This

means that the liftings $\tilde{\alpha}$ and $\tilde{\beta}$ end at the same point. Since E is simply connected, it follows that $\tilde{\alpha} \sim_p \tilde{\beta}$; but then $\alpha = p \circ \tilde{\alpha} \sim_p p \circ \tilde{\beta} = \beta$. □

We can now finish the computation of the fundamental group of $\mathbb{S}^1$.

Proof of Theorem 17.5. The covering space $p\colon \mathbb{R} \to \mathbb{S}^1$ has the property that $p^{-1}(b_0) = \mathbb{Z}$. Since $\mathbb{R}$ is simply connected, the lifting correspondence

$$\ell\colon \pi_1(\mathbb{S}^1, b_0) \to \mathbb{Z}$$

is bijective by Theorem 18.2. To conclude the proof, we have to show that ℓ is an isomorphism of groups, which is to say that

$$\ell(\alpha * \beta) = \ell(\alpha) + \ell(\beta).$$

Let $\tilde{\alpha}$ denote the unique lifting of α with $\tilde{\alpha}(0) = 0$; likewise for $\tilde{\beta}$. If we add the number $\tilde{\alpha}(1)$ to the path $\tilde{\beta}$, we obtain a new path

$$\tilde{\beta} + \tilde{\alpha}(1)\colon I \to \mathbb{R}, \quad s \mapsto \tilde{\beta}(s) + \tilde{\alpha}(1).$$

Since $\tilde{\alpha}(1) \in \mathbb{Z}$, this new path is still a lifting of β, but now starting at the point $\tilde{\alpha}(1)$. The composition

$$\tilde{\alpha} * \big(\tilde{\beta} + \tilde{\alpha}(1)\big)$$

is therefore well-defined, and clearly a lifting of the path $\alpha * \beta$; since it starts at the point 0 and ends at the point $\tilde{\alpha}(1) + \tilde{\beta}(1)$, we get

$$\ell(\alpha * \beta) = \tilde{\alpha}(1) + \tilde{\beta}(1) = \ell(\alpha) + \ell(\beta),$$

which is the result we were after. □

Applications. The fact that the fundamental group of the circle is nontrivial has several useful consequences. One is a topological proof for the fundamental theorem of algebra; for that, see this week's homework. Another one is Brouwer's fixed point theorem for the closed unit disk. Before we get to that, let me first recall the following definition.

Definition 18.3. Let $A \subseteq X$ be a subspace of a topological space X. A *retraction* is a continuous function $r\colon X \to A$ such that $r(x) = x$ for every $x \in A$. If such a retraction exists, one says that A is a *retract* of X.

If A is a retract of X, it is somehow an essential part of X; the following lemma shows how this manifests itself in the fundamental group.

Lemma 18.4. *If A is a retract of X, then the fundamental group of A embeds into the fundamental group of X.*

Proof. Let $j\colon A \to X$ denote the inclusion; a retraction is a continuous function $r\colon X \to A$ with $r \circ j = \mathrm{id}$. If we choose a base point $a_0 \in A$, we obtain two homomorphisms

$$j_*\colon \pi_1(A, a_0) \to \pi_1(X, a_0) \quad \text{and} \quad r_*\colon \pi_1(X, a_0) \to \pi_1(A, a_0).$$

By Lemma 17.2, we have $r_* \circ j_* = \mathrm{id}$, which of course means that j_* is injective (and that r_* is surjective). □

We can use this to show that the circle is not a retract of the closed disk B^2.

Corollary 18.5. *There is no retraction of B^2 onto $\mathbb{S}^1$.*

Proof. The group $\pi_1(\mathbb{S}^1, b_0)$ is nontrivial, whereas the group $\pi_1(B^2, b_0)$ is trivial (because B^2 is a convex subset of $\mathbb{R}^2$). In particular, there is no embedding of the former into the latter. □

We can now prove the following fixed point theorem for the closed unit ball

$$B^n = \left\{ x \in \mathbb{R}^n \;\middle|\; \|x\| \leq 1 \right\}$$

in Euclidean space, at least in the special case $n = 2$.

Theorem 18.6 (Brouwer's fixed point theorem)**.** *Every continuous function*

$$f \colon B^n \to B^n$$

has a fixed point: there is a point $x \in B^n$ with the property that $f(x) = x$.

Proof for $n = 2$. Let $f \colon B^2 \to B^2$ be an arbitrary continuous function; our goal is to prove that f must have a fixed point. Suppose that it does not; then $f(x) \neq x$ for every $x \in B^2$. We can exploit this to construct a retraction $r \colon B^2 \to \mathbb{S}^1$ as follows. Given any $x \in B^2$, consider the ray emanating from the point $f(x)$ and passing through the point x. Let $r(x)$ be the unique point where this ray meets the boundary of B^2; since $f(x) \neq x$, one can easily check that r is well-defined and continuous. For $x \in \mathbb{S}^1$, we obviously have $r(x) = x$, and so r is a retraction. This contradicts what we proved above, and so f must have a fixed point after all. □

Let me briefly describe how one uses algebraic topology to prove Brouwer's fixed point theorem for other values of n. Given a continuous function $f \colon B^n \to B^n$ without fixed points, one obtains a retraction $r \colon B^n \to \mathbb{S}^{n-1}$ by the same construction as for $n = 2$; so the point is to show that such retractions cannot exist. It is here that the *homology groups* $H_{n-1}(X, \mathbb{Z})$ come in: just as in the case of the fundamental group, the existence of a retraction would mean that $H_{n-1}(\mathbb{S}^{n-1}, \mathbb{Z})$ embeds into $H_{n-1}(B^n, \mathbb{Z})$; but this is not possible, because one can show that the first group is isomorphic to $\mathbb{Z}$ whereas the second group is trivial.

The fundamental group of the n-sphere. Our next example is the n-sphere

$$\mathbb{S}^n = \left\{ x \in \mathbb{R}^{n+1} \;\middle|\; x_1^2 + \cdots + x_{n+1}^2 = 1 \right\}$$

in $\mathbb{R}^{n+1}$. For $n \geq 2$, it looks like every closed loop in $\mathbb{S}^n$ can be contracted to a point, and indeed, we have the following theorem.

Theorem 18.7. *For $n \geq 2$, the n-sphere is simply connected.*

For the proof, we will exploit the fact that $\mathbb{S}^n$ can be covered by two open sets that are homeomorphic to $\mathbb{R}^n$. This is of course true for every n; what makes the case $n = 1$ special is that the intersection of these two open sets has two connected components. For $n \geq 2$, the intersection is path connected, and so Theorem 18.7 is a consequence of the following more general result.

Theorem 18.8. *Suppose that the topological space X is the union of two open sets U and V whose intersection $U \cap V$ is path connected. If $x_0 \in U \cap V$, then the image of $\pi_1(U, x_0)$ and $\pi_1(V, x_0)$ together generate $\pi_1(X, x_0)$ as a group.*

Recall that a nonempty subset $S \subseteq G$ *generates* the group G if every element $g \in G$ can be written (not necessarily in a unique way) as a product of elements of S and their inverses:

$$g = s_1 \cdots s_n,$$

where each $s_j \in S \cup S^{-1}$.

Proof. What we have to show is that every element $\alpha \in \pi_1(X, x_0)$ can be written (not necessarily uniquely) in the form

$$\alpha = \alpha_1 * \cdots * \alpha_n,$$

where the image of each path $\alpha_k \in \pi_1(X, x_0)$ is contained entirely inside U or entirely inside V. This is actually not very difficult.

The two open sets $\alpha^{-1}(U)$ and $\alpha^{-1}(V)$ cover I. According to Lemma 17.10 from last time, there is a subdivision

$$0 = s_0 < s_1 < \cdots < s_n = 1$$

of the interval I, in such a way that each little interval $[s_{i-1}, s_i]$ lies entirely inside $\alpha^{-1}(U)$ or entirely inside $\alpha^{-1}(V)$. Moreover, we can arrange that each of the points s_i actually belongs to $\alpha^{-1}(U \cap V)$. Indeed, if say $s_i \notin \alpha^{-1}(V)$, both intervals $[s_{i-1}, s_i]$ and $[s_i, s_{i+1}]$ are forced to lie entirely inside $\alpha^{-1}(U)$. But then the same is true for their union $[s_{i-1}, s_{i+1}]$, and so we can get rid of the point s_i. After thus removing finitely many points of our original subdivision, we end up with a subdivision where $s_i \in \alpha^{-1}(U) \cap \alpha^{-1}(V)$ for every $i = 0, \ldots, n$.

Choosing a subdivision of I with the above properties, we now define

$$\beta_i \colon I \to X, \quad \beta_i(t) = \alpha\big(s_{i-1} + t(s_i - s_{i-1})\big).$$

Each β_i is a path in X that starts and ends at a point of $U \cap V$ and whose image lies entirely inside U or entirely inside V; moreover,

$$\alpha \sim_p \beta_1 * \cdots * \beta_n.$$

This is almost what we want, except that the β_i are not elements of $\pi_1(X, x_0)$ because they do not start and end at the point x_0. But this we can achieve with the following trick. For each point $x_i = \alpha(s_i)$, choose a path

$$\gamma_i \colon I \to U \cap V$$

with $\gamma_i(0) = x_0$ and $\gamma_i(1) = x_i$; such a path exists because $U \cap V$ is path connected. For γ_0 and γ_n, we can choose the constant path e_{x_0}, which is okay because $x_n = x_0$. The path β_i starts at the point x_{i-1} and ends at the point x_i; therefore

$$\alpha_i = \gamma_{i-1} * \beta_i * \bar{\gamma}_i$$

is a loop based at the point x_0. The image of each α_i is still contained entirely inside U or entirely inside V; since one can easily show that

$$\alpha \sim_p \alpha_1 * \cdots * \alpha_n,$$

the theorem is proved. □

Theorem 18.8 is a special case of the *Seifert-van Kampen theorem*, which computes the fundamental group of X from that of U, V, and $U \cap V$. We will talk about the more general result next time. For the remainder of today's class, let us look at some more examples of fundamental groups of surfaces. Recall that a surface is the same thing as a (compact and connected) two-dimensional manifold: a second countable Hausdorff space that is locally homeomorphic to $\mathbb{R}^2$.

Example 18.9. Everyone's favorite example of a surface is the *torus* $T = \mathbb{S}^1 \times \mathbb{S}^1$. Its fundamental group can be computed with the help of one last week's homework problems: if we use the point $t_0 = (b_0, b_0)$ as a base point,

$$\pi_1(T, t_0) \simeq \pi_1(\mathbb{S}^1, b_0) \times \pi_1(\mathbb{S}^1, b_0) \simeq \mathbb{Z} \times \mathbb{Z}.$$

Note that this group is *abelian*: for any two elements g, h, one has $gh = hg$. On a picture of the torus, the two generators of the fundamental group are represented by loops going around the torus and around the hole in the center.

Example 18.10. Another example is the *torus with two holes* $T_2 = T \# T$. It is defined as the connected sum of two copies of T: from each of the two tori, we remove a small open disk, and then we glue the two spaces together along the boundary circles. After smoothing out the edges, we obtain a new surface T_2.

Since T_2 has two holes, it seems pretty obvious that T_2 is not homeomorphic to T. How can we prove this? One thing that distinguishes the two surfaces from each other is the fundamental group: for T, it is abelian, whereas for T_2, it is not abelian. This can be seen by the following geometric argument. T_2 contains a subspace X homeomorphic to the figure 8; the two circles in X go around the two holes in T_2. By drawing a picture, you can convince yourself that T_2 retracts onto X. This means that the fundamental group of X embeds into that of T_2, and so it is enough to show that $\pi_1(X, x_0)$ is not abelian (where x_0 denotes the center of the figure 8). In fact, if α and β are the two obvious loops in X, then $\alpha * \beta$ and $\beta * \alpha$ are not path homotopic: the reason is that there is a covering space of X where the liftings of $\alpha * \beta$ and $\beta * \alpha$ have different endpoints. One of the exercises on this week's homework will tell you more.

Lecture 19: November 3

Another example. Let us look at one more example of a fundamental group, namely that of the projective plane P^2. The projective plane is obtained from the 2-sphere $\mathbb{S}^2$ by identifying antipodal points; more precisely,

$$P^2 = \mathbb{S}^2/\sim,$$

where $x \sim -x$ for every $x \in \mathbb{S}^2$. The topology on P^2 is the quotient topology, of course. We can obtain P^2 by taking a closed disk and identifying antipodal points on the boundary circle. A short cut-and-paste argument shows that one can also get P^2 by gluing together a Möbius band (whose boundary is a circle) and a disk (whose boundary is also a circle).

Theorem 19.1. *P^2 is a compact surface, and the quotient map $q\colon \mathbb{S}^2 \to P^2$ is a two-sheeted covering space.*

Proof. Let us denote by $a\colon \mathbb{S}^2 \to \mathbb{S}^2$ the antipodal map $a(x) = -x$; it is a homeomorphism of $\mathbb{S}^2$ with itself. The equivalence class $[x]$ of a point $x \in \mathbb{S}^2$ is of course exactly the two-point set $\{x, a(x)\}$. Because of how the quotient topology is defined, q is continuous; let us show that it is also open. If $U \subseteq S^2$ is open, then

$$q^{-1}\big(q(U)\big) = U \cup a(U),$$

which is open in $\mathbb{S}^2$; but this means exactly that $q(U)$ is open in P^2. It follows that P^2 is a Hausdorff space: for any two points $[x] \neq [y]$, we can choose small neighborhoods $x \in U$ and $y \in V$ such that the four open sets $U, V, a(U), a(V)$ are disjoint; then $q(U)$ and $q(V)$ are open sets separating $[x]$ and $[y]$.

Now we can show that $\mathbb{S}^2$ is a two-sheeted covering space. Given any point $x \in \mathbb{S}^2$, we can choose a small neighborhood U such that $U \cap a(U) = \emptyset$; then

$$q\big|_U\colon U \to q(U)$$

is bijective, continuous, open, and therefore a homeomorphism; the same is true for the restriction of q to $a(U)$. This means that the neighborhood $q(U)$ of the point $[x] \in P^2$ is evenly covered by two open sets; thus q is a two-sheeted covering space.

This argument also proves that P^2 is locally homeomorphic to $\mathbb{R}^2$, because $q(U) \simeq U$ is homeomorphic to an open set in $\mathbb{R}^2$. We already know that P^2 is Hausdorff; since $\mathbb{S}^2$ is compact and second countable, it follows that P^2 is also compact and second countable, and therefore a compact two-dimensional manifold. □

Since $\mathbb{S}^2$ is simply connected, Theorem 18.2 shows that the fundamental group of P^2 is isomorphic to the cyclic group of order 2.

Deformation retracts. One can sometimes compute the fundamental group of a space X by deforming the space continuously into another (hopefully simpler) space A. Let me describe one particular case of this idea in detail.

Definition 19.2. Let X be a topological space. A subspace $A \subseteq X$ is *deformation retract* of X is there exists a homotopy

$$H\colon X \times I \to X$$

with the following properties: for every $x \in X$, one has $H(x, 0) = x$ and $H(x, 1) \in A$; for every $a \in A$ and every $t \in I$, one has $H(a, t) = a$.

For every point $x \in X$, we get a path $I \to X$, $t \mapsto H(x,t)$, that starts at the point x and ends at a point of A. If we define $r\colon X \to A$ by $r(x) = H(x,1)$, then r is clearly a retraction of X onto A. What H does is to give a homotopy between the identity function id and the retraction r – more precisely, between id and the composition $j \circ r$, where $j\colon A \to X$ is the inclusion. Note the subspace A has to stay fixed during the homotopy. A homotopy with these properties is called a *deformation retraction* of X onto A.

Example 19.3. The n-sphere $\mathbb{S}^n$ is a deformation retract of $\mathbb{R}^{n+1} \setminus \{0\}$. The reason is that we can move any point $x \neq 0$ in a straight line to a point on the $\mathbb{S}^n$;

$$H\colon \big(\mathbb{R}^{n+1} \setminus \{0\}\big) \times [0,1] \to \mathbb{R}^{n+1} \setminus \{0\}, \quad H(x,t) = (1-t)x + t\frac{x}{\|x\|}$$

is a formula for this deformation retraction.

Example 19.4. If we remove one point from the torus $\mathbb{T} = \mathbb{S}^1 \times \mathbb{S}^1$, the resulting space deformation retracts onto the union of two circles; this can be most easily seen by thinking of the torus as the square with opposite sides identified.

Deformation retractions do not change the fundamental group.

Theorem 19.5. *Let X be a topological space, $A \subseteq X$ a subspace, and $a_0 \in A$ a base point. If A is a deformation retract of X, then*

$$j_*\colon \pi_1(A, a_0) \to \pi_1(X, a_0)$$

is an isomorphism of groups.

Proof. Let $H\colon X \times I \to X$ be a deformation retraction. We already know that the function $r\colon X \to A$ given by $r(x) = H(x,1)$ is a retraction; according to Lemma 18.4, the group homomorphism

$$j_*\colon \pi_1(A, a_0) \to \pi_1(X, x_0)$$

is thus injective. It remains to prove that j_* is also surjective – concretely, this means that every loop in X (based at the point a_0) can be deformed continuously until it lies inside A. In fact, if $\alpha\colon I \to X$ is a path with $\alpha(0) = \alpha(1) = a_0$, we can easily show that $\alpha \sim_p r \circ \alpha$. Indeed, since H is a homotopy between id and $j \circ r$, the composition

$$F\colon I \times I \to X, \quad F(s,t) = H\big(\alpha(s), t\big)$$

is a homotopy between α and $j \circ r \circ \alpha$; it is even a path homotopy because $F(0,t) = F(1,t) = H(a_0,t) = a_0$. Thus

$$[\alpha] = [j \circ r \circ \alpha] = j_*[r \circ \alpha],$$

which shows that j_* is surjective. □

This means for example that $\mathbb{R}^n \setminus \{0\}$ is simply connected for $n \geq 3$ (because it deformation retracts onto $\mathbb{S}^{n-1}$). The same argument with homotopies that we just gave also proves the following useful lemma.

Lemma 19.6. *Let $f, g\colon (X, x_0) \to (Y, y_0)$ be two continuous functions. If f and g are homotopic by a homotopy $H\colon X \times I \to Y$ with $H(x_0, t) = y_0$, then*

$$f_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0) \quad \textit{and} \quad g_*\colon \pi_1(X, x_0) \to \pi_1(Y, y_0)$$

are equal.

Free products of groups. One of the most useful tools for computing fundamental groups is the Seifert-van Kampen theorem: it gives us a complete description of the fundamental group of X when $X = U \cup V$ is covered by two open sets such that U, V, and $U \cap V$ are all path connected. We have already seen (in Theorem 18.8) that $\pi_1(X, x_0)$ is generated as a group by the images of $\pi_1(U, x_0)$ and $\pi_1(V, x_0)$, which means that every element of $\pi_1(X, x_0)$ can be written as a product of elements in $\pi_1(U, x_0)$ and $\pi_1(V, x_0)$; the Seifert-van Kampen theorem will tell us exactly what the ambiguity is.

The statement of the theorem involves the *free product* of two groups, and so we have to discuss that idea first.

Example 19.7. Let X be the figure-eight space, meaning two circles touching in one point x_0. The fundamental group of each circle is $\mathbb{Z}$, and is generated by a loop going around the circle once. Let us denote by α and β the generators corresponding to the two circles. Any product of powers of α and β represents an element of $\pi_1(X, x_0)$: for instance,

$$\alpha^2\beta\alpha\beta^{-3}$$

is the loop that goes twice around the first circle, once around the second circle, then once around the first circle, and then three times around the second circle in the opposite direction. We can always write the product in the above form, where powers of α alternate with powers of β. We can therefore describe elements of $\pi_1(X, x_0)$ by such *words* in α and β. In fact, the set of all such words forms a group, usually denoted by $\mathbb{Z} * \mathbb{Z}$ and called the *free group on two generators*. Multiplication in this group is defined by juxtaposing two words and simplifying the result:

$$(\alpha^2\beta\alpha\beta^{-3})(\beta^2\alpha) = \alpha^2\beta\alpha\beta^{-1}\alpha.$$

The unit element is the empty word $\emptyset$; the inverse of a word is obtained simply by reverting the order of the letters and flipping the signs of the exponents, because

$$(\alpha^2\beta\alpha\beta^{-3})(\beta^3\alpha^1\beta^{-1}\alpha^{-2}) = \emptyset.$$

It would be nice if the fundamental group of the figure-eight space was exactly this group; the Seifert-van Kampen theorem will say that this is the case.

Now let me introduce the free product in general. Suppose that $(G_i)_{i\in I}$ is a family of groups, indexed by a (maybe infinite) set I. From the homework, we know that the Cartesian product

$$\prod_{i\in I} G_i$$

is again a group; when I is infinite, we can also consider the much smaller subgroup

$$\bigoplus_{i\in I} G_i \subseteq \prod_{i\in I} G_i,$$

consisting of all elements $(g_i)_{i\in I}$ such that $g_i = e$ for all but finitely many $i \in I$. Both groups contain all the original groups G_i as subgroups (by looking at those elements that are equal to e except in the i-th coordinate). Note that the elements of G_i and G_j commute with each other when $i \neq j$; sometimes, we may not want that. The free product

$$\ast_{i\in I} G_i$$

is also a group that contains all the G_i as subgroups, but in such a way that elements of G_i and G_j do not commute. Here is the construction.

As a set, the free product consists of all *words* of finite length

$$g_1 g_2 \cdots g_n,$$

where each g_k belongs to some $G_{i(k)}$. We also allow the empty word $\emptyset$ (for $n = 0$). We insist that adjacent letters should belong to different groups, meaning that $i(k) \neq i(k+1)$, and that none of the g_k is equal to the unit element e. Words with this property are called *reduced*: the idea is that we can take an arbitrary word and, by replacing $g_k g_{k+1}$ by their product in $G_{i(k)}$ if $i(k) = i(k+1)$ and by removing possible appearances of e, turn it into a reduced word in finitely many steps. Let W be the set of all reduced words.

Note. We could introduce an equivalence relation that does the same thing, but it is somehow easier to work with actual words instead of equivalence classes.

Now we define the product of two words

$$(g_1 \cdots g_n)(h_1 \cdots h_m)$$

by concatenating them and reducing the result: if g_n and h_1 happen to belong to the same group G_i, multiply them together; if the answer is e, remove it and continue with g_{n-1} and h_2. Note that there can be a lot of cancellation: for example,

$$(g_1 \cdots g_n)(g_n^{-1} \cdots g_1^{-1}) = \emptyset$$

produces the empty word. This procedure defines a binary operation on the set W; evidently, the empty word acts as the unit element, and every word $g_1 \cdots g_n$ has a unique inverse $g_n^{-1} \cdots g_1^{-1}$. To show that W is a group, it remains to prove associativity. Because of the reduction involved in computing the product of two words, trying to prove this directly is very painful; fortunately, there is a nice trick that gives us associativity almost for free.

Lemma 19.8. *The product on W is associative.*

Proof. The trick is to embed W into a space of functions, where associativity becomes obvious. For every $g \in G_i$, left multiplication by g defines a function

$$L_g \colon W \to W.$$

Concretely, $L_g(g_1 \cdots g_n)$ is reducing the word $g g_1 \cdots g_n$; convince yourself that there are five possibilities for what can happen (depending on whether $n = 0$ or $n \geq 1$, on whether $g = e$ or $g \neq e$, and so on). If $g, h \in G_i$, one can easily show that

$$L_g \circ L_h = L_{gh};$$

the point is that $(gh)(g_1 \cdots g_n)$ and $g(h g_1 \cdots g_n)$ reduce to the same word, because multiplication in the group G_i is associative. Since L_e is the identity on W, it follows that L_g is bijective with inverse $L_{g^{-1}}$, and therefore a permutation of W. Note that since composition of functions is associative, the set $P(W)$ of all permutations of W forms a group under composition.

Now we show that W can be embedded into the group $P(W)$. Define a function

$$L \colon W \to P(W), \quad g_1 \cdots g_n \mapsto L_{g_1} \circ \cdots \circ L_{g_n};$$

of course, we use the convention that $L(\emptyset) = \mathrm{id}$. The function L is injective, because if we evaluate $L(g_1 \cdots g_n)$ on the empty word, we get

$$\big(L_{g_1} \circ \cdots \circ L_{g_n}\big)(\emptyset) = g_1 \cdots g_n.$$

Moreover, L takes the product of two words to the composition of the corresponding functions. To see that this is true, recall how the product was defined: we concatenate

$$(g_1 \cdots g_n)(h_1 \cdots h_m)$$

and reduce the result. During the reduction, we either multiply two elements of the same group, or we eliminate an occurrence of e; but since $L_g \circ L_h = L_{gh}$ and $L_e = \mathrm{id}$, this does not change the value of

$$L_{g_1} \circ \cdots \circ L_{g_n} \circ L_{h_1} \circ \cdots \circ L_{h_m}.$$

Since the composition of functions in $P(W)$ is associative, we conclude that the product in W must also be associative. □

In the case of finitely many groups $G_1, \ldots, G_n$, we usually denote the free product by the symbol

$$G_1 * \cdots * G_n.$$

Example 19.9. The free product of a certain number of copies of $\mathbb{Z}$ is called the *free group* on so many generators; for instance, $\mathbb{Z} * \cdots * \mathbb{Z}$ (with n copies) is the free group on n generators. Of course, we have to use different symbols for the generators, such as $a, b, c, \ldots$

Example 19.10. Another interesting example is the free product $\mathbb{Z}_2 * \mathbb{Z}_2$. If we write a and b for the two generators, then $a^2 = b^2 = e$; consequently, the elements of the free product are

$$\emptyset, a, b, ab, ba, aba, bab, abab, baba, \ldots$$

Even though both factors have only two elements, the free product contains a subgroup isomorphic to $\mathbb{Z}$, namely the image of the homomorphism

$$\mathbb{Z} \to \mathbb{Z}_2 * \mathbb{Z}_2, \quad m \mapsto (ab)^m.$$

In fact, the entire group is generated by the two elements ab and a, subject to the two relations $a^2 = e$ and $a(ab)a = ba = (ab)^{-1}$. If we also had $(ab)^m = e$, this would be exactly the description of the dihedral group; for that reason, $\mathbb{Z}_2 * \mathbb{Z}_2$ is also called the infinite dihedral group.

Lecture 20: November 8

Let X be a topological space, and suppose that $X = U \cup V$ is the union of two open sets such that U, V, and $U \cap V$ are path connected. In this situation, one can describe the fundamental group of X in terms of the fundamental groups of the three open sets; this is the content of the Seifert-van Kampen theorem, the topic of today's class. Fix a base point $x_0 \in U \cap V$; to simplify the notation, we put

$$\pi_1(X) = \pi_1(X, x_0), \quad \pi_1(U) = \pi_1(U, x_0), \quad \text{etc.}$$

Recall from Theorem 18.8 that $\pi_1(X)$ is generated, as a group, by the images of the group homomorphisms $\pi_1(U) \to \pi_1(X)$ and $\pi_1(V) \to \pi_1(X)$. Concretely, this means that every element $[\alpha] \in \pi_1(X)$ can be written as a product

$$[\alpha] = [\alpha_1] \cdots [\alpha_n],$$

where each α_j is a loop based at the point x_0 and contained entirely inside U or entirely inside V. We can restate this fact using the free product construction from last time.

Lemma 20.1. *We have a surjective homomorphism* $\Phi\colon \pi_1(U) * \pi_1(V) \to \pi_1(X)$.

Proof. Let us start by defining Φ. The unit element in the free product is the empty word, so we define $\Phi(\emptyset) = e$, the homotopy class of the constant loop at x_0. The other elements of the free product are reduced words of the form $g_1 \cdots g_\ell$, where each g_j belongs either to $\pi_1(U)$ or to $\pi_1(V)$, and is therefore represented by a loop α_j in U or V. We then define $\Phi(g_1 \cdots g_\ell) = [\alpha_1 * \cdots * \alpha_\ell]$. It is an easy exercise to show that Φ is a group homomorphism. The result in Theorem 18.8 says that Φ is surjective. □

It remains to figure out by how much the groups $\pi_1(U) * \pi_1(V)$ and $\pi_1(X)$ differ from each other. This requires another small interlude on group theory.

Normal subgroups. Let $\varphi\colon G \to H$ be a homomorphism between two groups G and H. On the one hand, we can consider the image

$$\operatorname{im} \varphi = \{\, \phi(g) \mid g \in G \,\} \subseteq H.$$

It is a subgroup of H, in the following sense:

(1) It contains the unit element of H: indeed, $e = \varphi(e) \in \operatorname{im} \varphi$.
(2) For every $h \in \operatorname{im} \varphi$, we also have $h^{-1} \in \operatorname{im} \varphi$: indeed, if $h = \varphi(g)$, then $h^{-1} = \varphi(g^{-1})$.
(3) If $h_1, h_2 \in \operatorname{im} \varphi$, then also $h_1 h_2 \in \operatorname{im} \varphi$: indeed, if $h_1 = \varphi(g_1)$ and $h_2 = \varphi(g_2)$, then $h_1 h_2 = \varphi(g_1 g_2)$.

On the other hand, we can consider the kernel

$$\ker \varphi = \{\, g \in G \mid \varphi(g) = e \,\} \subseteq G.$$

For the same reasons as above, it is a subgroup of G; in fact, it is a normal subgroup.

Definition 20.2. A subgroup $K \subseteq G$ is called *normal* if $gkg^{-1} \in K$ for every $g \in G$ and every $k \in K$. In symbols, $K \trianglelefteq G$.

To see that $\ker \varphi$ is normal, note that if $\varphi(k) = e$, then we have

$$\varphi(gkg^{-1}) = \varphi(g)\varphi(k)\varphi(g)^{-1} = \varphi(g)\varphi(g)^{-1} = e,$$

which means that $gkg^{-1} \in \ker \varphi$. Now let me cite, without proof, two basic results from group theory. The first says that one can take the quotient of a grouy by a normal subgroup.

Proposition 20.3. *Given $K \trianglelefteq G$, let G/K be the set of all left cosets*

$$gK = \{\, gk \mid k \in K \,\}.$$

Then we have $gK \cdot hK = ghK$, and this operation makes G/K into a group with unit element eK and inverse $(gK)^{-1} = g^{-1}K$.

The second result is the so-called *first isomorphism theorem*, which relates the kernel and the image of a group homomorphism.

Proposition 20.4. *Let $\varphi\colon G \to H$ be a homomorphism with $K = \ker \varphi$. Then*

$$G/K \to \operatorname{im} \varphi, \quad gK \mapsto \varphi(g)$$

is a well-defined isomorphism of groups.

Analysis of the kernel. Recall that we have a surjective homomorphism $\Phi\colon \pi_1(U) * \pi_1(V) \to \pi_1(X)$. By Proposition 20.4,

$$\pi_1(X) \simeq \frac{\pi_1(U) * \pi_1(V)}{\ker \Phi},$$

and so the remaining task is to compute the kernel of Φ. Certain elements of $\pi_1(U) * \pi_1(V)$ are obviously contained in $\ker \Phi$: a loop α in the intersection $U \cap V$ gives rise to two elements

$$[\alpha]_U \in \pi_1(U) \quad \text{and} \quad [\alpha]_V \in \pi_1(V),$$

and because both have the same image in $\pi_1(X)$, we see that $[\alpha]_U [\alpha]_V^{-1} \in \ker \Phi$. Since $\ker \Phi$ is a normal subgroup of the free product, this proves one half of the following theorem.

Theorem 20.5. $\ker \Phi$ *is the smallest normal subgroup of $\pi_1(U) * \pi_1(V)$ containing every element of this kind.*

Let $N \trianglelefteq \pi_1(U) * \pi_1(V)$ be a normal subgroup containing every element of the form $[\alpha]_U [\alpha]_V^{-1}$. To prove Theorem 20.5, we have to show that $\ker \Phi \subseteq N$.

Suppose that $g_1 \cdots g_\ell$ is a word in $\ker \Phi$; we shall argue that $g_1 \cdots g_\ell \in N$. Each g_i is represented by a loop α_i based at x_0 and contained entirely inside one of the two open sets U or V. If we denote by

$$\alpha_1 * \cdots * \alpha_\ell \colon I \to X$$

the path obtained by transversing $\alpha_1, \ldots, \alpha_\ell$ at equal speed, i.e.,

$$(\alpha_1 * \cdots * \alpha_\ell)(s) = \alpha_j(\ell s - j + 1) \quad \text{for } s \in \left[\frac{j-1}{\ell}, \frac{j}{\ell}\right],$$

then $[\alpha_1 * \cdots * \alpha_\ell] = [\alpha_1] * \cdots * [\alpha_\ell] = e$, and so $\alpha_1 * \cdots * \alpha_\ell$ is path homotopic to the constant path at x_0. Thus there exists a path homotopy $H\colon I \times I \to X$ from $\alpha_1 * \cdots * \alpha_\ell$ to the constant path; it satisfies

$$H(s,1) = H(0,t) = H(1,t) = x_0 \quad \text{and} \quad H(s,0) = (\alpha_1 * \cdots * \alpha_\ell)(s) \tag{20.6}$$

for every $s, t \in I$. Just as in the proof of Theorem 18.8, we divide the unit square $I \times I$ into m^2 smaller squares of side length $\frac{1}{m}$; if we choose m sufficiently large, then

H maps every little square entirely into one of the two open sets (by Lemma 17.10). For the sake of convenience, we shall assume that m is a multiple of ℓ.

Now we have a total of $(m+1)^2$ points of the form

$$x_{a,b} = H\left(\frac{a}{m}, \frac{b}{m}\right) \in X.$$

For each $0 \leq a, b \leq m$, we choose a path $\gamma_{a,b}$ from the base point x_0 to the point $x_{a,b}$, in the following manner: if $x_{a,b} = x_0$, we let $\gamma_{a,b}$ be the constant path; if the point $x_{a,b}$ belongs to U, V, or $U \cap V$, we choose $\gamma_{a,b}$ in U, V, or $U \cap V$, respectively. This is possible because $x_0 \in U \cap V$ and all three open sets are path connected.

From the homotopy H, we now construct a large number of loops based at x_0. If we restrict H to the horizontal line segment

$$I \to I \times I, \quad s \mapsto \left(\frac{a+s}{m}, \frac{b}{m}\right),$$

we obtain a path from the point $x_{a,b}$ to the point $x_{a+1,b}$. By our choice of subdivision, this path stays inside U or V the whole time. We can compose it with $\bar{\gamma}_{a,b}$ and $\gamma_{a+1,b}$ to obtain a loop based at the point x_0; we shall denote it by $h_{a,b} \colon I \to X$. Because of how we choose the paths $\gamma_{a,b}$, the loop $h_{a,b}$ also stays inside U or V the whole time. Similarly, we can restrict H to the vertical line segment

$$I \to I \times I, \quad t \mapsto \left(\frac{a}{m}, \frac{b+t}{m}\right),$$

and by composing with $\bar{\gamma}_{a,b}$ and $\gamma_{a,b+1}$, obtain a loop $v_{a,b} \colon I \to X$ based at x_0 that stays inside U or V the whole time.

Lemma 20.7. *For each $0 \leq a, b \leq m-1$, the two loops*

$$h_{a,b} * v_{a+1,b} \quad \text{and} \quad v_{a,b} * h_{a,b+1}$$

are contained in the same open set U or V, and are path homotopic there.

Proof. Consider the restriction of H to the small square

$$(s,t) \mapsto \left(\frac{a+s}{m}, \frac{b+t}{m}\right).$$

By construction, the image lies inside one of the two open sets, say U. The four points $x_{a,b}$, $x_{a+1,b}$, $x_{a,b+1}$, and $x_{a+1,b+1}$ therefore all lie in U, and so the same is true for the four paths $\gamma_{a,b}$, $\gamma_{a+1,b}$, $\gamma_{a,b+1}$, and $\gamma_{a+1,b+1}$. This clearly implies that $h_{a,b}$, $v_{a+1,b}$, $v_{a,b}$, and $h_{a,b+1}$ are all contained in U as well. From the restriction of H, one can easily build a path homotopy in U between $h_{a,b} * v_{a+1,b}$ and $v_{a,b} * h_{a,b+1}$, and so the lemma is proved. □

Since our homotopy H satisfies (20.6), we have

$$[h_{a,m}] = [v_{0,b}] = [v_{m,b}] = e$$

for every $0 \leq a, b \leq m-1$. As an element of $\pi_1(X)$, the product

$$w_0 = [v_{0,0}] * \cdots * [v_{0,m-1}] * [h_{0,m}] * \cdots * [h_{m-1,m}]$$

is thus equal to the unit element e. By applying Lemma 20.7 for each of the m^2 small squares, we can gradually rewrite this product until it becomes equal to

$$w_{m^2} = [h_{0,0}] * \cdots * [h_{m-1,0}] * [v_{m,0}] * \cdots * [v_{m,m-1}].$$

In the first step, we use the relation $[h_{0,m-1}] * [v_{1,m-1}] = [v_{0,m-1}] * [h_{0,m}]$ to show that w_0 is equal to the product

$$w_1 = [v_{0,0}] * \cdots * [v_{0,m-2}] * [h_{0,m-1}] * [v_{1,m-1}] * [h_{1,m}] * \cdots * [h_{m-1,m}].$$

In the second step, we use the relation $[v_{1,m-1}] * [h_{1,m}] = [h_{1,m-1}] * [v_{2,m-1}]$ to show that w_1 is equal to the product

$$w_2 = [v_{0,0}] * \cdots * [v_{0,m-2}] * [h_{0,m-1}] * [h_{1,m-1}] * [v_{2,m-1}] * [h_{2,m}] * \cdots * [h_{m-1,m}],$$

and so on. After m^2 steps, we finally arrive at the expression for w_{m^2}.

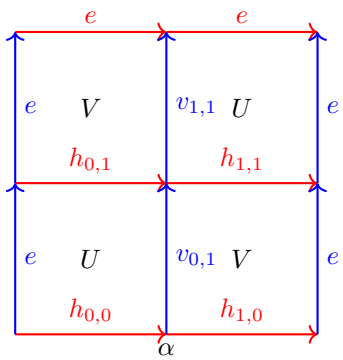
Now each loop $h_{a,b}$ and $v_{a,b}$ stays inside one of the two open sets, and so we can consider the words $w_0, w_1, \ldots, w_{m^2}$ (after reduction) as elements of the free product $\pi_1(U) * \pi_1(V)$. In going from w_{k-1} to w_k, we performed the following operations:

(1) If $h_{a,b}$ (or $v_{a,b}$) lies inside $U \cap V$, we may have viewed it as an element of one of the groups $\pi_1(U)$ or $\pi_1(V)$ in the word w_{k-1}, and are viewing it as an element of the other group in the word w_k.
(2) We use the relation from Lemma 20.7, which holds in the group $\pi_1(U)$ or $\pi_1(V)$, depending on where H takes the k-th small square.

What this means is that the composition $w_k^{-1} w_{k-1}$ is the conjugate of an element of the form $[h_{a,b}]_U [h_{a,b}]_V^{-1}$ or $[v_{a,b}]_U [v_{a,b}]_V^{-1}$, and therefore contained in the normal subgroup N. Since w_0 reduces to the empty word, it follows that the reduction of w_{m^2} belongs to N. But because the homotopy H satisfies (20.6), the reduction of w_{m^2} is equal to the word $g_1 \cdots g_\ell$, and so we conclude that $g_1 \cdots g_\ell \in N$. This finishes the proof of Theorem 20.5.

Example 20.8. The following special case may help you understand the proof. Suppose that we have a single element $g \in \pi_1(U)$ such that $\psi(g) = e$. It is represented by a loop $\alpha \colon I \to U$. Our goal is to prove that the word $[\alpha]_U$ belongs to the normal subgroup N; we use subscripts to indicate whether a given loop is to be viewed as an element of $\pi_1(U)$ or $\pi_1(V)$.

Let $H \colon I \times I \to X$ be a path homotopy between α and the constant path $e = e_{x_0}$, and suppose that $m = 2$, meaning that if we subdivide $I \times I$ into four squares, the image of each small square is contained entirely in U or V. The following schematic picture of the homotopy H shows the loops that we constructed during the proof:



To be specific, let us suppose that each of the four squares maps into the open set indicated by the label. Since α is contained in the open set U, we have

$$[\alpha]_U = [h_{0,0}]_U [h_{1,0}]_U$$

as elements in the group $\pi_1(U)$. The square in the bottom-right corner maps into V, and by Lemma 20.7, this gives us the relation

$$[h_{1,0}]_V = [v_{0,1}]_V [h_{1,1}]_V.$$

Now $h_{1,0}$ is a loop in $U \cap V$, and so the word $[h_{1,0}]_U [h_{1,0}]_V^{-1}$ belongs to the normal subgroup N. In the free product $\pi_1(U) * \pi_1(V)$, we then get

$$\begin{aligned}[\alpha]_U &= [h_{0,0}]_U [h_{1,0}]_U \cdot [h_{1,0}]_V^{-1} [h_{1,0}]_V \\ &= [h_{0,0}]_U [h_{1,0}]_U [h_{1,0}]_V^{-1} [h_{0,0}]_U^{-1} \cdot [h_{0,0}]_U [v_{0,1}]_V [h_{1,1}]_V.\end{aligned}$$

Since the first factor is an element of the normal subgroup N, it will be enough to show that $[h_{0,0}]_U [v_{0,1}]_V [h_{1,1}]_V \in N$.

The square in the bottom-left corner maps into U, and therefore

$$[h_{0,0}]_U [v_{0,1}]_U = [h_{0,1}]_U.$$

Since $v_{0,1}$ is a loop in $U \cap V$, the word $[v_{0,1}]_U^{-1} [v_{0,1}]_V$ belongs to N. In the free product, we can then rewrite the word from above as

$$\begin{aligned}[h_{0,0}]_U [v_{0,1}]_V [h_{1,1}]_V &= [h_{0,1}]_U [v_{0,1}]_U^{-1} [v_{0,1}]_V [h_{1,1}]_V \\ &= [h_{0,1}]_U [v_{0,1}]_U^{-1} [v_{0,1}]_V [h_{0,1}]_U^{-1} \cdot [h_{0,1}]_U [h_{1,1}]_V.\end{aligned}$$

This further reduces the problem to showing that $[h_{0,1}]_U [h_{1,1}]_V \in N$.

The square on the top-right gives $[h_{1,1}]_U = [v_{1,1}]_U$; we can use this to rewrite the word from the previous step as

$$\begin{aligned}[h_{0,1}]_U [h_{1,1}]_V &= [h_{0,1}]_U [h_{1,1}]_V [h_{1,1}]_U^{-1} [h_{1,1}]_U \\ &= [h_{0,1}]_U [h_{1,1}]_V [h_{1,1}]_U^{-1} [h_{0,1}]_U^{-1} \cdot [h_{0,1}]_U [v_{1,1}]_U\end{aligned}$$

The first factor belongs to N, and so it suffices to show that $[h_{0,1}]_U [v_{1,1}]_U \in N$.

Finally, the square on the top-left gives $[h_{0,1}]_V = [v_{1,1}]_V^{-1}$; once again, we use this to rewrite the word from the previous step as

$$[h_{0,1}]_U [v_{1,1}]_U = [h_{0,1}]_U [h_{0,1}]_V^{-1} \cdot [v_{1,1}]_V^{-1} [v_{1,1}]_U,$$

and now both factors belong to N. This proves that our original element $g \in \ker \psi$ lies in the normal subgroup N; in fact, we have shown that

$$\begin{aligned}g =& [h_{0,0}]_U [h_{1,0}]_U [h_{1,0}]_V^{-1} [h_{0,0}]_U^{-1} \cdot [h_{0,1}]_U [v_{0,1}]_U^{-1} [v_{0,1}]_V [h_{0,1}]_U^{-1} \\ &\cdot [h_{0,1}]_U [h_{1,1}]_V [h_{1,1}]_U^{-1} [h_{0,1}]_U^{-1} \cdot [h_{0,1}]_U [h_{0,1}]_V^{-1} \cdot [v_{1,1}]_V^{-1} [v_{1,1}]_U\end{aligned}$$

is the product of five elements in N.

The Seifert-van Kampen theorem. Let me now give the Seifert-van Kampen theorem in its final form. Suppose that X is the union of two path connected open sets U and V whose intersection $U \cap V$ is also path connected. Choose a base point $x_0 \in U \cap V$, and let $i \colon U \cap V \to U$ and $j \colon U \cap V \to V$ denote the two inclusions.

Theorem 20.9. *We have an isomorphism of groups*

$$\pi_1(X, x_0) \simeq \big(\pi_1(U, x_0) * \pi_1(V, x_0)\big) / N,$$

where N is the smallest normal subgroup of the free product containing all elements of the form $(i_* g)(j_* g)^{-1}$, for $g \in \pi_1(U \cap V, x_0)$.

There are two very useful special cases.

Corollary 20.10. *If the intersection $U \cap V$ is simply connected, then $\pi_1(X, x_0) \simeq \pi_1(U, x_0) * \pi_1(V, x_0)$.*

Example 20.11. The corollary shows that the fundamental group of the figure 8 is isomorphic to $\mathbb{Z} * \mathbb{Z}$. More precisely, let X be the union of two circles meeting in a point x_0. Note that we cannot take U and V to be simply the two circles, because U and V are supposed to be *open* sets. We can get around this problem with the following trick. Let U be the open set consisting of the first circle and a small open neighborhood of the point x_0 in the second circle; similarly, let V be the open set consisting of the second circle and a small open neighborhood of the point x_0 in the first circle. Then U deformation retracts onto the first circle, V deformation retracts onto the second circle, and $U \cap V$ deformation retracts on the one-point set $\{x_0\}$, and so we get

$$\pi_1(X, x_0) \cong \pi_1(U, x_0) * \pi_1(V, x_0) \cong \mathbb{Z} * \mathbb{Z}.$$

Corollary 20.12. *If V is simply connected, then $\pi_1(X, x_0) \simeq \pi_1(U, x_0)/N$, where N is the smallest normal subgroup of $\pi_1(U, x_0)$ containing the image of $\pi_1(U \cap V, x_0)$.*

Example 20.13. We can use the corollary to compute the fundamental group of the torus in a different way. The torus T can be covered by two open sets U and V, such that U deformation retracts onto the figure 8, and V deformation retracts to a point; the intersection $U \cap V$ deformation retracts to a circle. If we denote by a and b the two generators of $\pi_1(U) \simeq \mathbb{Z} * \mathbb{Z}$, the image of $\pi_1(U \cap V)$ consists of all powers of the element $aba^{-1}b^{-1}$. In the quotient group, the two cosets aN and bN commute, and so $\pi_1(T)$ is isomorphic to $\mathbb{Z} \times \mathbb{Z}$.

Lecture 21: November 10

Surfaces and their fundamental groups. Our topic today is surfaces. We are going to compute the fundamental groups of all surfaces, and sketch the proof of the classification theorem. To begin with, a *surface* will mean a (usually compact and connected) 2-dimensional topological manifold. Surfaces can be either orientable (such as the sphere or the torus) or non-orientable (such as the Klein bottle or the Möbius band). Let us first define this notion more precisely. One way to see that the Möbius band is non-orientable is by looking at the circle in the center: if we draw an arrow that is pointing up, and then we slide it along the circle, it will come back pointing down. This circle is an example of a one-sided curve: if we take a small neighborhood of the circle, and then remove the circle from it, the complement stays connected. Compare this with what happens to a small neighborhood of a circle going around the torus, for example.

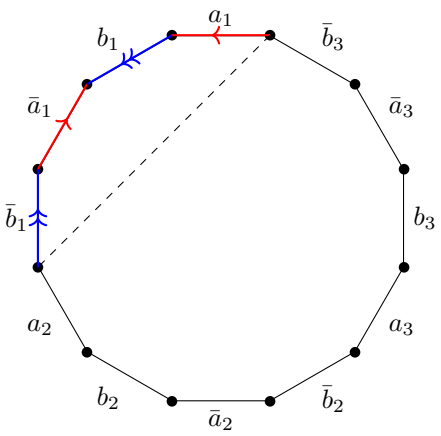
Definition 21.1. Let S be a surface, and $C \subseteq S$ a curve homeomorphic to $\mathbb{S}^1$. We say that C is *one-sided* if for every path connected open set U containing C, the complement $U \setminus C$ is path connected. Otherwise, we say that C is *two-sided*.

We can use this to define orientability.

Definition 21.2. A surface S is called *orientable* if every curve on S is two-sided. Otherwise, S is called *non-orientable*.

So any surface that contains a one-sided curve is non-orientable; a small neighborhood of such a curve will then be homeomorphic to a Möbius band. With this definition, it is not trivial to check that $\mathbb{S}^2$ or the torus are orientable, but we are going to ignore this technical point.

Example 21.3. We can construct orientable surfaces by taking the connected sum of $g \geq 1$ copies of the torus; for $g = 0$, we just take the sphere. An equivalent description is to start from the sphere $\mathbb{S}^2$ and to attach g handles: for each handle, we remove two small open disks from the surface, and then attach a cylindrical tube that joins the two boundary circles (in a way that keeps the surface orientable). A third description (for $g \geq 1$) is to take a regular $4g$-gon and to glue the edges together in pairs, as in the following picture (for $g = 3$):



Here we glue together edges with the same label (such as a_1 and $\bar{a}_1$); the bar means that we use the opposite orientation, as indicated by the arrows. All $4g$ corners

of the polygon therefore become a single point in the quotient space, and the $4g$ edges give us $2g$ different loops based at this point. It is easy to see that adding the region above the dashed line is the same thing as taking the connected sum with a copy of the torus; the quotient space is therefore another model for the connected sum of g copies of the torus. We denote the resulting surface by the symbol Σ_g.

The Seifert-van Kampen theorem makes it easy to compute the fundamental group of the surface Σ_g. We view Σ_g as the quotient of a regular polygon with $4g$ edges. Let $x_0 \in \Sigma_g$ be the image of $4g$ vertices (which all become one point in Σ_g). Let V be a small open disk in the center of the polygon, and let U be the complement of a slightly smaller closed disk. Then V is simply connected; $U \cap V$ deformation retracts onto a circle; and U deformation retracts onto the boundary of the polygon, hence on the union of $2g$ circles joined along the point x_0. It follows that $\pi_1(U, x_0)$ is isomorphic to the free group on $2g$ generators; we shall call these generators $a_1, \ldots, a_g, b_1, \ldots, b_g$, because they are the images of the $2g$ edges with those labels. The fundamental group of $U \cap V$ is isomorphic to $\mathbb{Z}$, and is generated by a circle going around the center of the polygon once in clockwise direction; in Σ_g, this is homotopic to the loop

$$a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}.$$

With the notation $[x, y] = xyx^{-1}y^{-1}$ for the commutator, we can abbreviate this loop as $[a_1, b_1] \cdots [a_g, b_g]$. According to Corollary 20.12, the fundamental group of Σ_g is isomorphic to the quotient of $\pi_1(U)$ by the smallest normal subgroup containing the image of $\pi_1(U \cap V)$. (The point x_0 is not contained in the intersection $U \cap V$, but this does not cause any problems because, up to isomorphism, the fundamental group does not depend on the choice of base point.)

We conclude that $\pi_1(\Sigma_g)$ is isomorphic to the quotient of the free group on the $2g$ letters $a_1, \ldots, a_g, b_1, \ldots, b_g$ by the smallest normal subgroup containing the element

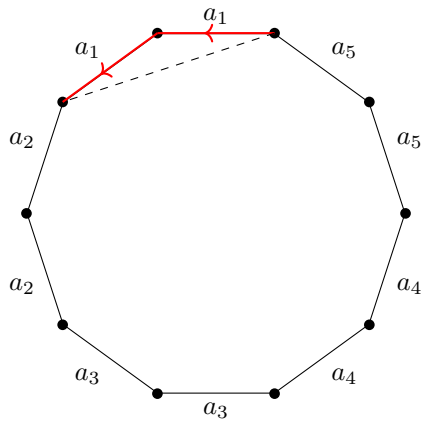
$$[a_1, b_1] \cdots [a_g, b_g] = a_1 b_1 a_1^{-1} b_1^{-1} \cdots a_g b_g a_g^{-1} b_g^{-1}.$$

So we have $2g$ generators and a single relation. In group-theoretic notation, we can express the result as follows.

Proposition 21.4. *We have* $\pi_1(\Sigma_g) \cong \langle\, a_1, \ldots, a_g, b_1, \ldots, b_g \mid [a_1, b_1] \cdots [a_g, b_g] \,\rangle$.

From this description of the fundamental group, we can read off the number of handles g, and thereby show that the number of handles is uniquely determined by the surface. For any group G, we can consider the *abelianization*, which is the quotient of G by the smallest normal subgroup containing all the commutators $[x, y] = xyx^{-1}y^{-1}$. Since $[a_1, b_1] \cdots [a_g, b_g]$ is a product of commutators, the effect of this operation on $\pi_1(\Sigma_g)$ is to make the $2g$ generators $a_1, \ldots, a_g, b_1, \ldots, b_g$ commute; the abelianization of $\pi_1(\Sigma_g)$ is therefore isomorphic to the product group $\mathbb{Z}^{2g}$. The conclusion is that Σ_g and Σ_h are not homeomorphic for $g \neq h$.

Example 21.5. To construct a non-orientable surface, we can take a connected sum of $n \geq 1$ copies of the projective plane $\mathbb{P}^2$. An equivalent description is to start from the sphere $\mathbb{S}^2$ and to attach n cross-caps: to attach a cross-cap, we remove a small disk and identify opposite points on the resulting circle. Since $\mathbb{P}^2$ with a disk removed is homeomorphic to the Möbius strip, this is the same thing as attaching n copies of the Möbius strip. A third description is to take a regular $2n$-gon and glue the edges together in pairs, as in the following picture (for $n = 5$):



Again, all $2g$ corners of the polygon become a single point in the quotient space, and the $2g$ edges give us g different loops based at this point. It is also easy to see that adding the region above the dashed line is the same thing as taking the connected sum with $\mathbb{P}^2$. We denote the resulting non-orientable surface by the symbol $\tilde{\Sigma}_n$.

The same argument as before proves the following result.

Proposition 21.6. *We have* $\pi_1(\tilde{\Sigma}_n) \cong \langle\, a_1, \ldots, a_n \mid a_1^2 \cdots a_n^2 \,\rangle$.

Can you see how to read off the number n from this group?

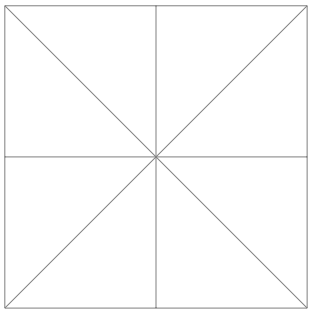
The classification of surfaces. The main result in the theory of surfaces is that every (compact and connected) surface is homeomorphic to one of the examples constructed above.

Theorem 21.7. *Let S be a compact and connected 2-dimensional topological manifold. If S is orientable, then it is homeomorphic to Σ_g for a unique $g \geq 0$; if S is non-orientable, then it is homeomorphic to $\tilde{\Sigma}_n$ for a unique $n \geq 1$.*

This is a pretty deep theorem. All known proofs have two parts: first one shows that a compact and connected surface has some additional structure; and then one uses this additional structure to classify the surfaces. The first part is always difficult; the second one is easier. The proof that I am going to present relies on the fact that any surface can be triangulated. This is fairly difficult, and so I am not going to talk about the proof; instead, I will show how one can classify surfaces given a triangulation. (The best reference for the existence of triangulations is the 1992 article "The Jordan-Schoenflies Theorem and the Classification of Surfaces" by Carsten Thomassen, see Amer. Math. Month. **99**, 116–131.)

What do we mean by a triangulation? Imagine a finite number of triangles that are glued together along their edges in order to make a compact and connected surface; at each edge, we are only allowed to glue together two triangles. The fact is that every (compact, connected) surface is homeomorphic to a surface made by gluing together finitely many triangles.

Example 21.8. The sphere is homeomorphic to a tetrehedron (which consists of 4 triangles) or to an octahedron (which consists of 8 triangles), and also to a cube in which we divide every face into two triangles (so that we get 12 triangles in total). Here is a triangulation of the torus:



In this picture, we need to identify opposite sides of the square in the usual way.

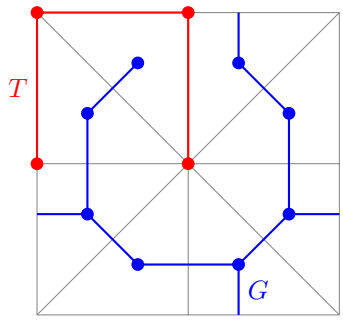
Let S be a (compact, connected) surface, and fix a triangulation. The triangulation gives us finitely many points and paths joining these points; it divides S into regions that are homeomorphic to a standard triangle in the plane. Let us say that the triangulation consists of t triangles, and that we get a total of e edges and v vertices. We can then define the *Euler characteristic* of the triangulation as

$$\chi = v - e + t \in \mathbb{Z}.$$

The name comes from Euler's formula, which says that for any triangulation of a convex polyhedron, the numbers v, e, and t are related by the equation $v-e+t = 2$. We will show in a little while that χ is actually the same for every triangulation of S, and is therefore an invariant of the surface itself.

Example 21.9. If we triangulate the sphere using the tetrahedron, we have $(v, e, t) = (4, 6, 4)$, and therefore $\chi = 2$. The triangulation of the torus has $(v, e, t) = (4, 12, 8)$, and therefore $\chi = 0$. Let us also compute the Euler characteristic of the projective plane $\mathbb{P}^2$. Here we first triangulate the sphere using an octahedron, so that $(v, e, t) = (6, 12, 8)$. To get $\mathbb{P}^2$, we take the quotient by the equivalence relation $x \sim -x$. This identifies every vertex/edge/triangle with the vertex/edge/triangle on the opposite side, and therefore gives us a triangulation of $\mathbb{P}^2$ in which $(v, e, t) = (3, 6, 4)$. Consequently, we get $\chi = 1$ for the projective plane.

We now use the triangulation to draw two graphs on the surface. The first graph T is a maximal tree in the triangulation, meaning a connected graph that contains all the vertices of the triangulations but has no cycles. This tree will have v vertices and some number $e_1 \leq e$ of edges. The second graph G is a sort of dual graph to the tree: we put one vertex in every triangle of the triangulation, and join two vertices by an edge if the two triangles have a common edge that is not contained in the tree T. The graph G will therefore have t vertices and some number $e_2 \leq e$ of edges; clearly, $e_1 + e_2 = e$. Here is what this looks like in the case of the torus:



Now we observe that the graph G is connected, due to the fact that S is connected and T is a tree. Indeed, suppose that G had more than one connected component. If we look at the union of all the triangles that are touched by a given component of G, then the outside boundary of this region would have to consist entirely of edges in T (because G cannot have any edges that cross this boundary), and this would mean that T contains a cycle; but this is not allowed because T is a tree.

For any graph G with v vertices and e edges, we define the *Euler characteristic* as $\chi(G) = v - e$. It is a simple exercise in graph theory to show that any connected graph satisfies $\chi(G) \leq 1$, and that $\chi(G) = 1$ happens if and only if G is a tree. We can therefore rewrite the Euler characteristic of the triangulation as

$$\chi = v - e + t = (v - e_1) + (t - e_2) = \chi(T) + \chi(G) = 1 + \chi(G) \leq 2.$$

Now there are two cases:

(1) The first case is $\chi = 2$. We will argue that S is homeomorphic to the sphere. From the inequality above, we get $\chi(G) = 1$, and so G is also a tree. It is easy to see that the tree T has a neighborhood U that is homeomorphic to the disk B^2. Since T and G are disjoint, we can also find a neighborhood V of G that is homeomorphic to B^2 and disjoint from U. Now we enlarge U and V until they fill up the entire surface and meet exactly along their boundaries. This sounds complicated, but it can easily be done by working in one triangle at a time. The conclusion is that S is the union of two disks glued along their boundary circles, and so $S \cong \mathbb{S}^2$.

(2) The second case is $\chi < 0$. Now $\chi(G) < 1$, and so G has to contain a cycle. Pick any cycle γ in G, and suppose that γ has k vertices and k edges. Since γ is contained in G, the complement of γ in S is connected: by construction, this complement contains T, and we can walk along T to any vertex of the triangulation.

Let us first assume that S is orientable. Then the curve described by γ is two-sided, and so if we make a cut along γ, we get a surface with boundary, whose boundary consists of two copies of γ. To each of these two boundaries, we attach a cap homeomorphic to the disk, consisting of k triangles meeting in one additional vertex. Let us call the new surface that we obtain in this way S'; it is still compact, connected, and orientable, and still has a triangulation. The new triangulation has $t' = t + 2k$ triangles; and since all the edges and vertices in γ got duplicated, it has $v' = v + k + 2$ vertices and $e' = e + 3k$ edges. The Euler characteristic of the triangulation on S' is therefore

$$\chi' = v' - e' + t' = (v + k + 2) - (e + 3k) + (t + 2k) = v - e + t + 2 = \chi + 2.$$

By induction (on the value of χ), we conclude that S' is homeomorphic to Σ_{g-1} for some $g \geq 1$. But S is obviously obtained from S' by attaching a single handle, and consequently $S \cong \Sigma_g$. This argument also shows that $\chi = 2 - 2g$. We already know that the number of handles is an invariant of the surface (up to homeomorphism), and so we may conclude that all triangulations of the surface Σ_g have the same Euler characteristic, namely $2 - 2g$. (Euler's formula is the special case $g = 0$.)

The analysis in the non-orientable case is basically the same. If the curve described by γ is two-sided, then we cut along γ and obtain a new surface S' with $\chi' = \chi + 2$ that is still non-orientable; S is obtained from S' by attaching a single handle. If the curve described by γ is one-sided, then cutting along γ results in a surface with a single boundary consisting of a cycle with $2k$ edges and $2k$ vertices.

After gluing on a cap homeomorphic to the disk, consisting of $2k$ triangles meeting in one additional vertex, we get a new surface S''. The new triangulation has $t'' = t + 2k$ triangles, $e'' = e + 3k$ edges, and $v'' = v + k + 1$ vertices, and therefore Euler characteristic

$$\chi'' = v'' - e'' + t'' = (v + k + 1) - (e + 3k) + (t + 2k) = \chi + 1.$$

Moreover, S is obtained from S'' by attaching a single cross-cap. Note that S'' could be orientable or non-orientable. Either way, by induction (on the value of χ), we find that S is homeomorphic to $\mathbb{S}^2$ with a certain number of handles (possibly zero) and a certain number of cross-caps (at least one) attached.

We can use a trick to transform each handle into two cross-caps. Since S is non-orientable, it contains a one-sided curve. For each handle, we take one of the two attaching circles, move it until it meets the one-sided curve, and then slide it along that one-sided curve until it comes back to the starting point. This has the effect of reversing the orientation of the attaching circle, and therefore transforms our handle, which is really the connected sum with a torus, into the connected sum with a Klein bottle. Because we know that the Klein bottle is homeomorphic to $\mathbb{P}^2\#\mathbb{P}^2$, this allows us to replace each handle by two cross-caps, and therefore proves that $S \cong \tilde{\Sigma}_n$ for some $n \geq 1$. From the formulas that we used during the proof, we can again see that $\chi = 2 - n$, and hence that every triangulation of the non-orientable surface $\tilde{\Sigma}_n$ has the same Euler characteristic $2 - n$.

Lecture 22: November 15

Covering spaces. We have already seen that there is a close relationship between covering spaces and the fundamental group. Recall that a covering space of a topological space B is a surjective continuous function $p\colon E \to B$ such that every point in B has a neighborhood U that is *evenly covered* by p. Under some relatively mild assumptions on B, one can classify all possible covering spaces in terms of the fundamental group $\pi_1(B, b_0)$; today and next time, I am going to explain how this classification works.

The starting point is the following simple observation.

Lemma 22.1. *Let $p\colon E \to B$ be a covering space. If $b_0 \in B$ and $e_0 \in E$ are base points satisfying $p(e_0) = b_0$, the the induced homomorphism*

$$p_*\colon \pi_1(E, e_0) \to \pi_1(B, b_0)$$

is injective.

Proof. Since p_* is a group homomorphism, it is enough to show that $\ker p_* = \{e\}$. Suppose that $[\alpha] \in \pi_1(E, e_0)$ is an element of $\ker p_*$, which means that $p \circ \alpha$ is path homotopic to the constant path at b_0. Now obviously α is a lifting of $p \circ \alpha$, and the constant path at e_0 is a lifting of the constant path at b_0; therefore Corollary 18.1 shows that α is path homotopic to the constant path. But this means exactly that $[\alpha] = e$, which is what we wanted to show. □

From a covering space $p\colon E \to B$ with $p(e_0) = b_0$, we therefore obtain a subgroup

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq \pi_1(B, b_0)$$

isomorphic to $\pi_1(E, e_0)$; we will see below that knowing this subgroup is more or less equivalent to knowing the original covering space. In fact, our goal will be to prove that there is a one-to-one correspondence between covering spaces of B (up to equivalence) and subgroups of $\pi_1(B, b_0)$ (up to conjugacy).

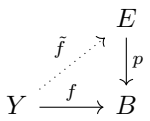
The lifting lemma. In order to get a satisfactory theory, we shall assume from now on that the space B is locally path connected. Now B is a disjoint union of path connected open subspaces: indeed, we proved earlier in the semester that the path components of a locally path connected space are open. Without any loss of generality, we may therefore assume in addition that B is path connected.

Definition 22.2. A topological space is called *nice* if it is path connected and locally path connected. A covering space $p\colon E \to B$ is called *nice* if E is nice; since p is surjective and a local homeomorphism, B is then automatically nice as well.

Some time ago, we showed that paths and homotopies can be uniquely lifted to a covering space. Those were special cases of the following lifting lemma.

Proposition 22.3. *Let $p\colon E \to B$ be a nice covering space with $p(e_0) = b_0$. Suppose that Y is a nice topological space, and that $f\colon Y \to B$ is a continuous function with $f(y_0) = b_0$. Then the following two statements are equivalent:*

(a) There exists a (unique) lifting $\tilde{f}\colon Y \to E$ with $\tilde{f}(y_0) = e_0$ and $p \circ \tilde{f} = f$.



(b) One has $f_*\big(\pi_1(Y,y_0)\big) \subseteq p_*\big(\pi_1(E,e_0)\big)$.

The proof that (a) implies (b) is a straightforward consequence of the fact that the fundamental group is a functor. Suppose we have a lifting $\tilde{f}\colon Y \to E$ with $\tilde{f}(y_0) = e_0$ and $p \circ \tilde{f} = f$. We obtain a commutative diagram

$$\begin{array}{ccc} & & \pi_1(E,e_0) \\ & \overset{\tilde{f}_*}{\nearrow} & \big\downarrow{\scriptstyle p_*} \\ \pi_1(Y,y_0) & \xrightarrow{f_*} & \pi_1(B,b_0) \end{array}$$

of groups and group homomorphisms. Since $p_* \circ \tilde{f}_* = f_*$, it is clear that

$$f_*\big(\pi_1(Y,y_0)\big) = (p_* \circ \tilde{f}_*)\big(\pi_1(Y,y_0)\big) \subseteq p_*\big(\pi_1(E,e_0)\big).$$

It is also not hard to see that the lifting $\tilde{f}$ is uniquely determined by f. Indeed, given $y_1 \in Y$, we can choose a path α from the base point y_0 to the point y_1; then $\gamma = \tilde{f} \circ \alpha$ is a path from the base point e_0 to the point $\tilde{f}(y_1)$. Now γ is clearly a lifting of the path $f \circ \alpha$; since $p\colon E \to B$ is a covering space, Theorem 17.9 tells us that the path $f \circ \alpha$ has a unique lifting with this property. This means that $\tilde{f}(y_1) = \gamma(1)$ is uniquely determined by the path $f \circ \alpha$, and so f can have at most one lifting $\tilde{f}$.

Now let us deal with the more interesting implication, namely (b) implies (a).

The first step is to construct a function $\tilde{f}\colon Y \to E$ with $\tilde{f}(y_0) = e_0$ and $p \circ \tilde{f} = f$. Here we take hint from the comments above, and define $\tilde{f}$ by lifting paths. Given $y_1 \in Y$, choose a path α from y_0 to y_1; such a path exists because Y is path connected. The image $f \circ \alpha$ is a path from $b_0 = f(y_0)$ to $f(y_1)$; by Theorem 17.9, it admits a unique lifting to a path $\gamma\colon I \to E$ with

$$\gamma(0) = e_0 \quad \text{and} \quad p \circ \gamma = f \circ \alpha.$$

We would like to define $\tilde{f}(y_1) = \gamma(1)$. This only makes sense if we can show that this point is independent of the path α.

Lemma 22.4. *If α' is another path from y_0 to y_1, then $\gamma'(1) = \gamma(1)$.*

Proof. The composition $\alpha' * \bar{\alpha}$ is a loop based at the point y_0, hence represents an element of $\pi_1(Y,y_0)$. Since we are assuming that (b) holds, we have

$$f_*[\alpha' * \bar{\alpha}] \in f_*\big(\pi_1(Y,y_0)\big) \subseteq p_*\big(\pi_1(E,e_0)\big);$$

consequently, there is an element $[\beta] \in \pi_1(E,e_0)$ such that $f_*[\alpha' * \bar{\alpha}] = p_*[\beta]$. Concretely, β is a loop based at the point e_0 and

$$(f \circ \alpha') * (f \circ \bar{\alpha}) \sim_p (p \circ \beta);$$

by the general properties of composition of paths, it follows that

$$(f \circ \alpha') \sim_p (p \circ \beta) * (f \circ \alpha).$$

Now γ and γ' are the unique liftings of $f \circ \alpha$ and $f \circ \alpha'$, and β is obviously a lifting of $p \circ \beta$; by Corollary 18.1, we get $\gamma' \sim_p \beta * \gamma$ and

$$\gamma'(1) = (\beta * \gamma)(1) = \gamma(1),$$

which is what we had set out to show. □

As a result, we have a well-defined function $\tilde{f}\colon Y \to E$. It is clear from the construction that $\tilde{f}(y_0) = e_0$ (take α to be the constant path at y_0); since

$$(p \circ \tilde{f})(y_1) = (p \circ \gamma)(1) = (f \circ \alpha)(1) = f(y_1),$$

we also obtain without difficulty that $p \circ \tilde{f} = f$. What is less clear is that $\tilde{f}$ is actually continuous – note that this is the only place where the assumption about local path connectedness is used.

Lemma 22.5. *The function $\tilde{f}\colon Y \to E$ is continuous.*

Proof. Let V be an arbitrary open set containing the point $\tilde{f}(y_1)$; to prove that $\tilde{f}$ is continuous, we have to find a neighborhood W of the point y_1 such that $\tilde{f}(W) \subseteq V$. Since $p\colon E \to B$ is a covering space, some neighborhood of $f(y_1)$ is evenly covered by p; inside of this neighborhood, we can find a smaller neighborhood U of $f(y_1)$ that is also path connected (because B is locally path connected). Then $p^{-1}(U)$ is a disjoint union of open subsets of E; let V_0 be the one containing the point $\tilde{f}(y_1)$. After shrinking U, if necessary, we may assume that $V_0 \subseteq V$. On the other hand, f is continuous, and so $f^{-1}(U)$ is an open set containing the point y_1; since Y is locally path connected, we can find a path connected neighborhood W of y_1 with $f(W) \subseteq U$.

Now I claim that $\tilde{f}(W) \subseteq V$. To prove this, let $y \in W$ be an arbitrary point. Since W is path connected, there is a path $\beta\colon I \to W$ with $\beta(0) = y_1$ and $\beta(1) = y$; then $\alpha * \beta$ is a path from y_0 to y. To compute $\tilde{f}(y)$, we have to lift

$$f \circ (\alpha * \beta) = (f \circ \alpha) * (f \circ \beta)$$

to a path in E. Recall that γ is the unique lifting of $f \circ \alpha$ with $\gamma(0) = e_0$. It is also not hard to find a lifting of the path $f \circ \beta$: since U is evenly covered by p, the restriction

$$p_0 = p\big|_{V_0}\colon V_0 \to U$$

is a homeomorphism, and so $\delta = p_0^{-1} \circ f \circ \beta\colon I \to V_0$ is a lifting of the path $f \circ \beta$ with $\delta(0) = \gamma(1)$. The composition $\gamma * \delta$ is defined, and

$$p \circ (\gamma * \delta) = (p \circ \gamma) * (p \circ \delta) = (f \circ \alpha) * (f \circ \beta);$$

hence $\gamma * \delta$ is the desired lifting of $f \circ (\alpha * \beta)$. But then

$$\tilde{f}(y) = (\gamma * \delta)(1) = \delta(1) \in V_0 \subseteq V,$$

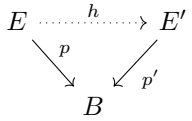
as claimed. □

Covering spaces and subgroups, Part 1. With the covering lemma in hand, we can now start looking at the correspondence between covering spaces of B and subgroups of $\pi_1(B, b_0)$. If $p\colon E \to B$ is a covering space with $p(e_0) = b_0$, we denote by

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq \pi_1(B, b_0)$$

the resulting subgroup of the fundamental group; note that it depends on the choice of base point $e_0 \in E$.

Definition 22.6. Two covering spaces $p\colon E \to B$ and $p'\colon E' \to B$ are called *equivalent* if there exists a homeomorphism $h\colon E \to E'$ with $p = p' \circ h$.

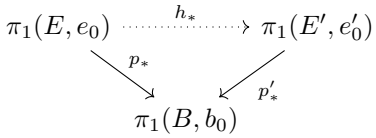


With this notion of equivalence, we are allowed to replace E by a homeomorphic space E', provided we also replace p by the function $p' = p \circ h^{-1}$. Here is a first result that relates covering spaces and subgroups.

Theorem 22.7. *Let $p\colon E \to B$ and $p'\colon E' \to B$ be two nice covering spaces, with base points $e_0 \in E$ and $e_0' \in E'$ satisfying $p(e_0) = p'(e_0') = b_0$. Then the following two statements are equivalent:*

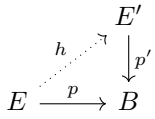
(a) There exists a homeomorphism $h\colon E \to E'$ with $p' \circ h = p$ and $h(e_0) = e_0'$.
(b) The two groups $H_0 = p_*\big(\pi_1(E, e_0)\big)$ and $H_0' = p_*'\big(\pi_1(E', e_0')\big)$ are equal.

Proof. As in the case of the lifting lemma, one implication is straightforward. Given a homeomorphism $h\colon E \to E'$ as in (a), we get the following commutative diagram of groups and group homomorphisms:



Now $p_*' \circ h_* = p_*$, and because h is a homeomorphism, h_* is bijective. From this, one easily deduces that $H_0 = H_0'$.

To prove the converse, we shall use the lifting lemma – as it turns out, four times. We are trying to find a continuous function $h\colon E \to E'$ with $h(e_0) = e_0'$ and $p' \circ h = p$, or in other words, a lifting of p to the covering space $p'\colon E' \to B$.

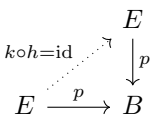


Since E and E' are both nice, Proposition 22.3 shows that a necessary and sufficient condition is that

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq p_*'\big(\pi_1(E', e_0')\big) = H_0'.$$

Because $H_0 = H_0'$, this condition is clearly satisfied, and so there is a (unique) lifting $h\colon E \to E'$. Since we also have $H_0' \subseteq H_0$, a similar argument shows that there is a (unique) lifting $k\colon E' \to E$ with $k(e_0') = e_0$ and $p \circ k = p'$.

Now I claim that both $k \circ h$ and $h \circ k$ are equal to the identity, which means that h is a homeomorphism with inverse k. This follows from the uniqueness statement in Proposition 22.3. Indeed, both $k \circ h$ and id are liftings of $p\colon E \to B$ to the covering space $p\colon E \to B$; since there can be at most one lifting, $k \circ h = \mathrm{id}$.



The same argument shows that $h \circ k = \mathrm{id}$, and so h is a homeomorphism. □

This result shows that if $H_0 = H_0'$, then the two covering spaces $p\colon E \to B$ and $p'\colon E' \to B$ are equivalent. The equivalence in the theorem was not arbitary, however, but was supposed to be compatible with a fixed choice of base points. We should therefore figure out what happens when we change the base point. Suppose that $e_1 \in E$ is another point with $p(e_1) = b_0$. Since E is path connected, the two groups $\pi_1(E, e_0)$ and $\pi_1(E, e_1)$ are isomorphic. This does not mean that their images H_0 and H_1 in $\pi_1(B, b_0)$ have to be equal; but they are related in the following way.

Definition 22.8. Two subgroups H_0 and H_1 of a group G are *conjugate* if there is an element $g \in G$ such that $H_1 = gH_0g^{-1}$; said differently, the isomorphism $G \to G$, $x \mapsto gxg^{-1}$, should take one subgroup to the other.

With this definition, we can state an improved version of Theorem 22.7.

Theorem 22.9. *Let $p\colon E \to B$ and $p'\colon E' \to B$ be two nice covering spaces, with base points $e_0 \in E$ and $e_0' \in E'$ satisfying $p(e_0) = p'(e_0') = b_0$. Then the following two statements are equivalent:*

(a) There exists a homeomorphism $h\colon E \to E'$ with $p' \circ h = p$.
(b) The two subgroups H_0 and H_0' of $\pi_1(B, b_0)$ are conjugate.

We begin the proof by analyzing how the group H_0 depends on the choice of base point $e_0 \in p^{-1}(b_0)$.

Lemma 22.10. *Let $p\colon E \to B$ be a nice covering space with $p(e_0) = b_0$, and define $H_0 = p_*\big(\pi_1(E, e_0)\big)$.*

(a) If $e_1 \in p^{-1}(b_0)$, then the subgroup $H_1 = p_*\big(\pi_1(E, e_1)\big)$ is conjugate to H_0.
(b) Conversely, if $H \subseteq G$ is any subgroup conjugate to H_0, then there is a point $e \in p^{-1}(b_0)$ such that $H = p_*\big(\pi_1(E, e)\big)$.

Proof. To prove (a), choose a path $\varphi\colon I \to E$ with $\varphi(0) = e_0$ and $\varphi(1) = e_1$. We already know from Lemma 17.1 that the function

$$\hat{\varphi}\colon \pi_1(E, e_0) \to \pi_1(E, e_1), \quad [\alpha] \mapsto [\bar{\varphi} * \alpha * \varphi]$$

is an isomorphism of groups. Since $p(e_0) = p(e_1) = b_0$, the path $\beta = p \circ \varphi$ is a loop based at the point b_0. Now we compute that

$$(p_* \circ \hat{\varphi})[\alpha] = p_*[\bar{\varphi} * \alpha * \varphi] = [p \circ \bar{\varphi}] * [p \circ \alpha] * [p \circ \varphi] = [\beta]^{-1} * \big(p_*[\alpha]\big) * [\beta],$$

and because $\hat{\varphi}$ is an isomorphism, it follows that

$$H_1 = p_*\big(\pi_1(E, e_1)\big) = (p_* \circ \hat{\varphi})\big(\pi_1(E, e_0)\big) = [\beta]^{-1} * H_0 * [\beta].$$

Because $[\beta] \in \pi_1(B, b_0)$, this means that H_1 and H_0 are conjugate subgroups.

To prove (b), let $[\alpha] \in \pi_1(B, b_0)$ be some element with the property that

$$H = [\alpha]^{-1} * H_0 * [\alpha].$$

By Theorem 17.9, α can be uniquely lifted to a path $\tilde{\alpha}\colon I \to E$ with $\tilde{\alpha}(0) = e_0$, and if we set $e = \tilde{\alpha}(1)$, the same argument as above shows that $H = p_*\big(\pi_1(E, e)\big)$. □

Proof of Theorem 22.9. Let us first show that (a) implies (b). The given homeomorphism $h\colon E \to E'$ may not take e_0 to e_0', so put $e_1' = h(e_0)$. Then Theorem 22.7 shows that

$$H_1' = p_*'\big(\pi_1(E', e_1')\big) = H_0.$$

By Lemma 22.10, H_0' and H_1' are conjugate subgroups of $\pi_1(B, b_0)$, and (b) follows.

To prove the converse, we apply Lemma 22.10 to the subgroup H_0 to conclude that there is a point $e \in E$ with $H_0 = p_*\big(\pi_1(E, e)\big)$. Now Theorem 22.7 shows that there is a homeomorphism $h\colon E \to E'$ with $p' \circ h = p$ and $h(e) = e_0'$; in particular, the two covering spaces are equivalent. □

Now an obvious question is whether every subgroup of $\pi_1(B, b_0)$ corresponds to some covering space. Under an additional minor assumption on B, the answer is yes; in particular, there is a simply connected covering space corresponding to the trivial subgroup $\{e\}$, called the *universal covering space*. Next time, I will explain how to construct covering spaces from subgroups.

Lecture 23: November 17

Last time, we showed that a nice covering space $p\colon E \to B$ is uniquely determined (up to homeomorphism) by the subgroup

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq \pi_1(B, b_0).$$

If we change the base point $e_0 \in E$, the subgroup may change, but only to another subgroup that is conjugate to H_0. Now the natural question is whether every subgroup of $\pi_1(B, b_0)$ is actually realized by some covering space.

Example 23.1. Since p_* is injective, the trivial subgroup $\{e\}$ would correspond to a simply connected covering space. Such a covering space, if it exists, is unique up to homeomorphism, and is called the *universal covering space* of B.

In this generality, the answer to the question is negative: not every nice (= path connected and locally path connected) space has a universal covering space. The following lemma shows that there is another necessary condition.

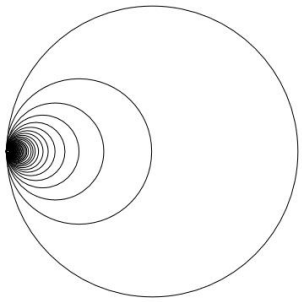
Lemma 23.2. *If a nice topological space has a simply connected covering space $p\colon E \to B$, then every point $b \in B$ has a neighborhood U such that the homomorphism $\pi_1(U, b) \to \pi_1(B, b)$ is trivial.*

Proof. Let U be any neighborhood of b that is evenly covered by p. Let $V \subseteq p^{-1}(U)$ be one of the disjoint open subsets homeomorphic to U, and let $e \in V \cap p^{-1}(b)$ be the unique point mapping to b. We get the following diagram of groups and group homomorphisms:

$$\begin{array}{ccc} \pi_1(V, e) & \longrightarrow & \pi_1(E, e) \\ \big\downarrow{\scriptstyle \simeq} & & \big\downarrow{\scriptstyle p_*} \\ \pi_1(U, b) & \longrightarrow & \pi_1(B, b) \end{array}$$

Since U and V are homeomorphic, the vertical arrow on the left is an isomorphism. Given that $\pi_1(E, e)$ is the trivial group, the image of $\pi_1(U, b) \to \pi_1(B, b)$ must be the trivial subgroup, too. □

Example 23.3. Here is an example of a space where this property fails to hold.



Let $B \subseteq \mathbb{R}^2$ be the union of the circles

$$C_n = \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; \left(x - \frac{1}{n}\right)^2 + y^2 = \frac{1}{n^2} \right\}$$

for $n = 1, 2, \ldots$. Since every one of these circles gives a different nontrivial element in $\pi_1\big(B, (0,0)\big)$, the point $(0,0)$ does not have any neighborhood as in the lemma, and therefore B cannot have a universal covering space.

Fortunately, it turns out that the property in Lemma 23.2 is also sufficient for the existence of a universal covering space.

Definition 23.4. A nice topological space B is called *semi-locally simply connected* if every point $b \in B$ has a neighborhood U such that $\pi_1(U, b) \to \pi_1(B, b)$ is the trivial homomorphism.

Of course, every neighborhood of b that is contained in U will have the same property. This notion is weaker than being *locally simply connected* (which would mean that U itself is simply connected).

Existence of covering spaces. The main existence result for covering spaces is the following.

Theorem 23.5. *Let B be nice and semi-locally simply connected. Then for every subgroup $H \subseteq \pi_1(B, b_0)$, there is a nice covering space $p\colon E \to B$ and a point $e_0 \in p^{-1}(b_0)$ such that $p_*\big(\pi_1(E, e_0)\big) = H$.*

The proof is somewhat long, but not very difficult. Given a subgroup H, we have to construct a suitable covering space $p\colon E \to B$; the main idea is to use the path lifting property of covering spaces. Suppose for a moment that we already had $p\colon E \to B$. Since E is supposed to be path connected, every point $e \in E$ can be joined to e_0 by a path $\tilde{\alpha}$. Its image $\alpha = p \circ \tilde{\alpha}$ is a path in B that starts at the base point b_0 and ends at the point $\alpha(1) = p(e)$. Now we observe two things:

(1) The path $\tilde{\alpha}$ is uniquely determined by α, according to the path lifting property of covering spaces in Theorem 17.9.
(2) If we choose a different path $\tilde{\beta}$ from e_0 to e, then $\beta = p \circ \tilde{\beta}$ ends at the same point $p(e)$; moreover $[\alpha * \bar{\beta}]$ is an element of the subgroup H, because $\tilde{\alpha}$ followed by the inverse of $\tilde{\beta}$ is an element of $\pi_1(E, e_0)$.

This means that points in E are in one-to-one correspondence with paths that start at the point b_0, subject to the equivalence relation

$$\alpha \sim \beta \quad \Longleftrightarrow \quad \alpha(1) = \beta(1) \text{ and } [\alpha * \bar{\beta}] \in H. \tag{23.6}$$

We now use this idea to construct $p\colon E \to B$ from the subgroup H; for the sake of clarity, I shall divide the proof into eight steps.

Step 1. We define $p\colon E \to B$ on the level of sets. Let $C_0(I, B)$ be the set of all paths $\alpha\colon I \to B$ with $\alpha(0) = b_0$. It is easy to see that the conditions in (23.6) define an equivalence relation on $C_0(I, B)$; we let $\alpha^\sharp$ denote the equivalence class of the path α. Now let

$$E = \big\{\, \alpha^\sharp \;\big|\; \alpha \in C_0(I, B) \,\big\}$$

be the set of equivalence classes. Since equivalent paths have the same endpoint, the function

$$p\colon E \to B, \quad p(\alpha^\sharp) = \alpha(1),$$

is well-defined; it is also surjective, due to the fact that B is path connected.

Step 2. We put a topology on E. There are two ways of doing this. One is to give the space $C(I, B)$ and its subspaces $C_0(I, B)$ the compact-open topology, and then to give $E = C_0(I, B)/\sim$ the quotient toplogy. A more concrete way is by writing down a basis. For any path $\alpha \in C_0(I, B)$ and for any path connected neighborhood U of the point $\alpha(1)$, let

$$B(U, \alpha) = \big\{\, (\alpha * \delta)^\sharp \;\big|\; \delta \text{ is a path in } U \text{ with } \delta(0) = \alpha(1) \,\big\}.$$

Taking $\delta = e_{\alpha(1)}$ shows that $\alpha^\sharp \in B(U,\alpha)$. We shall argue in a moment that these sets form a basis for a topology on E; to make that easier, let us first prove a small lemma.

Lemma 23.7. *If $\beta^\sharp \in B(U,\alpha)$, then $B(U,\alpha) = B(U,\beta)$.*

Proof. By assumption, $\beta^\sharp = (\alpha * \delta)^\sharp$ for some path $\delta\colon I \to U$ with $\delta(0) = \alpha(1)$. Since $\delta * \bar{\delta}$ is path homotopic to the constant path at $\alpha(1)$, we have

$$(\beta * \bar{\delta})^\sharp = \big((\alpha * \delta) * \bar{\delta}\big)^\sharp = \alpha^\sharp,$$

which means that $\alpha^\sharp \in B(U,\beta)$. Whenever γ is a path in U starting at the point $\beta(1)$, we compute that

$$(\beta * \gamma)^\sharp = \big((\alpha * \delta) * \gamma\big)^\sharp = \big(\alpha * (\delta * \gamma)\big)^\sharp \in B(U,\alpha),$$

and so $B(U,\beta) \subseteq B(U,\alpha)$; the other inclusion follows by symmetry. □

Step 3. We show that the sets $B(U,\alpha)$ form a basis for a topology on E. Since $\alpha^\sharp \in B(U,\alpha)$ and since B is locally path connected, it is clear that the union of all basic sets is E. It remains to show that $B(U_1,\alpha_1) \cap B(U_2,\alpha_2)$ is covered by basic sets. If $\beta^\sharp \in B(U_1,\alpha_1) \cap B(U_2,\alpha_2)$ is an arbitrary point in the intersection, let V be a path connected neighborhood of $\beta(1)$ contained in $U_1 \cap U_2$. Then

$$B(V,\beta) \subseteq B(U_1,\beta) \cap B(U_2,\beta) = B(U_1,\alpha_1) \cap B(U_2,\alpha_2),$$

where the last equality is due to the lemma. The criterion in Proposition 2.8 shows that E has a topology in which the $B(U,\alpha)$ are basic open sets.

Step 4. We argue that $p\colon E \to B$ is continuous and open. Openness is easy: it is enough to show that the image of every basic open sets is open, and $p\big(B(U,\alpha)\big) = U$ because U is path connected. To prove that p is continuous at the point $\alpha^\sharp \in E$, let $V \subseteq B$ be an arbitrary neighborhood of $p(\alpha^\sharp) = \alpha(1)$. Since B is locally path connected, we can find a path connected neighborhood $U \subseteq V$; then $p\big(B(U,\alpha)\big) = U \subseteq V$, and so p is continuous.

Step 5. We show that $p\colon E \to B$ is a covering space. Given any point $b_1 \in B$, let U be a path connected neighborhood with the property that $\pi_1(U,b_1) \to \pi_1(B,b_1)$ is trivial. We shall argue that U is evenly covered by p. First, we have

$$p^{-1}(U) = \bigcup_\alpha B(U,\alpha),$$

where the union is over all paths $\alpha\colon I \to B$ with $\alpha(0) = b_0$ and $\alpha(1) = b_1$. Since $p\big(B(U,\alpha)\big) = U$, one inclusion is clear. To prove the other one, suppose that $\beta(1) = p(\beta^\sharp) \in U$. Because U is path connected, we can choose a path δ in U from the point b_1 to the point $\beta(1)$; then

$$\alpha = \beta * \bar{\delta}$$

is a path from b_0 to b_1, and $\beta^\sharp = (\alpha * \delta)^\sharp \in B(U,\alpha)$. By the lemma, distinct sets of the form $B(U,\alpha)$ are disjoint; thus $p^{-1}(U)$ is a disjoint union of open sets.

Lemma 23.8. *p induces a homeomorphism between $B(U,\alpha)$ and U.*

Proof. We already know that p is continuous and open; moreover $p\big(B(U,\alpha)\big) = U$. It remains to show that the restriction of p to $B(U,\alpha)$ is injective. Suppose that

$$p\big((\alpha * \delta_1)^\sharp\big) = p\big((\alpha * \delta_2)^\sharp\big).$$

Then $\delta_1(1) = \delta_2(1)$, and so $\delta_1 * \bar{\delta_2}$ is a loop in U based at the point b_1. Because $\pi_1(U, b_1) \to \pi_1(B, b_1)$ is the trivial homomorphism, we have $[\delta_1 * \bar{\delta_2}] = e$; but then

$$[(\alpha * \delta_1) * \bar{\delta_2} * \bar{\alpha}] = [\alpha] * [\delta_1 * \bar{\delta_2}] * [\bar{\alpha}] = [\alpha] * [\bar{\alpha}] = e,$$

which proves that $(\alpha * \delta_1)^\sharp = (\alpha * \delta_2)^\sharp$. □

We conclude that $p\colon E \to B$ is a covering space. Let e_0 be the equivalence class of the constant path at b_0; obviously, $p(e_0) = b_0$.

Step 6. We give a formula for the lifting of paths from B to E. Suppose that $\alpha\colon I \to B$ is a path with $\alpha(0) = b_0$. Since $p\colon E \to B$ has the path lifting property, we know that there is a unique lifting $\tilde{\alpha}\colon I \to E$ with $\tilde{\alpha}(0) = e_0$. We shall validate the original idea for the construction by showing that this lifting ends at the point $\alpha^\sharp$. For any $c \in I$, define

$$\alpha_c\colon I \to B, \quad \alpha_c(t) = \alpha(ct);$$

it is simply the portion of the path α between $\alpha(0)$ and $\alpha(c)$. Note that α_0 is the constant path at b_0, whereas $\alpha_1 = \alpha$. Now define

$$\tilde{\alpha}\colon I \to E, \quad \tilde{\alpha}(c) = \alpha_c^\sharp.$$

Then $\tilde{\alpha}(0) = e_0$, and since $p\big(\tilde{\alpha}(c)\big) = \alpha_c(1) = \alpha(c)$, this must be the desired lifting of α, provided we can prove continuity.

Lemma 23.9. *$\tilde{\alpha}$ is continuous.*

Proof. To show that $\tilde{\alpha}$ is continuous at the point $c \in I$, let $B(U, \alpha_c)$ be one of the basic open sets containing the point $\alpha_c^\sharp$. Since α is continuous, we can certainly choose some $\varepsilon > 0$ so that $\alpha(d) \in U$ whenever $|c - d| < \varepsilon$; now it is not hard to see that $\alpha_d^\sharp \in B(U, \alpha_c)$ for every such d. □

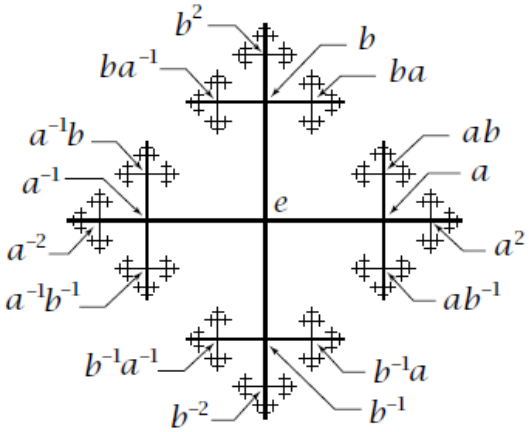
Step 7. We conclude that $p\colon E \to B$ is a nice covering space. This is immediate, because for any $\alpha^\sharp \in E$, the lifting $\tilde{\alpha}$ from Step 6 is a path that connects the base point e_0 to the point $\alpha^\sharp$. Consequently, E is path connected; it is also locally path connected (because it is locally homeomorphic to B by Step 5).

Step 8. We show that $H = p_*\big(\pi_1(E, e_0)\big)$. Let α be an arbitrary loop in B based at b_0, and let $\tilde{\alpha}\colon I \to E$ be the lifting constructed in Step 6. By the uniqueness of liftings, we have $\alpha \in p_*\big(\pi_1(E, e_0)\big)$ if and only if $\tilde{\alpha}$ starts and ends at the point e_0. Because $\tilde{\alpha}(1) = \alpha^\sharp$, this is the same as saying that $\alpha^\sharp = e_0$, which by (23.6) is equivalent to $[\alpha] \in H$.

This concludes the proof of Theorem 23.5. In the case $H = \{e\}$, the equivalence relation is just that of path homotopy; in the model we have constructed, the points of the universal covering space are therefore path homotopy classes of paths starting at the point b_0. Note that $p^{-1}(b_0)$ is exactly the set $\pi_1(B, b_0)$, as it should be.

Corollary 23.10. *If B is nice and semi-locally simply connected, then we have a bijective correspondence between nice covering spaces of B (up to equivalence) and subgroups of $\pi_1(B, b_0)$ (up to conjugacy).*

Example 23.11. The description of the covering space in terms of paths can be used to visualize the universal covering space. If X is the union of two circles a and b, touching at a single point x_0, the universal covering space has the following shape:



Every horizontal line goes around the circle a, every vertical line around the circle b, with left/right and up/down corresponding to the two possible orientations. The crossing points are naturally labeled by the elements of the free group on two generators; since these are exactly the points that map to the base point x_0, this shows again that the fundamental group of X is $\mathbb{Z} * \mathbb{Z}$.

Lecture 24: November 22

Deck transformations. When we studied the correspondence between subgroups of the fundamental group and covering space, we introduced the notion of equivalence for covering spaces. Our goal today is to describe all possible self-equivalences of a given covering space.

Definition 24.1. Let $p\colon E \to B$ be a nice covering space. A *deck transformation* is a homeomorphism $h\colon E \to E$ with $p \circ h = p$.

$$\begin{array}{ccc} E & \xrightarrow{\;h\;} & E \\ & {\scriptstyle p}\searrow \quad \swarrow{\scriptstyle p} & \\ & B & \end{array}$$

The set of all deck transformations of a given covering space is a group under composition; the unit element is the identity $\mathrm{id}_E\colon E \to E$. We denote this group by $\mathrm{Aut}_B(E)$. Our goal is to describe $\mathrm{Aut}_B(E)$ in terms of the fundamental group.

Deck transformations on the universal covering space. Let us begin by looking at the special case where $p\colon E \to B$ is the universal covering space of B. As usual, we choose base points $b_0 \in B$ and $e_0 \in E$ such that $p(e_0) = b_0$. Since E is simply connected, the subgroup

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq \pi_1(B, b_0)$$

is of course trivial.

Now if $e_1 \in E$ is a second point with $p(e_1) = b_0$, then the subgroup

$$H_1 = p_*\big(\pi_1(E, e_1)\big) \subseteq \pi_1(B, b_0)$$

is also trivial, and so $H_0 = H_1$. According to Theorem 22.7, there is a unique homeomorphism $h\colon E \to E$ such that $p \circ h = p$ and $h(e_0) = e_1$. Thus for every point $e_1 \in p^{-1}(b_0)$ in the fiber over b_0, there is a unique deck transformation that takes the base point e_0 to the point e_1. Said differently, the function

$$\varepsilon_0\colon \mathrm{Aut}_B(E) \to p^{-1}(b_0), \quad h \mapsto h(e_0)$$

is a bijection. The lifting correspondence in Theorem 18.2 gives us another bijection

$$\ell\colon \pi_1(B, b_0) \to p^{-1}(b_0), \quad \ell(\alpha) = \tilde{\alpha}(1);$$

here α is any loop based at the point b_0, and $\tilde{\alpha}$ is its unique lifting to a path starting at e_0. By composing ε_0 and ℓ^{-1}, we get a bijection between the group of deck transformations and the fundamental group:

$$\begin{array}{ccc} \mathrm{Aut}_B(E) & \xrightarrow{\;\varepsilon_0\;} & p^{-1}(b_0) \\ & {\scriptstyle \ell^{-1}\circ\varepsilon_0}\searrow & \uparrow{\scriptstyle \ell} \\ & & \pi_1(B, b_0) \end{array}$$

Proposition 24.2. *The bijection*

$$\ell^{-1} \circ \varepsilon_0\colon \mathrm{Aut}_B(E) \to \pi_1(B, b_0)$$

is an isomorphism of groups.

Proof. It suffices to show that $\ell^{-1} \circ \varepsilon_0$ is a group homomorphism. Let $h_1, h_2 \in \mathrm{Aut}_B(E)$ be two deck transformations, and put

$$e_1 = \varepsilon_0(h_1) = h_1(e_0) \quad \text{and} \quad e_2 = \varepsilon_0(h_2) = h_2(e_0).$$

Choose paths $\tilde{\alpha}_1$ from e_0 to e_1 and $\tilde{\alpha}_2$ from e_0 to e_2; then $\tilde{\alpha}_1$ is a lifting of the loop $\alpha_1 = p \circ \tilde{\alpha}_1$ and $\tilde{\alpha}_2$ is a lifting of $\alpha_2 = p \circ \tilde{\alpha}_2$; consequently,

$$\ell^{-1}(e_1) = [\alpha_1] \quad \text{and} \quad \ell^{-1}(e_2) = [\alpha_2].$$

To prove that $\ell^{-1} \circ \varepsilon_0$ is a homomorphism, we have to show that

$$\big(\ell^{-1} \circ \varepsilon_0\big)(h_1 \circ h_2) = [\alpha_1 * \alpha_2].$$

Now $\varepsilon_0(h_1 \circ h_2) = (h_1 \circ h_2)(e_0) = h_1(e_2)$, and so we will be done once we show that

$$\ell(\alpha_1 * \alpha_2) = h_1(e_2). \tag{24.3}$$

We already have a path $\tilde{\alpha}_2$ from the base point e_0 to the point e_2; if we apply the deck transformation $h_1 \colon E \to E$ to it, we obtain a path $h_1 \circ \tilde{\alpha}_2$ from the point $e_1 = h_1(e_0)$ to the point $h_1(e_2)$. The composed path

$$\tilde{\alpha}_1 * (h_1 \circ \tilde{\alpha}_2)$$

is still a lifting of the loop $\alpha_1 * \alpha_2$ because

$$p \circ \big(\tilde{\alpha}_1 * (h_1 \circ \tilde{\alpha}_2)\big) = (p \circ \tilde{\alpha}_1) * (p \circ h_1 \circ \tilde{\alpha}_2) = \alpha_1 * (p \circ \tilde{\alpha}_2) = \alpha_1 * \alpha_2.$$

Since it starts at the point e_0 and ends at the point $h_1(e_2)$, the definition of the lifting correspondence shows that (24.3) is satisfied. □

To summarize: The group of deck transformations of the universal covering space is isomorphic to the fundamental group; moreover, for any two points in the same fiber, there is a unique deck transformation taking one to the other. (One says that the group of deck transformations acts *simply transitively* on the fibers.)

Example 24.4. The universal covering space of the circle is $p \colon \mathbb{R} \to \mathbb{S}^1$; here the fundamental group is $\mathbb{Z}$, and since $p(x+n) = p(x)$, the group of deck transformations is also $\mathbb{Z}$.

Example 24.5. The real projective plane P^2 was the quotient of $\mathbb{S}^2$ by the equivalence relation $x \sim -x$. Here the fundamental group is $\mathbb{Z}_2$, and there are exactly two deck transformations: the identity and the antipodal map $x \mapsto -x$.

The fundamental group $\pi_1(B, b_0)$ therefore acts on the universal covering space by homeomorphisms; you can find some more information about this on the homework for next week.

Deck transformations and regular covering spaces. Now let us return to the general case where $p \colon E \to B$ is a nice covering space and $b_0 \in B$ and $e_0 \in E$ are base points with $p(e_0) = b_0$. Let

$$H_0 = p_*\big(\pi_1(E, e_0)\big) \subseteq \pi_1(B, b_0)$$

be the corresponding subgroup. We would like to relate the group of deck transformations to H_0, only this time, the relationship will not be quite as straightforward. We again consider the function

$$\varepsilon_0 \colon \mathrm{Aut}_B(E) \to p^{-1}(b_0), \quad \varepsilon_0(h) = h(e_0).$$

The uniqueness statement in Theorem 22.7 means that a deck transformation is completely determined by where it takes the point e_0; thus ε_0 is injective. As before, we also have the lifting correspondence

$$\ell\colon \pi_1(B, b_0) \to p^{-1}(b_0),$$

which is surjective since E is path connected.

Lemma 24.6. *Let α be a loop based at b_0. The point $\ell(\alpha)$ belongs to the image of ε_0 if and only if $[\alpha] * H_0 * [\alpha]^{-1} = H_0$.*

Proof. Let $\tilde{\alpha}$ denote the unique lifting of α to a path starting at e_0; then $e_1 = \tilde{\alpha}(1) = \ell(\alpha)$ belongs to the fiber $p^{-1}(b_0)$. By Theorem 22.7, there is a deck transformation h with $\varepsilon_0(h) = h(e_0) = e_1$ if and only if

$$p_*\big(\pi_1(E, e_1)\big) = H_0$$

Since $\tilde{\alpha}$ is a path from e_0 to e_1, we have

$$[\tilde{\alpha}] * \pi_1(E, e_1) = \pi_1(E, e_0) * [\tilde{\alpha}];$$

after applying the homomorphism p_*, this becomes

$$[\alpha] * p_*\big(\pi_1(E, e_1)\big) = H_0 * [\alpha],$$

and so a suitable deck transformation exists if and only if $[\alpha] * H_0 = H_0 * [\alpha]$. □

The property in the lemma already has a name in group theory.

Definition 24.7. Let H be a subgroup of a group G. The subgroup

$$N(H) = \big\{\, g \in G \;\big|\; gHg^{-1} = H \,\big\}$$

is called the *normalizer* of H in G.

Note that H is a normal subgroup of $N(H)$; in fact, it is not hard to see that $N(H)$ is the largest subgroup of G that contains H as a normal subgroup. For example, H is normal in G if and only if $N(H) = G$.

Returning to the case of covering spaces, Lemma 24.6 tells us that

$$\varepsilon_0\colon \operatorname{Aut}_B(E) \to \ell\big(N(H_0)\big)$$

is a bijection. To complete our description of the group $\operatorname{Aut}_B(E)$, it remains to compute $\ell\big(N(H_0)\big)$ in terms of the subgroup H_0.

Lemma 24.8. *The lifting correspondence ℓ induces a bijection*

$$\tilde{\ell}\colon N(H_0)/H_0 \to \ell\big(N(H_0)\big).$$

Proof. Recall that the elements of the quotient $N(H_0)/H_0$ are the left cosets $[\alpha]*H_0$ for $[\alpha] \in N(H_0)$; since $H_0 \trianglelefteq N(H_0)$ is normal, the quotient is itself a group.

Obviously, $\ell\colon N(H_0) \to \ell\big(N(H_0)\big)$ is surjective; thus it suffices to show that two elements $[\alpha], [\beta] \in N(H_0)$ have same image under ℓ if and only if $[\alpha]*H_0 = [\beta]*H_0$. Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the unique liftings to paths starting at e_0; then

$$\ell(\alpha_1) = \ell(\alpha_2) \quad\Longleftrightarrow\quad \tilde{\alpha}(1) = \tilde{\beta}(1).$$

If $\tilde{\alpha}(1) = \tilde{\beta}(1)$, then the composition of $\tilde{\alpha}$ and the inverse of $\tilde{\beta}$ is a loop based at e_0, and so $[\alpha] * [\beta]^{-1} \in H_0$. That being the case, we compute that

$$[\alpha] * H_0 = H_0 * [\alpha] = H_0 * [\alpha] * [\beta]^{-1} * [\beta] = H_0 * [\beta] = [\beta] * H_0,$$

where the first and last equalities hold because $[\alpha], [\beta] \in N(H_0)$. To prove the converse, run the same argument backwards. □

We can now consider the composition

$$\tilde{\ell}^{-1} \circ \varepsilon_0 \colon \operatorname{Aut}_B(E) \to N(H_0)/H_0,$$

and for the same reason as in Proposition 24.2, this bijection is actually a group isomorphism. Let me state the final conclusion in the form of a theorem.

Theorem 24.9. *Let $p \colon E \to B$ be a nice covering space. If $H_0 = p_*\big(\pi_1(E, e_0)\big)$ denotes the corresponding subgroup of $\pi_1(B, b_0)$, then*

$$\operatorname{Aut}_B(E) \simeq N(H_0)/H_0.$$

In words, the group of deck transformations is isomorphic to the quotient of the normalizer of H_0 by its normal subgroup H_0.

Unlike for the universal covering space, the group of deck transformations no longer acts transitively on the fibers in general. For example, if $N(H_0) = H_0$, then the only deck transformation of the corresponding covering space is the identity.

Definition 24.10. A nice covering space $p \colon E \to B$ is called *regular* if the group of deck transformations acts transitively on the fibers: for every two points $e_0, e_1 \in p^{-1}(b_0)$, there is a deck transformation taking e_0 to e_1.

This use of the word "regular" is different from the separation axioms, of course. The proof of the theorem shows that there is a deck transformation taking e_0 to e_1 if and only if $e_1 \in \ell\big(N(H_0)\big)$. If the subgroup H_0 is normal in G, then $N(H_0) = \pi_1(B, b_0)$, and so $\ell\big(N(H_0)\big) = p^{-1}(b_0)$; the proof of the converse is an easy exercise.

Corollary 24.11. *A nice covering space $p \colon E \to B$ is a regular covering space if and only if the corresponding subgroup H_0 of $\pi_1(B, b_0)$ is normal. In that case, the group of deck transformations is isomorphic to the quotient group $\pi_1(B, b_0)/H_0$.*

In particular, the fundamental group acts on any regular covering space.

Note. The correspondence between covering spaces and the fundamental group is basically the same as the correspondence between field extensions and subgroups of the Galois group. Since you may already know some Galois theory, let me briefly summarize what happens there. Given a field k, let $\bar{k}$ denote its (separable) algebraic closure; its analogue in topology is the universal covering space. The Galois group $G = \operatorname{Gal}(\bar{k}/k)$ consists of all field automorphisms of $\bar{k}$ that fix k; its analogue in topology is the group of deck transformations of the universal covering space (which, as we have seen, is isomorphic to the fundamental group). The Galois correspondence is a bijection between (separable) algebraic field extensions and subgroups of G: a field extension $k \subseteq L \subseteq \bar{k}$ corresponds to the subgroup $H = \operatorname{Gal}(\bar{k}/L)$ of all those field automorphisms that fix L. A field extension is called "normal" if every automorphism of $\bar{k}$ maps L into itself; in that case, L/k is itself a Galois extension. In terms of G, this is equivalent to $H \trianglelefteq G$, and one has the isomorphism $\operatorname{Gal}(L/k) \simeq G/H$. All of these facts have their analogues in topology.

Lecture 25: November 29

Scheduling. On December 6 (Tuesday), we are going to have an additional lecture (at 9:45 in P–117), to make up for the lecture I had to cancel due to COVID. The final exam is scheduled for the morning December 13 (Tuesday); more details about that to follow.

Homology. Our last topic this semester is homology theory. I will use the remaining three lectures to give a brief introduction to homology and discuss some applications (such as Brouwer's fixed point theorem and the Jordan curve theorem).

The most basic object in algebraic topology is the fundamental group. The problem with the fundamental group is that it is not that easy to compute, and that it only detects certain "low-dimensional" features of a space, basically because we are using loops (= images of $\mathbb{S}^1$) to probe the space. For instance, the fundamental group cannot distinguish spheres of dimension ≥ 2 from each other, because they are all simply connected. There is a higher-dimensional generalization of the fundamental group, the so-called higher homotopy groups; these are defined by looking at maps from $\mathbb{S}^n$ into a given space, up to homotopy. While relatively easy to define, the higher homotopy groups are very hard to compute: for example, people still do not know all the higher homotopy groups of spheres. The homology groups of a space are another invariant that is relatively easy to compute and that still contains a lot of higher-dimensional information about the space. (The drawback is that the definition of homology is more complicated.)

Today, I am going to describe how homology theory works, and what sort of properties it has, without saying anything about how it is defined. Next time, I will try to sketch the construction and the proofs of one or two of the basic theorems; and in the final lecture next week, I will give a few interesting applications.

Six properties of homology. Homology theory assigns to every topological space X a sequence of abelian groups $H_n(X)$, indexed by $n \in \mathbb{N}$, called its *homology groups*. The general idea is that the n-th homology group $H_n(X)$ contains some information about the n-dimensional features of X. I am now going to list six important properties of homology groups; when we talk about the definition of homology next time, we will see that each of these properties is really a (sometimes quite long) theorem.

The first property is *functoriality*. This is something that we already know from the fundamental group. It says that if $f\colon X \to Y$ is a continuous function, then one has induced group homomorphisms

$$f_*\colon H_n(X) \to H_n(Y),$$

one for each $n \in \mathbb{N}$. This assignment is compatible with composition: if $f\colon X \to Y$ and $g\colon Y \to Z$ are two continuous functions, then $(g \circ f)_* = g_* \circ f_*$. Moreover, the homomorphism assigned to the identity function $\mathrm{id}\colon X \to X$ is the identity homomorphism: $\mathrm{id}_* = \mathrm{id}$. (In other words, H_n is a functor from the category of topological spaces to the category of abelian groups.)

Example 25.1. Functoriality implies that homeomorphic spaces have isomorphic homology groups. Indeed, suppose that $f\colon X \to Y$ is a homeomorphism with inverse $g\colon Y \to X$. Then $g \circ f = \mathrm{id}$ and $f \circ g = \mathrm{id}$, and therefore

$$g_* \circ f_* = \mathrm{id}_* = \mathrm{id} \quad \text{and} \quad f_* \circ g_* = \mathrm{id}_* = \mathrm{id}\,.$$

This means that $f_*\colon H_n(X) \to H_n(Y)$ and $g_* H_n(Y) \to H_n(X)$ are inverse to each other, and so they are both isomorphisms.

The second property is a sort of *normalization*, to make sure that the theory does what it is supposed to do. It says that if X is the one-point space, then

$$H_n(X) \cong \begin{cases} \mathbb{Z} & \text{if } n = 0, \\ 0 & \text{if } n \neq 0. \end{cases}$$

Here 0 means the trivial group (with one element). This is sensible because a point is a 0-dimensional space with no higher-dimensional features.

Example 25.2. The second property implies that any nonempty topological space X has nontrivial 0-th homology. To see why, let $x \in X$ be an arbitrary point, and consider the inclusion $i\colon \{x\} \to X$ and the constant function $r\colon X \to \{x\}$. Then $r \circ i = \mathrm{id}$, and so by functoriality $r_* \circ i_* = \mathrm{id}$. Consequently, the homomorphism

$$i_*\colon H_0(\{x\}) \to H_0(X)$$

must be injective, and so $H_0(X)$ always contains a subgroup isomorphic to $\mathbb{Z}$.

The third property is called *additivity*. It says that if a topological space X is a disjoint union of open subspaces X_i, indexed by some set I, then

$$H_n(X) \cong \bigoplus_{i \in I} H_n(X_i)$$

for every $n \in \mathbb{N}$. The isomorphism is not just an abstract one, but it is given as follows: for each $i \in I$, the homomorphism $H_n(X_i) \to H_n(X)$ is the one associated to the inclusion $X_i \to X$ by functoriality, and the statement is that the sum of all these homomorphisms is an isomorphism.

Example 25.3. If X is a space with the discrete topology, we get

$$H_n(X) \cong \begin{cases} \bigoplus_X \mathbb{Z} & \text{if } n = 0, \\ 0 & \text{if } n \neq 0, \end{cases}$$

by combining the second and third property.

The fourth property is *homotopy invariance*, which is again something that we have seen already in the case of the fundamental group. It says that if $f, g\colon X \to Y$ are two continuous functions such that $f \sim g$, meaning f and g are homotopic, then $f_* = g_*$. Just like the fundental group, homology groups therefore only see a space up to homotopy.

Example 25.4. A contractible space has the same homology as a point. Recall that X is contractible if it there is a deformation retraction to a point $x_0 \in X$. If we let $i\colon \{x_0\} \to X$ be the inclusion, and $r\colon X \to \{x_0\}$ the retraction, then $i \circ r \sim \mathrm{id}$. Combining the first and fourth property, we get

$$i_* \circ r_* = (i \circ r)_* = \mathrm{id}_* = \mathrm{id},$$

and so $i_*\colon H_n(\{x_0\}) \to H_n(X)$ must be surjective. At the same time, r is a retraction, which means that $r \circ i = \mathrm{id}$. By the first property, this implies $r_* \circ i_* = \mathrm{id}$,

and so i_* is also injective, hence an isomorphism. Since we know the homology of a one-point space from the second property, we conclude that

$$H_n(X) \cong \begin{cases} \mathbb{Z} & \text{if } n = 0, \\ 0 & \text{if } n \neq 0. \end{cases}$$

This applies for example to $\mathbb{R}^d$, or to any convex subset of $\mathbb{R}^d$.

The remaining two properties are more complicated to state. They both involve the so-called "relative" homology groups. Let X be a topological space, and $A \subseteq X$ an arbitrary subset. In that situation, one also has a sequence of abelian groups $H_n(X, A)$, indexed by $n \in \mathbb{N}$, and called the *relative homology groups* of the pair (X, A). Roughly speaking, the relative homology groups only see the part of X that lies outside of A, and completely ignore what is happening inside the set A. (This is not the same as looking at the homology of the complement $X \setminus A$, though; a better approximation is to think of relative homology as the homology of the quotient space $X/\!\sim$, where the equivalence relation $\sim$ identifies all points of A with each other.)

Relative homology is again functorial: if we have two pairs (X, A) and (Y, B), and if $f\colon X \to Y$ is a continuous function with $f(A) \subseteq B$, then we get a sequence of group homomorphisms

$$f_*\colon H_n(X, A) \to H_n(Y, B),$$

compatible with composition. The relative homology groups include the usual (or "absolute") homology groups because $H_n(X) = H_n(X, \emptyset)$. In particular, functoriality for the identity function $(X, \emptyset) \to (X, A)$ gives us homomorphisms

$$H_n(X) \to H_n(X, A)$$

from homology to relative homology.

The fifth property is called *excision*. It says that if $A, B \subseteq X$ are two subsets such that $X = \operatorname{int} A \cup \operatorname{int} B$, then the inclusion $i\colon (B, A \cap B) \to (X, A)$ induces isomorphisms on relative homology, meaning that

$$i_*\colon H_n(B, A \cap B) \to H_n(X, A)$$

is an isomorphism for every $n \in \mathbb{N}$. Let me try to explain the intuition behind this. Since $X = \operatorname{int} A \cup \operatorname{int} B$, we have

$$X \setminus B \subseteq X \setminus \operatorname{int} B \subseteq \operatorname{int} A \subseteq A.$$

Excision is saying that we can excise (or cut out) the subset $X \setminus B$ from both A and X without changing the relative homology. This makes sense because relative homology is supposed to ignore what is going on inside the subset A. For technical reasons, we are not allowed to remove an arbitrary subset of A though – for example, removing all of A is forbidden because $H_n(X \setminus A) \neq H_n(X, A)$ in general. Instead, the chain of inclusions from above shows that the closure of the set $X \setminus B$ that we are removing needs to be contained in the interior $\operatorname{int} A$, and must therefore stay away from the boundary of A.

The sixth and final property relates the homology groups of X and A to the relative homology groups of the pair (X, A), but in a way that mixes together homology in different degrees. To state it properly, we first need to introduce "exact sequences", which is a very important concept in algebraic topology.

Definition 25.5. Suppose we have a sequence of homomorphisms of abelian groups

$$\cdots \longrightarrow C_{n+1} \xrightarrow{f_{n+1}} C_n \xrightarrow{f_n} C_{n-1} \longrightarrow \cdots$$

meaning that each C_n is an abelian group, and each $f_n \colon C_n \to C_{n-1}$ is a group homomorphism. We say that such a sequence is *exact* if $\ker f_n = \operatorname{im} f_{n+1}$ for all n.

In other words, if we have an element $x \in C_n$ with $f_n(x) = 0$, then there should exist some element $y \in C_{n+1}$ such that $x = f_{n+1}(y)$. Exactness implies also that compositions of successive homomorphisms are trivial: $f_n \circ f_{n+1} = 0$.

Example 25.6. Exactness can be used to restate many properties of group homomorphisms. For instance, consider a sequence of the form

$$0 \longrightarrow A \xrightarrow{f} B.$$

As the image of the trivial homomorphism $0 \to A$ is the trivial subgroup, exactness means that $\ker f = 0$, hence that f is injective. Similarly, a sequence of the form

$$B \xrightarrow{g} C \longrightarrow 0$$

is exact iff g is surjective, and a sequence of the form

$$0 \longrightarrow A \xrightarrow{f} B \longrightarrow 0$$

is exact iff f is an isomorphism.

Example 25.7. A exact sequence of the form

$$0 \longrightarrow A \xrightarrow{f} B \xrightarrow{g} C \longrightarrow 0$$

is called a *short exact sequence*. Exactness means that f is injective, g is surjective, and that $\ker g = \operatorname{im} f$. By the first isomorphism theorem, A is isomorphic to the subgroup $f(A) \subseteq B$, and $B/f(A) \cong C$, and so the short exact sequence expresses the fact that C is isomorphic to the quotient of B by a subgroup isomorphic to A.

The sixth property of homology is the *long exact sequence*. It says that if X is a topological space, and $A \subseteq X$ any subset, then the homology groups of X and A are related to the relative homology groups of (X, A) by an exact sequence

$$\cdots \to H_n(A) \to H_n(X) \to H_n(X, A) \to H_{n-1}(A) \to H_{n-1}(X) \to \cdots$$

This sequence is called a long exact sequence because it extends infinitely in both directions, with the convention that the (relative) homology groups are 0 for $n < 0$. Here $H_n(A) \to H_n(X)$ is induced by the inclusion $A \to X$ (by functoriality), and $H_n(X) \to H_n(X, A)$ is the homomorphism from absolute to relative homology. The remaining morphism $\delta \colon H_n(X, A) \to H_{n-1}(A)$ is new, and is called the *connecting homomorphism*. The connecting homomorphisms is functorial, in the sense that if $f \colon (X, A) \to (Y, B)$ is continuous function with $f(A) \subseteq B$, then $\delta \circ f_* = f_* \circ \delta$. This makes the entire long exact sequence functorial: in the diagram

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & H_n(A) & \longrightarrow & H_n(X) & \longrightarrow & H_n(X, A) & \xrightarrow{\delta} & H_{n-1}(A) & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle f_*} & & \downarrow{\scriptstyle f_*} & & \\
\cdots & \longrightarrow & H_n(B) & \longrightarrow & H_n(Y) & \longrightarrow & H_n(Y, B) & \xrightarrow{\delta} & H_{n-1}(B) & \longrightarrow & \cdots
\end{array}$$

all squares "commute", meaning that any compositions that start at the same group and end at the same group are equal.

Homology groups of spheres. We will see the definition of homology next time. In the remainder of today's class, I want to show you how to use the six properties to compute the homology groups of spheres. Here is the result.

Theorem 25.8. *Let $\mathbb{S}^d$ be the unit sphere in $\mathbb{R}^{d+1}$, and let B^d be the closed unit ball in $\mathbb{R}^d$. Then for $d \geq 1$, the following is true:*

(a) $H_n(\mathbb{S}^d) \cong \begin{cases} \mathbb{Z} & \text{if } n = 0 \text{ or } n = d, \\ 0 & \text{otherwise.} \end{cases}$

(b) $H_n(B^d, \mathbb{S}^{d-1}) \cong \begin{cases} \mathbb{Z} & \text{if } n = d, \\ 0 & \text{otherwise.} \end{cases}$

In (a), the case $d = 0$ is of course special: the 0-sphere has just two points, and so $H_0(\mathbb{S}^0) \cong \mathbb{Z} \oplus \mathbb{Z}$ (by properties two and three from above).

Proof. The proof is by induction on $d \geq 1$, using the long exact sequence and excision. We really care only about (a), but (b) is needed along the way to make the induction work out.

Let us start by showing that (a) in dimension $d-1$ implies (b) in dimension d. Consider the long exact sequence for the pair $(B^d, \mathbb{S}^{d-1})$. It reads

$$\cdots \to H_n(\mathbb{S}^{d-1}) \to H_n(B^d) \to H_n(B^d, \mathbb{S}^{d-1}) \to H_{n-1}(\mathbb{S}^{d-1}) \to H_{n-1}(B^d) \to \cdots$$

By induction on $d \geq 1$, we may assume that we already know all the homology groups of $\mathbb{S}^{d-1}$. Let us first consider the portion

$$\cdots \to H_0(\mathbb{S}^{d-1}) \to H_0(B^d) \to H_0(B^d, \mathbb{S}^{d-1}) \to 0$$

of the long exact sequence. As B^d is contractible, we know that $H_0(B^d) \cong \mathbb{Z}$. If we choose the point in the deformation retraction to lie on the boundary $\mathbb{S}^{d-1}$, then the argument we gave above shows that

$$H_0(\mathbb{S}^{d-1}) \to H_0(B^d)$$

must be surjective. The exactness of the squence now implies that the image of the homomorphism $H_0(B^d) \to H_0(B^d, \mathbb{S}^{d-1})$ must be trivial; since it is also surjective (because of the 0 at the right end), it follows that $H_0(B^d, \mathbb{S}^{d-1}) \cong 0$.

Next, we consider the portion

$$H_1(B^d) \to H_1(B^d, \mathbb{S}^{d-1}) \to H_0(\mathbb{S}^{d-1}) \to H_0(B^d) \to H_0(B^d, \mathbb{S}^{d-1})$$

of the long exact sequence. The two groups on the outside are both trivial, and we have $H_0(B^d) \cong \mathbb{Z}$ and $H_0(\mathbb{S}^{d-1}) \cong \mathbb{Z}$ if $d \geq 2$, and $\mathbb{Z} \oplus \mathbb{Z}$ if $d = 1$. In the first case, exactness of the sequence implies that $H_1(B^d, \mathbb{S}^{d-1}) \cong 0$; in the second case, it implies that $H_1(B^1, \mathbb{S}^0) \cong \mathbb{Z}$. This already proves (b) when $d = 1$.

To complete the proof of (b), we now consider (for $n \geq 2$) the portion

$$H_n(B^d) \to H_n(B^d, \mathbb{S}^{d-1}) \to H_{n-1}(\mathbb{S}^{d-1}) \to H_{n-1}(B^d)$$

of the long exact sequence. The two groups on the outside are again trivial, and so exactness, together with the inductive hypothesis, gives us

$$H_n(B^d, \mathbb{S}^{d-1}) \cong H_{n-1}(\mathbb{S}^{d-1}) \cong \begin{cases} \mathbb{Z} & \text{if } n = d, \\ 0 & \text{otherwise.} \end{cases}$$

We are done with the proof of the implication from (a) to (b).

The second half of the proof consists in showing that (b) in dimension d implies (a) in dimension d. We cover $\mathbb{S}^d$ by two closed sets A and B that are homeomorphic to B^d and that contain open neighborhoods of the upper respectively lower hemisphere. Since $\mathbb{S}^d = \operatorname{int} A \cup \operatorname{int} B$, excision gives us

$$H_n(\mathbb{S}^d, A) \cong H_n(B, A \cap B).$$

The intersection $A \cap B$ deformation retracts onto the boundary of B, which is homeomorphic to $\mathbb{S}^{d-1}$. This means that we can use exactly the same argument as in the proof that (b) implies (a) to show that

$$H_n(\mathbb{S}^d, A) \cong H_n(B, A \cap B) \cong \begin{cases} \mathbb{Z} & \text{if } n = d, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the long exact sequence for the pair $(\mathbb{S}^d, A)$, namely

$$\cdots \to H_n(A) \to H_n(\mathbb{S}^d) \to H_n(\mathbb{S}^d, A) \to H_{n-1}(A) \to H_{n-1}(\mathbb{S}^d) \to \cdots$$

Suppose first that $d = 1$. In that case, for $n \geq 2$, the sequence

$$H_n(A) \to H_n(\mathbb{S}^1) \to H_n(\mathbb{S}^1, A)$$

is exact, and because the two groups at the end are trivial, we get $H_n(\mathbb{S}^1) \cong 0$ for $n \geq 2$. The rest of the long exact sequence reads

$$0 \to H_1(\mathbb{S}^1) \to H_1(\mathbb{S}^1, A) \to H_0(A) \to H_0(\mathbb{S}^1) \to 0,$$

using the fact that $H_1(A)$ and $H_0(\mathbb{S}^d, A)$ are trivial. We know that $H_0(A) \cong \mathbb{Z}$, and that $H_0(\mathbb{S}^1)$ contains a subgroup isomorphic to $\mathbb{Z}$. Since the homomorphism $H_0(A) \to H_0(\mathbb{S}^1)$ is surjective (because of the 0 at the right end), it is then forced to be an isomorphism. By exactness, this means that $H_1(\mathbb{S}^1, A) \to H_0(A)$ must be the zero homomorphism; but then $H_1(\mathbb{S}^1) \to H_1(\mathbb{S}^1, A)$ is also surjective, and therefore also an isomorphism. This proves (a) when $d = 1$.

The remaining case $d \geq 2$ is much easier. The long exact sequence reads in part

$$H_{n+1}(\mathbb{S}^d, A) \to H_n(A) \to H_n(\mathbb{S}^d) \to H_n(\mathbb{S}^d, A) \to H_{n-1}(A)$$

For $n = 0$, both groups on the outside are trivial, and since $H_0(\mathbb{S}^d, A) \cong 0$, we get $H_0(\mathbb{S}^d) \cong H_0(A) \cong \mathbb{Z}$. For $n = d$, both groups on the outside are again trivial, and since $H_d(A) \cong 0$, we get $H_d(\mathbb{S}^d) \cong H_d(\mathbb{S}^d, A) \cong \mathbb{Z}$. For all other values of n, both $H_n(A)$ and $H_n(\mathbb{S}^d, A)$ are trivial, and therefore $H_n(\mathbb{S}^d)$ must also be trivial. This concludes the proof of (a) in all cases. □

Topological consequences. The computation of the groups $H_n(\mathbb{S}^d)$ has several nice consequences. First, it shows that homology is strong enough to distinguish spheres of different dimensions from each other. Even though $\mathbb{R}^n$ has no interesting homology, we can also use this fact to distinguish $\mathbb{R}^n$ for different values of n.

Corollary 25.9. *If $\mathbb{R}^n \cong \mathbb{R}^m$, then $n = m$.*

Proof. If $\mathbb{R}^n$ and $\mathbb{R}^m$ are homeomorphic, then their one-point compactifications $\mathbb{S}^n$ and $\mathbb{S}^m$ are also homeomorphic; by Theorem 25.8, this is only possible if $n = m$. □

Another application is that there are no retractions from the unit ball to its boundary; we had proved this earlier (in Corollary 18.5) for $d = 2$ with the help of the fundamental group.

Corollary 25.10. *There is no retraction of B^d onto $\mathbb{S}^{d-1}$ for $d \geq 1$.*

Proof. This is true for $d = 1$ because B^1 is connected. Assume now that $d \geq 2$. The homology group $H_{d-1}(\mathbb{S}^{d-1}) \cong \mathbb{Z}$ is nontrivial, whereas the group $H_{d-1}(B^d) \cong 0$ is trivial. If we had a retraction $r\colon B^d \to \mathbb{S}^{d-1}$, then $r \circ i = \mathrm{id}$, and so by functoriality, $r_* \circ i_* = \mathrm{id}$. This would mean that the homomorphism

$$r_*\colon H_{d-1}(B^d) \to H_{d-1}(\mathbb{S}^{d-1})$$

is surjective, which is clearly impossible. □

The same argument as in Lecture 18 now proves Brouwer's fixed point theorem in all dimensions.

Theorem 25.11 (Brouwer's fixed point theorem)**.** *Every continuous function*

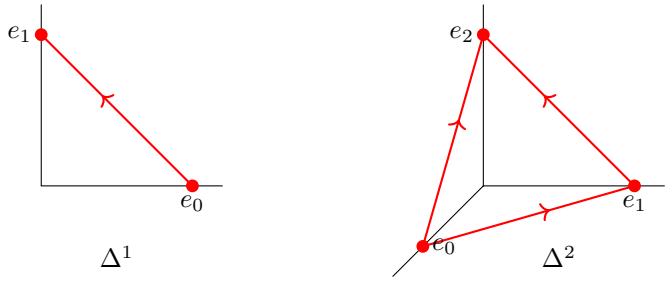
$$f\colon B^d \to B^d$$

has a fixed point: there is a point $x \in B^d$ with the property that $f(x) = x$.

Lecture 26: December 1

Definition of homology. Today, I want to show you the definition of (singular) homology and explain how some of the properties from last time are proved. Very vaguely, the idea is that the n-th homology group $H_n(X)$ is looking at continuous functions from n-dimensional shapes into X, but where we only use those shapes that can be assembled from simplices. For example, I mentioned in Lecture 21 that every compact surface can be triangulated; instead of mapping arbitrary compact surfaces into X, it is therefore enough to consider only triangles. As with the fundamental group, we also want to ignore those n-dimensional shapes that arise as the boundary of an $(n+1)$-dimensional shape.

Now let us see how homology theory makes this vague idea precise. We first define the standard n-simplex $\Delta^n \subseteq \mathbb{R}^{n+1}$ as the convex hull of the $n+1$ unit vectors $e_0, e_1, \dots, e_n$. We use the notation $\Delta^n = [e_0, e_1, \dots, e_n]$ to indicate that we are taking the convex hull, but that we are also remembering the order of the $n+1$ vectors; this gives us something like an orientation of the simplex.



The boundary of Δ^n consists of $n+1$ copies of the $(n-1)$-dimensional simplex Δ^{n-1}, up to identifying each of the $n+1$ coordinate hyperplanes with $\mathbb{R}^n$. If we use the induced ordering on each simplex in the boundary, we can write the i-th boundary simplex concisely as

$$[e_0, \dots, \hat{e}_i, \dots, e_n]$$

where the hat means that we drop the vector e_i from the list.

Definition 26.1. A *singular n-simplex* in a topological space X is just a continuous function $\sigma\colon \Delta^n \to X$.

The word "singular" comes from the fact that the image $\sigma(\Delta^n)$ may not look at all like a simplex: for example, a singular 1-simplex can be something like the Peano curve (that fills a whole square).

For each $n \in \mathbb{N}$, we then define $S_n(X)$ as the free abelian group generated by all singular n-simplices in X. In symbols,

$$S_n(X) = \bigoplus_{\sigma\colon \Delta^n \to X} \mathbb{Z}\sigma,$$

and so elements of $S_n(X)$ are finite sums of the form $a_1\sigma_1 + \cdots + a_k\sigma_k$ where $a_1, \dots, a_k$ are integers and $\sigma_1, \dots, \sigma_k$ are singular n-simplices. The idea is that if we have an n-dimensional space Y that is made by gluing together n-simplices, and a continuous function $f\colon Y \to X$, then the restriction of f to each simplex is a singular n-simplex in X, and instead of the function f, we consider the sum of all these singular n-simplices in the group $S_n(X)$.

Next, we have to deal with boundaries. Let $\sigma\colon \Delta^n \to X$ be a singular n-simplex. Restricting the function σ to the boundary of Δ^n gives us an element

$$\partial\sigma = \sum_{i=0}^{n} (-1)^i \sigma\big|_{[e_0,\ldots,\hat{e}_i,\ldots,e_n]} \in S_{n-1}(X).$$

Extended linearly, this defines the so-called *boundary operator*

$$\partial\colon S_n(X) \to S_{n-1}(X).$$

For example, the boundary of a singular 1-simplex $\sigma\colon [e_0, e_1] \to X$ is just the difference $\partial\sigma = \sigma(e_1) - \sigma(e_0)$ between the two end points; this is the reason for putting the sign factor $(-1)^i$. The key point is that the composition

$$S_n(X) \xrightarrow{\partial} S_{n-1}(X) \xrightarrow{\partial} S_{n-2}(X)$$

of two boundary operators is equal to zero (because of the signs).

Lemma 26.2. *We have $\partial \circ \partial = 0$.*

Proof. Since ∂ is a homomorphism, it is enough to show that $\partial(\partial\sigma) = 0$ for any singular n-simplex $\sigma\colon [e_0, \ldots, e_n] \to X$. But

$$\begin{aligned}
\partial(\partial\sigma) &= \partial\left(\sum_{i=0}^{n} (-1)^i \sigma\big|_{[e_0,\ldots,\hat{e}_i,\ldots,e_n]}\right) \\
&= \sum_{i=0}^{n} (-1)^i \left(\sum_{j=0}^{i-1} (-1)^j \sigma\big|_{[e_0,\ldots,\hat{e}_j,\ldots,\hat{e}_i,\ldots,e_n]} + \sum_{j=i+1}^{n} (-1)^{j-1} \sigma\big|_{[e_0,\ldots,\hat{e}_i,\ldots,\hat{e}_j,\ldots,e_n]}\right) \\
&= \sum_{j<i} (-1)^{i+j} \sigma\big|_{[e_0,\ldots,\hat{e}_j,\ldots,\hat{e}_i,\ldots,e_n]} - \sum_{i<j} (-1)^{i+j} \sigma\big|_{[e_0,\ldots,\hat{e}_i,\ldots,\hat{e}_j,\ldots,e_n]} = 0,
\end{aligned}$$

because the two sums are equal to each other. □

Definition 26.3. A sequence of homomorphisms of abelian groups

$$\cdots \xrightarrow{f_{n+2}} C_{n+1} \xrightarrow{f_{n+1}} C_n \xrightarrow{f_n} C_{n-1} \xrightarrow{f_{n-1}} \cdots$$

is called a *complex* if $f_n \circ f_{n+1} = 0$ for all n. The individual homomorphisms f_n are called the *differentials* in the complex, and we use the abbreviated notation $C_\bullet$ or $(C_\bullet, f_\bullet)$ to refer to the complex. The condition $f_n \circ f_{n+1} = 0$ means that $\operatorname{im} f_{n+1} \subseteq \ker f_n$, and we define the *$n$-th homology group* of the complex as

$$H_n(C_\bullet) = H_n(C_\bullet, f_\bullet) = \ker f_n / \operatorname{im} f_{n+1} = \frac{\ker f_n\colon C_n \to C_{n-1}}{\operatorname{im} f_{n+1}\colon C_{n+1} \to C_n}.$$

In particular, $H_n(C_\bullet) \cong 0$ if and only if the complex is exact at C_n.

The *singular chain complex* of a topological space X is the complex

$$\cdots \xrightarrow{\partial} S_n(X) \xrightarrow{\partial} S_{n-1}(X) \xrightarrow{\partial} \cdots \xrightarrow{\partial} S_1(X) \xrightarrow{\partial} S_0(X) \longrightarrow 0,$$

usually abbreviated as $S_\bullet(X)$. The *n-th singular homology group* of X is defined as the n-th homology group of this complex:

$$H_n(X) = H_n\big(S_\bullet(X), \partial\big) = \frac{\ker \partial\colon S_n(X) \to S_{n-1}(X)}{\operatorname{im} \partial\colon S_{n+1}(X) \to S_n(X)}$$

The elements in the kernel of ∂ are called n-cycles, and the elements in the image of ∂ are called n-boundaries.

Example 26.4. It is often easy to find elements in homology. For example

$$\sigma\colon \Delta^1 \to \mathbb{S}^1, \quad \sigma\big((1-t)e_0 + te_1\big) = \big(\cos(2\pi t), \sin(2\pi t)\big)$$

is a singular 1-simplex with $\partial\sigma = (1,0) - (1,0) = 0$, and so it defines an element in $H_1(\mathbb{S}^1)$. Similarly, the two-sphere $\mathbb{S}^2$ is homeomorphic to the surface of a cube, and if we decompose the six sides of the cube into a total of 12 triangles, we get a 2-cycle $\sigma_1 + \cdots + \sigma_{12} \in S_2(\mathbb{S}^2)$, and therefore an element in $H_2(\mathbb{S}^2)$. Of course, it is less easy to show that these elements are nonzero!

Even though the individual groups in the singular chain complex are gigantically large, the (singular) homology of reasonably nice spaces (such as manifolds) tends to be quite small. For nice spaces, singular homology is also very computable in practice, and often turns out to contain a lot of information. Very often, one can also find a much smaller complex whose homology groups are isomorphic to the singular homology groups: for example, if X can be triangulated, meaning glued together from simplices, then it is enough to use only the simplices that actually show up in the triangulation. The big advantage of the above definition is that it works for arbitrary topological spaces, which makes the theory very flexible.

About the only thing that we can compute directly from the definition is the 0-th homology group. It is related to the path components of the space X.

Lemma 26.5. *A topological space X is path connected iff $H_0(X) \cong \mathbb{Z}$.*

Proof. For the purposes of this proof, we are going to replace the 1-simplex Δ^1 by the unit interval $I = [0,1]$. Then $S_0(X)$ is the free abelian group generated by the points of X, and $S_1(X)$ is the free abelian group generated by paths $\sigma\colon I \to X$. The boundary operator is

$$\partial\colon S_1(X) \to S_0(X), \quad \partial(a_1\sigma_1 + \cdots + a_k\sigma_k) = \sum_{i=1}^{k} a_i\big(\sigma_i(1) - \sigma_i(0)\big),$$

and the 0-th homology group is therefore $H_0(X) = S_0(X)/\partial\big(S_1(X)\big)$. Let us first reformulate the condition that $H_0(X) \cong \mathbb{Z}$. Adding the coefficients of a 0-cycle defines a group homomorphism

$$\varepsilon\colon S_0(X) \to \mathbb{Z}, \quad \varepsilon(a_1x_1 + \cdots + a_kx_k) = a_1 + \cdots + a_k,$$

and it is easy to see that ε is surjective, and that $\varepsilon \circ \partial = 0$. It therefore induces a surjective homomorphism $H_0(X) \to \mathbb{Z}$. Now if $H_0(X) \cong \mathbb{Z}$, then this surjective homomorphism must be an isomorphism, which means that for any two points $x, y \in X$, the difference $y - x$ is zero in $H_0(X)$. But then

$$y - x = \sum_{i=1}^{k} a_i\big(\sigma_i(1) - \sigma_i(0)\big)$$

for some element $a_1\sigma_1 + \cdots + a_k\sigma_k \in S_1(X)$, and this implies that there is a path (made by joining some of the paths σ_i) connecting x and y. The converse is proved by running the same argument backwards. $\square$

Functoriality. Let us now take a look at some of the six properties from last time. The first property that we used was functoriality: A continuous function $f\colon X \to Y$ gives rise to group homomorphisms $f_*\colon H_n(X) \to H_n(Y)$ in a way that respects composition. This is very easy. If we have a singular n-simplex $\sigma\colon \Delta^n \to X$, then the composition $f \circ \sigma\colon \Delta^n \to Y$ is a singular n-simplex in Y. This defines

$$f_\sharp\colon S_n(X) \to S_n(Y), \quad f_\sharp(\sigma) = f \circ \sigma,$$

and it is easy to see from the formula for the boundary operator that $\partial \circ f_\sharp = f_\sharp \circ \partial$. Because of this identity, $f_\sharp$ takes n-cycles to n-cycles, and n-boundaries to n-boundaries, and so it descends to a group homomorphism

$$f_*\colon H_n(X) \to H_n(Y).$$

The identity $(f \circ g)_* = f_* \circ g_*$ is then just a consequence of fact that composition of functions is associative: $(f \circ g) \circ \sigma = f \circ (g \circ \sigma)$.

In fact, $f_\sharp$ is an example of a morphism of complexes.

Definition 26.6. Let $(A_\bullet, \partial)$ and $(B_\bullet, \partial)$ be two complexes. A *morphism of complexes* (or *chain map*) is a collection of group homomorphisms $f_n\colon A_n \to B_n$ that commute with the differentials in the two complexes: $\partial \circ f_{n+1} = f_n \circ \partial$ for every n. If this is the case, one says that the diagram

$$\begin{array}{ccccccccc}
\cdots & \xrightarrow{\partial} & A_{n+1} & \xrightarrow{\partial} & A_n & \xrightarrow{\partial} & A_{n-1} & \xrightarrow{\partial} & \cdots \\
 & & \downarrow{\scriptstyle f_{n+1}} & & \downarrow{\scriptstyle f_n} & & \downarrow{\scriptstyle f_{n-1}} & & \\
\cdots & \xrightarrow{\partial} & B_{n+1} & \xrightarrow{\partial} & B_n & \xrightarrow{\partial} & B_{n-1} & \xrightarrow{\partial} & \cdots
\end{array}$$

is commutative, meaning that the result of composing homomorphisms only depends on the source and the target group.

Any morphism of complexes $f\colon A_\bullet \to B_\bullet$ induces a homomorphism

$$f\colon H_n(A_\bullet) \to H_n(B_\bullet)$$

between the homology groups of the two complexes, again because $f(\ker \partial) \subseteq \ker \partial$ and $f(\operatorname{im} \partial) \subseteq \operatorname{im} \partial$.

The long exact sequence. I also want to talk about the sixth property from last time, namely the long exact sequence in homology. For that, we first need to define the relative homology groups $H_n(X, A)$, where X is a topological space and $A \subseteq X$ a subset. Let $i\colon A \to X$ be the inclusion. The induced homomorphism

$$i_\sharp\colon S_n(A) \to S_n(X)$$

is injective (because $A \subseteq X$), and we define

$$S_n(X, A) = S_n(X)/i_\sharp\big(S_n(A)\big)$$

as the quotient group. Since $\partial \circ i_\sharp = i_\sharp \circ \partial$, the boundary operator ∂ induces homomorphisms

$$\partial\colon S_n(X, A) \to S_{n-1}(X, A),$$

and of course we still have $\partial \circ \partial = 0$. The singular chain complex of the pair (X, A) is the complex

$$\cdots \xrightarrow{\partial} S_n(X, A) \xrightarrow{\partial} S_{n-1}(X, A) \xrightarrow{\partial} \cdots \xrightarrow{\partial} S_1(X, A) \xrightarrow{\partial} S_0(X, A) \longrightarrow 0,$$

and the n-th relative homology group is defined to be

$$H_n(X, A) = H_n\big(S_\bullet(X, A), \partial\big).$$

Because $S_n(X, A)$ is defined as the quotient of $S_n(X)$ by the subgroup $i_\sharp\big(S_n(A)\big)$, the sequence of group homomorphisms

$$0 \longrightarrow S_n(A) \xrightarrow{i_\sharp} S_n(X) \longrightarrow S_n(X, A) \longrightarrow 0$$

is a short exact sequence for every $n \in \mathbb{N}$. Together with three boundary operators, this gives us a big commutative diagram

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & S_{n+1}(A) & \xrightarrow{i_\sharp} & S_{n+1}(X) & \longrightarrow & S_{n+1}(X, A) & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & S_n(A) & \xrightarrow{i_\sharp} & S_n(X) & \longrightarrow & S_n(X, A) & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & S_{n-1}(A) & \xrightarrow{i_\sharp} & S_{n-1}(X) & \longrightarrow & S_{n-1}(X, A) & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
 & & \vdots & & \vdots & & \vdots & &
\end{array}$$

in which all the rows are short exact sequences, and in which the three columns are the three complexes $S_\bullet(A)$, $S_\bullet(X)$, and $S_\bullet(X, A)$. This is called a *short exact sequence of complexes*, and is usually abbreviated as

$$0 \longrightarrow S_\bullet(A) \xrightarrow{i_\sharp} S_\bullet(X) \longrightarrow S_\bullet(X, A) \longrightarrow 0.$$

The long exact sequence in homology is now a consequence of the following algebraic fact about complexes.

Theorem 26.7. *A short exact sequence of complexes*

$$0 \longrightarrow A_\bullet \xrightarrow{i} B_\bullet \xrightarrow{p} C_\bullet \longrightarrow 0$$

induces a long exact sequence

$$\cdots \longrightarrow H_n(A_\bullet) \xrightarrow{i_*} H_n(B_\bullet) \xrightarrow{p_*} H_n(C_\bullet) \xrightarrow{\delta} H_{n-1}(A_\bullet) \longrightarrow \cdots$$

among the homology groups of the three complexes.

Proof. The homomorphisms $i_*\colon H_n(A_\bullet) \to H_n(B_\bullet)$ and $p_*\colon H_n(B_\bullet) \to H_n(C_\bullet)$ are induced by the morphisms of complexes $i\colon A_\bullet \to B_\bullet$ and $p\colon B_\bullet \to C_\bullet$. Let me start by explaining the construction of $\delta\colon H_n(C_\bullet) \to H_{n-1}(A_\bullet)$, which is again called

the connecting homomorphism. For that, we need to draw another big diagram:

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & \vdots & & \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & A_{n+1} & \xrightarrow{i} & B_{n+1} & \xrightarrow{p} & C_{n+1} & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & A_n & \xrightarrow{i} & B_n & \xrightarrow{p} & C_n & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & A_{n-1} & \xrightarrow{i} & B_{n-1} & \xrightarrow{p} & C_{n-1} & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
0 & \longrightarrow & A_{n-2} & \xrightarrow{i} & B_{n-2} & \xrightarrow{p} & C_{n-2} & \longrightarrow & 0 \\
 & & \downarrow\partial & & \downarrow\partial & & \downarrow\partial & & \\
 & & \vdots & & \vdots & & \vdots & &
\end{array}$$

I repeat that the diagram is commutative, and that all the rows are short exact sequences. We need to construct a homomorphism

$$\delta\colon \frac{\ker\partial\colon C_n \to C_{n-1}}{\operatorname{im}\partial\colon C_{n+1} \to C_n} \to \frac{\ker\partial\colon A_{n-1} \to A_{n-2}}{\operatorname{im}\partial\colon A_n \to A_{n-1}}.$$

This is done by a method called "diagram chasing". Let $c \in C_n$ be an element such that $\partial c = 0$. Since $p\colon C_n \to B_n$ is surjective, we can find $b \in B_n$ with $p(b) = c$; because the n-th row is exact at B_n, the choice of b is unique up to elements of the form $i(a)$ for $a \in A_n$. Now $\partial b \in B_{n-1}$ satisfies

$$p(\partial b) = \partial p(b) = \partial c = 0,$$

and by exactness of the $(n-1)$-th row, there is a unique element $a' \in A_{n-1}$ with the property that $i(a') = \partial b$. We have

$$i(\partial a') = \partial i(a') = \partial(\partial b) = 0,$$

and since $i\colon A_{n-2} \to B_{n-2}$ is injective, it follows that $\partial a' = 0$, and so a' gives us an element in $H_{n-1}(A_\bullet)$. We now want to set

$$\delta[c] = [a'],$$

where the brackets stand for homology classes. Why is this well-defined? There are two choices involved in the construction. First, we can change the element $b \in B_n$ to something of the form $b + i(a)$, where $a \in A_n$. But then

$$\partial\big(b + i(a)\big) = \partial b + i(\partial a) = i(a' + \partial a),$$

and as $[a'] = [a' + \partial a]$, the resulting homology class is the same. Second, we could choose a different representative in the homology class $[c]$, adding to $c \in C_n$ any element of the form $c'' \in C_{n+1}$. As before, $c'' = p(b'')$ for some $b'' \in B_{n+1}$, and since

$$p(b + \partial b'') = c + \partial p(b'') = c + \partial c'',$$

we need to change $b \in B_n$ into $b + \partial b''$. But $\partial(b + \partial b'') = \partial b$, and so the element $a' \in A_{n-1}$ is unchanged. This shows that δ is independent of any choices.

One then has to check that the sequence of homology groups is exact in every place. As an example, let me show why

$$\ker \delta = \operatorname{im} p_*.$$

Suppose that we have a homology class $[c] \in H_n(C_\bullet)$ for which $\delta[c] = [a'] = 0$. This means that $a' = \partial a$ for some $a \in A_n$. We then have

$$\partial b = i(a) = i(\partial a) = \partial i(a),$$

and so the element $b - i(a) \in B_n$ is in the kernel of ∂, and therefore defines a homology class $[b - i(a)] \in H_n(B_\bullet)$. But

$$p\big(b - i(a)\big) = p(b) = c,$$

and this gives us $p_\bullet[b - i(a)] = [c]$, as required. □

The diagram chasing in the proof is typical for what is called "homological algebra", which studies algebraic properties of complexes and their homology groups.

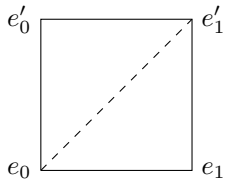
Homotopy invariance. In class, I did not have enough time to talk about the fourth property of homology, namely homotopy invariance. Recall that if $f, g\colon X \to Y$ are homotopic, then $f_* = g_*$. The idea of the proof is quite simple. By assumption, there is a homotopy $H\colon X \times I \to Y$ such that $H(x,0) = f(x)$ and $H(x,1) = g(x)$. Now for any singular n-simplex $\sigma\colon \Delta^n \to X$, the function

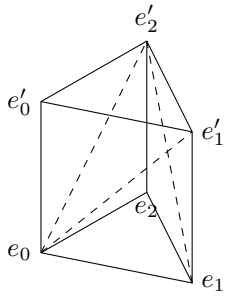
$$\Delta^n \times I \to Y, \quad (x,t) \mapsto H\big(\sigma(x), t\big),$$

continuously deforms $f \circ \sigma$ (at $t = 0$) into $g \circ \sigma$ (at $t = 1$), and this process should not change homology classes. But homology is defined in terms of simplices, and so we need to subdivide $\Delta^n \times I$ into $(n+1)$-simplices. This can be done as follows. In the product $\Delta^n \times I \subseteq \mathbb{R}^{n+1} \times \mathbb{R}$, we label the $n+1$ vertices of $\Delta^n \times \{0\}$ as $e_0, \dots, e_n$, and the $n+1$ vertices of $\Delta^n \times \{1\}$ as $e'_0, \dots, e'_n$. We then get a subdivision of $\Delta^n \times I$ into $n+1$ copies of Δ^{n+1}, which look like

$$[e_0, \dots, e_i, e'_i, \dots, e'_n] \quad \text{for } i = 0, \dots, n.$$

Here are two pictures of what this looks like for $n = 1$ and $n = 2$:





Using this subdivision of the prism $\Delta^n \times I$, we can now define the *prism operator*

$$P\colon S_n(X) \to S_{n+1}(Y)$$

with the help of the homotopy $H\colon X \times I \to Y$ by the formula

$$P(\sigma) = \sum_{i=0}^{n} (-1)^i H \circ (\sigma \times \mathrm{id})\big|_{[e_0,\dots,e_i,e'_i,\dots,e'_n]}.$$

Here $H \circ (\sigma \times \mathrm{id})$ is the composition $\Delta^n \times I \to X \times I \to Y$. A slightly lengthy computation then gives the fundamental identity

$$\partial \circ P = g_\sharp - f_\sharp - P \circ \partial.$$

Roughly speaking, the left-hand side represents the boundary of the prism, and the three summands on the right-hand side represent the top, bottom, and sides of the prism. The identity is saying that the two morphisms of complexes $f_\sharp, g_\sharp \colon S_\bullet(X) \to S_\bullet(Y)$ are chain homotopic, in the following sense.

Definition 26.8. Let $f \colon A_\bullet \to B_\bullet$ be a morphism of complexes. We say that f is *null homotopic*, in symbols $f \sim 0$, if there is a collection of homomorphisms $P \colon A_n \to B_{n+1}$ with the property that

$$f = \partial \circ P + P \circ \partial.$$

We say that two morphisms of complexes f and g are *chain homotopic* if their difference $f - g$ is null homotopic, meaning that $f - g \sim 0$.

Schematically, a null homotopy looks like the following:

$$\begin{array}{ccccccccc}
\cdots & \longrightarrow & A_{n+1} & \xrightarrow{\partial} & A_n & \xrightarrow{\partial} & A_{n-1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle f} & \swarrow{\scriptstyle P} & \downarrow{\scriptstyle f} & \swarrow{\scriptstyle P} & \downarrow{\scriptstyle f} & & \\
\cdots & \longrightarrow & B_{n+1} & \xrightarrow{\partial} & B_n & \xrightarrow{\partial} & B_{n-1} & \longrightarrow & \cdots
\end{array}$$

So the existence of a homotopy between $f, g \colon X \to Y$ means that the two morphisms of complexes $f_\sharp, g_\sharp \colon S_\bullet(X) \to S_\bullet(Y)$ are chain homotopic. The identity $f_* = g_*$ is then a consequence of the following algebraic lemma.

Lemma 26.9. *Let $f \colon A_\bullet \to B_\bullet$ be a morphism of complexes. If $f \sim 0$, then the induced homomorphism on homology satisfies $f_* = 0$.*

Proof. Let $[a] \in H_n(A_\bullet)$ be an arbitrary homology class, represented by an element $a \in A_n$ with $\partial a = 0$. Since $f \sim 0$, there are homomorphisms $P \colon A_n \to B_{n+1}$ for which $f = \partial \circ P + P \circ \partial$. Consequently,

$$f(a) = \partial P(a) + P(\partial a) = \partial P(a),$$

and on the level of homology, this gives $f_*[a] = [f(a)] = [\partial P(a)] = 0$. □

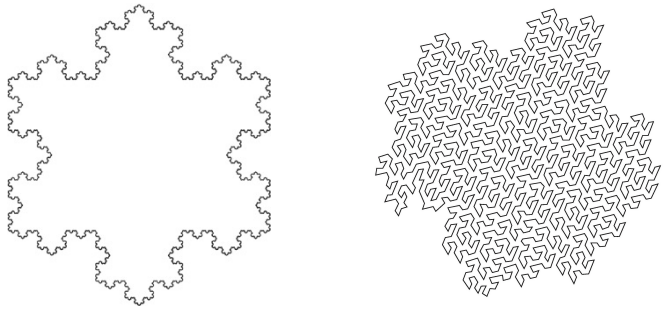
Lecture 27: December 6

The Jordan curve theorem. In this final lecture, I want to show you how homology can be used to prove two classical results from topology: the Jordan curve theorem and Brouwer's theorem on the invariance of domain. Let us start with the Jordan curve theorem. I am sure all of you have heard about this result at some point: any closed curve in $\mathbb{R}^2$ divides the plane into two regions, one inside the curve, the other outside the curve.

More precisely, a closed curve in the plane is a continuous function $f\colon [0,1] \to \mathbb{R}^2$ with $f(0) = f(1)$; if we identify the two endpoints of the interval, we obtain a circle, and so we can just as well say that a closed curve is a continuous function $f\colon \mathbb{S}^1 \to \mathbb{R}^2$. This definition includes space-filling curves: the image of the Peano curve, for example, would be the entire unit square, which obviously fails to divide $\mathbb{R}^2$ into two regions. Jordan discovered that this problem goes away if we consider *simple* closed curves (also called *Jordan curves*), where the function $f\colon \mathbb{S}^1 \to \mathbb{R}^2$ is injective. Note that for such curves, the image $f(\mathbb{S}^1)$ is homeomorphic to $\mathbb{S}^1$, due to the compactness of $\mathbb{S}^1$. Jordan curves are closer to our intuitive notion of curve.

Theorem 1 (Jordan)**.** *If $C \subseteq \mathbb{R}^2$ is a simple closed curve, then $\mathbb{R}^2 \setminus C$ has exactly two connected components, each of which has C as its boundary.*

This theorem was proved by Jordan in the late 19th century. One often reads that his proof was unsatisfactory and that the first correct one is due to Veblen; but Hales, who wrote the first computer-checkable proof of the Jordan curve theorem, says that: "In view of the heavy criticism of Jordan's proof, I was surprised when I sat down to read his proof to find nothing objectionable about it."



The tricky thing is that most simple closed curves do not look at all like a circle. There are fractal curves (like the Koch snowflake, pictured above), and even for curves that are not fractals, it can be very hard to decide by looking at the curve whether a given point lies inside or outside the curve.

Here are some basic observations. Let $h\colon \mathbb{S}^1 \to \mathbb{R}^2$ be a simple closed curve, and denote by $C = h(\mathbb{S}^1)$ its image in $\mathbb{R}^2$.

(1) The curve C is compact, and h is a homeomorphism between $\mathbb{S}^1$ and C; in particular, we can forget about the parametrization and remember only the subset $C \subseteq \mathbb{R}^2$. This follows immediately from Corollary 7.2.
(2) The complement $\mathbb{R}^2 \setminus C$ is open and locally path connected; every connected component of $\mathbb{R}^2 \setminus C$ is thus open and path connected (Proposition 6.7).
(3) Exactly one of the connected components of $\mathbb{R}^2 \setminus C$ is unbounded. The reason is that C is contained in a large disk, whose complement is path connected, hence contained in exactly one connected component of $\mathbb{R}^2 \setminus C$.

(4) After replacing $\mathbb{R}^2$ by its one-point compactification $\mathbb{S}^2$, we may assume that $h\colon \mathbb{S}^1 \to \mathbb{S}^2$ is a simple closed curve on the sphere; the unbounded component of $\mathbb{R}^2 \setminus C$ then becomes the component containing the point at infinity.

The Mayer-Vietoris sequence. The proof I am going to present uses the so-called "Mayer-Vietoris sequence", which is an analogue of the Seifert-van Kampen theorem for homology groups. Let X be a topological space, and let $U, V \subseteq X$ be two open subsets such that $X = U \cup V$. Let us denote by $i\colon U \to X$ and $j\colon V \to X$ the two inclusions. The excision property for homology says that

$$i_*\colon H_n(U, U \cap V) \cong H_n(X, V)$$

is an isomorphism for every $n \geq 0$. From the two pairs $(U, U \cap V)$ and (X, V), we also get two long exact sequences

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & H_n(U \cap V) & \xrightarrow{j_*} & H_n(U) & \longrightarrow & H_n(U, U \cap V) & \xrightarrow{\delta} & H_{n-1}(U \cap V) & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle i_*} & & \downarrow{\scriptstyle i_*} & & \downarrow{\scriptstyle i_*} & & \downarrow{\scriptstyle i_*} & & \\
\cdots & \longrightarrow & H_n(V) & \xrightarrow{j_*} & H_n(X) & \longrightarrow & H_n(X, V) & \xrightarrow{\delta} & H_{n-1}(V) & \longrightarrow & \cdots
\end{array}$$

in homology; here I am using the same letters $i\colon U \cap V \to V$ and $j\colon U \cap V \to U$ for the inclusions of $U \cap V$ into U and V. The diagram above is again commutative, because of the functoriality of the long exact sequence.

Theorem 27.1. *In the situation above, one gets a long exact sequence*

$$\cdots \longrightarrow H_n(U \cap V) \xrightarrow{\alpha} H_n(U) \oplus H_n(V) \xrightarrow{\beta} H_n(X) \xrightarrow{\gamma} H_{n-1}(U \cap V) \longrightarrow \cdots$$

where the individual morphisms are

$$\alpha\colon H_n(U \cap V) \to H_n(U) \oplus H_n(V), \quad \alpha(x) = \big(j_*(x), i_*(x)\big),$$
$$\beta\colon H_n(U) \oplus H_n(V) \to H_n(X), \quad \beta(y, z) = j_*(y) - i_*(z),$$

and where $\gamma\colon H_n(X) \to H_{n-1}(U \cap V)$ is another connecting homomorphism.

The exact sequence in the theorem is called the *Mayer-Vietoris sequence*.

Proof. This is another general result in homological algebra. Suppose that we have a commutative diagram of abelian groups

$$\begin{array}{ccccccccccc}
\cdots & \longrightarrow & A'_n & \xrightarrow{i'} & B'_n & \xrightarrow{p'} & C'_n & \xrightarrow{\delta'} & A'_{n-1} & \longrightarrow & \cdots \\
 & & \downarrow{\scriptstyle f} & & \downarrow{\scriptstyle g} & & \downarrow{\scriptstyle h} & & \downarrow & & \\
\cdots & \longrightarrow & A_n & \xrightarrow{i} & B_n & \xrightarrow{p} & C_n & \xrightarrow{\delta} & A_{n-1} & \longrightarrow & \cdots
\end{array}$$

in which both rows are long exact sequences, and in which all the homomorphisms $h\colon C'_n \to C_n$ are isomorphisms. In this situation, we get a long exact sequence

$$\cdots \longrightarrow A'_n \xrightarrow{\alpha} A_n \oplus B'_n \xrightarrow{\beta} B_n \xrightarrow{\gamma} A'_{n-1} \longrightarrow \cdots \tag{27.2}$$

in which $\alpha(x) = \big(f(x), i'(x)\big)$ and $\beta(y, z) = i(y) - g(z)$. The proof is another application of "diagram chasing".

Let me explain how the connecting homomorphism $\gamma\colon B_n \to A'_{n-1}$ is constructed in this case. Take an element $b \in B_n$. Then $p(b) \in C_n$, and since $h\colon C'_n \to C_n$ is an isomorphism, one has $p(b) = h(c')$ for a unique element $c \in C'_n$, and we define

$\gamma(b) = \delta'(c) \in A'_{n-1}$. The rest of the proof consists in checking that the sequence in (27.2) is exact in all places.

For the sake of illustration, let me show you why $\ker\gamma = \operatorname{im}\beta$. Suppose that $b \in B_n$ satisfies $\gamma(b) = \delta'(c) = 0$. Since the top row is exact, we have $\ker\delta' = \operatorname{im}p'$, and so there is an element $b' \in B'_n$ with $p'(b') = c'$. The difference $b - g(b')$ satisfies

$$p\big(b - g(b')\big) = p(b) - p\big(g(b')\big) = c - h\big(p'(b')\big) = c - h(c') = 0,$$

and therefore lies in $\ker p = \operatorname{im} i$ (by exactness of the bottom row). This gives us an element $a \in A_n$ such that $b - g(b') = i(a)$; but then

$$b = i(a) + g(b') = \beta(a, -b')$$

belongs to the image of β, as claimed. □

The generalized Jordan curve theorem. We can use the Mayer-Vietoris sequence to prove the following generalization of the Jordan curve theorem.

Theorem 27.3. *Let $n \geq 1$ be an integer.*

(a) If $h\colon B^{n-1} \to \mathbb{S}^n$ is injective and continuous, then

$$H_i\big(\mathbb{S}^n \setminus h(B^{n-1})\big) \cong \begin{cases} \mathbb{Z} & \text{for } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

(b) If $h\colon \mathbb{S}^{n-1} \to \mathbb{S}^n$ is injective and continuous, then

$$H_i\big(\mathbb{S}^n \setminus h(\mathbb{S}^{n-1})\big) \cong \begin{cases} \mathbb{Z} \oplus \mathbb{Z} & \text{for } i = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Since $\mathbb{S}^{n-1}$ is compact and $\mathbb{S}^n$ is Hausdorff, every injective continuous function $h\colon \mathbb{S}^{n-1} \to \mathbb{S}^n$ is a homeomorphism between $\mathbb{S}^{n-1}$ and its image $h(\mathbb{S}^{n-1})$. The image $h(\mathbb{S}^{n-1})$ is therefore a higher-dimensional version of a Jordan curve, and the theorem is saying that the complement of $h(\mathbb{S}^{n-1})$ has exactly two connected components (by Lemma 26.5).

Proof. In the interest of time, I will give the proof only for $n = 2$; the general case is similar. Let us first prove (a). When $n = 2$, the closed unit ball is $B^1 = [-1, 1]$. Consider then an injective continuous function $h\colon [-1, 1] \to \mathbb{S}^2$. The image $h(B^1)$ is an arc, and the main point is that the complement of an arc in $\mathbb{S}^2$ remains connected.

The proof uses a nice trick. By dividing the interval $[-1, 1]$ into the two halves $[-1, 0]$ and $[0, 1]$, we get two open sets

$$U = \mathbb{S}^2 \setminus h([-1, 0]) \quad \text{and} \quad V = \mathbb{S}^2 \setminus h([0, 1]).$$

The intersection is $U \cap V = \mathbb{S}^2 \setminus h(B^1)$, and the union is $U \cup V = \mathbb{S}^2 \setminus \{h(0)\} \cong \mathbb{R}^2$. For $i \geq 0$, the Mayer-Vietoris sequence gives us an exact sequence

$$H_{i+1}(U \cup V) \to H_i(U \cap V) \to H_i(U) \oplus H_i(V),$$

and since we know the homology of $\mathbb{R}^2$, this becomes

$$0 \to H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_i\big(\mathbb{S}^2 \setminus h([-1, 0])\big) \oplus H_i\big(\mathbb{S}^2 \setminus h([0, 1])\big).$$

In other words, the homomorphism in the center is injective. If $x \in H_i\big(\mathbb{S}^2 \setminus h(B^1)\big)$ is an arbitrary element, then its image under at least one of the two homomorphisms

$$H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_i\big(\mathbb{S}^2 \setminus h([-1,0])\big)$$
$$H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_i\big(\mathbb{S}^2 \setminus h([0,1])\big)$$

must be nonzero, unless $x = 0$.

We can repeat this process, and in this way, we obtain a nested chain

$$I_0 = [-1,1] \supseteq I_1 \supseteq I_2 \supseteq \cdots$$

in which I_k is a closed interval of length 2^{1-k}, such that the image of our fixed element $x \in H_i\big(\mathbb{S}^2 \setminus h(B^1)\big)$ under the homomorphism

$$H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_i\big(\mathbb{S}^2 \setminus h(I_k)\big)$$

is nonzero for every $k \geq 0$, except if $x = 0$. Let $t \in I_0$ be the unique point in the intersection of all the intervals I_k. Note that $\mathbb{S}^2 \setminus h(t)$ is homeomorphic to $\mathbb{R}^2$.

We can now use what we know about homology groups to deduce (a). Consider first the case $i > 0$. Our homology class x is represented by an i-cycle $x \in S_i\big(\mathbb{S}^2 \setminus h(B^1)\big)$. Since $H_i(\mathbb{R}^2) \cong 0$, the image of x under the homomorphism

$$H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_i\big(\mathbb{S}^2 \setminus h(t)\big)$$

is zero, and so there is some $y \in S_{i+1}\big(\mathbb{S}^2 \setminus h(t)\big)$ such that $x = \partial y$. But the image of y is compact, and since $\mathbb{S}^2 \setminus h(t)$ is the union of the open sets $\mathbb{S}^2 \setminus h(I_k)$, it follows that $y \in S_{i+1}\big(\mathbb{S}^2 \setminus h(I_k)\big)$ for some $k \geq 0$. But then $x = \partial y$ means that the image of x in $H_i\big(\mathbb{S}^2 \setminus h(I_k)\big)$ is zero, and by construction, this implies that $x = 0$. The conclusion is that $H_i\big(\mathbb{S}^2 \setminus h(B^1)\big) \cong 0$ for $i > 1$.

In the remaining case $i = 0$, we know from Lemma 26.5 that the 0-th homology of $\mathbb{S}^2 \setminus h(B^1)$ is always at least $\mathbb{Z}$. After subtracting from $x \in S_0\big(\mathbb{S}^2 \setminus h(B^1)\big)$ a suitable multiple of the homology class of a point, we may therefore assume that the coefficients of the 0-cycle x add up to zero. Now $H_0(\mathbb{R}^2) \cong \mathbb{Z}$, and so the image of x under the homomorphism

$$H_0\big(\mathbb{S}^2 \setminus h(B^1)\big) \to H_0\big(\mathbb{S}^2 \setminus h(t)\big)$$

is equal to zero. We can then argue as before to show that $x = 0$, and hence that $H_0\big(\mathbb{S}^2 \setminus h(B^1)\big) \cong \mathbb{Z}$. This proves (a) for $n = 2$.

Now let us deal with (b). Suppose that $h\colon \mathbb{S}^1 \to \mathbb{S}^2$ is continuous and injective, and denote by $C = h(\mathbb{S}^1)$ the image curve. Pick two distinct points $p, q \in C$, and let A be the portion of the curve from p to q, and B the remaining portion from q to p. Then $C = A \cup B$ and $A \cap B = \{p, q\}$. This gives us two open sets

$$U = \mathbb{S}^2 \setminus A \quad \text{and} \quad V = \mathbb{S}^2 \setminus B$$

with $U \cap V = \mathbb{S}^2 \setminus C$ and $U \cup V = \mathbb{S}^2 \setminus \{p, q\}$. Observe that $U \cup V$ is homeomorphic to $\mathbb{R}^2$ minus a point, and therefore deformation retracts onto $\mathbb{S}^1$. From the Mayer-Vietoris sequence, we get an exact sequence

$$H_1(U) \oplus H_1(V) \to H_1(U \cup V) \to H_0(U \cap V) \to H_0(U) \oplus H_0(V) \to H_0(U \cup V) \to 0,$$

and since we know the homology of U and B by part (a), this becomes

$$0 \to \mathbb{Z} \to H_0(\mathbb{S}^2 \setminus C) \to \mathbb{Z} \oplus \mathbb{Z} \to \mathbb{Z} \to 0.$$

It is then a simple algebra exercise to deduce that $H_0(\mathbb{S}^2 \setminus C) \cong \mathbb{Z} \oplus \mathbb{Z}$. For $i > 0$, the Mayer-Vietoris sequence gives the exactness of

$$H_{i+1}(U \cup V) \to H_i(U \cap V) \to H_i(U) \oplus H_i(V),$$

and since the two groups on the outside are trivial by (a), we get $H_i(\mathbb{S}^2 \setminus C) \cong 0$ for $i > 0$. □

The two connected components in the Jordan curve theorem are therefore explained by the fact that $H_0(\mathbb{S}^2 \setminus C) \cong \mathbb{Z} \oplus \mathbb{Z}$. Homology theory gives a very natural proof of this fact.

Invariance of domain. Another classical result that can be proved in this manner is Brouwer's theorem about the invariance of domain. It is a generalization of the fact that $\mathbb{R}^n$ and $\mathbb{R}^m$ are not homeomorphic when $m \cong n$, but Brouwer's result is much stronger.

Theorem 27.4 (Invariance of Domain)**.** *Let $U \subseteq \mathbb{R}^n$ be an open subset. If $h\colon U \to \mathbb{R}^n$ is injective and continuous, then $h(U)$ is also an open subset of $\mathbb{R}^n$.*

Proof. After replacing $\mathbb{R}^n$ by its one-point compactification, we may assume that $h\colon U \to \mathbb{S}^n$ is injective and continuous. It is enough to show that for any point $x \in U$, the image $h(U)$ contains an open neighborhood of $h(x)$. Let $B \subseteq U$ be a closed ball of positive radius containing the point x, and let $S = \partial B$ be its boundary; then $B \cong B^n$ and $S \cong \mathbb{S}^{n-1}$. We have

$$\mathbb{S}^n \setminus h(S) = h(B \setminus S) \sqcup \mathbb{S}^n \setminus h(B).$$

Now $h(B \setminus S)$ is path connected (because $B \setminus S$ is path connected), and $\mathbb{S}^n \setminus h(B)$ is path connected (because $H_0\big(\mathbb{S}^n \setminus h(B)\big) \cong \mathbb{Z}$). Therefore $\mathbb{S}^n \setminus h(S)$ has exactly two path connected components, and since $\mathbb{S}^n$ is locally path connected, both of them must be open. Therefore $h(B \setminus S)$ is an open neighborhood of the point $h(x)$ that is contained in $h(U)$, and so $h(U)$ is open. □

Recall that the word "domain" is used in analysis to refer to open subsets of $\mathbb{R}^n$. Brouwer's result tells us that if we take an open subset in $\mathbb{R}^n$ and embed it into $\mathbb{R}^n$ in a possibly different way, the image will again be an open subset.

Corollary 27.5. *If $U \subseteq \mathbb{R}^n$ is a nonempty open subset, then U is not homeomorphic to a subset of $\mathbb{R}^m$ for $m < n$.*

Proof. Suppose to the contrary that we had a homeomorphism $h\colon U \to V$ for some $V \subseteq \mathbb{R}^m$. Now $\mathbb{R}^m$ is a proper linear subspace of $\mathbb{R}^n$, and so we obtain an injective continuous function

$$f\colon U \to \mathbb{R}^n, \quad f(x) = \big(h_1(x), \dots, h_m(x), 0, \dots, 0\big).$$

The image is obviously not an open subset of $\mathbb{R}^n$, in contradiction to Brouwer's theorem. □

Extra Lecture: Invariance of Domain

Invariance of domain. The next topic I wish to discuss is a famous result called the *invariance of domain*; roughly speaking, it says that $\mathbb{R}^m$ and $\mathbb{R}^n$ are not homeomorphic unless $m = n$. This result is of great importance in the theory of manifolds, because it means that the dimension of a topological manifold is well-defined. Recall that an m-dimensional manifold is a (second countable, Hausdorff) topological space in which every point has a neighborhood homeomorphic to an open subset of $\mathbb{R}^m$. If an open set in $\mathbb{R}^m$ could be homeomorphic to an open set in $\mathbb{R}^n$, it would not make sense to speak of the dimension of a manifold.

The first person to show that this cannot happen was the Dutch mathematician Brouwer (who later became one of the founders of "intuitionist mathematics"); in fact, he proved the following stronger theorem.

Theorem 2 (Invariance of Domain)**.** *Let $U \subseteq \mathbb{R}^n$ be an open subset. If $f\colon U \to \mathbb{R}^n$ is injective and continuous, then $f(U)$ is also an open subset of $\mathbb{R}^n$.*

Recall that the word "domain" is used in analysis to refer to open subsets of $\mathbb{R}^n$. Brouwer's result tells us that if we take an open subset in $\mathbb{R}^n$ and embed it into $\mathbb{R}^n$ in a possibly different way, the image will again be an open subset.

Corollary 3. *If $U \subseteq \mathbb{R}^n$ is a nonempty open subset, then U is not homeomorphic to a subset of $\mathbb{R}^m$ for $m < n$. In particular, $\mathbb{R}^n$ is not homeomorphic to $\mathbb{R}^m$ for $m < n$.*

Proof. Suppose to the contrary that we had a homeomorphism $h\colon U \to V$ for some $V \subseteq \mathbb{R}^m$. Now $\mathbb{R}^m$ is a proper linear subspace of $\mathbb{R}^n$, and so we obtain an injective continuous function

$$f\colon U \to \mathbb{R}^n, \quad f(x) = \big(h_1(x), \dots, h_m(x), 0, \dots, 0\big).$$

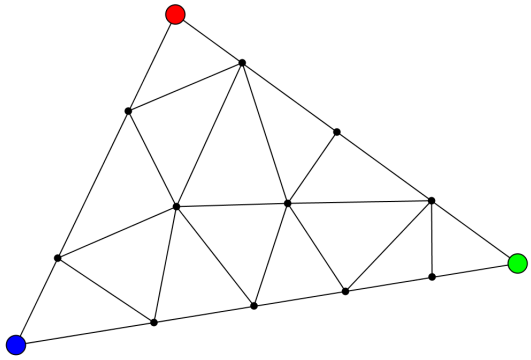
The image is obviously not an open subset of $\mathbb{R}^n$, in contradiction to Brouwer's theorem. □

Intuitively, it seems quite obvious that there cannot be a continuous injective function from $\mathbb{R}^n$ to $\mathbb{R}^m$ when $m < n$; the problem is that it is equally obvious that there cannot be a surjective continuous function from $\mathbb{R}^m$ to $\mathbb{R}^n$, but the Peano curve in analysis does exactly that! (The moral is that there are a lot more continuous functions than one might expect.)

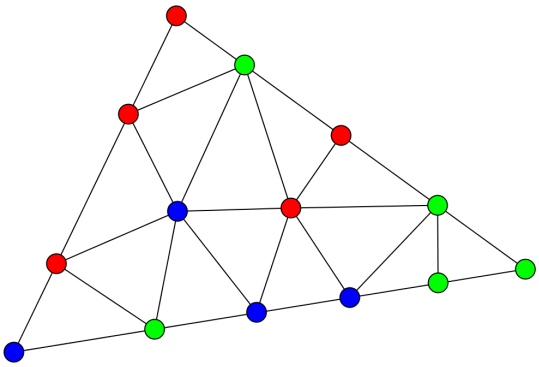
We know how to prove that $\mathbb{R}$ is not homeomorphic to $\mathbb{R}^n$ unless $n = 1$, using connectedness. Traditionally, the invariance of domain theorem in higher dimensions is proved by using methods from algebraic topology; but my plan is to present an elementary proof that only requires the theorems and definitions that we have talked about so far. I learned the details for some of the steps from Terry Tao's blog. The broad outline is the following:

(1) We prove a combinatorial result about triangulations of simplices, called Sperner's lemma.
(2) From Sperner's lemma, we deduce Brouwer's fixed point theorem: every continuous function from the closed ball in $\mathbb{R}^n$ to itself has a fixed point.
(3) The fixed point theorem can then be used to prove the invariance of domain theorem.

Sperner's lemma. Our starting point is a combinatorial result about colorings of simplices, due to Sperner. Let Δ^n be an n-dimensional simplex: a closed interval if $n = 1$, a triangle if $n = 2$, etc. Such a simplex has $n + 1$ vertices $v_1, \ldots, v_{n+1}$, and $n + 1$ faces $F_1, \ldots, F_{n+1}$; we label the faces in such a way that F_k is the face opposite the vertex v_k. We also take $n+1$ different colors, and color the k-th vertex using the k-th color. Now we consider a *triangulation*, that is to say, a subdivision of Δ^n into smaller n-simplices. The following picture shows an example with $n = 2$: the first color is red, the second color green, the third color blue.



Suppose that all the vertices in the triangulation are also colored, in such a way that on every face, we only use the colors from the n vertices on that face; another way to say this is that we do *not* use the k-th color for vertices that lie on the face F_k. The following picture shows an example of such a coloring.

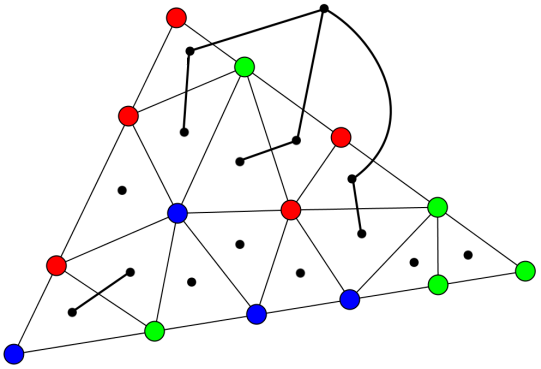


Theorem 4 (Sperner's lemma)**.** *In this situation, there is at least one small simplex all of whose vertices have different colors; in fact, the number of simplices with this property is always odd.*

In the example above, there are five such simplices.

Proof. The proof is by induction on $n \geq 0$; because a 0-simplex is just a single point, the case $n = 0$ is trivial. Suppose then that we already know the result in dimension $n - 1 \geq 0$, and let us prove it in dimension n. Given a triangulation of Δ^n and a coloring as above, let N be the number of small simplices all of whose vertices have different colors. Our goal is to show that $N \equiv 1 \mod 2$.

Put a dot in every simplex of the triangulation, and connect two dots by an edge iff the two simplices in question have a common face whose n vertices are colored using each of the colors $1, \dots, n$ exactly once. Also put an additional dot outside of Δ^n, and connect this dot with a dot in a small simplex iff one of the faces of the small simplex lies on the boundary of Δ^n and the n vertices of that face are colored using each of the colors $1, \dots, n$ exactly once. (By our assumptions on the coloring, every such face has to lie on F_n.) In this way, we obtain a graph; here is what it looks like in the example from above.



The *degree* of a dot is by definition the number of edges going into the dot. In our graph, every dot inside a small simplex has degree 0, 1, or 2, because a simplex can have at most two faces whose vertices use all the colors $1, \dots, n$ exactly once; the degree is 1 precisely when all vertices of the corresponding simplex have different colors. What this means is that N is congruent, modulo 2, to the sum of the degrees of all the dots inside Δ^n.

Now a basic fact in graph theory is that, in every graph, the sum of all the degrees must be an even number; this is due to the fact that every edge has exactly two endpoints. Consequently, N is also congruent, modulo 2, to the degree of the outside dot. That degree is nothing but the number of $(n-1)$-simplices on F_n all of whose vertices have different colors. Since F_n also satisfies the assumptions of Sperner's lemma (in dimension $n-1$), the inductive hypothesis shows that the number of such simplices is odd; but then N must be an odd number, too. □

The case $n = 1$ of Sperner's lemma is a sort of discrete version of the intermediate value theorem: it says that if the two endpoints of an interval have different colors, then there must be at least one small interval where the color changes from one endpoint to the other. In fact, for many purposes in topology, Sperner's lemma can serve as a replacement for results that can otherwise be proved only with the help of algebraic topology.

Brouwer's fixed point theorem. Now we come back to topology: with the help of Sperner's lemma and some basic analysis, we can prove the following fixed point theorem for the closed unit ball

$$B^n = \left\{ x \in \mathbb{R}^n \;\middle|\; \|x\| \leq 1 \right\}$$

in Euclidean space.

Theorem 5 (Brouwer's fixed point theorem)**.** *Every continuous function*

$$f\colon B^n \to B^n$$

has a fixed point: there is a point $x \in B^n$ with the property that $f(x) = x$.

Somebody asked whether there are fixed point theorems in infinite-dimensional spaces. The answer is yes: there is a generalization of Brouwer's theorem, due to Schauder, for closed unit balls in arbitrary Banach spaces.

The fact that every continuous function from B^n to itself has a fixed point is a purely topological property of B^n; any topological space homeomorphic to B^n (such as the closed unit cube or a closed n-simplex) has exactly the same property.

Proof. Let Δ^n denote the closed convex hull of the $n+1$ coordinate vectors in $\mathbb{R}^{n+1}$; this is of course an n-dimensional simplex. To prove the fixed point theorem, it is enough to show that every continuous function $f\colon \Delta^n \to \Delta^n$ has a fixed point; this is because Δ^n and B^n are homeomorphic.

If we use coordinates $x = (x_1, \dots, x_{n+1})$ on $\mathbb{R}^{n+1}$, points $x \in \Delta^n$ are characterized by the conditions $x_i \geq 0$ for all i, and $x_1 + \cdots + x_{n+1} = 1$. These also hold for the coordinates $f_1(x), \dots, f_{n+1}(x)$ of the point $f(x)$. Writing F_k for the face opposite the vertex e_k, points $x \in F_k$ satisfy the additional equation $x_k = 0$.

The idea of the proof is to consider subdivisions of Δ^n and to apply Sperner's lemma to them. To begin with, let $\mathcal{T}$ be an arbitray triangulation. We color its vertices with $n+1$ colors, taking care to use the i-th color for a vertex v only if it satisfies $f_i(v) \leq v_i$. This requirement can always be satisfied: it cannot be the case that $f_i(v) > v_i$ for every i, because then

$$1 = \sum_{i=1}^{n+1} f_i(v) > \sum_{i=1}^{n+1} v_i = 1.$$

We are also allowed to use the k-th color for the vertex e_k, because the k-th coordinate of e_k is equal to 1. In fact, it is possible to choose the coloring in such a way that the assumptions of Sperner's lemma are satisfied: if $v \in F_k$ a vertex on the k-th face, then $v_k = 0$, and because

$$\sum_{i=1}^{n+1} f_i(v) = \sum_{i=1}^{n+1} v_i,$$

we must have $f_i(v) \leq v_i$ for at least one $i \neq k$. Sperner's lemma therefore guarantees the existence of a small simplex in $\mathcal{T}$, all of whose vertices have different colors.

To find the desired fixed point of f, we now apply this result to triangulations by smaller and smaller simplices. To be precise, let us define the *mesh* of a triangulation to be maximum of the diameters of all the simplices in the triangulation. Pick a sequence $\mathcal{T}^m$ of triangulations of Δ^n whose mesh converges to zero. For each triangulation, we choose an arbitrary coloring subject to the condition above; let δ^m be a small simplex in $\mathcal{T}^m$, all of whose vertices have different colors. Inside each δ^m, choose a point x^m. These points form a sequence in Δ^n; because Δ^n is compact, we may pass to a convergent subsequence (that we denote by the same symbol) with limit $x \in \Delta^n$.

Now I claim that x is a fixed point of f. By our choice of coloring, the simplex δ^m has, for each $i = 1, \dots, n+1$, at least one vertex v with $f_i(v) \leq v_i$. Since we assumed that the mesh of the triangulations goes to zero, the vertices of δ^m converge

to the point x as well; because f is continuous, this implies that $f_i(x) \leq x_i$ for all i. But since

$$\sum_{i=1}^{n+1} f_i(x) = \sum_{i=1}^{n+1} x_i,$$

all those inequalities must be equalities, and so $f(x) = x$. □

Proof of the invariance of domain theorem. It is pretty easy to see that Theorem 2 is a consequence of the following result.

Theorem 6. *If $f\colon B^n \to \mathbb{R}^n$ is continuous and injective, then $f(0)$ is an interior point of $f(B^n)$.*

Indeed, to show that $f(U)$ is an open subset whenever $f\colon U \to \mathbb{R}^n$ is continuous and injective, we have to prove that $f(x)$ is an interior point of $f(U)$ whenever $x \in U$. For some $r > 0$, the closed ball $\overline{B_r(x)}$ is contained in U; by applying the theorem to the restriction of f to this closed ball (which is clearly homeomorphic to B^n), we obtain that $f(x)$ is an interior point of $f\big(B_r(x)\big)$, and hence also of $f(U)$.

The proof of Theorem 6 takes several steps; the fixed point theorem will make an appearance in the second step.

Step 1. By assumption, the function $f\colon B^n \to f(B^n)$ is continuous and bijective. Since B^n is compact and $\mathbb{R}^n$ is Hausdorff, f must be a homeomorphism (by Corollary 7.2). In other words, the inverse function $f^{-1}\colon f(B^n) \to B^n$ is also continuous. Now the set $f(B^n)$ may be very complicated, and so we don't really know what the domain of f^{-1} looks like – but we can always extend f^{-1} to a continuous function $G\colon \mathbb{R}^n \to \mathbb{R}^n$ by appealing to the Tietze extension theorem! Indeed, each of the n components of f^{-1} is a continuous function from $f(B^n)$ to the closed interval $[-1, 1]$; because $f(B^n)$ is compact Hausdorff and hence normal, Theorem 13.4 allows us to extend them to continuous functions from $\mathbb{R}^n$ to $[-1, 1]$. Putting these n functions together, we obtain the desired extension G. By construction, we have

$$G\big(f(x)\big) = f^{-1}\big(f(x)\big) = x$$

for every $x \in B^n$.

Step 2. Next, we observe that the function G has precisely one zero on the compact set $f(B^n)$, namely at the point $f(0)$. This is a consequence of the identity $G\big(f(x)\big) = x$. We can use the fixed point theorem to prove that this zero is "stable" under small perturbations of G, in the following sense.

Proposition 7. *If $\tilde{G}\colon f(B^n) \to \mathbb{R}^n$ is a continuous function such that*

$$\|G(y) - \tilde{G}(y)\| \leq 1 \quad \textit{for every } y \in f(B^n),$$

then $\tilde{G}$ also has at least one zero on the compact set $f(B^n)$.

What this means is that, even if we jiggle the function G slightly, the zero of G cannot disappear into nowhere.

Proof. Consider the continuous function

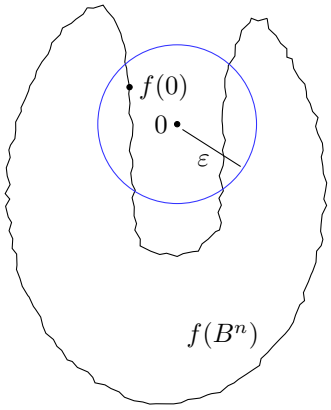
$$B^n \to B^n, \quad x \mapsto x - \tilde{G}\big(f(x)\big) = G\big(f(x)\big) - \tilde{G}\big(f(x)\big);$$

the assumptions on $\tilde{G}$ guarantee that it maps B^n into itself. According to Theorem 5, there must be at least one fixed point $x \in B^n$; but then

$$x = x - \tilde{G}\big(f(x)\big),$$

which obviously means that the point $f(x)$ is a zero of $\tilde{G}$. □

Step 3. Let me now explain the strategy for proving Theorem 6. Suppose that the result was false, in other words, suppose that $f(0)$ was not an interior point of $f(B^n)$. Then $f(0)$ has to lie on the boundary of $f(B^n)$, and so the function G has a zero on the boundary. The idea is that a zero on the boundary is unstable: by jiggling things slightly, one can push the zero outside of $f(B^n)$, which then contradicts Proposition 7. More precisely, we would like to construct a small perturbation $\tilde{G}$ that no longer has any zeros on $f(B^n)$. Let us start turning this idea into a rigorous proof.



Recall that $G\big(f(0)\big) = 0$; because G is continuous, we can choose $\varepsilon > 0$ such that $\|G(y)\| \leq \frac{1}{10}$ whenever $\|y - f(0)\| \leq 2\varepsilon$. Since we are assuming that $f(0)$ is not an interior point of $f(B^n)$, there has to be some point $c \in \mathbb{R}^n$ with $\|c - f(0)\| < \varepsilon$ but $c \notin f(B^n)$. We can translate the whole picture so that c gets moved to the origin, which allows us to assume that $c = 0$. We have arranged that

$$0 \notin f(B^n), \qquad \|f(0)\| < \varepsilon, \qquad \|G(y)\| \leq \tfrac{1}{10} \text{ if } \|y\| \leq \varepsilon.$$

The third item is because $\|y\| \leq \varepsilon$ means that $\|y - f(0)\| \leq \|y\| + \|f(0)\| \leq 2\varepsilon$. The picture above should help you visualize what is going on.

Step 4. Let me now explain how to push the set $f(B^n)$ away from the point $f(0)$. Define two closed sets

$$\Sigma_1 = \big\{\, y \in f(B^n) \;\big|\; \|y\| \geq \varepsilon \,\big\} \quad \text{and} \quad \Sigma_2 = \big\{\, y \in \mathbb{R}^n \;\big|\; \|y\| = \varepsilon \,\big\}.$$

Here Σ_2 is the boundary of the ball of radius ε around the origin, and Σ_1 is the part of $f(B^n)$ that lies outside that ball; since $f(B^n)$ is compact, both Σ_1 and Σ_2 are also compact. Note that the function G has no zeros on the compact set Σ_1. The important thing is that we can write down a continuous function that maps $f(B^n)$ into the set $\Sigma_1 \cup \Sigma_2$, namely

$$\phi\colon f(B^n) \to \Sigma_1 \cup \Sigma_2, \quad \phi(y) = \max\left(\frac{\varepsilon}{\|y\|}, 1\right) \cdot y.$$

This function is well-defined and continuous, due to the fact that $0 \notin f(B^n)$. For a point $y \in \Sigma_1$, we have $\|y\| \geq \varepsilon$, and so $\phi(y) = y$; therefore ϕ does nothing to the

points of $f(B^n)$ outside the ε-ball. For a point $y \in f(B^n)$ with $\|y\| \le \varepsilon$, we get

$$\phi(y) = \varepsilon \cdot \frac{y}{\|y\|} \in \Sigma_2,$$

and so ϕ takes those points to the boundary of the ε-ball. Intuitively, the effect of ϕ is to push the part of $f(B^n)$ that lies inside the ε-ball to the boundary. In particular, the point $f(0)$ does not belong to the image of $f(B^n)$ under ϕ.

Step 5. We would like to use the continuous function $G \circ \phi \colon f(B^n) \to \mathbb{R}^n$ as our perturbation of G. For $y \in \Sigma_1$, we have $(G \circ \phi)(y) = G(y) \neq 0$, and so this function has no zeros on Σ_1. It is however possible that $G \circ \phi$ might vanish at some point of $f(B^n)$ inside the ε-ball; this can happen for instance if G has zeros on Σ_2. There is not much we can do about this: G was an arbitrary continuous extension of f^{-1}, and so we have no control over its zeros on Σ_2. Now the idea is to replace G by a function that is better-behaved: a polynomial function. This can be done with the help of the Weierstrass approximation theorem and some measure theory, both results from analysis. (The homework for this week explains a purely topological argument for achieving the same thing.)

Recall that G has no zeros on the compact set Σ_1, which means that $\|G(y)\| > 0$ for every $y \in \Sigma_1$. By compactness, we can therefore find a real number $\delta > 0$ such that $\|G(y)\| \ge \delta$ for every $y \in \Sigma_1$; without loss of generality, we may assume that $\delta \le \frac{1}{10}$. According to the Weierstrass approximation theorem, there is a polynomial function $P \colon \mathbb{R}^n \to \mathbb{R}^n$ with the property that

$$\|G(y) - P(y)\| < \delta \quad \text{for every } y \in \Sigma_1 \cup \Sigma_2.$$

Observe that P still does not have any zeros on $f(B^n)$: if we had $P(y) = 0$, then $\|G(y)\| < \delta$, contradicting the fact that $\|G(y)\| \ge \delta$. If we are unlucky, it may happen that P has a zero somewhere on Σ_2, but we can easily fix this by the following measure-theoretic argument. The set Σ_2 obviously has measure zero; because P is a polynomial function, one can prove that the image $P(\Sigma_2)$ also has measure zero. By choosing a sufficiently small vector $v \in \mathbb{R}^n \setminus P(\Sigma_2)$, and by replacing P by the polynomial function $P - v$, we can arrange that P does not have any zeros on $\Sigma_1 \cup \Sigma_2$.

Step 6. Now we are basically done. Let us define a function

$$\tilde{G} \colon f(B^n) \to \mathbb{R}^n, \quad \tilde{G}(y) = P\big(\phi(y)\big).$$

This function is continuous, and by what we said above, has no zeros on the compact set $f(B^n)$. Let us show that $\tilde{G}$ is a small perturbation of G. Let $y \in f(B^n)$ be an arbitrary point; there are two cases. (1) If $\|y\| \ge \varepsilon$, then $\phi(y) = y$, and so

$$\|G(y) - \tilde{G}(y)\| = \|G(y) - P(y)\| < \delta \le \frac{1}{10}.$$

(2) If $\|y\| \le \varepsilon$, then $\phi(y) \in \Sigma_2$ and $\|\phi(y)\| = \varepsilon$, and so we get

$$\begin{aligned} \|G(y) - \tilde{G}(y)\| &= \|G(y) - P(\phi(y))\| = \|G(y) - G(\phi(y)) + G(\phi(y)) - P(\phi(y))\| \\ &\le \|G(y)\| + \|G(\phi(y))\| + \|G(\phi(y)) - P(\phi(y))\| \\ &\le \frac{1}{10} + \frac{1}{10} + \delta \le \frac{3}{10}. \end{aligned}$$

In both cases, the distance between $G(y)$ and $\tilde{G}(y)$ is less than 1, and so Proposition 7 says that $\tilde{G}(y)$ has a zero somewhere on $f(B^n)$. This is a contradiction, and so we have proved Theorem 6.

Exercises.

1. Use the invariance of domain theorem to show that an n-dimensional manifold can never be homeomorphic to an m-dimensional manifold with $m \neq n$.

2. This exercise will tell you how, in proving the invariance of domain theorem, one can bypass the use of Weierstrass approximation.

Theorem (Kulpa)**.** *Let $f\colon X \to \mathbb{R}^n \setminus \{0\}$ be a continuous function defined on a compact subset $X \subseteq \mathbb{R}^n$. Then for each $\varepsilon > 0$ and every compact subset $Y \subseteq \mathbb{R}^n$ with empty interior, there is a continuous function $F\colon X \cup Y \to \mathbb{R}^n \setminus \{0\}$ such that $\|F(x) - f(x)\| < \varepsilon$ for all $x \in X$.*

Convince yourself that this result can take the place of Weierstrass approximation in our proof of the invariance of domain theorem. (*Hint*: In the notation from class, take $X = \Sigma_1$ and $Y = \Sigma_2$.) Then prove the following lemma.

Lemma. *Let $\Delta \subseteq \mathbb{R}^n$ be an n-dimensional simplex with vertices $v_1, \ldots, v_n$. Given a continuous function $g\colon \Delta \to \mathbb{R}^n$, show that the formula*

$$h(x) = \sum_{i=1}^{n} t_i g(v_i), \quad \text{where } x = \sum_{i=1}^{n} t_i v_i,$$

defines a continuous function $h\colon \Delta \to \mathbb{R}^n$. We call it the linearization *of g on Δ.*

The idea behind the proof of Kulpa's theorem is to approximate the given function by piecewise linear functions instead of by polynomials. Now prove Kulpa's theorem in the following steps.

(a) Since $X \cup Y$ is bounded, it is contained in the n-dimensional box $I^n = [-R, R]^n$ for some $R > 0$. Show that f can be extended to a continuous function $g\colon I^n \to \mathbb{R}^n$.

(b) Choose $\delta > 0$ with $2\delta < \varepsilon$ such that $f(X) \cap B_{2\delta}(0) = \emptyset$. Show that one can subdivide I^n into n-dimensional simplices $\Delta_1, \ldots, \Delta_N$, in such a way that the image $g(\Delta_k)$ of every simplex Δ_k has diameter at most δ.

(c) Let $h\colon I^n \to \mathbb{R}^n$ be the piecewise linear function whose restriction to each Δ_k is the linearization of g on Δ_k. Show that h is continuous, and that $\|f(x) - h(x)\| \leq \delta$ for every $x \in X$.

(d) Show that $h(Y) \subseteq \mathbb{R}^n$ is a compact set with empty interior. Conclude that there is a point $c \in B_\delta(0) \setminus h(X \cup Y)$.

(e) Now define $F\colon X \cup Y \to \mathbb{R}^n$ as $F(z) = h(z) - c$, and prove that it has all the required properties.

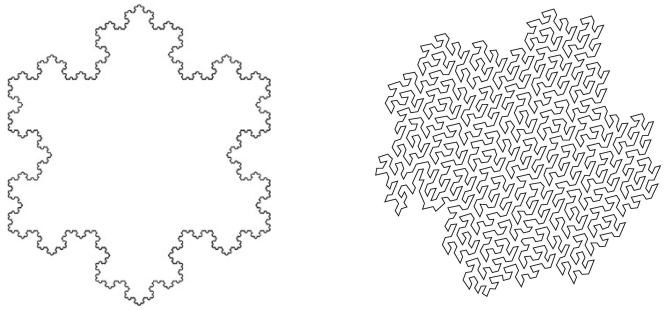
Extra Lecture: The Jordan Curve Theorem

The Jordan curve theorem. Instead of homology, one can also use some of the ideas about paths and fundamental groups to prove the Jordan curve theorem. I am sure all of you have heard about this result at some point in your life: any closed curve in $\mathbb{R}^2$ divides the plane into two regions, one inside the curve, the other outside the curve.

More precisely, a closed curve in the plane is a continuous function $h\colon [0,1] \to \mathbb{R}^2$ with $h(0) = h(1)$; if we identify the two endpoints of the interval, we obtain a circle, and so we can just as well say that a closed curve is a continuous function $h\colon \mathbb{S}^1 \to \mathbb{R}^2$. This definition includes space-filling curves: the image of the Peano curve, for example, would be the entire unit square, which obviously fails to divide $\mathbb{R}^2$ into two regions. Jordan discovered that this problem goes away if we consider *simple* closed curves (also called *Jordan curves*), where the function $h\colon \mathbb{S}^1 \to \mathbb{R}^2$ is injective. Note that for such curves, the image $h(\mathbb{S}^1)$ is homeomorphic to $\mathbb{S}^1$, due to the compactness of $\mathbb{S}^1$. Jordan curves are closer to our intuitive notion of curve.

Theorem 27.1 (Jordan)**.** *If $C \subseteq \mathbb{R}^2$ is a simple closed curve, then $\mathbb{R}^2 \setminus C$ has exactly two connected components, each of which has C as its boundary.*

This theorem was proved by Jordan in the late 19th century. One often reads that his proof was unsatisfactory and that the first correct one is due to Veblen; but Hales, who wrote the first computer-checkable proof of the Jordan curve theorem, says that: "In view of the heavy criticism of Jordan's proof, I was surprised when I sat down to read his proof to find nothing objectionable about it."



The tricky thing is that most simple closed curves do not look at all like a circle. There are fractal curves (like the Koch snowflake, pictured above), and even for curves that are not fractals, it can be very hard to decide by looking at the curve whether a given point lies inside or outside the curve.

Some observations. The argument I am going to present uses basic algebraic topology, and especially some ideas from our proof of the Seifert-van Kampen theorem. Let $f\colon \mathbb{S}^1 \to \mathbb{R}^2$ be a simple closed curve, and denote by $C = f(\mathbb{S}^1)$ its image in $\mathbb{R}^2$. To get started, here are some observations:

(1) The curve C is compact, and f is a homeomorphism between $\mathbb{S}^1$ and C; in particular, we can forget about the parametrization and remember only the subset $C \subseteq \mathbb{R}^2$. This follows immediately from Corollary 7.2.
(2) The complement $\mathbb{R}^2 \setminus C$ is open and locally path connected; every connected component of $\mathbb{R}^2 \setminus C$ is thus open and path connected (Proposition 6.7).

(3) Exactly one of the connected components of $\mathbb{R}^2 \setminus C$ is unbounded. The reason is that C is contained in a large disk, whose complement is path connected, hence contained in exactly one connected component of $\mathbb{R}^2 \setminus C$.

Let us first show that C is the common boundary of every connected component of $\mathbb{R}^2 \setminus C$. For the time being, we assume that the complement of C has at least two connected components; we will show later that this is always the case.

Lemma 1. *Let $C \subseteq \mathbb{R}^2$ be a simple closed curve. If $\mathbb{R}^2 \setminus C$ is not connected, then every connected component has C as its boundary.*

Proof. Let $U \subseteq \mathbb{R}^2 \setminus C$ be one of the connected components. Since the other connected components are open, we have $\partial U = \bar{U} \setminus U \subseteq C$. Suppose that $\partial U \neq C$; we shall argue that this leads to a contradiction. Since C is homeomorphic to a circle, ∂U is homeomorphic to a proper closed subset of the circle, and therefore contained in a subset homeomorphic to a closed interval. In other words, we have $\partial U \subseteq A$, where $A \subseteq C$ is a closed subspace homeomorphic to $[0,1]$; a subspace homeomorphic to the closed unit interval is called an *arc*.

By assumption, $\mathbb{R}^2 \setminus C$ has at least two connected components, and so at least one of its connected components must be bounded. After translating C by some finite distance, if necessary, we can arrange that the origin lies in a bounded component; if U itself happens to be bounded, we may assume that $0 \in U$. Let D be a closed disk of sufficiently large radius centered at 0, such that the curve C is contained in D; its boundary ∂D is clearly contained in the unbounded component of $\mathbb{R}^2 \setminus C$.

Now we use the Tietze extension theorem to extend the identity $\mathrm{id}_A \colon A \to A$ to a continuous function $g \colon D \to A$. Recall that A is homeomorphic to $[0,1]$; since D is normal, Theorem 13.4 says that any homeomorphism $h \colon A \to [0,1]$ extends to a continuous function $h' \colon D \to [0,1]$, and then $g = h^{-1} \circ h'$ does the job. So far, nothing seems to be wrong – but in fact, such a function g cannot exist, because one can use it to build a retraction from the disk D onto its boundary circle!

There are two cases. If U is bounded, define

$$r \colon D \to D \setminus \{0\}, \quad r(x) = \begin{cases} g(x) & \text{if } x \in \overline{U}, \\ x & \text{if } x \in \mathbb{R}^2 \setminus U; \end{cases}$$

this is continuous because both sets are closed and $g(x) = x$ for every $x \in \partial U \subseteq A$. If U is unbounded, define

$$r \colon D \to D \setminus \{0\}, \quad r(x) = \begin{cases} g(x) & \text{if } x \in \mathbb{R}^2 \setminus U, \\ x & \text{if } x \in \overline{U}, \end{cases}$$

which is continuous for the same reason. Because ∂D is contained in the unbounded component of $\mathbb{R}^2 \setminus C$, we have $r(x) = x$ for every $x \in \partial D$. If we now compose r with the obvious retraction of $D \setminus \{0\}$ onto its boundary, we obtain a retraction of D onto ∂D; but such a retraction cannot exist by Corollary 18.5.

The conclusion is that $\partial U = C$, which is what we wanted to show. □

If you look at the proof carefully, you will see that we have actually proved the following additional result.

Corollary 27.2. *The complement of any arc $A \subseteq \mathbb{R}^2$ is path connected.*

Proof. The complement is open and locally path connected; if it has more than one connected component, we can use the same argument as above to build a retraction from a disk onto its boundary. □

The complement is not connected. The next step is to show that the complement $\mathbb{R}^2 \setminus C$ is not connected. It will be more convenient to replace $\mathbb{R}^2$ by its one-point compactification $\mathbb{S}^2$. It is easy to see that $\mathbb{R}^2 \setminus C$ and $\mathbb{S}^2 \setminus C$ have exactly the same number of connected components: the unique unbounded component of $\mathbb{R}^2 \setminus C$ corresponds to the unique component of $\mathbb{S}^2 \setminus C$ containing the extra point at infinity. We may therefore look at a simple closed curve C on the sphere.

Choose two points $p, q \in C$; they divide C into two arcs A and B, both of which have p and q as their endpoints. Consider the two open sets $U = \mathbb{S}^2 \setminus A$ and $V = \mathbb{S}^2 \setminus B$; since they are both complements of arcs, Corollary 27.2 shows that they are both path connected. Now

$$U \cap V = \mathbb{S}^2 \setminus C$$
$$U \cup V = \mathbb{S}^2 \setminus \{p, q\}$$

and so if $\mathbb{S}^2 \setminus C$ was connected (and hence path connected), the fundamental group of $X = U \cup V$ would be governed by the Seifert-van Kampen theorem. But X is just the complement of two points on the sphere, and so

$$\pi_1(X, x_0) \simeq \mathbb{Z}$$

for any choice of base point $x_0 \in U \cap V$.

Lemma 2. *The homomorphisms $\pi_1(U, x_0) \to \pi_1(X, x_0)$ and $\pi_1(V, x_0) \to \pi_1(X, x_0)$ are both trivial.*

Proof. Given any loop $\alpha \colon I \to U$ based at the point x_0, we have to show that it is path homotopic, in $X = \mathbb{S}^2 \setminus \{p, q\}$, to the constant path at x_0. Let $g \colon \mathbb{S}^1 \to U$ be the induced continuous function from the circle; then $[\alpha] \in \pi_1(X, x_0)$ is the image of the generator of $\pi_1(\mathbb{S}^1, b_0)$ under the homomorphism g_*; by Lemma 19.6, it will be enough to show that f is homotopic to a constant function.

By construction, $g(\mathbb{S}^1)$ is disjoint from the arc A; because $p, q \in A$, this means that p and q belong to the same connected component of $\mathbb{S}^2 \setminus g(\mathbb{S}^1)$. Now $\mathbb{S}^2 \setminus \{q\}$ is homeomorphic to $\mathbb{R}^2$, and so we may assume without loss of generality that $g \colon \mathbb{S}^1 \to \mathbb{R}^2 \setminus \{p\}$ is a continuous function with the property that $p \in \mathbb{R}^2$ lies in the unbounded component of $\mathbb{R}^2 \setminus g(\mathbb{S}^1)$. After translating everything by $-p$, we can arrange furthermore that $p = 0$.

Now let D be a disk of sufficiently large radius containing $g(\mathbb{S}^1)$, and choose a point $v \in \mathbb{R}^2 \setminus D$. Let $\gamma \colon I \to \mathbb{R}^2 \setminus g(\mathbb{S}^1)$ be a path from the origin to the point v; such a path exists because the origin lies in the unbounded component. Consider the homotopy

$$G \colon \mathbb{S}^1 \times I \to \mathbb{R}^2 \setminus \{0\}, \quad G(x, t) = g(x) - \gamma(t);$$

note that $G(x, t) \neq 0$ because the path γ is contained in the complement of $g(\mathbb{S}^1)$. This shows that g is homotopic to the function $g - v$; the point is that the image of $g - v$ stays far away from the origin. We can then use a second homotopy

$$H \colon \mathbb{S}^1 \times I \to \mathbb{R}^2 \setminus \{0\}, \quad H(x, t) = tg(x) - v$$

to deform $g - v$ continuously into the constant function $-v$. This shows that g is homotopic to a constant function, as needed. □

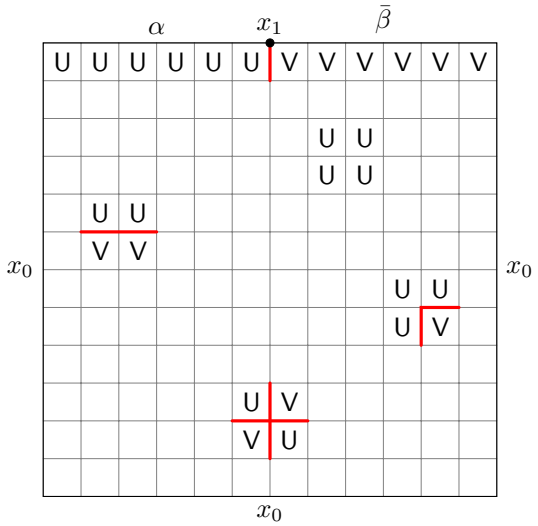
The lemma shows that $\pi_1(U, x_0)$ and $\pi_1(V, x_0)$ generate the trivial subgroup of $\pi_1(X, x_0)$. If $U \cap V$ was path connected, Theorem 18.8 would imply that $\pi_1(X, x_0)$ is itself trivial; since this is not the case, it follows that $U \cap V = \mathbb{S}^2 \setminus C$ must have at least two connected components.

Exactly two connected components. The remainder of the proof consists in exploiting the fact that $\pi_1(X, x_0) \simeq \mathbb{Z}$ to prove that $\mathbb{S}^2 \setminus C$ can have at most two connected components. The idea is that each additional component gives a new element in $\pi_1(X, x_0)$; since the group has rank 1, there can be only 2 components.

Let me first explain how the existence of more than one connected component produces nontrivial elements in the fundamental group of X. We already have a base point $x_0 \in U \cap V$; let $x_1 \in U \cap V$ be a point in a different connected component. Since U and V are both path connected, we can choose a path α in U and a path β in V that join the two points x_0 and x_1. Then $\alpha * \bar{\beta}$ is a loop in X based at the point x_0.

Lemma 3. *The element $[\alpha * \bar{\beta}] \in \pi_1(X, x_0)$ is nontrivial.*

Proof. Suppose to the contrary that the element was trivial. Then there would be a path homotopy $H\colon I \times I \to X$ from $\alpha * \bar{\beta}$ to the constant path at x_0. As in the proof of the Seifert-van Kampen theorem, we choose a sufficiently large even integer N and divide $I \times I$ into N^2 little boxes, each of which is mapped by H into one of the two open sets U or V. We then label each box either with the letter U or the letter V, depending on whether its image under H lies in U or V; if the image happens to lie in $U \cap V$, we simply choose one of the two. Since $\alpha(I) \subseteq U$, we can clearly arrange that the first $N/2$ boxes in the topmost row are labeled U; similarly, we can arrange that the remaining $N/2$ boxes are labeled V.



Now we construct a graph whose vertices are the $(N + 1)^2$ points on the grid; we include only those edges into the graph whose two adjacent boxes have different labels. It is easy to see that the degree of any vertex inside (= not on the boundary of) the square has degree 0, 2, or 4. By construction, there is exactly one vertex of degree 1 on the top boundary, namely the midpoint (which is mapped to the

point x_1 by H). Consider the connected component of our graph containing this vertex. For degree reasons, it has to meet the boundary of the square in an even number of vertices, so it has to contain at least one vertex on the left, bottom, or right boundary of the square (which are mapped to the point x_0 by H). But then the image of this component under H is a connected subset of $U \cap V$ containing the two points x_0 and x_1, contradicting the fact that these two points belonged to different connected components. □

We can use the same strategy to prove that $U \cap V = \mathbb{S}^2 \setminus C$ can have at most two connected components. Suppose by way of contradiction that there were three or more components. We could then choose a point $x_2 \in U \cap V$ so that x_0, x_1, and x_2 belong to three different connected components; if γ and δ are paths in U and V connecting the two points x_0 and x_2, then $\gamma * \bar{\delta}$ is a second loop based at x_0. The two classes $[\alpha * \bar{\beta}]$ and $[\gamma * \bar{\delta}]$ are linearly independent, in the following sense.

Lemma 4. *If* $[\alpha * \bar{\beta}]^m = [\gamma * \bar{\delta}]^n$ *in* $\pi_1(X, x_0)$*, then* $m = n = 0$.

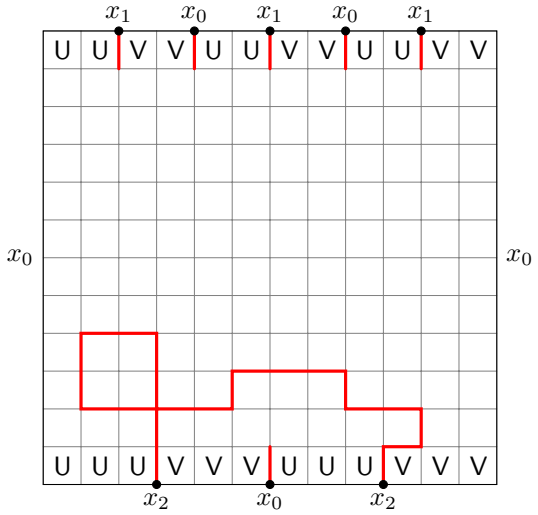
Proof. Suppose that there was such a relation with $m, n \neq 0$. Then there would be a homotopy $H\colon I \times I \to X$ from the loop

$$\alpha * \bar{\beta} * \cdots * \alpha * \bar{\beta},$$

where each of the $2m$ individual loops is parametrized by an interval of length $\frac{1}{2m}$, to the loop

$$\gamma * \bar{\delta} * \cdots * \gamma * \bar{\delta},$$

where each of the $2n$ individual loops is parametrized by an interval of length $\frac{1}{2n}$. As before, we choose a sufficiently large integer N that is divisible by both m and n, and subdivide $I \times I$ into N^2 little boxes that we label either with U or with V, depending on where they are mapped under H. We also construct a graph by exactly the same procedure.



Again, every vertex inside the square has even degree, and so any connected component of the graph has to meet the boundary of the square in an even number of vertices. Now there are two cases. If m is even (as in the picture above), there is

an odd number of vertices of degree 1 that get mapped to the point x_1; for degree reasons, at least one of the connected components containing these vertices has to meet the boundary of the square in a vertex that gets mapped to the point x_0 or to the point x_2; but then we get a path in $U \cap V$ joining these points, contradiction. The same argument works of course when n is even.

If n is odd (as in the picture above), there is an even number of vertices of degree 1 that get mapped to the point x_2. We get the same contradiction as above, except in the case where all such vertices belong to one connected component of the graph. But then every vertex of degree 1 on the bottom boundary has to belong to that same component, because a connected component that starts at such a vertex still has to meet the boundary of the square in an even number of vertices. We thus get a path in $U \cap V$ joining the point x_0 to the point x_2, which is again a contradiction. The same argument works when m is odd. □

Since $\pi_1(X, x_0) \simeq \mathbb{Z}$, both elements $[\alpha * \bar{\beta}]$ and $[\gamma * \bar{\delta}]$ are nonzero multiples of the generator; this means that there is a relation of the form

$$[\alpha * \bar{\beta}]^m = [\gamma * \bar{\delta}]^n$$

for some $(m, n) \neq (0, 0)$. Since the lemma tells us that this is not possible, the only conclusion is that $U \cap V$ must have exactly two connected components. This finishes the proof of Theorem 1.