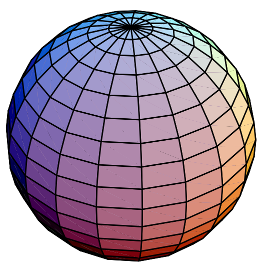
ALGEBRAIC TOPOLOGY

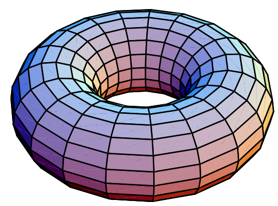
N. P. STRICKLAND

1. Introduction

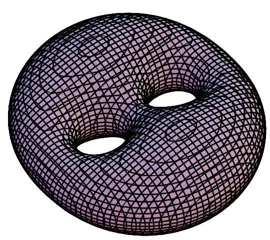
In this course, we will study metric spaces (which will often be subspaces of $\mathbb{R}^n$ for some n) with interesting topological structure. Here are some examples of such subsets of $\mathbb{R}^3$.

1. Closed orientable surfaces, as studied in the Knots and Surfaces course. For example, the sphere and the torus:

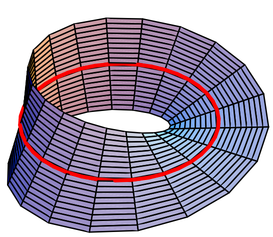




or the following surface of genus two:

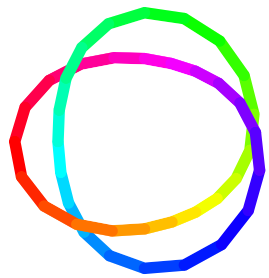


2. The Möbius strip:

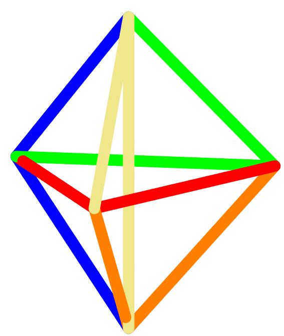


Date: February 18, 2003.

3. Knots, such as the trefoil:



4. Graphs (as studied in the Graph Theory course) such as K_5 (the complete graph on five vertices):



5. The letters of the alphabet can be thought of as subsets of the plane and thus as metric spaces; these will be used as examples in various places. The topology depends on the precise way in which the letters are written. We will always use the form indicated below:

We will also be interested in spaces that cannot be visualised in three dimensions. You may already have met the example of non-orientable surfaces, such as the Klein bottle and the projective plane, but there are many others.

For example, consider the set $GL_3(\mathbb{R})$ of all 3×3 invertible matrices over $\mathbb{R}$. There is a reasonably natural notion of what it means for a matrix to vary continuously, which can be formalised to make $GL_3(\mathbb{R})$ into a metric space. A matrix $A \in GL_3(\mathbb{R})$ has nine different entries, which can vary independently, so $GL_3(\mathbb{R})$ is a nine-dimensional space. We will try to understand spaces like this "geometrically", and ask whether they have holes in them, how any holes are linked together, and so on. It turns out, in fact, that the topological structure of $GL_3(\mathbb{R})$ plays an important rôle in understanding the quantum-mechanical behaviour of electrons. In the general theory of relativity there are interesting questions about the topology of the universe as a whole, and some very complicated topologies enter less directly into more recent theories of fundamental physics.

2. Metric spaces

Definition 2.1. A *metric* on a set X is a notion of the distance between any two points $x, y \in X$ (written $d(x, y)$) such that

M1: $d(x, y) = 0$ if and only if $x = y$.
M2: $d(y, x) = d(x, y)$.
M3: $d(x, z) \leq d(x, y) + d(y, z)$ (the *Triangle Inequality*).

A *metric space* is a set equipped with a specified metric.

Definition 2.2. A sequence (x_n) in a metric space X converges to a point $x \in X$ if for all $\epsilon > 0$ there exists $N \geq 0$ such that $d(x_n, x) < \epsilon$ for all $n \geq N$. This is equivalent to the condition that $d(x_n, x) \longrightarrow 0$ in the usual sense of convergence of sequences of real numbers.

Definition 2.3. Let d and d' be two metrics on the same set X. We say that d and d' are *strongly equivalent* if there are constants $A, A' > 0$ such that $d(x, y) \leq Ad'(x, y)$ and $d'(x, y) \leq A'd(x, y)$ for all $x, y \in X$.

Remark 2.4. If d and d' are equivalent, we see using the sandwich test that $d(x_n, x) \longrightarrow 0$ iff $d'(x_n, x) \longrightarrow 0$. This means that a sequence (x_n) converges to a point x with respect to d iff it converges to x with respect to d'.

Example 2.5. Let x and y be points in $\mathbb{R}^n$. Define

$$\begin{aligned} d_1(x,y) &= \|x-y\|_1 \quad \text{where} \quad \|z\|_1 = |z_1| + \ldots + |z_n| \\ d_2(x,y) &= \|x-y\|_2 \quad \text{where} \quad \|z\|_2 = \sqrt{z_1^2 + \ldots + z_n^2} \\ d_\infty(x,y) &= \|x-y\|_\infty \quad \text{where} \quad \|z\|_\infty = \max\{|z_1|, \ldots, |z_n|\}. \end{aligned}$$

These give three different metrics on $\mathbb{R}^n$. More generally, if X is any subset of $\mathbb{R}^n$ then we can use any of these metrics to make X into a metric space. We'll use the metric d_2 unless otherwise specified.

Lemma 2.6. *The metrics d_1, d_2 and d_∞ on $\mathbb{R}^n$ are all strongly equivalent to each other.*

Proof. Suppose $x, y \in \mathbb{R}^n$. Put $z_i = |x_i - y_i|$ (for $i = 1, \ldots, n$) and $r_k = \|z\|_k$ (for $k = 1, 2, \infty$). It will be enough to show that

$$r_2 \leq \sqrt{n} r_\infty \leq \sqrt{n} r_1 \leq n r_2.$$

Clearly $z_i \leq \max\{z_1, \ldots, z_n\} = r_\infty$ for all i, so

$$r_2^2 = \sum_i z_i^2 \leq \sum_i r_\infty^2 = n r_\infty^2,$$

so $r_2 \leq \sqrt{n} r_\infty$.

Next, note that $r_\infty = z_j$ for some j, so r_∞ is one of the terms in the sum that defines r_1, and all these terms are nonnegative, so $r_\infty \leq r_1$, so $\sqrt{n} r_\infty \leq \sqrt{n} r_1$.

Finally, put $w_i = z_i - r_1/n$, so $w_i^2 = z_i^2 - 2z_i r_1/n + r_1^2/n^2$. We then have

$$\begin{aligned}\sum_{i=1}^n w_i^2 &= \sum_{i=1}^n z_i^2 - 2r_1 n^{-1}\sum_{i=1}^n z_i + r_1^2 n^{-2}\sum_{i=1}^n 1\\ &= r_2^2 - 2r_1 n^{-1} r_1 + r_1^2 n^{-2}.n\\ &= r_2^2 - r_1^2/n.\end{aligned}$$

As $\sum_i w_i^2$ is clearly nonnegative, we conclude that $r_2^2 - r_1^2/n \geq 0$ so $r_1^2/n \leq r_2^2$ so $r_1 \leq \sqrt{n} r_2$ so $\sqrt{n} r_1 \leq n r_2$, as claimed. □

Example 2.7. Let X be any set, and define

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

This is a metric on X, called the *discrete metric*.

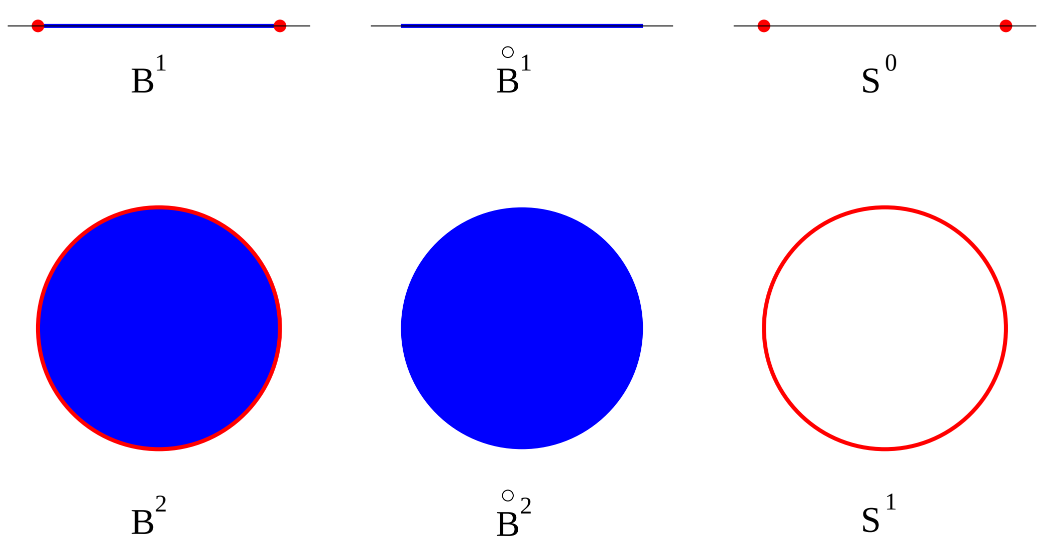
Example 2.8. Let X be the set of continuous functions from $[0,1]$ to $\mathbb{R}$, and define

$$d(f,g) = \left(\int_0^1 (f(x) - g(x))^2 dx \right)^{1/2}.$$

This is a metric on X.

We next define some important examples of subsets of $\mathbb{R}^n$, that are defined for all n.

$$\begin{aligned} I &= [0,1] = \{x \in \mathbb{R} \;:\; 0 \leq x \leq 1\} \text{ (the \textit{unit interval})}\\ B^n &= \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \ldots + x_n^2 \leq 1\} \text{ (the \textit{closed unit n-ball})}\\ \mathring{B}^n &= \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_1^2 + \ldots + x_n^2 < 1\} \text{ (the \textit{open unit n-ball})}\\ S^n &= \{(x_0, \ldots, x_n) \in \mathbb{R}^{n+1} : x_0^2 + \ldots + x_n^2 = 1\} \text{ (the \textit{unit n-sphere})} \end{aligned}$$



Remark 2.9. We will often identify $\mathbb{R}^2$ with $\mathbb{C}$ (via $(x,y) \leftrightarrow x + iy$), and thus S^1 with $\{z \in \mathbb{C} : |z| = 1\}$.

Example 2.10. Let $M_n(\mathbb{R})$ be the set of all $n \times n$ matrices over $\mathbb{R}$. This can be identified with $\mathbb{R}^{n^2}$ and thus considered as a metric space using any of the metrics in Example 2.5. We will also consider the subspace $GL_n(\mathbb{R})$ consisting of the invertible matrices (or equivalently, matrices A with $\det(A) \neq 0$).

The metric d_2 for matrices has a nice interpretation that will be useful later.

Definition 2.11. Let A be an $n \times n$ matrix with entries A_{ij} for $i, j = 1, \dots, n$. Recall that the *transpose* of A is the matrix A^T whose (i,j)'th entry is A_{ji}, and the *trace* of A is the number $\text{trace}(A) = \sum_{i=1}^n A_{ii}$.

Proposition 2.12. *For any $n \times n$ matrix A we have $\|A\|_2 = \sqrt{\text{trace}(AA^T)}$. Thus, if A and B are $n \times n$ matrices we have*

$$d_2(A, B) = \|A - B\|_2 = \sqrt{\text{trace}((A - B)(A^T - B^T))}.$$

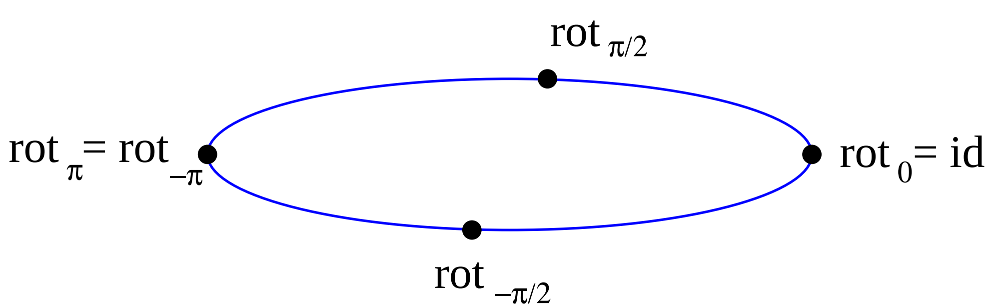
Proof. If A and B are any $n \times n$ matrices, then the (i,k)'th entry in AB is just $\sum_{j=1}^n A_{ij}B_{jk}$. In particular, the (i,i)'th entry is $\sum_j A_{ij}B_{ji}$, so the trace of AB is $\sum_{i,j} A_{ij}B_{ji}$. Now take $B = A^T$, so $B_{ji} = A_{ij}$. We find that $\text{trace}(AA^T) = \sum_{i,j} A_{ij}^2 = \|A\|_2^2$, as claimed. □

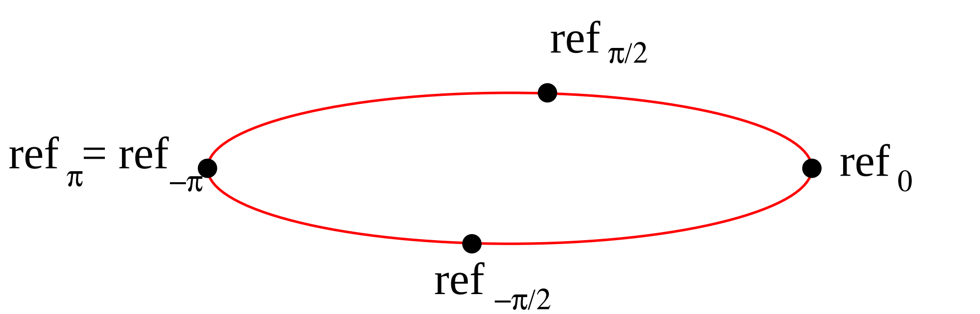
Definition 2.13. An $n \times n$ matrix A is *orthogonal* if $AA^T = I$. (Here A^T is the transpose of A, so the (i,j)'th entry in A^T is the (j,i)'th entry in A; and I is the identity matrix.) We write $O(n)$ for the set of all $n \times n$ orthogonal matrices, and $SO(n)$ for $\{A \in O(n) \; : \; \det(A) = 1\}$. We regard these as subspaces of $M_n(\mathbb{R})$.

The space $O(2)$ consists of all the rotation matrices and reflection matrices:

$$\text{rot}_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$$
$$\text{ref}_\theta = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix}$$

The subspace $SO(2)$ contains only the rotations. As we will explain in more detail in example 3.12, the space $SO(2)$ can be identified with the circle S^1, and $O(2)$ can be identified with a pair of disjoint circles.





The set $SO(3)$ consists of all the rotations (about all possible axes passing through the origin) of three dimensional space. It has a more complicated topology, which we shall study later.

Definition 2.14. Let A be a 2×2 matrix of complex numbers. Let $A^\dagger$ be the result of transposing A and taking the complex conjugate of all the entries. Thus, if $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then $A^\dagger = \left(\begin{smallmatrix} \overline{a} & \overline{c} \\ \overline{b} & \overline{d} \end{smallmatrix}\right)$. We define

$$SU(2) = \{A \in M_2\mathbb{C} \; : \; \det(A) = 1 \text{ and } A^\dagger = A^{-1}\}.$$

Recall that

$$S^3 = \{(w,x,y,z) \in \mathbb{R}^4 \ : \ w^2+x^2+y^2+z^2 = 1\}.$$

It can be shown that the correspondence

$$(w,x,y,z) \leftrightarrow \begin{pmatrix} w+ix & y+iz \\ -y+iz & w-ix \end{pmatrix}$$

identifies S^3 with $SU(2)$.

Definition 2.15. Let A be an $(n+1)\times(n+1)$ matrix, with entries A_{ij} for $i,j=0,\dots,n$. Recall that the *trace* of A is the sum of the diagonal entries, so $\text{trace}(A) = \sum_{i=0}^n A_{ii}$. We define

$$\mathbb{R}P^n = \{A \in M_{n+1}(\mathbb{R}) \ : \ A^T = A^2 = A \text{ and } \text{trace}(A) = 1\},$$

and we call this the *real projective space* of dimension n.

You may have previously seen a rather different definition of the projective plane $\mathbb{R}P^2$; later we will show that it is homeomorphic to the space described above.

Definition 2.16. Let X and Y be metric spaces. We define

$$d\colon (X\times Y)\times(X\times Y) \longrightarrow [0,\infty)$$

by $d((x,y),(x',y')) = d(x,x') + d(y,y')$.

Proposition 2.17. *The function d is a metric on $X\times Y$ (called the* product metric*).*

Proof. For M1, note that

$$\begin{aligned} d((x,y),(x',y')) = 0 &\Leftrightarrow d(x,x') + d(y,y') = 0 \\ &\Leftrightarrow d(x,x') = 0 \text{ and } d(y,y') = 0 \\ &\Leftrightarrow x = x' \text{ and } y = y' \\ &\Leftrightarrow (x,y) = (x',y'). \end{aligned}$$

At the second stage we used the fact that $d(x,x') \geq 0$ and $d(y,y') \geq 0$, and at the third stage we used M1 for X and Y.

Axiom M2 for $X \times Y$ clearly follows from M2 for X and Y.

For M3, suppose we have three points (x,y), (x',y') and (x'',y'') in $X\times Y$. The triangle inequalities for X and Y tell us that

$$\begin{aligned} d(x,x'') &\leq d(x,x') + d(x',x'') \\ d(y,y'') &\leq d(y,y') + d(y',y''). \end{aligned}$$

By adding these two inequalities, we obtain

$$d((x,y),(x'',y'')) \leq d((x,y),(x',y')) + d((x',y'),(x'',y'')),$$

as required. ∎

Remark 2.18. We could instead have used the functions

$$d'((x,y),(x',y')) = \sqrt{d(x,x')^2 + d(y,y')^2}$$

or

$$d''((x,y),(x',y')) = \max(d(x,x'), d(y,y')).$$

These give two more metrics on $X\times Y$, both of them strongly equivalent to the metric considered above.

Proposition 2.19. *A sequence (x_n,y_n) converges to (x,y) with respect to the product metric on $X\times Y$ iff $x_n \to x$ in X and $y_n \to y$ in Y.*

Proof. Suppose that $x_n \to x$ in X and $y_n \to y$ in Y. Then $d(x_n, x) \to 0$ and $d(y_n, y) \to 0$ so

$$d((x_n, y_n), (x, y)) = d(x_n, x) + d(y_n, y) \to 0 + 0 = 0,$$

so $(x_n, y_n) \to (x, y)$ in $X \times Y$.

Conversely, suppose that $(x_n, y_n) \to (x, y)$ in $X \times Y$, so $d((x_n, y_n), (x, y)) \to 0$ in $\mathbb{R}$. Clearly we have

$$0 \leq d(x_n, x) \leq d(x_n, x) + d(y_n, y) = d((x_n, y_n), (x, y)) \to 0,$$

so by the sandwich test we see that $d(x_n, x) \to 0$, so that $x_n \to x$ in X. By a similar argument we see that $y_n \to y$ in Y. □

Remark 2.20. Suppose we have a sequence $\underline{x}_1, \underline{x}_2, \dots$ in $\mathbb{R}^n$, where $\underline{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ say. The same argument shows that this sequence converges to a vector $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ iff $x_{ik} \to y_i$ as $k \to \infty$ for $i = 1, \dots, n$.

3. Continuous maps

Definition 3.1. Let X and Y be metric spaces, and let f be a function from X to Y. We say that f is *continuous* if whenever we have a sequence (x_n) converging to $x \in X$, the sequence $(f(x_n))$ in Y converges to the point $f(x)$.

Remark 3.2. If we replace the metrics on X and Y by strongly equivalent ones, this does not change the notion of convergence, and thus does not change the notion of continuity.

Example 3.3. If $X \subseteq \mathbb{R}^n$ and $Y \subseteq \mathbb{R}^m$ and $f(x) = (f_1(x), \dots, f_m(x))$ and each function $f_i \colon X \to \mathbb{R}$ is a polynomial in $(x_1, \dots, x_n)$ then f is continuous. Similarly if $f_i = g_i/h_i$ where g_i and h_i are polynomials and h_i is never zero on X.

Example 3.4. We can define a map $f \colon \mathbb{R}^3 \to \mathbb{R}^3$ by $f(x, y, z) = (x + y + z, xy + yz + zx, xyz)$. As $x + y + z$, $xy + yz + zx$ and xyz are polynomial functions of x, y and z, we see that this is continuous.

Now put $U = \mathbb{R}^3 \setminus \{0\}$ and define $g \colon U \to U$ by

$$g(x, y, z) = (x/(x^2 + y^2 + z^2), y/(x^2 + y^2 + z^2), z/(x^2 + y^2 + z^2)).$$

As $x^2 + y^2 + z^2$ is a polynomial that is nonzero everywhere on U, we see that g is also continuous.

Example 3.5. Define $f, g \colon M_2(\mathbb{R}) \to M_2(\mathbb{R})$ by $f(A) = A^2$ and $g(A) =$ the reduced echelon form of A, and define $h \colon GL_2(\mathbb{R}) \to GL_2(\mathbb{R})$ by $h(A) = A^{-1}$. Then

$$f \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2 + bc & ab + bd \\ ac + cd & bc + d^2 \end{pmatrix},$$

so the entries in $f(A)$ are polynomials in the entries of A, so f is continuous. Similarly,

$$h \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} d/(ad - bc) & -b/(ad - bc) \\ c/(ad - bc) & a/(ad - bc) \end{pmatrix},$$

and $ad - bc = \det(A)$ is a polynomial that is nonzero everywhere on $GL_2(\mathbb{R})$, so h is continuous.

However g is not continuous. To see this, put $A_n = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1/n \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. Clearly $A_n \to B$, but $g(A_n) = I$ and $g(B) = B \neq I$ so $g(A_n) \not\to g(B)$.

Proposition 3.6. *If $X \xrightarrow{f} Y \xrightarrow{g} Z$ and f and g are continuous then gf is continuous.*

Proof. Suppose we have a convergent sequence $x_n \to x$ in X. As f is continuous, we deduce that $f(x_n) \to f(x)$ in Y. As g is continuous, we deduce that $gf(x_n) \to gf(x)$ in Z. This means that gf is continuous as claimed. □

Proposition 3.7. *Let X, Y and Z be metric spaces, and give $Y \times Z$ the product metric. Let $f \colon X \to Y \times Z$ be a function, so $f(x) = (g(x), h(x))$ for some functions $g \colon X \to Y$ and $h \colon X \to Z$. Then f is continuous iff g and h are continuous.*

Proof. Suppose that g and h are continuous. For any convergent sequence $x_n \longrightarrow x$ in X, we have $g(x_n) \longrightarrow g(x)$ in Y (because g is continuous) and $h(x_n) \longrightarrow h(x)$ in Z (because h is continuous). We deduce from Proposition 2.19 that $(g(x_n), h(x_n)) \longrightarrow (g(x), h(x))$, or in other words that $f(x_n) \longrightarrow f(x)$. This shows that f is continuous, as required. The argument can be reversed to prove the converse. □

Definition 3.8. A function $f\colon X \longrightarrow Y$ between metric spaces is *Lipschitz* if there is a number $C \geq 0$ (called a *Lipschitz constant* for f) such that $d(f(x), f(x')) \leq Cd(x, x')$ for all $x, x' \in X$. It is easy to see that Lipschitz functions are continuous.

The following alternative characterisation of continuity is also useful.

Proposition 3.9. *Let $f\colon X \longrightarrow Y$ be a function between metric spaces. Then f is continuous iff for each $\epsilon > 0$ and $x \in X$ there exists $\delta > 0$ such that $d(f(x), f(x')) < \epsilon$ whenever $d(x, x') < \delta$.*

Proof. Suppose that f is continuous, and we are given $\epsilon > 0$ and $x \in X$. Suppose for a contradiction that there is no δ with the stated property. Then in particular, for any $n > 0$, we cannot take $\delta = 1/n$, so there must exist $x_n \in X$ with $d(x, x_n) < 1/n$ but $d(f(x), f(x_n)) \geq \epsilon$. It follows from this that $x_n \longrightarrow x$ but $f(x_n) \not\longrightarrow f(x)$, which is impossible because f is continuous. Thus there must exist a δ after all.

Conversely, suppose that f satisfies the condition in the proposition. Consider a convergent sequence $x_n \longrightarrow x$ in X; we must show that $f(x_n) \longrightarrow f(x)$. In other words, given $\epsilon > 0$, we must show that there exists N such that $d(f(x_n), f(x)) < \epsilon$ whenever $n \geq N$. By assumption we can choose δ such that $d(f(x'), f(x)) < \epsilon$ whenever $d(x', x) < \delta$. As $x_n \longrightarrow x$, we can choose N such that $d(x_n, x) < \delta$ whenever $n \geq N$. It follows that when $n \geq N$ we have $d(f(x_n), f(x)) < \epsilon$ as required. □

Definition 3.10. A *homeomorphism* between spaces X and Y is a continuous map $f\colon X \longrightarrow Y$ which has an inverse map $g\colon Y \longrightarrow X$ (so $f(g(y)) = y$ for all y and $g(f(x)) = x$ for all x) that is also continuous.

Example 3.11. If $a < b$ and $c < d$ then the map $f(t) = c + (t - a)(d - c)/(b - a)$ gives a homeomorphism $f\colon [a, b] \longrightarrow [c, d]$, with inverse $f^{-1}(s) = a + (s - c)(b - a)/(d - c)$. The same formulae give homeomorphisms $(a, b) \simeq (c, d)$ and $[a, b) \simeq [c, d)$ and $(a, b] \simeq (c, d]$.

Example 3.12. Define $f\colon S^1 \times \{1, -1\} \longrightarrow O(2)$ by

$$f(x, y, z) = \begin{pmatrix} x & -yz \\ y & xz \end{pmatrix}$$

so

$$f(\cos(\theta), \sin(\theta), +1) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} = \mathrm{rot}_\theta$$

and

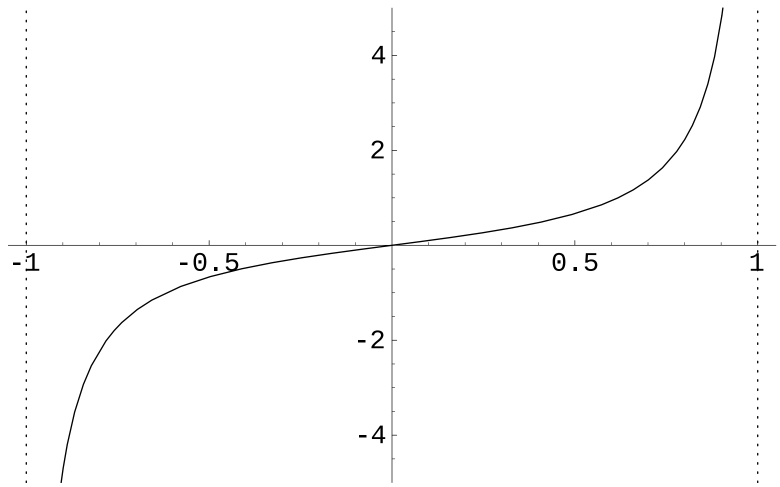
$$f(\cos(\theta), \sin(\theta), -1) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix} = \mathrm{ref}_\theta\,.$$

It is easy to see that this is a homeomorphism, with inverse

$$f^{-1}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = (a, c, ad - bc).$$

Proposition 3.13. *The map $f(x) = x/(1 - x^2)$ gives a homeomorphism $f\colon (-1, 1) \longrightarrow \mathbb{R}$.*

Proof. This is essentially clear from the graph of f:



You can check that the derivative is $f'(x) = (1+x^2)/(1-x^2)^2$, which is always strictly positive, so $f(x)$ is strictly increasing, so it never takes the same value more than once. In other words, it is an injective function from $(-1,1)$ to $\mathbb{R}$. As x approaches ± 1 it is easy to see that $f(x)$ tends to $\pm\infty$. Given any real number y, this means that $f(x) < y$ when x is close to -1, and $f(x) > y$ when x is close to 1. Thus, the intermediate value theorem tells us that $f(x) = y$ for some $x \in (-1,1)$, so f is surjective, and thus bijective.

We now give a more algebraic argument, which enables us to check that the inverse is continuous. For any $y \neq 0$, we define $g(y) = (-1+\sqrt{1+4y^2})/(2y)$; this is a continuous function $g\colon \mathbb{R}\setminus\{0\} \longrightarrow \mathbb{R}$. One can check using L'Hôpital's rule that $g(y) \longrightarrow 0$ as $y \longrightarrow 0$, so we can define $g(0) = 0$ to get a continuous map $g\colon \mathbb{R} \longrightarrow \mathbb{R}$. For $y \neq 0$ we have

$$1 < 1 + 4y^2 < 1 + 4|y| + 4y^2 = (1 + 2|y|)^2,$$

so (by taking square roots)

$$1 < \sqrt{1+4y^2} < 1 + 2|y|.$$

If we subtract 1 and divide by $2|y|$ we find that

$$0 < |g(y)| = \frac{-1+\sqrt{1+4y^2}}{2|y|} < 1,$$

so the range of g is contained in $(-1,1)$, so we can regard g as a map $\mathbb{R} \longrightarrow (-1,1)$. (We need to remark separately that $g(0) \in (-1,1)$, just because $g(0) = 0$.)

The definition of $g(y)$ comes from the quadratic formula: it is one of the solutions of the equation $yx^2 + x - y = 0$. Thus, $yg(y)^2 + g(y) - y = 0$, which can be rearranged to give $y = g(y)/(1-g(y)^2)$, or in other words $y = f(g(y))$.

In the other direction, we have

$$1 + 4f(x)^2 = \frac{(1-x^2)^2}{(1-x^2)^2} + \frac{4x^2}{(1-x^2)^2} = \frac{1-2x^2+x^4+4x^2}{(1-x^2)^2} = \left(\frac{1+x^2}{1-x^2}\right)^2.$$

When $x \in (-1,1)$ we see that the term in brackets is positive, so

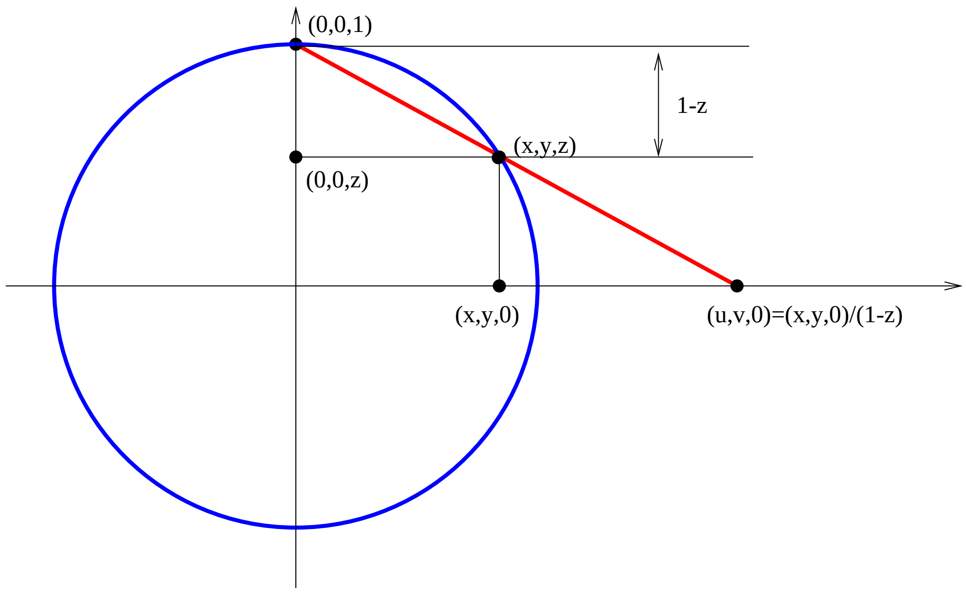
$$\sqrt{1+4f(x)^2} = \frac{1+x^2}{1-x^2}$$

so

$$\frac{-1+\sqrt{1+4f(x)^2}}{2f(x)} = \left(-\frac{1-x^2}{1-x^2} + \frac{1+x^2}{1-x^2}\right) \Big/ \frac{2x}{1-x^2} = x,$$

or in other words $g(f(x)) = x$. Thus f and g are inverses of each other, and they give a homeomorphism $(-1,1) \simeq \mathbb{R}$. □

Example 3.14. Let N be the "north pole" of the sphere S^2, in other words the point $(0,0,1)$. The map $f(x,y,z) = (x/(1-z), y/(1-z))$ gives a homeomorphism $S^2 \setminus \{N\} \to \mathbb{R}^2$, with inverse $g(u,v) = (2u, 2v, u^2+v^2-1)/(u^2+v^2+1)$. This is called *stereographic projection*. Geometrically, $f(x,y,z)$ is the unique point where the line joining N to (x,y,z) meets the plane $z=0$, and $f(u,v)$ is the unique point where the line joining N to $(u,v,0)$ meets S^2.



To validate all this algebraically, first note that $u^2+v^2+1>0$ for all $(u,v)\in\mathbb{R}^2$, so g is defined and continuous everywhere on $\mathbb{R}^2$. Next, note that N is the only point in S^2 where $z=1$, so f is defined and continuous on $S^2\setminus\{N\}$. Moreover, we have

$$\begin{aligned}\|g(u,v)\|^2 &= (u^2+v^2+1)^{-2}((2u)^2+(2v)^2+(u^2+v^2-1)^2) \\ &= \frac{4u^2+4v^2+u^4+v^4+1+2u^2v^2-2u^2-2v^2}{u^4+v^4+1+2u^2v^2+2u^2+2v^2} \\ &= 1,\end{aligned}$$

so g really is a map from $\mathbb{R}^2$ to S^2. As $u^2+v^2-1 < u^2+v^2+1$ for all (u,v) we see that $g(u,v)\neq N$ so g is a map from $\mathbb{R}^2$ to $S^2\setminus\{N\}$. Also, if $g(u,v)=(x,y,z)$ then $x = 2u/(u^2+v^2+1)$ and

$$1-z = 1-\frac{u^2+v^2-1}{u^2+v^2+1} = \frac{(u^2+v^2+1)-(u^2+v^2-1)}{u^2+v^2+1} = \frac{2}{(u^2+v^2+1)},$$

so $x/(1-z)=u$. Similarly $y/(1-z)=v$, and it follows that $fg(u,v)=(u,v)$.

Finally, suppose that $(x,y,z)\in S^2$ and $f(x,y,z)=(u,v)$. Then $x^2+y^2+z^2=1$ and $u=x/(1-z)$ and $v=y/(1-z)$. This gives

$$u^2+v^2+1 = \frac{x^2+y^2+(1-z)^2}{(1-z)^2} = \frac{x^2+y^2+z^2+1-2z}{(1-z)^2} = \frac{2-2z}{(1-z)^2} = \frac{2}{1-z}.$$

Thus

$$\frac{2u}{u^2+v^2+1} = \frac{2x}{1-z}\Big/\frac{2}{1-z} = x.$$

Similarly, we see that $(2v)/(u^2+v^2+1) = y$. We also have $x^2+y^2 = 1-z^2$, so

$$\begin{aligned} u^2+v^2-1 &= \frac{x^2+y^2-(1-z)^2}{(1-z)^2} \\ &= \frac{x^2+y^2-z^2-1+2z}{(1-z)^2} \\ &= \frac{1-z^2-z^2-1+2z}{(1-z)^2} \\ &= \frac{2z(1-z)}{1-z} = \frac{2z}{1-z}. \end{aligned}$$

From this we see that $(u^2+v^2-1)/(u^2+v^2+1) = z$, and it follows that $gf(x,y,z) = (x,y,z)$ as required.

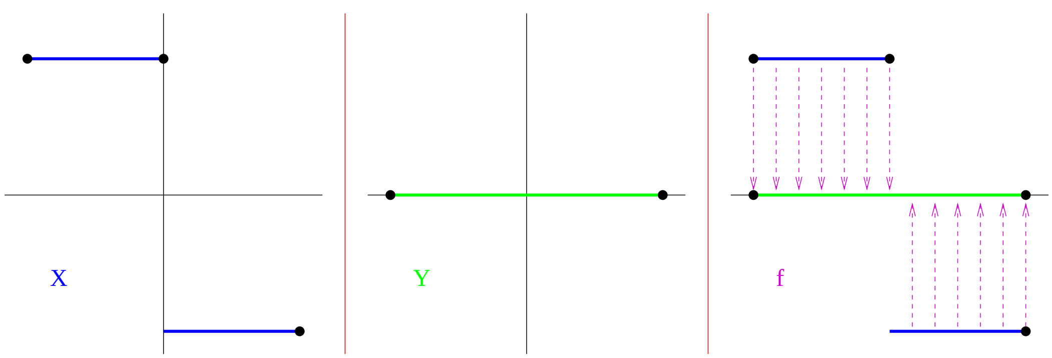
Example 3.15. The previous example can be generalised: if a is any point in S^n for any n, then $S^n \setminus \{a\}$ is homeomorphic to $\mathbb{R}^n$. To see this, put $V = \{x \in \mathbb{R}^{n+1} \ : \ \langle a, x\rangle = 0\}$, the set of vectors orthogonal to a. This is a vector subspace of dimension n, so it is isomorphic to $\mathbb{R}^n$, and one can check that any linear isomorphism is in fact a homeomorphism. It is thus enough to show that $S^n \setminus \{a\}$ is homeomorphic to V. As before, for $x \in S^n \setminus \{a\}$ we let $f(x)$ be the point where the line from a through x meets v. The formulae are

$$\begin{aligned} f(x) &= \frac{x - \langle x, a\rangle a}{1 - \langle x, a\rangle} \\ f^{-1}(x) &= \frac{2v + (\|v\|^2 - 1)a}{\|v\|^2 + 1}. \end{aligned}$$

Example 3.16. Put $X = [-1,0] \times \{1\} \cup (0,1] \times \{0\} \subset \mathbb{R}^2$; note that the point $(0,1)$ is included in X, but $(0,-1)$ is not. Put $Y = [-1,1] \times \{0\}$, and define $f\colon X \to Y$ by $f(x,y) = (x,0)$. This is evidently continuous, and in fact is a bijection, with inverse

$$g(x,0) = \begin{cases} (x,1) & \text{if } x \leq 0 \\ (x,-1) & \text{if } x > 0. \end{cases}$$

However, the map g is discontinuous at the point $(0,0)$, so f is not a homeomorphism.



Example 3.17. The map $f(x) = (\|x\|, x/\|x\|)$ is a homeomorphism $\mathbb{R}^n \setminus \{0\} \simeq (0,\infty) \times S^{n-1}$, with inverse $g(t,y) = ty$.

Example 3.18. Recall the torus

$$T = S^1 \times S^1 = \{(u,v,x,y) \in \mathbb{R}^4 : u^2+v^2 = 1 = x^2+y^2\}.$$

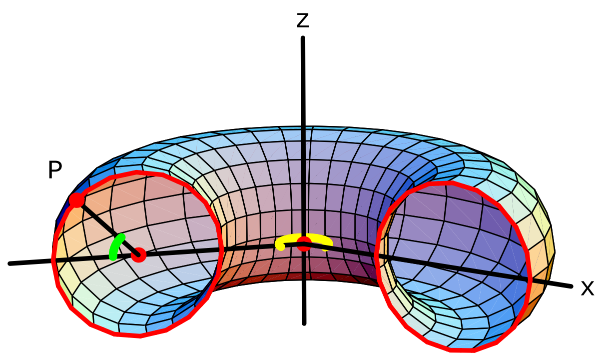
This is homeomorphic to a sort of inner-tube shape, which can be embedded in $\mathbb{R}^3$. To be more precise, let C be the vertical circle of radius 1 lying in the (x,z) plane and centred at the point $(2,0,0)$. The formula is

$$C = \{(x,0,z) \in \mathbb{R}^3 \ : \ (x-2)^2 + z^2 = 1\}.$$

Let T' be the surface obtained by rotating C around the z-axis. One can define a homeomorphism $f\colon T \longrightarrow T'$ by

$$\begin{aligned} f(u,v,x,y) &= 2(u,v,0) + x(u,v,0) + y(0,0,1) \\ &= ((2+x)u, (2+x)v, y). \end{aligned}$$

More geometrically, $f(\cos(\theta), \sin(\theta), \cos(\phi), \sin(\phi))$ is the point P in the diagram below, where the angle marked in yellow is θ, and the angle marked in green is ϕ.



The inverse map $f^{-1}\colon T' \longrightarrow T$ is given by

$$f^{-1}(X,Y,Z) = \left(\frac{X}{\sqrt{X^2+Y^2}}, \frac{Y}{\sqrt{X^2+Y^2}}, \frac{X^2+Y^2+Z^2-5}{4}, Z \right).$$

Example 3.19. Consider the spaces

$$X = \{(u,v,w) \in \mathbb{C}^3 : u \neq v \text{ and } v \neq w \text{ and } w \neq u\}$$

and

$$Y = \mathbb{C} \times (\mathbb{C} \setminus \{0\}) \times (\mathbb{C} \setminus \{0,1\}) = \{(a,b,c) \in \mathbb{C}^3 : b \neq 0 \text{ and } c \neq 0 \text{ and } c \neq 1\}.$$

I claim that X is homeomorphic to Y. Indeed, we can define a homeomorphism $f\colon X \longrightarrow Y$ by $\phi(u,v,w) = (u, v-u, (w-u)/(v-u))$; the inverse is $f^{-1}(a,b,c) = (a, a+b, a+bc)$.

Example 3.20. We next consider an example involving a space of functions. Specifically, let X be the set of all functions $f\colon \mathbb{R} \longrightarrow \mathbb{C}$ that satisfy the differential equations

$$\begin{aligned} f'' + f &= 0 \\ |f|^2 + |f'|^2 &= 2. \end{aligned}$$

One can check that there is a homeomorphism $\phi\colon S^3 \longrightarrow X$ given by

$$\phi(a,b,c,d)(x) = (a+ib)e^{ix} + (c+id)e^{-ix}.$$

4. Compactness

Definition 4.1. A metric space X is *compact* if every sequence in X has a subsequence that converges in X.

Example 4.2. The space $(0,1]$ is not compact, because the sequence $1, 1/2, 1/3, \dots$ has no convergent subsequence. Indeed, the sequence converges in $\mathbb{R}$ to 0, so every subsequence converges in $\mathbb{R}$ to 0, and thus cannot converge in $(0,1]$ to anything.

Example 4.3. The space $\mathbb{R}$ is not compact, because the sequence $(1,2,3,\dots)$ has no convergent subsequence. Indeed, we have $d(n,m) \geq 1$ for any two distinct terms in the sequence, which clearly means that no subsequence can be Cauchy.

Example 4.4. Let X be a space with only finitely many points; then I claim that X is compact. Indeed, suppose that $X = \{a_1, \dots, a_m\}$. Given a sequence (x_n) in X, put $S_k = \{n : x_n = a_k\}$. Then $\mathbb{N} = S_1 \cup \dots \cup S_k$, so at least one of the sets S_k must be infinite, say $S_k = \{n_1, n_2, \dots\}$ with $n_1 < n_2 < \dots$. Each term in the subsequence $x_{n_1}, x_{n_2}, \dots$ is equal to a_k, so the subsequence obviously converges to a_k.

Example 4.5. Let X be the space of all continuous functions from I to I, with the max metric. Then X is not compact. Indeed, consider the sequence $f_1, f_2, \dots$, where $f_n(x) = x^n$. This converges pointwise to the discontinuous function g, where $g(x) = 0$ for $x < 1$ and $g(1) = 1$. It follows that every subsequence converges pointwise to g and thus does not converge at all in X.

Proposition 4.6. *The space I is compact.*

Proof. Consider a sequence (x_n) in I. Suppose that there is an increasing subsequence (y_n). This subsequence is clearly bounded above by 1 (and below by 0) so it converges to some point z. As $0 \leq y_n \leq 1$, we see that $0 \leq z \leq 1$, so some subsequence of (x_n) converges to a point in I as required. A similar argument works if (x_n) has a decreasing subsequence. Moreover, the "Spanish hotels theorem" guarantees that every sequence in $\mathbb{R}$ has either an increasing subsequence or a decreasing one, so the argument works in general. □

Definition 4.7. A metric space X is *bounded* if there is a constant C such that $d(x, y) \leq C$ for all $x, y \in X$.

Remark 4.8. The bounded space $(-1, 1)$ is homeomorphic to the unbounded space $\mathbb{R}$ (see Proposition 3.13). Thus, unlike most of the other concepts that we have defined, boundedness is not homeomorphism-invariant.

Exercise 4.9. Show that a nonempty metric space X is bounded iff there is a point $x_0 \in X$ and a constant $C \in \mathbb{R}$ such that $d(x, x_0) \leq C$ for all $x \in X$.

Proposition 4.10. *If X is compact and there exists a continuous surjective map $f: X \to Y$ then Y is compact. In particular this holds if Y is homeomorphic to X.*

Proof. Given a sequence (y_n) in Y, choose points $x_n \in X$ such that $f(x_n) = y_n$ (which is possible because f is surjective). As X is compact, we have $x_{n_k} \to x$ for some subsequence (x_{n_k}) of (x_n) and some point $x \in X$. As f is continuous we have $y_{n_k} = f(x_{n_k}) \to f(x)$, which shows that some subsequence of (y_n) converges in Y. Thus Y is compact, as claimed. □

Corollary 4.11. *If $f: X \to Y$ is a continuous map and X' is a compact subspace of X, then the space $Y' := f(X')$ is a compact subspace of Y (because we can regard f as a continuous surjection from X' to Y').*

Proposition 4.12. *A closed subspace of a compact space is compact.*

Proof. Let X be a compact space, and let Y be a closed subspace. Suppose we have a sequence (y_n) in Y. We can regard this a a sequence in X, so by compactness there is a subsequence (y_{n_k}) that converges to some point $y \in X$. However, Y is closed and (y_{n_k}) is a sequence in Y that converges in X to y so y must also lie in Y. Thus our original sequence has a subsequence that converges in Y, showing that Y is compact. □

Proposition 4.13. *If X and Y are compact metric spaces, then so is $X \times Y$.*

Proof. Suppose we have a sequence (z_n) in $X \times Y$, where $z_n = (x_n, y_n)$ say. As X is compact, some subsequence (x_{n_k}) of (x_n) converges in X to some point x say. Write $x'_k = x_{n_k}$ and $y'_k = y_{n_k}$ and $z'_k = (x'_k, y'_k)$, so $x'_k \to x$.

Next, as Y is compact, some subsequence of the sequence (y'_k) converges in Y, say $y'_{k_j} \to y$. Put $x''_j = x'_{k_j}$ and $y''_j = y'_{k_j}$ and $z''_j = (x''_j, y''_j)$. Then (z''_j) is a subsequence of (z_n), and $y''_j \to y$. Moreover, (x''_j) is a subsequence of the sequence (x'_k), which converges to x, so $x''_j \to x$. Using Proposition 2.19 we deduce that $z''_j \to (x, y)$, so our original sequence has a convergent subsequence as required. □

Corollary 4.14. *If $X_1, \dots, X_n$ are compact metric spaces then so is $X_1 \times \dots \times X_n$.*

Proof. Induction. □

Theorem 4.15. *A subspace $X \subseteq \mathbb{R}^n$ is compact iff it is closed and bounded.*

Proof. First suppose that X is compact. We will use the same arguments as in Examples 4.2 and 4.3 to show that X is closed and bounded. First, suppose we have a sequence (x_n) in X that converges to some point $x \in \mathbb{R}^n$. By compactness, some subsequence (x_{n_k}) converges to some point x' in X. However, as the sequence (x_n) converges to x we see that every subsequence also converges to x so $x = x'$. As $x' \in X$ we see that $x \in X$, which proves that X is closed.

Next, suppose for a contradiction that X is unbounded (and thus nonempty). Choose a point $x_0 \in X$. Using Exercise 4.9, we see that there is no constant C such that $d(x, x_0) \leq C$ for all x. We can thus choose a sequence (x_k) such that $d(x_{k+1}, x_0) \geq d(x_k, x_0) + 1$ for all $k \geq 0$. The triangle inequality implies that

$$d(x_n, x_m) \geq |d(x_n, x_0) - d(x_m, x_0)| \geq 1$$

whenever $n \neq m$. Using this, we see that no subsequence of (x_n) can be Cauchy, so no subsequence can converge. This contradicts compactness, so X must be bounded after all.

For the converse, we now assume that X is bounded and closed. As X is bounded, it is contained in some set of the form $[-R, R]^n$. The space $[-R, R]$ is homeomorphic to I, and thus is compact (by Propositions 4.6 and 4.10). It follows from Corollary 4.14 that $[-R, R]^n$ is compact, and thus from Proposition 4.12 that X is compact. □

Corollary 4.16. *If X is a compact metric space and $f : X \to \mathbb{R}$ is continuous then f is bounded. (In other words, there is a constant C such that $|f(x)| \leq C$ for all $x \in X$.)*

Proof. The space $f(X) \subseteq \mathbb{R}$ is compact by Proposition 4.10, so the Theorem tells us that $f(X) \subseteq [-C, C]$ for some C, as required. □

Theorem 4.17. *If $f : X \to Y$ is a continuous bijection between metric spaces and X is compact, then f^{-1} is also continuous (so f is a homeomorphism).*

Remark 4.18. In Example 3.16 we saw that the projection map $[-1, 0] \times \{1\} \cup (0, 1] \times \{-1\} \to [-1, 1]$ is a continuous bijection but not a homeomorphism. This does not contradict the Theorem, because the set $[-1, 0] \times \{1\} \cup (0, 1] \times \{-1\}$ is not closed in $\mathbb{R}^2$ and thus (by Theorem 4.15) is not compact.

Proof. Suppose not. Then we can find a convergent sequence $y_n \to y$ in Y such that $f^{-1}(y_n) \not\to f^{-1}(y)$. Write $x_n = f^{-1}(y_n)$ and $x = f^{-1}(y)$, so $x_n \not\to x$. By the definition of convergence, there exists $\epsilon > 0$ such that there is no $N \geq 0$ for which $d(x_n, x) < \epsilon$ for all $n \geq N$. This means that we can choose a strictly increasing sequence $n_1 < n_2 < \dots$ such that $d(x_{n_k}, x) \geq \epsilon$ for all k. Put $x'_k = x_{n_k}$, so (x'_k) is a subsequence of (x_n). As X is compact, we can choose a subsequence (x''_j) of (x'_k) such that $x''_j \to x''$ for some $x'' \in X$. As f is continuous, we conclude that $f(x''_j) \to f(x'')$. However, $(f(x''_j))$ is a subsequence of $(f(x_n))$ and $f(x_n) = y_n \to y$ so $f(x''_j) \to y$. As limits are unique, we conclude that $f(x'') = y$, so $x'' = f^{-1}(y) = x$. This means that $x''_j \to x$, which is a contradiction because $d(x''_j, x) \geq \epsilon$ for all j. Thus f^{-1} must be continuous after all. □

5. Uniform continuity

Definition 5.1. Let $f : X \to Y$ be a map between metric spaces. We say that f is *uniformly continuous* if given $\epsilon > 0$ we can choose $\delta > 0$ such that $d(f(x), f(x')) < \epsilon$ whenever $d(x, x') < \delta$.

At first sight, Proposition 3.9 might seem to say that all continuous maps are uniformly continuous, but there is a subtle point. In Proposition 3.9, the number δ is allowed to depend on the point x; in the definition above, the number δ is the same for all points $x \in X$.

Example 5.2. Consider the map $f\colon \mathbb{R} \longrightarrow \mathbb{R}$ given by $f(x) = x^2$, which is of course continuous. We claim that it is not uniformly continuous; in fact, if we take $\epsilon = 1$, then there is no $\delta > 0$ such that $d(f(x), f(x')) = |f(x) - f(x')| < \epsilon = 1$ whenever $|x - x'| < \delta$. Indeed, for any $\delta > 0$ we can take $x = 1/\delta$ and $x' = 1/\delta + \delta/2$ and we find that $|x - x'| = \delta/2 < \delta$ and $|f(x) - f(x')| = 1 + \delta^2/4 > 1$.

On the other hand, it is an important fact that when X is compact, continuity implies uniform continuity.

Theorem 5.3. *Let $f\colon X \longrightarrow Y$ be a continuous map of metric spaces, and suppose that X is compact. Then f is uniformly continuous.*

Proof. Suppose $\epsilon > 0$; we must find $\delta > 0$ such that $d(f(x), f(x')) < \delta$ whenever $d(x, x') < \epsilon$. Suppose for a contradiction that there is no such δ. In particular, for any $n > 0$, we cannot take $\delta = 1/n$, so there must exist $x_n, x'_n \in X$ with $d(x_n, x'_n) < 1/n$ but $d(f(x_n), f(x'_n)) \geq \epsilon$. As X is compact, we can choose a convergent subsequence (x_{n_k}) of (x_n), with limit y say. Note that

$$d(x'_{n_k}, y) \leq d(x'_{n_k}, x_{n_k}) + d(x_{n_k}, y) < 1/n_k + d(x_{n_k}, y) \longrightarrow 0 + 0 = 0.$$

Thus, the sequence (x'_{n_k}) also converges to y. As f is continuous we conclude that $f(x_{n_k}) \longrightarrow f(y)$ and also $f(x'_{n_k}) \longrightarrow f(y)$, so $d(f(x_{n_k}), f(x'_{n_k})) \longrightarrow 0$. This contradicts the fact that $d(f(x_n), f(x'_n)) \geq \epsilon$ for all n. □

6. Projective space

Proposition 6.1. *There is a homeomorphism $f\colon S^1 \longrightarrow \mathbb{R}P^1$, given by*

$$f(x, y) = \frac{1}{2}\begin{pmatrix} 1 + x & y \\ y & 1 - x \end{pmatrix}.$$

Proof. Recall that $\mathbb{R}P^1$ is the set of all 2×2 matrices A that satisfy the conditions $A^2 = A^T = A$ and $\operatorname{trace}(A) = 1$. If $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ then $A^T = \left(\begin{smallmatrix} a & c \\ b & d \end{smallmatrix}\right)$ and $\operatorname{trace}(A) = a + d$. Thus, if $A^T = A$ and $\operatorname{trace}(A) = 1$ then we must have $b = c$ and $d = 1 - a$, so A has the form $\left(\begin{smallmatrix} a & b \\ b & 1-a \end{smallmatrix}\right)$. We then find that

$$A^2 = \begin{pmatrix} a^2 + b^2 & b \\ b & 1 - 2a + a^2 + b^2 \end{pmatrix},$$

so $A^2 = A$ if and only if $a^2 + b^2 = a$. In summary, $\mathbb{R}P^1$ is the set of matrices $\left(\begin{smallmatrix} a & b \\ b & 1-a \end{smallmatrix}\right)$ for which $a^2 + b^2 = a$.

On the other hand S^1 is the set of vectors $(x, y) \in \mathbb{R}^2$ for which $x^2 + y^2 = 1$.

If $(x, y) \in S^1$ and we put $a = (1 + x)/2$ and $b = y/2$ then $1 - a = (1 - x)/2$ and

$$a^2 + b^2 = (1 + 2x + x^2 + y^2)/4 = (2 + 2x)/4 = a$$

so the matrix

$$f(x, y) = \begin{pmatrix} a & b \\ b & 1 - a \end{pmatrix} = \frac{1}{2}\begin{pmatrix} 1 + x & y \\ y & 1 - x \end{pmatrix}$$

lies in $\mathbb{R}P^1$. This defines a continuous map $f\colon S^1 \longrightarrow \mathbb{R}P^1$.

Conversely, suppose we start with a matrix $A = \left(\begin{smallmatrix} a & b \\ b & 1-a \end{smallmatrix}\right) \in \mathbb{R}P^1$, so $a^2 + b^2 - a = 0$. Then

$$(2a - 1)^2 + (2b)^2 = 4a^2 - 4a + 1 + 4b^2 = 4(a^2 + b^2 - a) + 1 = 1,$$

so the vector $g(A) = (2a - 1, 2b)$ lies in S^1. This gives a continuous map $g\colon \mathbb{R}P^1 \longrightarrow S^1$. It is easy to check that $f(g(A)) = A$ and $g(f(x, y)) = (x, y)$, so f and g are homeomorphisms. □

When $n > 1$, the topology of $\mathbb{R}P^n$ is more complicated, and is related to the topology of S^n in a more subtle way.

Proposition 6.2. *There is a surjective continuous map $q\colon S^n \longrightarrow \mathbb{R}P^n$, such that $q(x) = q(y)$ if and only if $y = x$ or $y = -x$.*

Proof. Combine Lemma 6.3, Lemma 6.4 and Corollary 6.6. □

Lemma 6.3. *There is a continuous map $q\colon S^n \to \mathbb{R}P^n$, such that $q(-x) = q(x)$, and $q(x)y = \langle x, y\rangle x$ for all vectors $y \in \mathbb{R}^{n+1}$. (Here $q(x)y$ is the usual product of the matrix $q(x)$ with the vector y.)*

Proof. Suppose that $x \in S^n$, so $x = (x_0, \dots, x_n) \in \mathbb{R}^n$ and $\sum_i x_i^2 = 1$. We define $A_{ij} = x_i x_j$, which gives a square matrix $A \in M_{n+1}(\mathbb{R})$. Clearly $A_{ji} = A_{ij}$, so $A = A^T$. We also have

$$\mathrm{trace}(A) = \sum_{i=0}^{n} A_{ii} = \sum_{i=0}^{n} x_i^2 = 1.$$

Finally, recall the definition of matrix multiplication: the (i, j)'th entry in AB is $\sum_k A_{ik} B_{kj}$. Taking $B = A$, we see that the (i, j)'th entry in A^2 is

$$\sum_k A_{ik} A_{kj} = \sum_k x_i x_k x_k x_j = x_i x_j \sum_k x_k^2 = x_i x_j = A_{ij}.$$

This shows that $A^2 = A$ and thus that $A \in \mathbb{R}P^n$.

Similarly, for any vector y, the i'th entry in the vector Ay is

$$\sum_j A_{ij} y_j = x_i \sum_j x_j y_j = \langle x, y\rangle x_i,$$

so $Ay = \langle x, y\rangle x$ as stated.

As A depends (only) on x, we can define a function $q\colon S^n \to \mathbb{R}P^n$ by $q(x) = A$. This is clearly continuous. As $(-x_i)(-x_j) = x_i x_j$, we see that $q(-x) = q(x)$. □

Lemma 6.4. *The map $q\colon S^n \to \mathbb{R}P^n$ is surjective.*

Proof. Suppose that $B \in \mathbb{R}P^n$, so $B^2 = B^T = B$ and $\mathrm{trace}(B) = 1$. We must find x such that $B = q(x)$. Because $\mathrm{trace}(B) = 1$ we have $B \neq 0$ so $Bu \neq 0$ for some vector u. Put $x = Bu/\|Bu\|$, so $x \in S^n$, so we can put $A = q(x)$. Using $B^2 = B$ we see that $Bx = B^2 u/\|Bu\| = x$. The previous lemma tells us that for any vector y we have $Ay = \langle y, x\rangle x$ so $BAy = \langle y, x\rangle Bx = \langle y, x\rangle x = Ay$, so $BA = A$. Taking transposes gives $A^T B^T = (BA)^T = A^T$, but $A^T = A$ and $B^T = B$ so we see that $AB = A$ also. It follows that

$$\begin{aligned}(A - B)(A^T - B^T) &= (A - B)(A - B)\\ &= A^2 - AB - BA + B^2\\ &= A - A - A + B\\ &= B - A.\end{aligned}$$

Thus, $\mathrm{trace}((A - B)(A^T - B^T)) = \mathrm{trace}(B) - \mathrm{trace}(A) = 1 - 1 = 0$. Thus, Proposition 2.12 shows that $d_2(A, B) = 0$, so $B = A = q(x)$ as required. □

Lemma 6.5. *If $x, y \in S^n$ then $\langle x, y\rangle^2 = \mathrm{trace}(q(x)q(y))$ and $d_2(q(x), q(y)) = \sqrt{2(1 - \langle x, y\rangle^2)}$.*

Proof. In general, if we have matrices A and B, with entries A_{ij} and B_{ij}, then the (i, k)'th entry in the product matrix AB is $\sum_j A_{ij} B_{jk}$. In particular, the (i, i)'th entry is $\sum_j A_{ij} B_{ji}$. Thus, the trace of AB is $\sum_i \sum_j A_{ij} B_{ji}$. Now take $A = q(x)$ and $B = q(y)$, so $A_{ij} = x_i x_j$ and $B_{ji} = y_j y_i$, so

$$\mathrm{trace}(AB) = \sum_{ij} x_i x_j y_j y_i = (\sum_i x_i y_i)(\sum_j x_j y_j) = \langle x, y\rangle\langle x, y\rangle = \langle x, y\rangle^2.$$

Similarly, we have $\mathrm{trace}(BA) = \langle x, y\rangle^2$.

Next observe that $A^2 = A^T = A$ and $B^2 = B^T = B$ so

$$(A - B)(A^T - B^T) = (A - B)(A - B) = A^2 - AB - BA + B^2 = A + B - AB - BA.$$

As $\mathrm{trace}(A) = \mathrm{trace}(B) = 1$ and $\mathrm{trace}(AB) = \mathrm{trace}(BA) = \langle x, y\rangle^2$ we find that $\mathrm{trace}((A - B)(A^T - B^T)) = 2(1 - \langle x, y\rangle^2)$. Proposition 2.12 tells us that this is the same as $d_2(A, B)^2$, which completes the proof. □

Corollary 6.6. *If $x, y \in S^n$ and $q(x) = q(y)$ then either $x = y$ or $x = -y$.*

Proof. It is clear from the lemma that $q(x) = q(y)$ iff $d_2(q(x), q(y)) = 0$ iff $\langle x, y\rangle^2 = 1$ iff $\langle x, y\rangle = \pm 1$. As x and y are unit vectors, the only way we can have $\langle x, y\rangle = \pm 1$ is if $y = \pm x$. □

Proposition 6.2 has the following corollary.

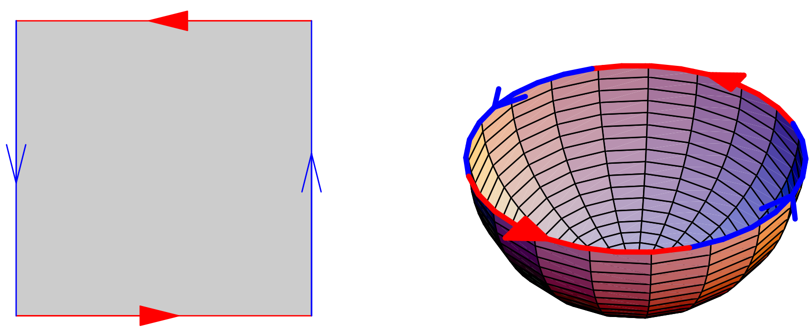
Corollary 6.7. *If we define an equivalence relation $\sim$ on S^n by $x \sim y$ if and only if $y = x$ or $y = -x$, then q induces a bijection $(S^n/\sim) \longrightarrow \mathbb{R}P^n$.* □

In the case $n = 1$, Proposition 6.1 gives a homeomorphism $f\colon S^1 \longrightarrow \mathbb{R}P^1$, and Proposition 6.2 gives a map $q\colon S^1 \longrightarrow \mathbb{R}P^1$ that satisfies $q(-x) = q(x)$ and thus cannot be a homeomorphism. To understand the relationship between f and q, consider $f^{-1} \circ q\colon S^1 \longrightarrow S^1$. Using the formulae in the two proofs and the fact that $x^2 + y^2 = 1$ we find that

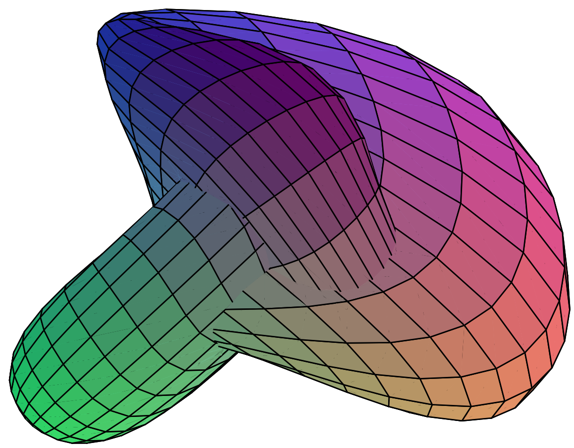
$$f^{-1}q(x,y) = f^{-1}\begin{pmatrix} x^2 & xy \\ xy & y^2 \end{pmatrix} = (2x^2 - 1, 2xy) = (x^2 - y^2, 2xy).$$

For comparison, observe that $(x + iy)^2 = (x^2 - y^2) + 2xyi$. Thus, if we identify S^1 with the set $\{z \in \mathbb{C} \,:\, |z| = 1\}$ then $f^{-1}q(z) = z^2$.

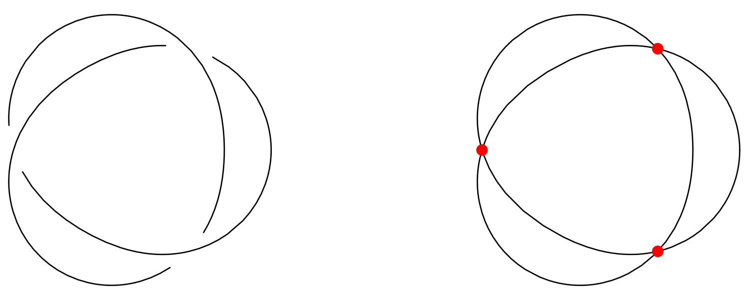
Next, we relate our definition of $\mathbb{R}P^2$ to that given in the Knots and Surfaces course (for example). There, one starts with a square, and attaches the edges together as shown on the left below.



Equivalently, we can start with the lower hemisphere as shown on the right, and attach the boundary circle to itself as shown. Of course, it is not possible to attach the circle to itself in this way in three dimensions without making the surface cross over itself, so $\mathbb{R}P^2$ cannot be realised as a subset of $\mathbb{R}^3$. If we allow self-intersections we get the following picture, called the Boy surface:



The relationship between $\mathbb{R}P^2$ and the Boy surface is like the relationship between (for example) the trefoil knot (a curve in $\mathbb{R}^3$ without self-intersections) and its plane projection (a curve in $\mathbb{R}^2$ that crosses itself in three places).



7. Open and closed sets

Definition 7.1. Let X be a metric space, x a point of X, and r a strictly positive real number. Define

$$\overset{\circ}{B}_r(x) = \{y \in X \ : \ d(x,y) < r\}$$
$$B_r(x) = \{y \in X \ : \ d(x,y) \leq r\};$$

These are called the open and closed balls of radius r around x.

Definition 7.2. Let X be a metric space. A subset $U \subseteq X$ is *open* if for each $x \in U$ there exists $\epsilon > 0$ such that $\overset{\circ}{B}_\epsilon(x) \subseteq U$. A subset $F \subseteq X$ is *closed* if the complement $F^c = X \setminus F$ is open.

Remark 7.3. Yet again, this depends only on the strong equivalence class of the metric on X.

Proposition 7.4. *A subspace $F \subseteq X$ is closed iff the following condition holds: if x_n is a sequence in F that converges in X to a point $x \in X$, then $x \in F$.*

Proof. Put $U = X \setminus F$. Suppose that F satisfies the stated condition; we need to prove that F is closed, or equivalently that U is open. If not, then there must exist $x \in U$ such that no ball $\mathring{B}_\epsilon(x)$ is contained in U. In particular, the ball $\mathring{B}_{1/n}(x)$ is not contained in U, so we can choose $x_n \in X$ such that $d(x_n, x) < 1/n$ and $x_n \notin U$. As $d(x_n, x) < 1/n$ we clearly have $x_n \to x$. As $U = X \setminus F$ and $x_n \notin U$ we must have $x_n \in F$. Our condition on F now tells us that $x \in F$, contradicting the assumption that $x \in U$. This contradiction shows that U must be open after all.

Conversely, suppose that F is closed, or equivalently U is open. We need to show that F satisfies the stated condition. Suppose that $x_n \to x$ in X and $x_n \in F$ for all n; we need to show that $x \in F$. If not, then $x \in X \setminus F = U$. As U is open we know that some ball $\mathring{B}_\epsilon(x)$ is contained in U. As $x_n \to x$, there exists $N \geq 0$ such that $d(x_n, x) < \epsilon$ whenever $n \geq N$. This means that $x_N \in B_\epsilon(x) \subseteq U$, so $x_N \in U$, so $x_N \notin F$, which contradicts our assumption about the sequence (x_n). This contradiction shows that we must have $x \in F$ after all. □

Proposition 7.5.

1. *If $U_1, \dots, U_n$ is a finite sequence of open sets, then $\bigcap_{i=1}^n U_i = U_1 \cap \dots \cap U_n$ is also open.*
2. *If $U_1, U_2, \dots$ is a (possibly infinite) sequence of open sets then $\bigcup_i U_i$ is open.*
3. *If $F_1, \dots, F_n$ is a finite sequence of closed sets, then $\bigcup_{i=1}^n F_i = F_1 \cup \dots \cup F_n$ is also closed.*
4. *If $F_1, F_2, \dots$ is a (possibly infinite) sequence of closed sets then $\bigcap_i F_i$ is closed.*

Proof.

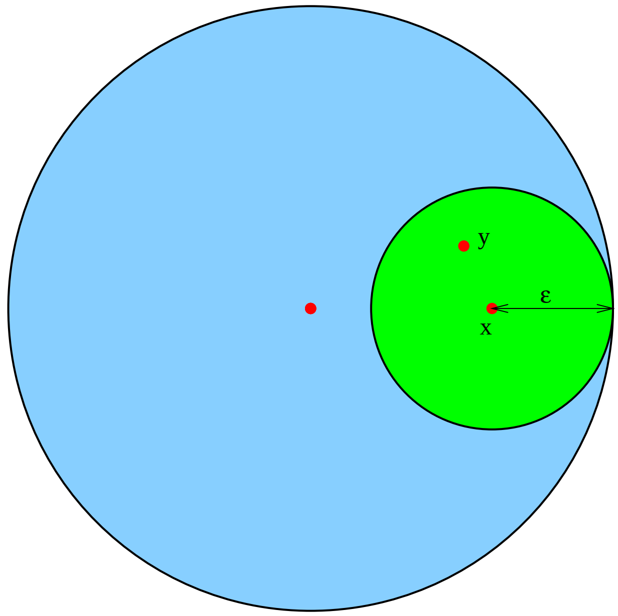
1. Suppose $x \in U_1 \cap \dots \cap U_n$. As U_i is open and $x \in U_i$ we see that $\mathring{B}_{\epsilon_i}(x) \subseteq U_i$ for some $\epsilon_i > 0$. Now put $\epsilon = \min(\epsilon_1, \dots, \epsilon_n)$, so that $\mathring{B}_\epsilon(x) \subseteq \mathring{B}_{\epsilon_i}(x) \subseteq U_i$ for all i. Thus $\mathring{B}_\epsilon(x) \subseteq \bigcap_i U_i$, and this shows that $\bigcap_i U_i$ is open as required.
2. Suppose $x \in \bigcup_i U_i$, so $x \in U_j$ for some j. As U_j is open we have $B_\epsilon(x) \subseteq U_j$ for some $\epsilon > 0$. Clearly $U_j \subseteq \bigcup_i U_i$, so $B_\epsilon(x) \subseteq \bigcup_i U_i$. This shows that $\bigcup_i U_i$ is open as required.
3. This follows from (1) by taking complements, noting that $X \setminus \bigcap_i U_i = \bigcup_i (X \setminus U_i)$.
4. This follows from (2) by taking complements, noting that $X \setminus \bigcup_i U_i = \bigcap_i (X \setminus U_i)$.

□

Remark 7.6. The intersection of an infinite sequence of open sets need not be open; for an example, consider $U_n = (-1/n, 1/n) \subset \mathbb{R}$, so $\bigcap_n U_n = \{0\}$. Similarly, the union of an infinite sequence of closed sets need not be closed; for an example, consider $F_n = [-1 + 1/n, 1 - 1/n]$, so $\bigcup_n F_n = (-1, 1)$.

Example 7.7. The set $\mathring{B}^n = \{x \in \mathbb{R}^n \ : \ \|x\| < 1\}$ is open in $\mathbb{R}^n$. Indeed, suppose that $x \in \mathring{B}^n$, and put $\epsilon = 1 - \|x\| > 0$. If $y \in \mathring{B}_\epsilon(x)$ then $\|y - x\| < \epsilon$ so $\|y\| = \|x + (y - x)\| \leq \|x\| + \|y - x\| <$

$\|x\| + \epsilon = 1$, so $y \in \mathring{B}^n$. This shows that $\mathring{B}_\epsilon(x) \subseteq \mathring{B}^n$, so $\mathring{B}^n$ is open as claimed.



Example 7.8. The set $B^n = \{x \in \mathbb{R}^n \ : \ \|x\| \leq 1\}$ is closed in $\mathbb{R}^n$. Indeed, if $x_k \in B^n$ for all k and $x_k \longrightarrow x$ in $\mathbb{R}^n$ then $\|x_k\| \longrightarrow \|x\|$ but $\|x_k\| \leq 1$ so $\|x\| \leq 1$ so $x \in B^n$, as required. Similar proofs show that S^{n-1} and Δ_n are closed in $\mathbb{R}^n$.

Proposition 7.9. *Let X and Y be metric spaces, and let $F \subseteq X$ and $G \subseteq Y$ be closed sets. Then $F \times G$ is closed in $X \times Y$.*

Proof. Suppose we have a convergent sequence $(x_n, y_n) \longrightarrow (x, y)$ in $X \times Y$, with $(x_n, y_n) \in F \times G$ for all n. This means that $x_n \in F$ and $y_n \in G$ for all n. Moreover, Proposition 2.19 tells us that $x_n \longrightarrow x$ in X and $y_n \longrightarrow y$ in Y. As F and G are closed, we deduce that $x \in F$ and $y \in G$, or equivalently that $(x, y) \in F \times G$. This proves that $F \times G$ is closed, as claimed. □

It is an important fact that continuity of functions can be defined in terms of open and closed sets, without explicit mention of the metric. To prove this, we need to recall a few facts about preimages.

Definition 7.10. Let $f \colon X \longrightarrow Y$ be any function. For any set $A \subseteq X$, we write $f(A) = \{f(x) \ : \ x \in A\} \subseteq Y$, and call this the *image of* A under f. For any set $B \subseteq Y$, we write $f^{-1}B = \{x \in X \ : \ f(x) \in B\}$, and call this the *preimage* of B under f.

Proposition 7.11. *Let $f \colon X \longrightarrow Y$ be a function. Then for any sets $A, B \subseteq Y$ we have*

(a) $f^{-1}(A \cap B) = f^{-1}A \cap f^{-1}B$
(b) $f^{-1}(A \cup B) = f^{-1}A \cup f^{-1}B$
(c) $f^{-1}(Y \setminus A) = X \setminus f^{-1}A$.

Proof. This is all trivial once one has untangled the notation. For (a), we note that

$$\begin{aligned} x \in f^{-1}(A \cap B) &\Leftrightarrow f(x) \in A \cap B \\ &\Leftrightarrow (f(x) \in A) \text{ and } (f(x) \in B) \\ &\Leftrightarrow (x \in f^{-1}(A)) \text{ and } (x \in f^{-1}(B)) \\ &\Leftrightarrow x \in f^{-1}(A) \cap f^{-1}(B). \end{aligned}$$

The other two parts are similar. □

Remark 7.12. Contrary to what one might expect, it is not always true that $f(f^{-1}(B)) = B$ or $f^{-1}(f(A)) = A$. However, we do always have $f(f^{-1}(B)) \subseteq B$ and $f^{-1}(f(A)) \supseteq A$; the proof is an easy exercise.

Proposition 7.13. *Let $f\colon X \to Y$ be a function between metric spaces. Then the following are equivalent.*

(a) *f is continuous.*
(b) *For every open set $U \subseteq Y$, the preimage $f^{-1}U$ is an open subset of X.*
(c) *For every closed set $F \subseteq Y$, the preimage $f^{-1}F$ is a closed subset of X.*

Proof. We first show that (b)⇒(c). Suppose that (b) holds, and that $F \subseteq Y$ is closed. Then the set $U := Y \setminus F$ is open in Y, so $f^{-1}(U)$ is open in X by (b). However, we know that $f^{-1}(U) = f^{-1}(Y \setminus F) = X \setminus f^{-1}(F)$, so $f^{-1}(F) = X \setminus f^{-1}(U)$. This is the complement of an open set, so it is closed as required.

The same argument works the other way around to show that (c)⇒(b).

Next, we prove that (b)⇒(a) (using the characterisation of continuity given in Proposition 3.9). Suppose that (b) holds, and we are given $x \in X$ and $\epsilon > 0$. Put $U = \overset{\circ}{B}_\epsilon(f(x)) = \{y \in Y : d(y, f(x)) < \epsilon\}$, which is open by Example 7.7. As (b) holds, we know that $f^{-1}U$ is open. Also, $f(x) \in U$, so $x \in f^{-1}U$. As $f^{-1}U$ is open, there exists $\delta > 0$ such that $\overset{\circ}{B}_\delta(x) \subseteq f^{-1}U$. Thus

$$\begin{aligned} d(x, x') < \delta &\Rightarrow x' \in \overset{\circ}{B}_\delta(x) \\ &\Rightarrow x' \in U = f^{-1}\overset{\circ}{B}_\epsilon(f(x)) \\ &\Rightarrow f(x') \in \overset{\circ}{B}_\epsilon(f(x)) \\ &\Rightarrow d(f(x), f(x')) < \epsilon, \end{aligned}$$

as required. This means that f is continuous, so (a) holds.

Finally, we prove that (a)⇒(b). Suppose that f is continuous, and that $U \subseteq Y$ is open; we must prove that $f^{-1}U$ is open in X. Suppose that $x \in f^{-1}U$; we must prove that $\overset{\circ}{B}_\delta(x) \subseteq f^{-1}U$ for some $\delta > 0$. As $x \in f^{-1}U$, we have $f(x) \in U$. As U is open, we have $B_\epsilon(f(x)) \subseteq U$ for some $\epsilon > 0$. As f is continuous, there exists $\delta > 0$ such that whenever $d(x', x) < \delta$, we have $d(f(x'), f(x)) < \epsilon$. In other words, whenever $x' \in \overset{\circ}{B}_\delta(x)$, we have $f(x') \in \overset{\circ}{B}_\epsilon(f(x)) \subseteq U$, and thus $x' \in f^{-1}U$. This means that $\overset{\circ}{B}_\delta(x) \subseteq f^{-1}U$, as required. □

Proposition 7.14. *Let X and Y be metric spaces, and let F and G be closed subsets of X such that $X = F \cup G$. Suppose we have continuous maps $f\colon F \to Y$ and $g\colon G \to Y$ such that $f(x) = g(x)$ for all $x \in F \cap G$. Then there is a unique function $h\colon X \to Y$ such that*

$$h(x) = \begin{cases} f(x) & \text{if } x \in F \\ g(x) & \text{if } x \in G. \end{cases}$$

Moreover, this function is continuous.

Proof. Only the continuity needs proof. Suppose we have a convergent sequence $x_n \to x$ in X; we need to show that $h(x_n) \to h(x)$. Put $S = \{n : x_n \in F\}$ and $T = \{n : x_n \in G\}$. As $X = F \cup G$ we have $\mathbb{N} = S \cup T$, so at least one of S and T must be infinite. Suppose first that both S and T are infinite, say

$$\begin{aligned} S &= \{n_1, n_2, \dots\} \\ T &= \{m_1, m_2, \dots\} \end{aligned}$$

with $n_1 < n_2 < \dots$ and $m_1 < m_2 < \dots$. As $x_n \to x$ we also have $x_{n_j} \to x$ and $x_{m_k} \to x$. As $n_j \in S$ we have $x_{n_j} \in F$ for all j, and F is closed so $x = \lim_{j\to\infty} x_{n_j} \in F$. As $f\colon F \to Y$ is continuous, we have $h(x_{n_j}) = f(x_{n_j}) \to f(x) = h(x)$. By a similar argument, we have $h(x_{m_k}) \to h(x)$.

Now suppose we are given $\epsilon > 0$. As $h(x_{n_j}) \to h(x)$, there exists J such that $d(h(x_{n_j}), h(x)) < \epsilon$ whenever $j \geq J$. Similarly, as $h(x_{m_k}) \to h(x)$ there exists K such that $d(h(x_{m_k}), h(x)) < \epsilon$ for $k \geq K$. Now put $N = \max(n_J, m_K)$. If $i \geq N$ then either $i = n_j$ for some $j \geq J$ or $i = m_k$ for some $k \geq K$ (or possibly both). Either way we have $d(h(x_n), h(x)) < \epsilon$. This shows that $h(x_n) \to h(x)$, as required. Similar but easier arguments show that $h(x_n) \to h(x)$ if S is infinite and T is finite, or if S is finite and T is infinite. Thus $h(x_n) \to h(x)$ in all cases, proving that h is continuous. □

8. Path connectedness

Definition 8.1. A *path* in a metric space X is a continuous map $u\colon I \to X$. If $u(0) = x$ and $u(1) = y$ we say that u is a path from x to y. The *reverse* of u is the path $\overline{u}\colon I \to X$ defined by $\overline{u}(t) = u(1-t)$; this goes from y to x. If u is a path from x to y and v is a path from y to z then we define a map $u * v\colon I \to X$ by

$$(u * v)(t) = \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ v(2t-1) & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

This is a path from x to z, called the *join* of u and v. Finally, for any $x \in X$ we can define a constant map $c_x\colon I \to X$ by $c_x(t) = x$ for all t; this is a path from x to x.

Example 8.2. If X is a vector space and $x, y \in X$ we can define a map $u\colon I \to X$ by $u(t) = (1-t)x + ty$, called the *linear path* from x to y. This is continuous with respect to any reasonable metric on X; certainly this works for any of the metrics d_1, d_2 or d_∞ on $\mathbb{R}^n$. If X is merely a subset of a vector space then we can still use linear paths, but we need to check that they actually lie in X.

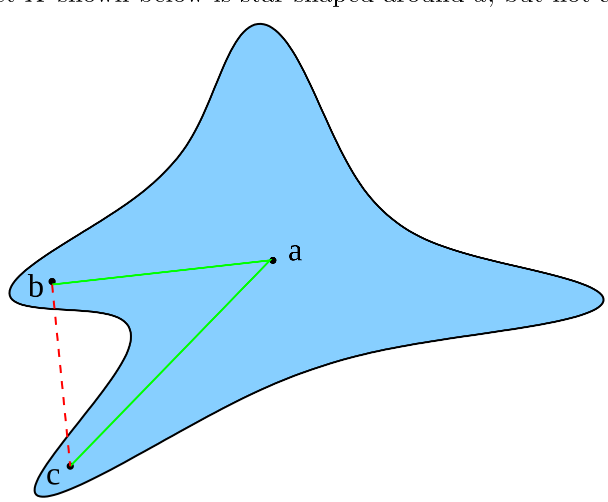
Example 8.3. Recall that $S^1 = \{z \in \mathbb{C} \;:\; |z| = 1\}$. Given two points $z, w \in S^1$ we can choose $\theta, \phi \in \mathbb{R}$ such that $z = e^{i\theta}$ and $w = e^{i\phi}$. We can then define $u(t) = e^{i((1-t)\theta + t\phi)}$ to get a path from z to w in S^1.

Definition 8.4. Let X be a metric space. We define an equivalence relation on X by $x \sim y$ iff there exists a path in X from x to y. (Constant paths show that this relation is reflexive, path reversal shows that it is symmetric, and path join shows that it is transitive.) The equivalence classes are called the *path components* of X. We write $\pi_0 X$ for the set of path components (in other words $\pi_0 X = X/\sim$). We say that X is *path-connected* if any two points in X can be joined by a path, or equivalently $|\pi_0(X)| = 1$.

Example 8.5. We see using linear paths that $\mathbb{R}^n$ is connected.

Definition 8.6. Let X be a subspace of a vector space V, and let a be a point in X. We say that X is *star-shaped* round a if for every $x \in X$, the linear path joining a to x in V lies wholly in X.

Example 8.7. The set X shown below is star-shaped around a, but not around b or c.

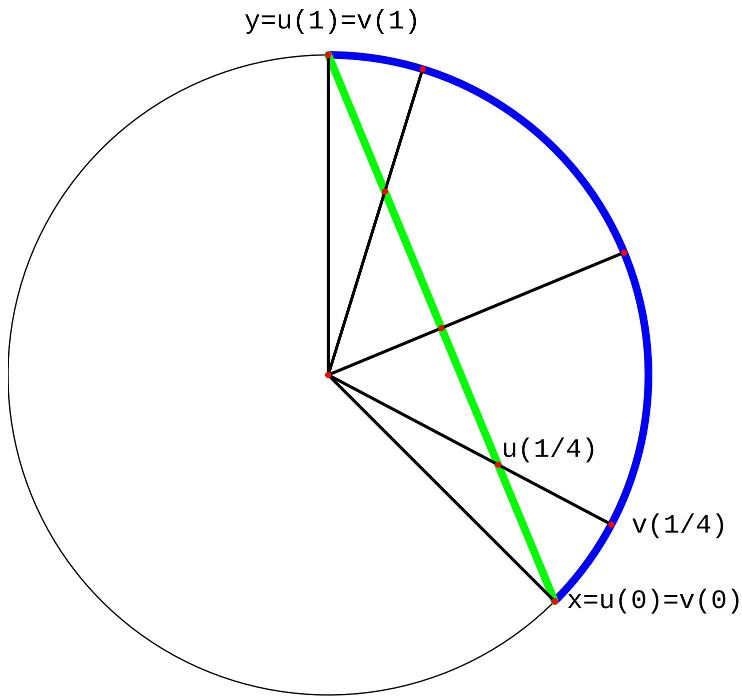


Proposition 8.8. *If X is star-shaped around some point $a \in X$, then X is path-connected.*

Proof. If $x, y \in X$ then by using linear paths we see that $x \sim a$ and $y \sim a$ but $\sim$ is an equivalence relation so $x \sim y$. More concretely, we can move from x to y within X by going in a straight line from x to a, and then in a straight line from a to y. □

Proposition 8.9. *If $n > 0$ then the sphere S^n is path-connected.*

Proof. Suppose that $x, y \in S^n$. We first assume that $y \neq -x$, so the linear path from x to y does not pass through zero, in other words the function $u(t) = (1-t)x + ty$ has $\|u(t)\| \neq 0$ for $0 \leq t \leq 1$. We may thus define $v(t) = u(t)/\|u(t)\|$ to get a continuous function $v\colon I \longrightarrow S^n$ with $v(0) = u(0)/\|u(0)\| = x/\|x\| = x$, and similarly $v(1) = y$, so $x \sim y$.



If $y = -x$ then we just choose a third point z such that $z \notin \{x, y\} = \{x, -x\}$, and note that $x \sim z \sim y$ so $x \sim y$. (Note that when $n = 0$, the space $S^0 = \{1, -1\}$ has only two points so we cannot choose z; this is why the proposition needs $n > 0$.) □

Proposition 8.10. *Let X be a metric space, and let $f\colon X \longrightarrow \mathbb{R}$ be a continuous map such that $f(z) \neq 0$ for all $z \in X$. Suppose that $x, y \in X$ and $f(x) < 0 < f(y)$. Then $x \not\sim y$.*

Proof. If $x \sim y$, we can choose a path $u\colon I \longrightarrow X$ with $u(0) = x$ and $u(1) = y$. Define $g = f \circ u\colon I \longrightarrow \mathbb{R}$, so g is continuous and $g(0) = f(u(0)) = f(x) < 0$ and $g(1) = f(u(1)) = f(y) > 0$. The Intermediate Value Theorem now tells us that $g(t) = 0$ for some $t \in I$. Thus $u(t) \in X$ and $f(u(t)) = 0$, contradicting our assumption about f. This contradiction shows that in fact $x \not\sim y$, as claimed. □

Corollary 8.11. *Let X be a metric space, and let U and V be open subsets of X such that $U \cup V = X$ and $U \cap V = \emptyset$. If $x \in U$ and $y \in V$ then $x \not\sim y$.*

Proof. Define a map $f\colon X \longrightarrow \mathbb{R}$ by

$$f(x) = \begin{cases} -1 & \text{if } x \in U \\ +1 & \text{if } x \in V. \end{cases}$$

Note that f is defined everywhere (because $U \cup V = X$) and is well-defined (because $U \cap V = \emptyset$). If $W \subseteq \mathbb{R}$ is open then $f^{-1}(W)$ is either $\emptyset$, U, V or $U \cup V$ (depending on whether $W \cap \{-1, 1\}$ is $\emptyset$, $\{-1\}$, $\{1\}$ or $\{-1, 1\}$). All four of these sets are open, so f is continuous. As $f(z) \neq 0$ for all $z \in X$ and $f(x) < 0 < f(y)$, we conclude that $x \not\sim y$. □

Proposition 8.12. *Let $f \colon X \to Y$ be a continuous map between metric spaces. If $x_0 \sim x_1$ in X then $f(x_0) \sim f(x_1)$ in Y.*

Proof. If $x_0 \sim x_1$ then we can choose a path $u \colon I \to X$ with $u(0) = x_0$ and $u(1) = x_1$. Define $v = f \circ u \colon I \to Y$ (which is again continuous by Proposition 3.6). We then have $v(0) = f(u(0)) = f(x_0)$ and $v(1) = f(u(1)) = f(x_1)$, so v is a path from $f(x_0)$ to $f(x_1)$ as required. □

This allows us to make the following definition.

Definition 8.13. Let $f \colon X \to Y$ be a continuous map between metric spaces. We then define a map $f_* \colon \pi_0 X \to \pi_0 Y$ by $f_*[x] = [f(x)]$, which is well-defined by the Proposition.

Proposition 8.14. *The maps f_* have the following properties.*

(a) *If $f \colon X \to X$ is the identity map, then $f_* \colon \pi_0(X) \to \pi_0(X)$ is the identity map.*
(b) *If $f \colon X \to Y$ and $g \colon Y \to Z$ are continuous then $(gf)_* = g_* f_* \colon \pi_0(X) \to \pi_0(Z)$.*
(c) *If f is surjective then f_* is surjective.*
(d) *If f is a homeomorphism then f_* is a bijection.*

Remark 8.15. Parts (a) and (b) say that the construction π_0 is a *functor*. They are easy to check, but we state them explicitly because the idea of a functor turns out to be very important.

Proof. (a) In this case $f(x) = x$ for all x, so $f_*[x] = [x]$.
(b) Here we have

$$(gf)_*[x] = [(gf)(x)] = [g(f(x))] = g_*[f(x)] = g_* f_*[x],$$

as required.
(c) Suppose we have $b \in \pi_0 Y$. Choose $y \in Y$ such that $b = [y]$. As f is surjective, we can choose $x \in X$ such that $f(x) = y$. Now put $a = [x]$, and observe that $f_*(a) = b$; this shows that f_* is surjective.
(d) By parts (b) and (a) we have $f_*(f^{-1})_* = (ff^{-1})_* = 1_* = 1$, and similarly $(f^{-1})_* f_* = 1$. Thus $(f^{-1})_*$ is an inverse for the function $f_* \colon \pi_0 X \to \pi_0 Y$, so f_* is a bijection.

□

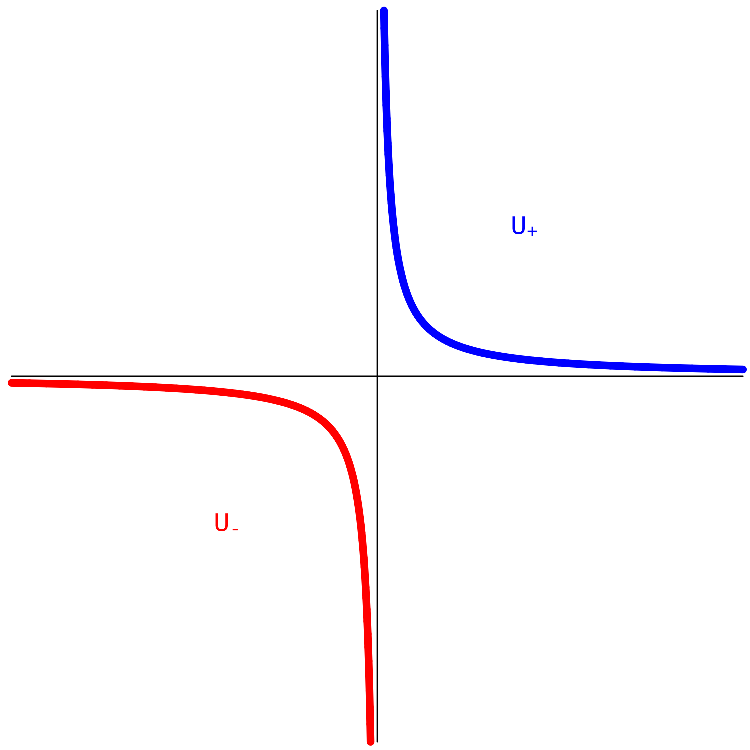
Exercise 8.16. If f is injective, then f_* *need not* be injective.

Proposition 8.17. *If X is path-connected and there exists a continuous surjective map $f \colon X \to Y$ then Y is path-connected. In particular this holds if Y is homeomorphic to X.*

Proof. Suppose $y_0, y_1 \in Y$. As f is surjective, we can choose $x_0, x_1 \in X$ such that $f(x_0) = y_0$ and $f(x_1) = y_1$. As X is path-connected, we can choose a path $u \colon I \to X$ with $u(0) = x_0$ and $u(1) = x_1$. Now define $v = f \circ u \colon I \to Y$. This is continuous, because f and u are. Clearly $v(0) = f(u(0)) = f(x_0) = y_0$, and similarly $v(1) = y_1$, so $y_0 \sim y_1$. Thus any two points in Y can be connected by a path, in other words Y is path-connected. □

Corollary 8.18. *If $f \colon X \to Y$ is a continuous map and X' is a path-connected subspace of X, then the space $Y' := f(X')$ is a path-connected subspace of Y (because we can regard f as a continuous surjection from X' to Y').*

Example 8.19. Put $X = \{(x,y) \in \mathbb{R}^2 \ : \ xy = 1\}$.



We can divide this into two subsets as follows:

$$U_+ = \{(x,y) \in X \ : \ x > 0\}$$
$$U_- = \{(x,y) \in X \ : \ x < 0\}$$

These are both open in X, and $X = U_+ \cup U_-$ and $U_+ \cap U_- = \emptyset$. Thus no point in U_+ can be connected to any point in U_-. We can define a continuous map $f_+ \colon \mathbb{R} \longrightarrow U_+$ by $f(a) = (e^a, e^{-a})$. This is easily seen to be surjective, and $\mathbb{R}$ is path-connected, so U_+ is path-connected. We see by a similar argument that U_- is path-connected as well. This shows that U_+ and U_- are the two path-components of X, so $|\pi_0(X)| = 2$.

Example 8.20. Take

$$X = \mathbb{R} \setminus \mathbb{Z} = \{x \in \mathbb{R} \ : \ x \text{ is not an integer } \}$$
$$= \ldots \cup (-1,0) \cup (0,1) \cup (1,2) \cup \ldots$$

There is a bijection $\phi \colon \mathbb{Z} \longrightarrow \pi_0(X)$ given by $\phi(n) = [n + \frac{1}{2}]$.

Example 8.21. Take

$$X = GL_n(\mathbb{R}) = \{A \in M_n(\mathbb{R}) \ : \ A \text{ is invertible } \}$$
$$= \{A \in M_n(\mathbb{R}) \ : \ \det(A) \neq 0\}.$$

We can define $U_+ = \{A \in X \ : \ \det(A) > 0\}$ and $U_- = \{A \in X \ : \ \det(A) < 0\}$. These are both open and nonempty (why?) and $X = U_- \cup U_+$ and $U_+ \cap U_- = \emptyset$; this proves that X has at least two path components. In fact, it can be shown that U_+ is path-connected. Moreover, if B is any fixed element of U_- then one checks that the map $A \mapsto AB$ gives a homeomorphism $U_+ \longrightarrow U_-$, so U_- is also path-connected. We conclude that U_+ and U_- are the path-components of X, so $|\pi_0(X)| = 2$.

9. Cutting

In this section we use the functor π_0 to prove that certain spaces are not homeomorphic to each other.

The most obvious way to do this is part (d) of Proposition 8.14: this implies that if X is homeomorphic to Y then there is a bijection from $\pi_0 X$ to $\pi_0 Y$, so $|\pi_0 X| = |\pi_0 Y|$.

Example 9.1. If $X = \mathbb{R}\setminus\mathbb{Z} = \bigcup_{n\in\mathbb{Z}}(n, n+1)$ and $Y = \mathbb{R}\setminus\{0\}$ then $\pi_0 X$ is infinite but $|\pi_0 Y| = 2$ so X is not homeomorphic to Y.

Example 9.2. We saw in Example 8.21 that $GL_n(\mathbb{R})$ is not path-connected, but $M_n(\mathbb{R})$ is a vector space and thus is path-connected, so $GL_n(\mathbb{R})$ is not homeomorphic to $M_n(\mathbb{R})$.

This method is unfortunately inadequate to prove many other visually obvious facts such as that S^1 is not homeomorphic to $\mathbb{R}$ or that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^3$. At least the first of these can be proved by a small adaptation of the method, however.

Proposition 9.3. *S^1 is not homeomorphic to $\mathbb{R}$.*

Proof. Suppose for a contradiction that $f\colon \mathbb{R} \to S^1$ is a homeomorphism. Put $a = f(0) \in S^1$. Because f is injective we have $f(t) \neq a$ when $t \neq 0$ so f gives a continuous map $f\colon \mathbb{R}\setminus\{0\} \to S^1\setminus\{a\}$. Similarly, f^{-1} gives a continuous map $f^{-1}\colon S^1\setminus\{a\} \to \mathbb{R}\setminus\{0\}$. These maps are clearly inverse to each other, so $\mathbb{R}\setminus\{0\}$ is homeomorphic to $S^1\setminus\{a\}$. This is a contradiction, because it is easy to see that $S^1\setminus\{a\}$ is path-connected but $\mathbb{R}\setminus\{0\}$ is not. □

More generally, if X is homeomorphic to Y and $a_1,\dots,a_n$ are n distinct points in X then there exist n distinct points $b_1,\dots,b_n$ in Y such that $X\setminus\{a_1,\dots,a_n\}$ is homeomorphic to $Y\setminus\{b_1,\dots,b_n\}$. Indeed, if $f\colon X \to Y$ is a homeomorphism, we can just take $b_i = f(a_i)$. Using this, we can prove a number of other non-homeomorphism results.

Proposition 9.4. *I is not homeomorphic to S^1.*

Proof. If I were homeomorphic to S^1, then $(0,1) = I\setminus\{0,1\}$ would be homeomorphic to $S^1\setminus\{b_1,b_2\}$ for some b_1, b_2. This is a contradiction, because $(0,1)$ is path-connected whereas $S^1\setminus\{b_1,b_2\}$ is always disconnected for any pair $\{b_1,b_2\}$ of distinct points. □

Proposition 9.5. *$\mathbb{R}$ is not homeomorphic to $\mathbb{R}^2$, because $\mathbb{R}$ is disconnected by the removal of any point, whereas there is no point in $\mathbb{R}^2$ whose removal disconnects the space.* □

Proposition 9.6. *$(0,1)$ is not homeomorphic to $[0,1)$, because $[0,1)\setminus\{0\}$ is connected and thus not homeomorphic to $(0,1)\setminus\{b\}$ for any b.* □

Proposition 9.7. *Put $X = \{(x,y)\in\mathbb{R}^2 \;:\; xy = 0\} = (\mathbb{R}\times\{0\})\cup(\{0\}\times\mathbb{R})$. Then X is not homeomorphic to $\mathbb{R}$, because $X\setminus\{(0,0)\}$ has four path components and thus is not homeomorphic to $\mathbb{R}\setminus\{b\}$ for any b.* □

These ideas can be extended to give some numerical invariants of spaces.

Definition 9.8. Let X be a metric space. We write

$$\begin{aligned} a(X) &= \max\{|Y| \;:\; Y\subseteq X \text{ and } X\setminus Y \text{ is path-connected }\}\\ &= \text{ the greatest number of points that can be removed without disconnecting } X\\ b(X) &= \min\{|Y| \;:\; Y\subseteq X \text{ and } X\setminus Y \text{ is disconnected }\}\\ &= \text{ the least number of points that need to be removed to disconnect the space } X. \end{aligned}$$

These invariants can be infinite: for example $a(\mathbb{R}^2) = \infty$, because the plane remains connected after the removal of any finite set of points. However, we will principally be interested in cases in which they are finite.

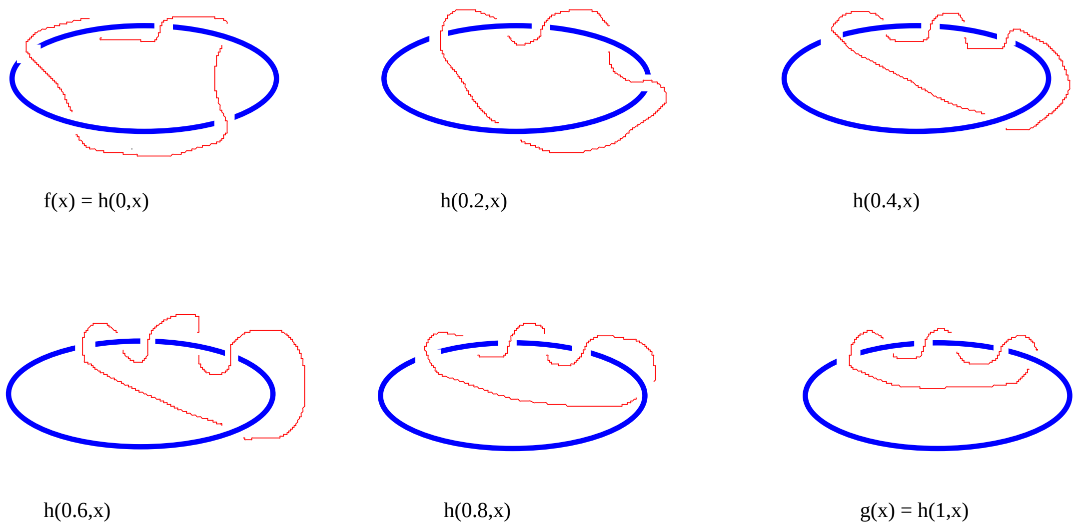
Proposition 9.9. *If X is homeomorphic to X' then $a(X) = a(X')$ and $b(X) = b(X')$.*

Proof. Let $f\colon X \longrightarrow X'$ be a homeomorphism. If $Y \subseteq X$ and $X \setminus Y$ is path-connected then we put $Y' = f(Y) \subseteq X'$. As f is a bijection we have $|Y'| = |Y|$. We also note that f gives a homeomorphism $X \setminus Y \longrightarrow X' \setminus Y'$, so $X' \setminus Y'$ is path-connected. We must therefore have $a(X') \geq |Y'| = |Y|$. By taking the maximum over all possible Y's, we see that $a(X') \geq a(X)$. By applying this argument to $f^{-1}\colon X' \longrightarrow X$ instead of $f\colon X \longrightarrow X'$, we also see that $a(X) \geq a(X')$. This means that $a(X) = a(X')$. The argument for b is similar. □

10. Homotopy

Definition 10.1. We say that two continuous maps $f, g\colon X \longrightarrow Y$ are *homotopic* if there exists a continuous map $h\colon [0,1] \times X \longrightarrow Y$ such that $h(0,x) = f(x)$ for all x, and $h(1,x) = g(x)$ for all x. Such a map h is called a *homotopy* between f and g. The notation $f \simeq g$ means that f is homotopic to g.

Example 10.2. Consider the space $X = \mathbb{R}^3 \setminus S^1$. In the diagram below, the copy of S^1 that has been removed is the thick circle. The image of a map from S^1 to X can be pictured as a thin loop that does not touch that thick one. The diagram is a sketch of a homotopy from a certain map $f\colon S^1 \longrightarrow X$ to a certain other map $g\colon S^1 \longrightarrow X$.



Example 10.3. Define $f, g\colon S^1 \longrightarrow S^1$ by $f(z) = z$ and $g(z) = -z$. These maps turn out to be homotopic. Indeed, if we regard S^1 as a subset of $\mathbb{C}$ in the usual way and define $h(t,z) = e^{i\pi t}z$ then h is a continuous map $I \times S^1 \longrightarrow S^1$ with $h(0,z) = f(z)$ and $h(1,z) = -z$ (because $e^{i\pi} = -1$).

Proposition 10.4. *The relation of being homotopic is an equivalence relation.*

Proof. This is essentially the same as the proof that the relation in Definition 8.4 is an equivalence relation.

(a) The function $h(t,x) = f(x)$ gives a homotopy from f to itself, proving that the relation is reflexive.

(b) Suppose that f is homotopic to g. Choose a homotopy $h\colon I \times X \longrightarrow Y$ with $h(0,x) = f(x)$ and $h(1,x) = g(x)$. Define a map $\overline{h}\colon I \times X \longrightarrow Y$ by $\overline{h}(t,x) = h(1-t,x)$. This is easily seen to be a homotopy from g to f, proving that the relation is symmetric.

(c) Suppose that e is homotopic to f and f is homotopic to g. We can then choose a homotopy k from e to f, and a homotopy h from f to g. Now define a map $k * h\colon I \times X \longrightarrow Y$ by

$$(k*h)(t,x) = \begin{cases} k(2t,x) & \text{if } 0 \leq t \leq 1/2 \\ h(2t-1,x) & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

Note that the two clauses of the definition coincide when $t = 1/2$, because $k(1,x) = f(x) = h(0,x)$. It is easy to see that $k * h$ is a homotopy from e to g, proving that the relation is transitive as required.

□

Proposition 10.5. *Suppose we have continuous maps $e\colon W \to X$ and $f, g\colon X \to Y$ and $h\colon Y \to Z$, and that $f \simeq g$. Then $fe \simeq ge$ and $hf \simeq hg$ and $hfe \simeq hge$.*

Proof. Let $u\colon I \times X \to Y$ be a homotopy from f to g. Define $v\colon I \times W \to Z$ by $v(t, w) = h(v(t, e(w)))$. Then v is continuous and

$$v(0, w) = h(v(0, e(w))) = hfe(w)$$
$$v(1, w) = h(v(1, e(w))) = hge(w),$$

which shows that $hfe \simeq hge$. By taking $h = 1$ we see that $fe \simeq ge$; by taking $e = 1$ instead we find that $hf \simeq hg$. □

Corollary 10.6. *Suppose we have $f_0, f_1\colon X \to Y$ and $g_0, g_1\colon Y \to Z$ and $f_0 \simeq f_1$ and $g_0 \simeq g_1$. Then $g_0 f_0 \simeq g_1 f_1$.*

Proof. Using one half of the proposition, we see that $f_0 g_0 \simeq f_1 g_0$. Using the other half, we see that $f_1 g_0 \simeq f_1 g_1$. As $\simeq$ is an equivalence relation, we see that $f_0 g_0 \simeq f_1 g_1$. □

Definition 10.7. A map $f\colon X \to Y$ is a *homotopy equivalence* if there exists a map $g\colon Y \to X$ (called a *homotopy inverse* for f) such that $gf \simeq 1_X$ and $fg \simeq 1_Y$. Such a map g is called a *homotopy inverse* for f. We say that spaces X and Y are *homotopy equivalent* if there exists a homotopy equivalence from X to Y.

Proposition 10.8. *The relation of being homotopy equivalent is an equivalence relation.*

Proof. The identity function from X to itself is clearly a homotopy equivalence, so the relation is reflexive. Suppose that X is homotopy equivalent to Y, so we can choose a homotopy equivalence $f\colon X \to Y$ and a homotopy inverse $g\colon Y \to X$, so $fg\colon Y \to Y$ and $gf\colon X \to X$ are homotopic to the respective identity maps. This means that g is a homotopy equivalence with homotopy inverse f, so Y is homotopy equivalent to X. This proves that our relation is reflexive. Finally, suppose that X is homotopy equivalent to Y and Y is homotopy equivalent to Z. Let d and e be homotopy inverses for f and g, so $df \simeq 1_X$ and $fd \simeq 1_Y \simeq eg$ and $ge \simeq 1_Z$. Using Proposition 10.5, we deduce that

$$degf = d(eg)f \simeq d1_Y f = df \simeq 1_X$$

$$gfde = g(fd)e \simeq g1_Y e = ge \simeq 1_Z.$$

This proves that $gf\colon X \to Z$ is a homotopy equivalence, with homotopy inverse $de\colon Z \to X$. This in turn shows that our relation is transitive, as required. □

Exercise 10.9. If X is homotopy equivalent to X' and Y is homotopy equivalent to Y' then $X \times Y$ is homotopy equivalent to $X' \times Y'$.

Example 10.10. Let V be a vector space, and let $X \subseteq V$ be a subspace that is star-shaped round some point $a \in X$. Then X is homotopy equivalent to the one-point space $\{0\}$. To see this, define $f(0) = a$ and $g(x) = 0$ for all $x \in X$ and $h(t, x) = (1 - t)x + ta$. Then $gf = 1_{\{0\}}$ (so certainly $gf \simeq 1$) and h is a homotopy from 1_X to fg, so f is a homotopy equivalence as required.

Definition 10.11. We say that a space X is *contractible* if it is homotopy equivalent to a one-point space.

Example 10.12. More generally, suppose that Y is a subspace of a vector space. Suppose also that $f, g\colon X \to Y$ are two maps such that for every point $x \in X$, the linear path from $f(x)$ to $g(x)$ lies wholly within Y. We can then define a homotopy $h\colon f \simeq g$ by $h(t, x) = (1 - t)f(x) + tg(x)$. We call this the *linear homotopy* between f and g, and we say that f and g are *linearly homotopic.*

Example 10.13. The space $\mathbb{R}^n \setminus \{0\}$ is homotopy equivalent to S^{n-1}. Indeed, by example 3.17 we know that $\mathbb{R}^n \setminus \{0\}$ is homeomorphic to $S^{n-1} \times (0, \infty)$, and $(0, \infty)$ is homotopy equivalent to a point so $\mathbb{R}^n \setminus \{0\} \simeq S^{n-1} \times \text{point} = S^{n-1}$.

Example 10.14. The solid torus, the Möbius band, and $\mathbb{C}\setminus\{0\}$ are all homotopy equivalent to S^1. To explain this in more detail, let D be the vertical disc in the xz plane of radius 1 centred at $(2,0,0)$. The "solid torus" is the space obtained by revolving D around the z-axis; this is easily seen to be homeomorphic to $S^1\times D^2$. As in the previous example, we see that this is homotopy equivalent to S^1.

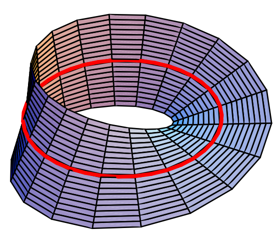
Next, for $\theta\in[0,2\pi]$, let P_θ be the vertical plane through the z-axis in $\mathbb{R}^3$ that has angle θ with the xz-plane. Let D_θ be the intersection of P_θ with the solid torus, which is a vertical disc of radius 1 centred at $(2\cos(\theta),2\sin(\theta),0)$. Let I_θ be the diameter of D_θ that makes an angle of $\theta/4$ to the vertical, and let M be the union of all the sets D_θ. This is a version of the Möbius band. It is homeomorphic to the space

$$M'=\{(z,w)\in S^1\times B^2\ :\ w^2/z \text{ is real and nonnegative }\}.$$

Define $f\colon S^1\longrightarrow M'$ and $g\colon M'\longrightarrow S^1$ and $h\colon I\times M'\longrightarrow M'$ by

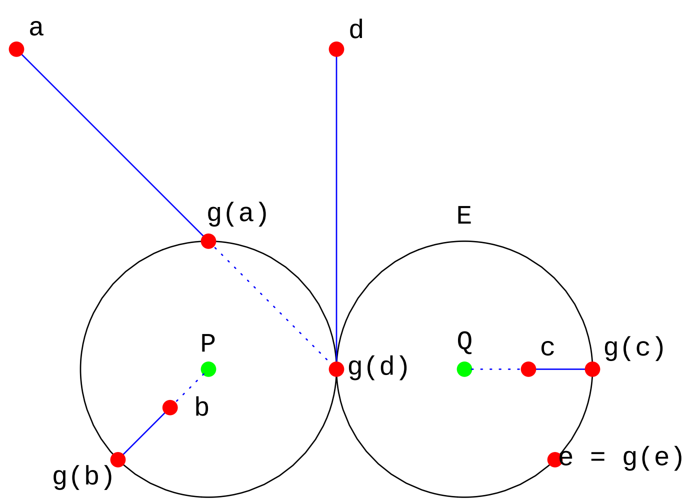
$$\begin{aligned} f(z)&=(z,0)\\ g(z,w)&=z\\ h(t,(z,w))&=(z,tw). \end{aligned}$$

Then $gf=1$ and h is a homotopy from fg to 1, so g is a homotopy inverse for f, so M' is homotopy equivalent to S^1 as claimed.



Finally, it is a special case of Example 10.13 that $\mathbb{C}\setminus\{0\}$ is homotopy equivalent to S^1.

Example 10.15. Put $P=(-1,0)$ and $Q=(+1,0)$. Let E be the union of the two circles in $\mathbb{R}^2$ of radius one centred at P and Q (so E is a "figure eight"). Then $\mathbb{R}^2\setminus\{P,Q\}$ is homotopy equivalent to E. To see this, we define a function $g\colon\mathbb{R}^2\setminus\{P,Q\}\longrightarrow E$ as in the following diagram.



In other words, if u is a point in the left hand disc then we draw a line from P to u and then extend it outwards until it meets E, and we define $g(u)$ to be the point of intersection. (We would have problems drawing such a line if u were equal to P, but the point P has been excluded from our space $\mathbb{R}^2 \setminus \{P, Q\}$, so this does not matter.) Note that if u lies on the left hand circle then $g(u) = u$. If u lies in the right hand disc, we define $g(u)$ by a similar procedure, but with a line drawn from Q. If u lies outside both discs, then we draw a line from u to 0, and let $g(u)$ be the point where it first crosses E. In particular, if u lies on the y-axis then $g(u) = 0$.

This gives a map $g\colon \mathbb{R}^2 \setminus \{P, Q\} \to E$, which can be shown to be continuous. We let $f\colon E \to \mathbb{R}^2 \setminus \{P, Q\}$ be the inclusion. As $g(u) = u$ for all $u \in E$, we have $gf = 1$. As the line segment from u to $g(u)$ never meets P or Q, we have a linear homotopy between fg and 1. Thus $\mathbb{R}^2 \setminus \{P, Q\}$ is homotopy equivalent to E.

The map g can also be described by the following formulae. Given $(x, y) \in \mathbb{R}^2 \setminus \{P, Q\}$ and $x \neq 0$, we have

$$\begin{aligned} r &= (x^2 + y^2)/|x| \\ s &= x/|x| = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases} \\ t &= \sqrt{(x-s)^2 + y^2} \\ g(x, y) &= \begin{cases} (x - s, y)/t + (s, 0) & \text{if } r \leq 1 \\ (x, y)/r & \text{if } r \geq 1 \end{cases} \end{aligned}$$

In the case $x = 0$ we just have $g(0, y) = (0, 0)$.

Example 10.16. Let T be the torus, and let a be a point of T. Then $T \setminus \{a\}$ is also homotopy equivalent to the figure eight space E. To see this, recall that $T = S^1 \times S^1$, and regard S^1 as a subset of $\mathbb{C}$ in the usual way. If $a = (b, c)$ then the map $(z, w) \mapsto (z/a, w/b)$ gives a homeomorphism $T \setminus \{a\} \to T \setminus \{(1, 1)\}$, so we may as well assume that $b = c = 1$. Define

$$E' = (S^1 \times \{-1\}) \cup (\{-1\} \times S^1) \subset T \setminus \{(1, 1)\}.$$

Note that the two circles in E' meet only at the single point $(-1, -1)$, which suggests that E' should be homeomorphic to E. Indeed, we can define a map $e\colon E' \to E$ by $e(z, -1) = 1 + z$ and $e(-1, w) = -1 - w$ (or equivalently by $e(z, w) = z - w$). It can be checked that this is a homeomorphism, as required.

Any point in $T \setminus \{(0, 0)\}$ can be expressed as $(e^{i\theta}, e^{i\phi})$ with $-\pi \leq \theta, \phi \leq \pi$ and $(\theta, \phi) \neq (0, 0)$. This expression is almost unique, except for the ambiguity that arises because $e^{i\pi} = e^{-i\pi} = -1$. Define

$$f(e^{i\theta}, e^{i\phi}) = (e^{i\overline{\theta}}, e^{i\overline{\phi}}),$$

where

$$\begin{aligned} \overline{\theta} &= \theta / \max(|\theta|, |\phi|) \\ \overline{\phi} &= \phi / \max(|\theta|, |\phi|). \end{aligned}$$

One checks that this is well-defined and lies in E', so we have a function $f\colon T \setminus \{(0, 0)\} \to E'$. It takes some work to prove that this is continuous, and we shall not do so here. We then let $g\colon E' \to T \setminus \{(0, 0)\}$ be the inclusion and define a map $h\colon I \times T \setminus \{(1, 1)\} \to T \setminus \{(1, 1)\}$ by

$$h(t, (e^{i\theta}, e^{i\phi})) = (e^{i((1-t)\theta + t\overline{\theta})}, e^{i((1-t)\phi + t\overline{\phi})}).$$

We then have $fg = 1$ and $h\colon 1 \simeq gf$ so f is a homotopy equivalence as required.

Example 10.17. $\mathbb{R}^3 \setminus S^1$ is homotopy equivalent to $S^2 \cup L$, where L is the line segment in $\mathbb{R}^3$ joining $(0, 0, 1)$ (the "North pole") to $(0, 0, -1)$ (the "South pole").

Proposition 10.18. *If $f \simeq g\colon X \to Y$ then $f_* = g_*\colon \pi_0(X) \to \pi_0(Y)$.*

Proof. Choose a homotopy h from f to g. For any point $x \in X$, define a map $u\colon I \longrightarrow Y$ by $u(t) = h(t,x)$. Then $u(0) = h(0,x) = f(x)$ and $u(1) = h(1,x) = g(x)$, so $f(x) \sim g(x)$ in Y. This means that $f_*[x] = [f(x)] = [g(x)] = g_*[x]$. As x was arbitrary, we see that $f_* = g_*$ as claimed. □

Corollary 10.19. *If $f\colon X \longrightarrow Y$ is a homotopy equivalence then $f_*\colon \pi_0(X) \longrightarrow \pi_0(Y)$ is a bijection.*

Proof. Let $g\colon Y \longrightarrow X$ be a homotopy inverse to f. Then $gf \simeq 1_X$ so $g_*f_* = (gf)_* = (1_X)_* = 1_{\pi_0 X}$, and similarly we have $f_*g_* = (fg)_* = (1_Y)_* = 1_{\pi_0 Y}$. This shows that g_* is an inverse for f_*, so f_* is a bijection. □

11. Homotopy of paths

Let X be a space, and fix points $x_0, x_1 \in X$. Suppose we have two paths u and v from x_0 to x_1, so $u, v\colon I \longrightarrow X$ and $u(0) = v(0) = x_0$ and $u(1) = v(1) = x_1$. We say that u and v are *homotopic relative to endpoints* (and write $u \simeq_{\text{re}} v$) if there is a continuous map $h\colon I \times I \longrightarrow X$ such that

(a) $h(0,t) = u(t)$
(b) $h(1,t) = v(t)$
(c) $h(s,0) = x_0$
(d) $h(s,1) = x_1$.

This means that for every s, the map $t \mapsto h(s,t)$ is a path from x_0 to x_1. We thus have a continuous family of such paths, starting with u and ending with v.

Just as in Proposition 10.4 we see that this notion gives an equivalence relation on the set of all paths from x_0 to x_1. We define $\pi_1(X; x_0, x_1)$ for the set of equivalence classes. Thus, every element $a \in \pi_1(X; x_0, x_1)$ has the form $a = [u]$ for some path u from x_0 to x_1, and $[u] = [v]$ if and only if $u \simeq_{\text{re}} v$.

Remark 11.1. The ordinary notion of homotopy (not relative to endpoints) is not very interesting in this context. It is not hard to see that if X is a path-connected space, then any two maps $u, v\colon I \longrightarrow X$ are homotopic to each other.

We next want to check that the notion of homotopy relative to endpoints is compatible with the operations of joining and reversing paths.

Lemma 11.2. *Let u, u' be paths from x_0 to x_1 such that $u \simeq_{re} u'$, and let v, v' be paths from x_1 to x_2 such that $v \simeq_{re} v'$. Then $u * v \simeq_{re} u' * v'$ (as paths from x_0 to x_2).*

Proof. Let $h\colon u \simeq_{\text{re}} u'$ and $k\colon v \simeq_{\text{re}} v'$ be homotopies relative to endpoints, so

(a) $h(0,t) = u(t)$
(b) $h(1,t) = u'(t)$
(c) $k(0,t) = v(t)$
(d) $k(1,t) = v'(t)$
(e) $h(s,0) = x_0$
(f) $h(s,1) = x_1$
(g) $k(s,0) = x_1$
(h) $k(s,1) = x_2$.

Recall that

$$(u * v)(t) = \begin{cases} u(2t) & \text{if } 0 \le t \le 1/2 \\ v(2t-1) & \text{if } 1/2 \le t \le 1. \end{cases}$$

$$(u' * v')(t) = \begin{cases} u'(2t) & \text{if } 0 \le t \le 1/2 \\ v'(2t-1) & \text{if } 1/2 \le t \le 1. \end{cases}$$

Define

$$m(s,t) = \begin{cases} h(s,2t) & \text{if } 0 \le t \le 1/2 \\ k(s,2t-1) & \text{if } 1/2 \le t \le 1 \end{cases}$$

Note that when $t = 1/2$ we have

$$h(s, 2t) = h(s, 1) = x_1 = k(s, 0) = k(s, 2t - 1)$$

(by (f) and (g)) so the two clauses in the definition of m are consistent. As h and k are continuous, we deduce from Proposition 7.14 that m is continuous. Using (a) and (c) we see that $m(0,t) = (u * v)(t)$. Using (b) and (d) we see that $m(1,t) = (u' * v')(t)$, so m is a homotopy from $u * v$ to $u' * v'$. Using (e) and (h) we see that $m(s, 0) = x_0$ and $m(s, 1) = x_2$, so it is a based homotopy, as required. □

Lemma 11.3. *Let u, u' be paths from x_0 to x_1 such that $u \simeq_{re} u'$. Then $\overline{u} \simeq_{re} \overline{u'}$ as paths from x_1 to x_0.*

Proof. Let $h\colon u \simeq_{\text{re}} u'$ be a homotopy relative to endpoints, and put $k(s,t) = h(s, 1-t)$. Then k is the required homotopy relative to endpoints between $\overline{u}$ and $\overline{u'}$. □

Now suppose we have elements $a \in \pi_1(X; x_0, x_1)$ and $b \in \pi_1(X; x_1, x_2)$. We can choose paths u and v such that $a = [u]$ and $b = [v]$ and then define $ab = [u * v] \in \pi_1(X; x_0, x_2)$; Lemma 11.2 tells us that this is well-defined. Similarly, we can define $a^{-1} = [\overline{u}] \in \pi_1(X; x_1, x_0)$. Finally, we define $e_x = [c_x] \in \pi_1(X; x, x)$, where as usual c_x is the constant path with $c_x(t) = x$ for all t.

Proposition 11.4. *If $a \in \pi_1(X; x_0, x_1)$ and $b \in \pi_1(X; x_1, x_2)$ and $c \in \pi_1(X; x_2, x_3)$ then $(ab)c = a(bc) \in \pi_1(X; x_0, x_3)$.*

Proof. We can choose paths u, v, w with $a = [u]$ and $b = [v]$ and $c = [w]$ (so u runs from x_0 to x_1, v runs from x_1 to x_2 and w runs from x_2 to x_3). We then have $ab = [u * v]$ and thus $(ab)c = [(u * v) * w]$, and similarly $a(bc) = [u * (v * w)]$. Next, we have

$$\begin{aligned} ((u*v)*w)(t) &= \begin{cases} (u*v)(2t) & \text{if } 0 \le t \le 1/2 \\ w(2t-1) & \text{if } 1/2 \le t \le 1 \end{cases} \\ &= \begin{cases} u(4t) & \text{if } 0 \le t \le 1/4 \\ v(4t-1) & \text{if } 1/4 \le t \le 1/2 \\ w(2t-1) & \text{if } 1/2 \le t \le 1 \end{cases} \end{aligned}$$

whereas

$$\begin{aligned} (u*(v*w))(t) &= \begin{cases} u(2t) & \text{if } 0 \le t \le 1/2 \\ (v*w)(2t-1) & \text{if } 1/2 \le t \le 1 \end{cases} \\ &= \begin{cases} u(2t) & \text{if } 0 \le t \le 1/2 \\ v(4t-2) & \text{if } 1/2 \le t \le 3/4 \\ w(4t-3) & \text{if } 3/4 \le t \le 1. \end{cases} \end{aligned}$$

Informally, for $(u * v) * w$ we go along u at speed 4 in the first 1/4 second, then along v in the second 1/4 second, then along w at speed 2 in the last 1/2 second. For $u * (v * w)$ we go along u at speed 2 in the first 1/2 second, then along v and w in 1/4 second each.

Now define $h\colon I \times I \longrightarrow X$ by

$$h(s,t) = \begin{cases} u(4t/(2-s)) & \text{if } 0 \le t \le (2-s)/4 \\ v(4t-2+s) & \text{if } (2-s)/4 \le t \le (3-s)/4 \\ w((4t-3+s)/(1+s)) & \text{if } (3-s)/4 \le t \le 1. \end{cases}$$

To see that this is well-defined, we must check that the first and second clauses give the same answer when $t = (2-s)/4$, and that the second and third clauses give the same answer when $t = (3-s)/4$. The first check works because $u(1) = x_1 = v(0)$, and the second one because $v(1) = x_2 = w(0)$. In view of this, Proposition 7.14 tells us that h is continuous.

It is easy to check that $h(0,t) = (u*(v*w))(t)$ and $h(1,t) = ((u*v)*w)(t)$ and $h(s,0) = u(0) = x_0$ and $h(s, 1) = w(1) = x_3$, so h is a homotopy relative to endpoints between $u*(v*w)$ and $(u*v)*w$. This means that $[u * (v * w)] = [(u * v) * w]$, or in other words $a(bc) = (ab)c$. □

Proposition 11.5. *For any $a \in \pi_1(X; x_0, x_1)$ we have $e_{x_0}a = a = ae_{x_1} \in \pi_1(X, x_0, x_1)$.*

Proof. Choose a path u such that $a = [u]$. We then have $e_{x_0}a = [c_{x_0} * u]$ and $ae_{x_1} = [u * c_{x_1}]$. We have

$$
\begin{aligned}
(u * c_{x_1})(t) &= \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ c_{x_1}(2t-1) & \text{if } 1/2 \leq t \leq 1 \end{cases} \\
&= \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ x_1 & \text{if } 1/2 \leq t \leq 1. \end{cases}
\end{aligned}
$$

We can define a map $h\colon I \times I \longrightarrow X$ by

$$
h(s,t) = \begin{cases} u((1+s)t) & \text{if } 0 \leq t \leq 1/(1+s) \\ x_1 & \text{if } 1/(1+s) \leq t \leq 1. \end{cases}
$$

We find that this is continuous, and

(a) $h(0,t) = u(t)$
(b) $h(1,t) = (u * c_{x_1})(t)$
(c) $h(s,0) = u(0) = x_0$
(d) $h(s,1) = x_1$.

Thus h is a based homotopy between u and $u * c_{x_1}$, showing that $[u] = [u * c_{x_1}]$, or equivalently $a = ae_{x_1}$. A similar argument shows that $a = e_{x_0}a$. □

Proposition 11.6. *For any $a \in \pi_1(X; x_0, x_1)$ we have $a^{-1}a = e_{x_0} \in \pi_1(X; x_0, x_0)$ and $aa^{-1} = e_{x_1} \in \pi_1(X, x_1, x_1)$.*

Proof. Choose a path u such that $a = [u]$. We then have $a^{-1} = [\overline{u}]$ and $aa^{-1} = [u * \overline{u}]$. Note that

$$
\begin{aligned}
(u * \overline{u})(t) &= \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ \overline{u}(2t-1) & \text{if } 1/2 \leq t \leq 1 \end{cases} \\
&= \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ u(2-2t) & \text{if } 1/2 \leq t \leq 1. \end{cases}
\end{aligned}
$$

We can define a map $h\colon I \times I \longrightarrow X$ by

$$
h(s,t) = \begin{cases} u(2st) & \text{if } 0 \leq t \leq 1/2 \\ u(2s(1-t)) & \text{if } 1/2 \leq t \leq 1. \end{cases}
$$

Note that when $t = 1/2$ we have $u(2st) = u(2s(1-t)) = u(s)$, so the two clauses are compatible and h is a continuous map. We have $h(0,t) = u(0) = x_0 = c_{x_0}(t)$ and $h(1,t) = (u * \overline{u})(t)$ so h is a homotopy between c_{x_0} and $u * \overline{u}$. Moreover, $h(s,0) = u(0) = x_0$ and $h(s,1) = u(0) = x_0$ so h is a homotopy relative to endpoints. Thus $[u * \overline{u}] = [c_{x_0}]$ and so $aa^{-1} = e_{x_0}$ as claimed. A similar proof shows that $a^{-1}a = e_{x_1}$. □

12. The fundamental group

A *based space* is a space X equipped with a chosen point $x_0 \in X$, which we will call the *basepoint*. A *based loop* in X is a path from x_0 to itself, or equivalently a continuous map $u\colon I \longrightarrow X$ with $u(0) = u(1) = x_0$. We write $\pi_1 X$ for the set $\pi_1(X; x_0, x_0)$ of equivalence classes of based loops under the relation of homotopy relative to endpoints.

If u and v are based loops then so are $u * v$ and $\overline{u}$. Thus, if $a, b \in \pi_1 X$ then ab and a^{-1} are also elements of $\pi_1 X$. We will write e for e_{x_0}, which is again an element of $\pi_1 X$.

We can specialise Propositions 11.4, 11.5 and 11.6 to the case where $x_3 = x_2 = x_1 = x_0$ to see that $a(bc) = (ab)c$ and $ae = a = ea$ and $aa^{-1} = e = a^{-1}a$, so $\pi_1 X$ is a group. It is called the *fundamental group* of X.

In this section we will just investigate some examples where $\pi_1 X$ is the trivial group. Cases where $\pi_1 X$ is nontrivial will be studied later, but they require more work.

Proposition 12.1. *Let X be a based space, and let X_0 be the set of all points in X that can be connected by a path to x_0. Then $\pi_1 X = \pi_1 X_0$.*

Proof. Let $u\colon I \longrightarrow X$ be a based loop. For any $t \in I$ we can define a path $u_t\colon I \longrightarrow X$ by $u_t(s) = u(st)$, and this connects $x_0 = u_t(0)$ to $u(t) = u_t(1)$, so $u(t) \in X_0$. This holds for all t, so u can be regarded as a map $I \longrightarrow X_0$. Similarly, if we have two based loops $u, v\colon I \longrightarrow X_0$ and a map $h\colon I \times I \longrightarrow X$ giving a based homotopy between them, we can define a path $h_{st}\colon I \longrightarrow X$ joining x_0 to $h(s,t)$ by $h_{st}(r) = h(s, rt)$, so h can be regarded as a map $I \times I \longrightarrow X_0$. The claim is immediate from this. □

Corollary 12.2. *If X is a discrete based space, then $\pi_1 X$ is the trivial group.*

Proof. In this case, the space X_0 is just the single point x_0, so it is clear that $\pi_1 X = \pi_1 X_0 = \{e\}$. □

Definition 12.3. A *based contraction* of a based space X is a map $h\colon I \times X \longrightarrow X$ such that

(a) $h(0,x) = x$ for all $x \in X$
(b) $h(1,x) = x_0$ for all $x \in X$
(c) $h(s,x_0) = x_0$ for all $s \in I$.

We say that X is *based-contractible* if there exists such a contraction.

Proposition 12.4. *If X is based-contractible, then $\pi_1 X$ is the trivial group.*

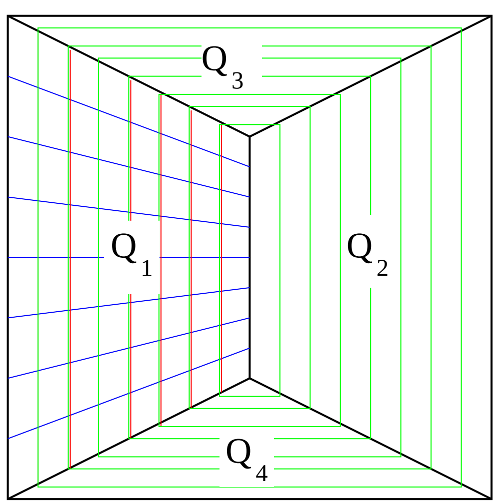
Proof. Let h be a based contraction of X. For any based loop u in X, the function $k(s,t) = h(s,u(t))$ satisfies

(a) $k(0,t) = h(0,u(t)) = u(t)$
(b) $k(1,t) = h(1,u(t)) = x_0 = c_{x_0}(t)$
(c) $k(s,0) = h(s,u(0)) = h(s,x_0) = x_0$
(d) $k(s,1) = h(s,u(1)) = h(s,x_0) = x_0$,

so $h\colon u \simeq_{\text{re}} c_{x_0}$ and $[u] = e$. □

Remark 12.5. Suppose that X is merely contractible in the unbased sense, so there is a point x_1 (possibly different from x_0) and a map $h\colon I \times X \longrightarrow X$ such that $h(0,x) = x$ and $h(1,x) = x_1$ for all x (with no condition on $h(s,x_0)$). It is still true that $\pi_1 X$ is trivial, but rather harder to prove. To see this, we divide the square $I \times I$ into four parts as follows:

$$Q_1 = \{(s,t) \,:\, 0 \leq s \leq 1/2 \,,\, s/2 \leq t \leq 1 - s/2\}$$
$$Q_2 = \{(s,t) \,:\, 1/2 \leq s \leq 1 \,,\, (1-s)/2 \leq t \leq (1+s)/2\}$$
$$Q_3 = \{(s,t) \,:\, 3/4 \leq t \leq 1 \,,\, 2 - 2t \leq s \leq 2t - 1\}$$
$$Q_4 = \{(s,t) \,:\, 0 \leq t \leq 1/4 \,,\, 2t \leq s \leq 1 - 2t\}$$



Then, given a based loop u, we define $k\colon I \times I \longrightarrow X$ by

$$k(s,t) = \begin{cases} h(2s, u((t-s/2)/(1-s))) & \text{if } (s,t) \in Q_1 \\ h(2-2s, x_0) & \text{if } (s,t) \in Q_2 \\ h(4-4t, x_0) & \text{if } (s,t) \in Q_3 \\ h(4-4t, x_0) & \text{if } (s,t) \in Q_4. \end{cases}$$

One can check that this is well-defined and continuous, and that it gives a homotopy relative to endpoints between u and c_{x_0}.

Finally, we will prove a result about changing basepoints. Suppose we have a space X and two different points $x_0, x_1 \in X$. If we use x_0 as the basepoint, then the fundamental group is $\pi_1(X; x_0, x_0)$. If we use x_1 as the basepoint instead, then the fundamental group is $\pi_1(X; x_1, x_1)$. If x_0 and x_1 lie in different path components of X, then these two fundamental groups can be completely different. However, if x_0 and x_1 can be joined by a path, then it turns out that the two groups are isomorphic.

Proposition 12.6. *If x_0 and x_1 can be joined by a path, then the group $\pi_1(X; x_0, x_0)$ is isomorphic to $\pi_1(X; x_1, x_1)$. In particular, if X is a path-connected space, then the group $\pi_1 X$ does not depend (up to isomorphism) on the choice of basepoint.*

Proof. Choose a path u from x_0 to x_1, and put $q = [u] \in \pi_1(X; x_0, x_1)$, so $q^{-1} \in \pi_1(X; x_1, x_0)$. Given $a \in \pi_1(X; x_0, x_0)$ we note that q^{-1} runs from x_1 to x_0, a runs from x_0 to x_0 and q runs from x_0 to x_1, so we can join them together to get an element $q^{-1}aq$ running from x_1 to x_1. We can thus define a function $f\colon \pi_1(X; x_0, x_0) \longrightarrow \pi_1(X; x_1, x_1)$ by $f(a) = q^{-1}aq$. Note that

$$f(a)f(b) = q^{-1}aqq^{-1}bq = q^{-1}ae_{x_0}bq = q^{-1}abq = f(ab),$$

so f is a homomorphism. Similarly, we can define a homomorphism $g\colon \pi_1(X; x_1, x_1) \longrightarrow \pi_1(X; x_0, x_0)$ by $g(b) = qbq^{-1}$. This satisfies $f(g(b)) = q^{-1}qbq^{-1}q = e_{x_1}be_{x_1} = b$ and $g(f(a)) = qq^{-1}aqq^{-1} = e_{x_0}ae_{x_0} = a$, so f and g are inverse to each other and are isomorphisms. □

Definition 12.7. A space X is *simply connected* if it is path-connected, and the fundamental group is trivial. (By the proposition, it does not matter which basepoint we use.)

13. $\pi_1 S^1$

In this section we will calculate $\pi_1 S^1$, which is the simplest case in which π_1 can be calculated and is not just the trivial group.

Recall that $S^1 = \{z \in \mathbb{C} \;:\; |z| = 1\}$, and take 1 as the basepoint in S^1. Define $\eta\colon \mathbb{R} \longrightarrow S^1$ by $\eta(t) = \exp(2\pi i t)$, and note that $\eta(s+t) = \eta(s)\eta(t)$ and $\eta(k) = 1$ for $k \in \mathbb{Z}$. In fact, we have $\eta(t) = 1$ if and only if t is an integer.

For $n \in \mathbb{Z}$ define $u_n\colon I \longrightarrow S^1$ by $u_n(t) = \eta(nt)$. This satisfies $u_n(0) = \eta(0) = 1$ and $u_n(1) = \eta(n) = 1$ so u_n is a based loop and thus defines an element $a_n = [u_n] \in \pi_1 S^1$.

Theorem 13.1. *We have $\pi_1 S^1 = \{a_n \;:\; n \in \mathbb{Z}\}$, and $a_n \neq a_m$ when $n \neq m$. Moreover, $a_n a_m = a_{n+m}$, and $a_n = a_1^n$.*

The proof will follow after some lemmas and related discussion.

Remark 13.2. Consider the group $\mathbb{Z}$ of integers under addition. Because the group law in $\mathbb{Z}$ is written additively and the group law in $\pi_1 X$ is written multiplicatively, a *homomorphism* from $\mathbb{Z}$ to $\pi_1 X$ is a function $\phi\colon \mathbb{Z} \longrightarrow \pi_1 X$ such that $\phi(n+m) = \phi(n)\phi(m)$. The theorem can be restated as follows: the function $\phi(n) = a_n$ is a bijective homomorphism, or equivalently an isomorphism, from $\mathbb{Z}$ to $\pi_1 S^1$.

Lemma 13.3. *$a_n a_m = a_{n+m}$.*

Proof. We must show that $u_n * u_m$ is based-homotopic to u_{n+m}. Define $f\colon I \longrightarrow \mathbb{R}$ by

$$f(t) = \begin{cases} 2tn & \text{if } 0 \leq t \leq 1/2 \\ n + (2t-1)m & \text{if } 1/2 \leq t \leq 1. \end{cases}$$

For $t \leq 1/2$ we have $\eta(f(t)) = \eta(2nt) = u_n(2t)$. For $t \geq 1/2$ we have

$$\begin{aligned} \eta(f(t)) &= \eta(n)\eta((2t-1)m) \\ &= \eta((2t-1)m) \\ &= u_m(2t-1). \end{aligned}$$

This means that $\eta(f(t)) = (u_n * u_m)(t)$ for all t. Note also that $f(0) = 0$ and $f(1) = n + m$.

Now define $h\colon I \times I \longrightarrow \mathbb{R}$ by

$$h(s,t) = s(n+m)t + (1-s)f(t).$$

This satisfies

(a) $h(0,t) = f(t)$
(b) $h(1,t) = (n+m)t$
(c) $h(s,0) = 0$
(d) $h(s,1) = n+m$.

Now put $k(s,t) = \eta(h(s,t))$; this gives a map $k\colon I \times I \longrightarrow S^1$ satisfying

(a) $k(0,t) = (u_n * u_m)(t)$
(b) $k(1,t) = u_{n+m}(t)$
(c) $k(s,0) = 1$
(d) $k(s,1) = 1$.

Thus, k is the required based homotopy between $u_n * u_m$ and u_{n+m}. □

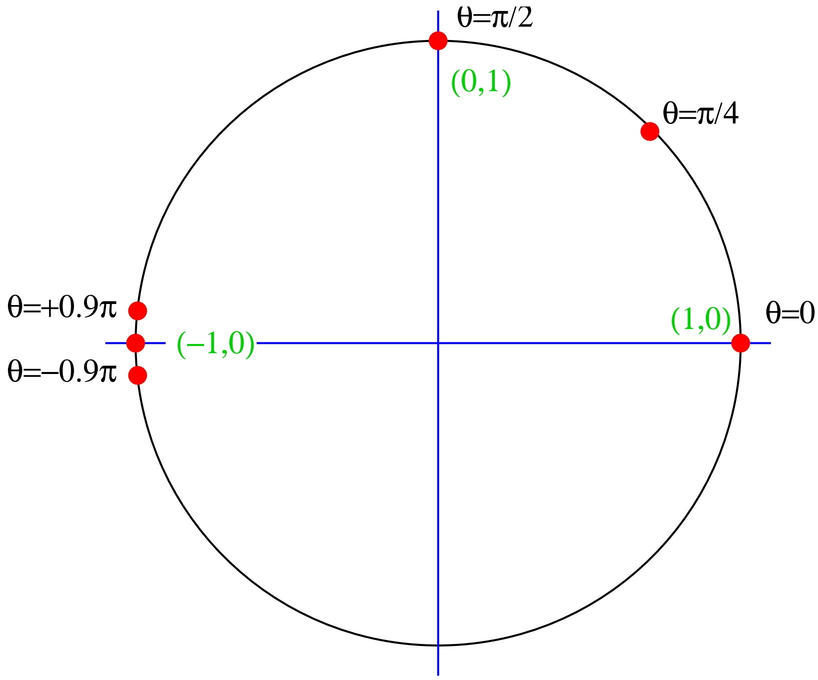
Corollary 13.4. *$a_n = a_1^n$ for all $n \in \mathbb{Z}$.*

Proof. It is clear that u_0 is the constant loop so $a_0 = e = a_1^0$. When $n > 0$, it follows easily by induction from the lemma that $a_n = a_1^n$. The lemma also tells us that $a_{-n}a_n = a_0 = e$, so $a_{-n} = a_n^{-1} = a_1^{-n}$, which completes the proof. □

Lemma 13.5. *For each $t \in \mathbb{R}$, put $U_t = (t - \frac{1}{2}, t + \frac{1}{2})$ and $V_t = S^1 \setminus \{-\eta(t)\}$. Then there is a continuous map $\lambda_t\colon V_t \longrightarrow U_t$ such that $\eta(\lambda_t(z)) = z$ for all $z \in V_t$, and $\lambda_t(\eta(t)) = t$.*

Proof. Note that every element $z \in S^1$ can be written in the form $z = \exp(i\theta)$ for some θ, which can be chosen in the range $-\pi < \theta \leq \pi$. As z moves across the negative x-axis through the point

$z = -1$, the angle θ jumps between $-\pi$ and $+\pi$, but otherwise θ varies continuously with z.



Thus, we can define a continuous map $\lambda\colon S^1 \setminus \{-1\} \longrightarrow (-1/2, +1/2)$ by $\lambda(\exp(i\theta)) = \theta/(2\pi)$. This means that $z = \eta(\lambda(z))$ for all $z \in S^1 \setminus \{-1\}$.

Now, if $z \in V_t$ then $z/\eta(t) \neq -1$ so $\lambda(z/\eta(t))$ is defined and lies in $(-\frac{1}{2}, \frac{1}{2})$. We can thus define a continuous map $\lambda_t\colon V_t \longrightarrow U_t$ by $\lambda_t(z) = t + \lambda(z/\eta(t))$. This satisfies

$$\eta(\lambda_t(z)) = \eta(t)\eta(\lambda(z/\eta(t)) = \eta(t)z/\eta(t) = z,$$

as required. We also have $\lambda_t(\eta(t)) = t + \lambda(1) = t$. □

Lemma 13.6. *Let $w\colon I \longrightarrow S^1$ be a based loop. Then there is a unique continuous map $\tilde{w}\colon I \longrightarrow \mathbb{R}$ such that $\tilde{w}(0) = 0$ and $w(t) = \eta(\tilde{w}(t))$ for all t. Moreover, $\tilde{w}(1)$ is an integer.*

Proof. As I is compact, Theorem 5.3 tells us that w is uniformly continuous. Taking $\epsilon = 2$ in the definition of uniform continuity, we see that there is a number $\delta > 0$ such that $|w(t) - w(t')| < 2$ whenever $|t - t'| < \delta$. Now choose a natural number $N > 1/\delta$ and put $I_j = [j/N, (j+1)/N]$ for $j = 0, \ldots, N-1$. Note that if $t \in I_j$ then $|t - j/N| \leq 1/N < \delta$ so $|w(t) - w(j/N)| < 2$, so $w(t) \neq -w(j/N)$.

Put $r_0 = 0$, so $w(0) = 1 = \eta(r_0)$. For $t \in I_0$ we deduce that $w(t) \neq -w(0) = -\eta(r_0)$, so $w(t) \in V_{r_0}$ and $\lambda_{r_0}(w(t))$ is defined. We can thus define $\tilde{w}_0\colon I_0 \longrightarrow \mathbb{R}$ by $\tilde{w}_0(t) = \lambda_{r_0}(w(t))$, and note that $\eta(\tilde{w}_0(t)) = w(t)$ for $t \in I_0$.

Now put $r_1 = \tilde{w}_0(1/N)$, so $w(1/N) = \eta(r_1)$. For $t \in I_1$ we deduce that $w(t) \neq -w(1/N) = -\eta(r_1)$, so $w(t) \in V_{r_1}$ and $\lambda_{r_0}(w(t))$ is defined. We can thus define $\tilde{w}_1\colon I_1 \longrightarrow \mathbb{R}$ by $\tilde{w}_1(t) = \lambda_{r_1}(w(t))$, and note that $\eta(\tilde{w}_1(t)) = w(t)$ for $t \in I_0$. Note also that $\tilde{w}_1(1/N) = \lambda_{r_1}(w(1/N)) = \lambda_{r_1}(\eta(r_1)) = r_1$.

We now put $r_2 = \tilde{w}_1(2/N)$ and continue in the same way. This gives numbers $r_0, \ldots, r_N$ and maps $\tilde{w}_j\colon I_j \longrightarrow \mathbb{R}$ (defined by $\tilde{w}_j(t) = \lambda_{r_j}(w(t))$) such that $\tilde{w}_j(j/N) = r_j$ and $\tilde{w}_j((j+1)/N) = r_{j+1}$. In particular, the value of $\tilde{w}_j$ at the end of the interval I_j is r_{j+1}, which is the same as the value of $\tilde{w}_{j+1}$ at the start of I_{j+1}. Thus, the functions $\tilde{w}_j$ can be combined to give a continuous map $\tilde{w}\colon I \longrightarrow \mathbb{R}$, with $w(t) = \tilde{w}_j(t)$ for all $t \in I_j$. As $\eta(\tilde{w}_j(t)) = w(t)$ for all $t \in I_j$, we see that $\eta(\tilde{w}(t)) = w(t)$ for all $t \in I$, as required. Clearly we also have $\tilde{w}(0) = \tilde{w}_0(0) = \lambda_0(w(0)) = \lambda_0(1) = 0$.

We next need to check that $\tilde{w}$ is unique. Suppose that $v\colon I \longrightarrow \mathbb{R}$ is another continuous map with $v(0) = 0$ and $\eta(v(t)) = w(t)$ for all $t \in I$. Put $f(t) = \tilde{w}(t) - v(t)$, so $\eta(f(t)) = \eta(\tilde{w}(t))/\eta(v(t)) =$

$w(t)/w(t) = 1$. We know that $\eta(r)$ is only equal to 1 when r is an integer, so $f(t)$ is an integer for all t. Any continuous function from I to $\mathbb{Z}$ must be constant (by the Intermediate Value Theorem), so f is constant. We also know that $f(0) = \tilde{w}(0) - v(0) = 0 - 0 = 0$, and f is constant, so $f(t) = 0$ for all t. This means that $\tilde{w}(t) = v(t)$, as required.

Finally, we observe that $\eta(\tilde{w}(1)) = w(1) = 1$, so $\tilde{w}(1)$ must be an integer. □

Definition 13.7. Given a based loop w in S^1, we define $\nu(w) = \tilde{w}(1)$, where $\tilde{w}$ is as above. This is an integer, and is called the *winding number* of w.

Remark 13.8. If $u_n(t) = \eta(nt)$ as before, we see that $\tilde{u}_n(t) = nt$, and thus that $\nu(u_n) = n$.

Corollary 13.9. *Any element $b \in \pi_1 S^1$ has $b = a_n$ for some integer n.*

Proof. We can choose a based loop w such that $b = [w]$, and then find n and $\tilde{w}$ as in the lemma. Now define $h: I \times I \to \mathbb{R}$ and $k: I \times I \to S^1$ by $h(s,t) = (1-s)\tilde{w}(t) + nst$ and $k(s,t) = \eta(h(s,t))$. We then have

(a) $h(0,t) = \tilde{w}(t)$
(b) $h(1,t) = nt$
(c) $h(s,0) = 0$ (because $\tilde{w}(0) = 0$)
(d) $h(s,1) = (1-s)n + ns = n$ (because $\tilde{w}(1) = n$)

so

(a) $k(0,t) = w(t)$
(b) $k(1,t) = u_n(t)$
(c) $k(s,0) = 1$
(d) $k(s,1) = 1$

so k is a based homotopy between w and u_n, so $b = a_n$. □

Lemma 13.10. *Let v and w be based loops in S^1 such that $d(v(t), w(t)) < 2$ for all t. Then $\nu(v) = \nu(w)$.*

Proof. As $d(v(t), w(t)) < 2$ we have $w(t) \neq -v(t)$ and so $v(t)/w(t) \neq -1$. This means that the function $g(t) = \lambda(v(t)/w(t))$ is defined and continuous for all t, and satisfies $\eta(g(t)) = v(t)/w(t)$. Note that $v(0) = w(0) = v(1) = w(1) = 1$, so $g(0) = g(1) = 0$.

Now find $\tilde{w}$ as in Lemma 13.6, so $\nu(w) = \tilde{w}(1)$. Put $\tilde{v}(t) = \tilde{w}(t) + g(t)$. Then $\tilde{v}(0) = \tilde{w}(0) + g(0) = 0 + 0 = 0$, and

$$\eta(\tilde{v}(t)) = \eta(\tilde{w}(t))\eta(g(t)) = w(t)v(t)/w(t) = v(t).$$

This means that $\tilde{v}$ is the function corresponding to v as in Lemma 13.6, so $\nu(v) = \tilde{v}(1) = \tilde{w}(1) + g(1) = \tilde{w}(1) = \nu(w)$, as claimed. □

Corollary 13.11. *Let v and w be based loops in S^1, and suppose that $v \simeq_{re} w$. Then $\nu(v) = \nu(w)$.*

Proof. Choose a map $h: I \times I \to S^1$ giving a homotopy relative to endpoints between v and w. As $I \times I$ is compact, h must be uniformly continuous, so we can find $\delta > 0$ such that $d(h(s,t), h(s',t')) < 2$ whenever $d((s,t),(s',t')) < \delta$. Choose an integer N such that $N > 1/\delta$, and define maps $v_0, \dots, v_N: I \to S^1$ by $v_i(t) = h(i/N, t)$. As h is a homotpy relative to endpoints, we see that each v_i is a based loop. As h is a homotopy from v to w, we see that $v_0 = v$ and $v_N = w$. As $d((i/N,t),((i+1)/N,t)) = 1/N < \delta$, we see that $d(h(i/N,t), h((i+1)/N,t)) < 2$, or in other words $d(v_i(t), v_{i+1}(t)) < 2$ for all t. Thus, the lemma tells us that $\nu(v_i) = \nu(v_{i+1})$ for $i = 0, \dots, N-1$, which implies that $\nu(v_0) = \nu(v_N)$, or in other words $\nu(v) = \nu(w)$. □

Corollary 13.12. *If $n \neq m$ then $u_n \not\simeq_{re} u_m$ (because $\nu(u_n) \neq \nu(u_m)$), so $a_n \neq a_m$.* □

This completes the proof of the theorem.

14. The Fundamental Theorem of Algebra

Theorem 14.1. *Let $p(z)$ be a monic complex polynomial of degree $n > 0$, so $p(z) = z^n + a_{n-1}z^{n-1} + \ldots + a_1 z + a_0$ for some $a_0, \ldots, a_{n-1} \in \mathbb{C}$. Then p has a complex root. (In other words, there exists $z \in \mathbb{C}$ such that $p(z) = 0$.)*

There are many different proofs of this important theorem. Here we give one that uses algebraic topology.

Proof. We will suppose that $p(z) \neq 0$ for all z and derive a contradiction.

First define $q(z) = p(z)/|p(z)|$, which gives a continuous map $q\colon \mathbb{C} \to S^1$, because $p(z) \neq 0$ for all $z \in \mathbb{C}$. Then define $r(z) = q(z)/q(|z|)$, which is again a continuous map $r\colon \mathbb{C} \to S^1$. Note that if R is a nonnegative real number then $R = |R|$ and so $r(R) = 1$. Now define $v_R\colon I \to S^1$ by $v_R(t) = r(Re^{2\pi it})$; this is easily seen to be a based loop in S^1. Clearly v_0 is just the constant loop c_1. Moreover, using the map $h(s,t) = r(Rse^{2\pi it})$ we see that v_0 is homotopic relative to endpoints to v_R, so $[v_R] = [v_0] = e$ in $\pi_1 S^1$.

On the other hand, when $|z|$ is large, the term z^n in $p(z)$ dominates. Thus, when R is large, $p(Re^{2\pi it})$ is approximately $R^n e^{2n\pi it}$, so $q(Re^{2\pi it})$ is approximately $e^{2n\pi it} = u_n(t)$, so $v_R(t)$ is approximately $u_n(t)$. In particular, when R is sufficiently large we should have $|v_R(t) - u_n(t)| < 2$ for all t, so $v_R \simeq_{\text{re}} u_n$ by Lemma 13.10. It follows that $a_0 = e = [v_R] = [u_n] = a_n$, which contradicts Corollary 13.12.

To make this rigorous, we would need to give more precise estimates. Put $A = 1 + |a_0| + \ldots + |a_{n-1}|$ (so in particular $A \geq 1$). The basic fact is that when $|w| = R > A$ we have

$$\begin{aligned} |p(w) - w^n| &= |a_0 + a_1 w + \ldots + a_{n-1}w^{n-1}| \\ &\leq |a_0| + |a_1|R + \ldots + |a_{n-1}|R^{n-1} \\ &\leq (|a_0| + \ldots + |a_{n-1}|)R^{n-1} \\ &< AR^{n-1}, \end{aligned}$$

so $|p(w)/w^n - 1| = |p(w) - w^n|/R^n < A/R < 1$. Using this one can check that for $R > A$ we have $|v_R(t) - u_n(t)| < 2$ as required; we omit the details. □

15. Functorality

To proceed any further, we need a way to compare the fundamental groups of different spaces.

Definition 15.1. Let X and Y be metric spaces, with chosen basepoints x_0 and y_0. A *based map* from X to Y is a continuous map $f\colon X \to Y$ with $f(x_0) = y_0$. If $f, g\colon X \to Y$ are based maps, a *based homotopy* between them is a map $h\colon I \times X \to Y$ such that $h(s, x_0) = y_0$ for all $s \in I$ and $h(0,x) = f(x)$ and $h(1,x) = g(x)$ for all $x \in X$. We say that f and g are based-homotopic if there is a based homotopy between them.

We also say that X and Y are based-homotopy equivalent if there are based maps $f\colon X \to Y$ and $g\colon Y \to X$ such that fg is based-homotopic to 1_Y and gf is based-homotopic to 1_X.

The results of Section 10 can easily be adapted to the based situation, so the relation of being based-homotopic is an equivalence relation, and so on.

Let $f\colon X \to Y$ be a based map. If $u\colon I \to X$ is a based loop in X, then $f \circ u\colon I \to Y$ is a based loop in Y. If $h\colon I \times I \to X$ is a based homotopy between two based loops u and u', then we see that the map $f \circ h\colon I \times I \to Y$ is a based homotopy between $f \circ u$ and $f \circ u'$. It follows that there is a well-defined function $f_*\colon \pi_1 X \to \pi_1 Y$ given by $f_*[u] = [f \circ u]$.

Proposition 15.2. *The function $f_*\colon \pi_1 X \to \pi_1 Y$ is a group homomorphism.*

Proof. Suppose we have two elements $a = [u]$ and $b = [v]$ of $\pi_1 X$, so $ab = [u * v]$. Then $f_*(ab) = f_*[u * v] = [f \circ (u * v)]$, and $f_*(a)f_*(b) = [f \circ u][f \circ v] = [(f \circ u) * (f \circ v)]$. Now,

$$(u * v)(t) = \begin{cases} u(2t) & \text{if } 0 \leq t \leq 1/2 \\ v(2t-1) & \text{if } 1/2 \leq t \leq 1, \end{cases}$$

so

$$\begin{aligned}(f\circ(u*v))(t) &= f((u*v)(t)) \\ &= \begin{cases} f(u(2t)) & \text{if } 0\le t\le 1/2 \\ f(v(2t-1)) & \text{if } 1/2\le t\le 1, \end{cases} \\ &= \begin{cases} (f\circ u)(2t) & \text{if } 0\le t\le 1/2 \\ (f\circ v)(2t-1) & \text{if } 1/2\le t\le 1, \end{cases} \\ &= ((f\circ u)*(f\circ v))(t),\end{aligned}$$

so $f_*(ab) = f_*(a)f_*(b)$ as required. □

Proposition 15.3. *The maps f_* have the following properties.*

(a) *If $f\colon X\longrightarrow X$ is the identity map, then $f_*\colon \pi_0(X)\longrightarrow\pi_0(X)$ is the identity map.*
(b) *If $f\colon X\longrightarrow Y$ and $g\colon Y\longrightarrow Z$ are based maps then $(gf)_* = g_*f_*\colon\pi_0(X)\longrightarrow\pi_0(Z)$.*

In other words, $\pi_1 X$ is a functor for based maps.

Proof. This is easy. □

Proposition 15.4. *If $f,g\colon X\longrightarrow Y$ are based maps that are based-homotopic to each other, then $f_* = g_*\colon \pi_1X\longrightarrow\pi_1Y$.*

Proof. Let $h\colon I\times X\longrightarrow Y$ be a based homotopy between f and g. Then for any based loop $u\colon I\longrightarrow X$, we see that the map $k(s,t) = h(s,u(t))$ gives a based homotopy between $f\circ u$ and $g\circ u$, so $f_*[u] = g_*[u]$ as claimed. □

Corollary 15.5. *If $f\colon X\longrightarrow Y$ is a based homotopy equivalence, then $f_*\colon\pi_1X\longrightarrow\pi_1Y$ is an isomorphism of groups.*

Proof. As f is a based homotopy equivalence, there is a based map $g\colon Y\longrightarrow X$ such that fg is based homotopic to 1_Y and gf is based homotopic to 1_X. By the above results, the maps $f_*\colon\pi_1X\longrightarrow\pi_1Y$ and $g_*\colon\pi_1Y\longrightarrow\pi_1X$ are group homomorphisms such that $f_*g_* = (fg)_* = (1_Y)_* = 1_{\pi_1Y}$ and $g_*f_* = (gf)_* = (1_X)_* = 1_{\pi_1X}$. Thus, g_* is an inverse for f_* and so f_* is an isomorphism. □

Example 15.6. Consider $X = S^1$ and $Y = \mathbb{C}\setminus\{0\}$, and take 1 as the basepoint in both X and Y. Define $f\colon X\longrightarrow Y$ and $g\colon Y\longrightarrow X$ by $f(z) = z$ and $g(z) = z/|z|$. Clearly $f(1) = 1$ and $g(1) = 1$ so these are based maps. We have $g\circ f = 1_X$, and the map $h\colon I\times Y\longrightarrow Y$ given by $h(t,z) = (1-t)z + tz/|z|$ is a based homotopy between $f\circ g$ and 1_Y. Thus S^1 and $\mathbb{C}\setminus\{0\}$ are based homotopy equivalent, and $\pi_1(\mathbb{C}\setminus\{0\})\simeq\pi_1S^1\simeq\mathbb{Z}$.

Example 15.7. The letters A,D,O,P,Q and R are all homotopy equivalent to S^1, and with appropriate choice of basepoints, one can ensure that they are based-homotopy equivalent to S^1. Thus, in each case, the fundamental group is $\mathbb{Z}$.

16. Products

Theorem 16.1. *Let X and Y be spaces with basepoints x_0 and y_0, and give the product space $X\times Y$ the basepoint (x_0,y_0). Then $\pi_1(X\times Y) = \pi_1(X)\times\pi_1(Y)$.*

Proof. The basic idea of the proof is very simple: everything just works independently in the two different "coordinates" of $X\times Y$. Here are the details.

Define $X\xleftarrow{p} X\times Y\xrightarrow{q} Y$ by $p(x,y) = x$ and $q(x,y) = y$. This gives homomorphisms

$$\pi_1(X)\xleftarrow{p_*}\pi_1(X\times Y)\xrightarrow{q_*}\pi_1(Y),$$

which we can combine to define a homomorphism $\phi\colon\pi_1(X\times Y)\longrightarrow\pi_1(X)\times\pi_1(Y)$ by $\phi(a) = (p_*(a), q_*(a))$. We claim that this is an isomorphism.

Suppose we have an element $(b,c)\in\pi_1(X)\times\pi_1(Y)$. Then $b\in\pi_1(X)$ so $b = [v]$ for some based loop $v\colon I\longrightarrow X$. Similarly $c = [w]$ for some based loop $w\colon I\longrightarrow Y$. We can define a continuous map $u\colon I\longrightarrow X\times Y$ by $u(t) = (v(t), w(t))$. This satisfies $u(0) = (v(0), w(0)) = (x_0, y_0)$ and similarly

$u(1) = (x_0, y_0)$, so u is a based loop in $X \times Y$. Clearly $p \circ u = v$ and $q \circ u = w$ so $p_*[u] = [v] = b$ and $q_*[u] = [w] = c$ so $\phi([u]) = (b, c)$. This shows that ϕ is surjective.

Now suppose we have two based loops $u, u' \colon I \longrightarrow X \times Y$ such that $\phi([u]) = \phi([u'])$. We have $u(t) = (v(t), w(t))$ for some $v(t) \in X$ and $w(t) \in Y$, and it is easy to see that the maps $v = p \circ u$ and $w = q \circ u$ are based loops in X and Y respectively. Similarly $u'(t) = (v'(t), w'(t))$ for some based loops $v' = p \circ u'$ and $w' = q \circ u'$ in X and Y. We have $\phi([u]) = ([v], [w])$ and $\phi([u']) = ([v'], [w'])$. These are equal by assumption, so $[v] = [v']$ in $\pi_1 X$ and $[w] = [w']$ in $\pi_1 Y$. Thus, there are based homotopies $h \colon v \simeq v'$ in X, and $k \colon w \simeq w'$ in Y. We can define $m \colon I \times I \longrightarrow X \times Y$ by $m(s, t) = (h(s, t), k(s, t))$ and we find that this is a based homotopy between u and u'. Thus $[u] = [u']$. We conclude that ϕ is injective, and thus an isomorphism as claimed. □

Corollary 16.2. *The torus T is homeomorphic to $S^1 \times S^1$, so $\pi_1 T$ is isomorphic to $\mathbb{Z} \times \mathbb{Z}$.* □

17. The weak van Kampen theorem

Suppose we have a based space X, and we can write X as the union of two open subsets U and V, with $x_0 \in U \cap V$. We use x_0 as the base point in U and V as well as X. Suppose we already understand the fundamental groups of U and V. What can we deduce about the fundamental group of X itself? Firstly, we have inclusion maps $i \colon U \longrightarrow X$ and $j \colon V \longrightarrow X$, which gives homomorphisms $i_* \colon \pi_1 U \longrightarrow \pi_1 X$ and $j_* \colon \pi_1 V \longrightarrow \pi_1 X$. The following theorem tells us that in some sense these give everything, provided that $U \cap V$ is connected.

Theorem 17.1. *Suppose that X is a based space, and that U and V are open subsets of X such that*

(a) $x_0 \in U \cap V$
(b) $U \cap V$ *is path-connected*
(c) $X = U \cup V$.

Let $U \xrightarrow{i} X \xleftarrow{j} V$ be the inclusions. Then every element $a \in \pi_1 X$ can be written in the form $a = b_1 b_2 \dots b_n$, where each element b_k is either in the image of i_* or in the image of j_*.

Remark 17.2. The actual van Kampen theorem gives a complete determination of $\pi_1 X$, and can be proved by an elaboration of the argument given below. However, the answer is somewhat unwieldy and involves a certain amount of extra group theory, so we omit it.

Remark 17.3. Hypothesis (b) is necessary. To see this, take $X = S^1$ and $U = S^1 \setminus \{i\}$ and $V = S^1 \setminus \{-i\}$. Both U and V are contractible, so the images of i_* and j_* are the trivial group. If the theorem applied we could conclude that every element in $\pi_1 S^1$ is trivial, which we know is not the case. Of course the theorem does *not* apply, because the space $U \cap V = S^1 \setminus \{i, -i\}$ is disconnected.

The Theorem will be proved after some preliminary results.

Lemma 17.4. *Let $u \colon I \longrightarrow X$ be a path from x_0 to x_1, and suppose that $0 < r < 1$. Put $y = u(r)$ and define $u_0, u_1 \colon I \longrightarrow X$ by $u_0(t) = u(rt)$ and $u_1(t) = u(r + (1-r)t)$, so u_0 is a path from x_0 to y, and u_1 is a path from y to x_1. Then $[u] = [u_0][u_1]$*

Proof. Put $v = u_0 * u_1$, so

$$v(t) = \begin{cases} u_0(2t) & \text{if } 0 \le t \le 1/2 \\ u_1(2t-1) & \text{if } 1/2 \le t \le 1 \end{cases}$$

$$= \begin{cases} u(2rt) & \text{if } 0 \le t \le 1/2 \\ u(r + (1-r)(2t-1)) & \text{if } 1/2 \le t \le 1 \end{cases}$$

In other words, if we define $f \colon I \longrightarrow I$ by $f(t) = 2rt$ for $t \le 1/2$ and $f(t) = r + (1-r)(2t-1)$ for $t \ge 1/2$, then $v(t) = u(f(t))$. Moreover, we see that f is continuous and $f(0) = 0$ and $f(1) = 1$. Now define $h \colon I \times I \longrightarrow X$ by $h(s, t) = u((1-s)t + sf(t))$. One can easily check that this is a homotopy relative to endpoints between u and v. □

Corollary 17.5. *Let $u\colon I \longrightarrow X$ be a path from x_0 to x_n, and suppose that $0 = r_0 < r_1 < \ldots < r_n = 1$. Put $x_i = u(r_i)$ and define $u_i\colon I \longrightarrow X$ by $u_i(t) = u(r_i + t(r_{i+1} - r_i))$, so u_i is a path from x_i to x_{i+1}. Then $[u] = [u_0][u_1]\ldots[u_{n-1}]$.*

Proof. This can be deduced from the lemma by induction, or proved by the same method as the lemma. □

Lemma 17.6. *Let A and B be open subsets of $I = [0,1]$ such that $I = A \cup B$. Then there exists $N > 0$ such that every subinterval $[a,b] \subseteq [0,1]$ with $b - a \leq 1/N$ has either $[a,b] \subseteq A$ or $[a,b] \subseteq B$.*

Proof. Suppose not. Put $F = I \setminus A$ and $G = I \setminus B$. As A and B are open, the sets F and G are closed. As $I = A \cup B$, we see that $F \cap G = \emptyset$. By assumption, for any $n > 0$ we can choose an interval $[a_n, b_n]$ with $b_n - a_n \leq 1/n$ that is not contained in A or in B. As $[a_n, b_n] \not\subseteq A$ we see that $[a_n, b_n] \cap F \neq \emptyset$, so we can choose $x_n \in [a_n, b_n] \cap F$. Similarly, we can choose $y_n \in [a_n, b_n] \cap G$. Because both x_n and y_n lie in the interval $[a_n, b_n]$ which has length at most $1/n$, we see that $|x_n - y_n| \leq 1/n$. As I is compact, we can choose a subsequence (x_{n_k}) of (x_n) that converges to some point $z \in I$ say. As $x_{n_k} \in F$ and F is closed, we see that $z \in F$. As $|y_{n_k} - x_{n_k}| \leq 1/n_k$, we see that (y_{n_k}) also converges to z. As $y_{n_k} \in G$ and G is closed, it follows that $z \in G$. Thus $z \in F \cap G$, contradicting the fact that $F \cap G = \emptyset$. □

Corollary 17.7. *Let X be a space, and let U and V be open subsets of X such that $X = U \cup V$. Let $u\colon I \longrightarrow X$ be a path. Then there exists $N > 0$ such that for each $i \in \{0, 1, \ldots, N-1\}$ either $u([i/N, (i+1)/N]) \subseteq U$ or $u([i/N, (i+1)/N]) \subseteq V$.*

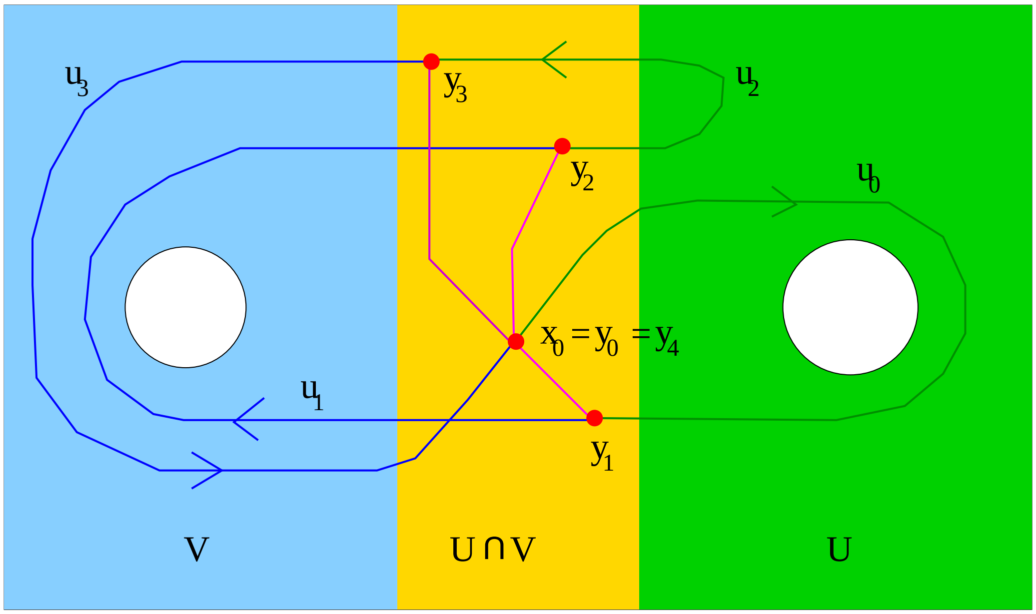
Proof. Put $A = \{t \in I \;:\; u(t) \in U\}$ and $B = \{t \in I \;:\; u(t) \in V\}$. Then A and B are open subsets of I and $I = A \cup B$. Thus, by the lemma, there exists N such that every subinterval of length at most $1/N$ is either contained in A or contained in B. In particular, the interval $[i/N, (i+1)/N]$ is either contained in A (in which case $u([i/N, (i+1)/N]) \subseteq U$) or contained in B (in which case $u([i/N, (i+1)/N]) \subseteq V$). □

Proof of Theorem 17.1. Any element $a \in \pi_1 X$ can be represented by a based loop $u\colon I \longrightarrow X$. We can then choose N as in Corollary 17.7 and put $I_i = [i/N, (i+1)/N]$, so for each i we have either $u(I_i) \subseteq U$ or $u(I_i) \subseteq V$. We will suppose that $u(I_0) \subseteq U$; essentially the same argument works if $u(I_0) \subseteq V$. Put $i_0 = t_0 = 0$ and $y_0 = x_0 \in U \cap V$, so $y_0 = u(t_0)$. Then let i_1 be the first index such that $u(I_{i_1}) \not\subseteq U$, and put $t_1 = i_1/N$ and $y_1 = u(t_1)$. We then have $u([t_0, t_1]) \subseteq U$, so in particular $y_1 \in U$. As $u(I_{i_1}) \not\subseteq U$, we must have $u(I_{i_1}) \subseteq V$ instead, so in particular $y_1 \in V$ as well, so $y_1 \in U \cap V$. Let i_2 be the first index after i_1 such that $u(I_{i_2}) \not\subseteq V$, and put $t_2 = i_2/N$ and $y_2 = u(t_2)$. Then $u([t_1, t_2]) \subseteq V$ and $y_2 \in U \cap V$. Continuing in this way, we get a sequence $0 = t_0 < \ldots < t_n = 1$ such that the points $y_i = u(t_i)$ lie in $U \cap V$ and each set $u([y_i, y_{i+1}])$ lies either in U or in V.

Now put $u_i(t) = u(t_i + t(t_{i+1} - t_i))$, which is a path from y_i to y_{i+1} lying either in U or in V. Corollary 17.5 tells us that $[u] = [u_0]\ldots[u_{n-1}]$. Next, for $i = 1, \ldots, n-1$ we choose a path v_i in $U \cap V$ from x_0 to y_i, which is possible because $U \cap V$ is assumed to be path-connected. We define $u'_0 = u_0 * \overline{v}_1$ and $u'_{n-1} = v_{n-1} * u_{n-1}$ and $u'_k = (v_k * u_k) * \overline{v}_{k+1}$ for $0 < k < n-1$. We find that u'_k is a based loop either in U or in V, so the element $b_k = [u'_k]$ lies either in the image of i_* or in the image of j_*. Moreover, we have

$$b_0 \ldots b_{n-1} = [u_0][v_1]^{-1}[v_1][u_1][v_2]^{-1} \ldots [v_{n-2}][u_{n-2}][v_{n-1}]^{-1}[v_{n-1}][u_{n-1}] = [u_0] \ldots [u_{n-1}] = [u] = a,$$

as required.



□

We will often consider the following special case.

Corollary 17.8. *Suppose that X is a based space, and that U and V are open subsets of X such that*

(a) $x_0 \in U \cap V$
(b) $U \cap V$ *is path-connected*
(c) $X = U \cup V$.
(d) $\pi_1 U$ *and* $\pi_1 V$ *are trivial.*

Then $\pi_1 X$ is also trivial.

Proof. Every element of $\pi_1 X$ can be written as a product of elements in the image of $i_* \colon \pi_1 U \longrightarrow \pi_1 X$ or $j_* \colon \pi_1 V \longrightarrow \pi_1 X$, but assumption (d) implies that these images are trivial, so $\pi_1 X$ is trivial. □

Theorem 17.9. *For $n > 1$, the space S^n is simply connected.*

Proof. We saw in Proposition 8.9 that S^n is path-connected, so we just need to show that the fundamental group is trivial. We use the point $x_0 = (1, 0, \dots, 0)$ as the basepoint. We also consider the points $x_1 = (0, \dots, 0, 1)$ and $x_2 = (0, \dots, 0, -1)$, and we put $U = S^n \setminus \{x_1\}$ and $V = S^n \setminus \{x_2\}$. We know that S^n with a point removed is homeomorphic (by stereographic projection) to $\mathbb{R}^n$, so U and V are homeomorphic to $\mathbb{R}^n$ and thus are contractible, so $\pi_1 U$ and $\pi_1 V$ are trivial. Moreover, $U \cap V = U \setminus \{x_2\}$ is homeomorphic to $\mathbb{R}^n$ with a point removed, which is path-connected because $n > 1$. Thus, Corollary 17.8 applies and tells us that $\pi_1 S^n$ is trivial. □

Example 17.10. For another example, consider the figure eight space E as in Example 10.15, so E is the union of the unit circle centred at $(-1, 0)$ with the unit circle centred at $(1, 0)$. Put $x_0 = (0, 0)$ and $x_1 = (-2, 0)$ and $x_2 = (2, 0)$, and take x_0 as the basepoint. Put $U = E \setminus \{x_1\}$ and $V = E \setminus \{x_2\}$, so U and V are open and $E = U \cup V$. It is not hard to see that $U \cap V$ is path-connected and that U and V are both homotopy equivalent to S^1. It follows that $\pi_1 U$ and $\pi_1 V$ are isomorphic to $\mathbb{Z}$. More explicitly, define $u \colon I \longrightarrow U$ by $u(t) = (1 - \cos(t), \sin(t))$, define $v \colon I \longrightarrow V$ by $v(t) = (\cos(t) - 1, \sin(t))$, and put $a = [u] \in \pi_1 U$ and $b = [v] \in \pi_1 V$. Then every element of $\pi_1 U$ is a^j for a unique integer j, and every element of $\pi_1 V$ is b^k for a unique integer k.

Now put $c = i_* a \in \pi_1 E$ and $d = j_* b \in \pi_1 E$. The weak van Kampen theorem tells us that every element in $\pi_1 E$ can be written as a product of powers of c and d. For example, some of the shortest elements are

$$e, c, c^{-1}, d, d^{-1}, c^2, cd, cd^{-1}, c^{-2}, c^{-1}d, c^{-1}d^{-1}, dc, dc^{-1}, d^2, d^{-1}c, d^{-1}c^{-1}, d^{-2}, \dots$$

There are some obvious identities relating expressions of this type, such as $cdd^{-1}c = c^2$. However, it turns out that there are no non-obvious identities. More precisely, every element $g \in \pi_1 E$ can be written in a *unique* way in the form $g = g_1 g_2 \dots g_r$, where each element g_i is either c, c^{-1}, d or d^{-1}, and c is never adjacent to c^{-1}, and d is never adjacent to d^{-1}. (The case $r = 0$ is allowed, corresponding to the identity element of $\pi_1 E$.)

Note that the uniqueness clause means that $cd \neq dc$, so $\pi_1 E$ is nonabelian.

The usual terminology is that $\pi_1 E$ is *freely generated* by c and d, or that $\pi_1 E$ is a *free group on two generators*.

We saw in Example 10.15 that $\mathbb{R}^2 \setminus$ (two points) is homotopy equivalent to E, so $\pi_1(\mathbb{R}^2 \setminus$ (two points)) is also a free group on two generators. More generally, it can be shown that $\pi_1(\mathbb{R}^2 \setminus$ (n points)) is a free group on n generators.

18. The fundamental group of projective space

We next calculate $\pi_1 \mathbb{R}P^n$. Note that $\mathbb{R}P^0$ is just a point, so $\pi_1 \mathbb{R}P^0$ is trivial; and $\mathbb{R}P^1 \simeq S^1$, so $\pi_1 \mathbb{R}P^1 \simeq \mathbb{Z}$. From now on, we assume that $n > 1$. We take the point $x_0 = (1, 0, \dots, 0)$ as our basepoint in S^n, and the point $y_0 = q(x_0)$ as the basepoint in $\mathbb{R}P^n$.

Definition 18.1. We define $u \colon I \longrightarrow S^n$ by $u(t) = (\cos(\pi t), \sin(\pi t), 0, \dots, 0)$, so $u(0) = x_0$ and $u(1) = -x_0$. We then define $v(t) = q(u(t))$, so $v \colon I \longrightarrow \mathbb{R}P^n$. Because $q(x_0) = q(-x_0) = y_0$, we see that v is a based loop in $\mathbb{R}P^n$, so we have an element $a = [v] \in \pi_1 \mathbb{R}P^n$.

Theorem 18.2. *When $n > 1$ we have $\pi_1 \mathbb{R}P^n = \{e, a\}$, where $a \neq e$ but $a^2 = e$. Thus, $\pi_1 \mathbb{R}P^n$ is isomorphic to the cyclic group C_2.*

Proof. Combine Lemma 18.3, Proposition 18.9 and Corollary 18.12. □

The argument is similar in many respects to our calculation of $\pi_1 S^1$. In a more abstract treatment, we would prove a general theorem about fundamental groups and "covering spaces", and deduce our calculations of $\pi_1 S^1$ and $\pi_1 \mathbb{R}P^n$ as corollaries.

Lemma 18.3. *$a^2 = e$ in $\pi_1 \mathbb{R}P^n$.*

Proof. Define

$$u'(t) = -u(t) = (-\cos(\pi t), -\sin(\pi t)) = (\cos(\pi(t+1)), \sin(\pi(t+1))),$$

so u' is a path from $-x_0$ to x_0. Because $q(-x) = q(x)$ for all x, we have $q \circ u' = q \circ u = v$. Thus $v * v = (q \circ u) * (q \circ u') = q \circ (u * u')$. On the other hand, $u * u'$ is a path from x_0 to x_0, in other words a based loop in the space S^n. We know that S^n is simply connected (as $n > 1$) so $u * u' \simeq_{\mathrm{re}} c_{x_0}$. It follows that $v * v = q \circ (u * u') \simeq_{\mathrm{re}} q \circ c_{x_0} = c_{y_0}$, so $a^2 = [v * v] = e$ as claimed. □

Definition 18.4. Suppose $x \in S^n$. Put $U_x = \{y \in S^n \;:\; \langle x, y \rangle > 0\}$ and $V_x = \{A \in \mathbb{R}P^n \;:\; d_2(q(x), A) < \sqrt{2}\}$. These are clearly open sets.

Lemma 18.5. *The map $q \colon S^n \longrightarrow \mathbb{R}P^n$ restricts to give a homeomorphism $q_x \colon U_x \longrightarrow V_x$.*

Proof. First, Lemma 6.5 tells us that $q(y) \in V_x$ if and only if $\langle x, y \rangle \neq 0$. In particular, we see that $q(U_x) \subseteq V_x$, so q restricts to give a map $q_x \colon U_x \longrightarrow V_x$.

Suppose that $A \in V_x$. As $q \colon S^n \longrightarrow \mathbb{R}P^n$ is surjective, we have $A = q(y)$ for some y. As $q(y) \in V_x$, we have $\langle x, y \rangle \neq 0$. If $\langle x, y \rangle > 0$ we put $z = y$, otherwise we put $z = -y$. Either way we have $\langle x, z \rangle > 0$ (so $z \in U_x$) and $q(z) = A$. Thus, $q_x \colon U_x \longrightarrow V_x$ is surjective. We note for future reference that $Ax = q(z)x = \langle x, z \rangle z$, which is nonzero.

Now suppose that $y, z \in U_x$ and $q(y) = q(z)$. Then $z = \pm y$, so $\langle x, z \rangle = \pm \langle x, y \rangle$. The negative case is impossible, because $\langle x, z \rangle$ and $\langle y, z \rangle$ are both strictly positive by assumption. Thus $y = z$. This proves that $q_x \colon U_x \longrightarrow V_x$ is a bijection.

Now define $r_x\colon V_x \longrightarrow S^n$ by $r(A) = Ax/\|Ax\|$. This is legitimate because we observed above that $Ax \neq 0$ for all $A \in V_x$. If $y \in U_x$ then $q_x(y)x = \langle x, y\rangle y$ so $\|q_x(y)x\| = \langle x, y\rangle$ so $r_x(q_x(y)) = y$. It follows that $r_x = q_x^{-1}$. Our formula for r_x shows that it is continuous, so q_x is a homeomorphism. □

Proposition 18.6. *Let $w\colon I \longrightarrow \mathbb{R}P^n$ be a based loop. Then there is a unique path $\tilde{w}\colon I \longrightarrow S^n$ such that $\tilde{w}(0) = x_0$ and $q \circ \tilde{w} = w\colon I \longrightarrow \mathbb{R}P^n$. Moreover, we have $\tilde{w}(1) = \pm x_0$*

Proof. As I is compact, the map w is uniformly continuous. We can therefore choose $\delta > 0$ such that $d_2(w(s), w(t)) < \sqrt{2}$ whenever $|s - t| < \delta$. Choose an integer N such that $N > 1/\delta$ and put $I_j = [j/N, (j+1)/N]$ for $j = 0, \dots, N-1$. We also put $y_j = w(j/N) \in \mathbb{R}P^n$.

Note that for $t \in I_0$ we have $|t - 0| \leq 1/N < \delta$, so $d_2(q(x_0), w(t)) = d_2(w(0), w(t)) < \sqrt{2}$, so $w(t) \in V_{x_0}$. We can thus use our homeomorphism $q_{x_0}\colon U_{x_0} \longrightarrow V_{x_0}$ to get a map $\tilde{w}_0 = q_{x_0}^{-1} \circ w\colon I_0 \longrightarrow U_{x_0} \subset S^n$, with $\tilde{w}_0(0) = q_{x_0}^{-1}(q(x_0)) = x_0$. We put $x_1 = \tilde{w}_0(1/N)$, and note that $q(x_1) = w(1/N) = y_1$.

Next, note that for $t \in I_1$ we have $|t - 1/N| < \delta$ so $d_2(q(x_0), w(t)) = d_2(w(1/N), w(t)) < \sqrt{2}$ so $w(t) \in V_{x_1}$. We can thus use our homeomorphism $q_{x_1}: U_{x_1} \longrightarrow V_{x_1}$ to get a map $\tilde{w}_1 = q_{x_1}^{-1} \circ w\colon I_1 \longrightarrow U_{x_1} \subset S^n$, with $\tilde{w}_1(1/N) = q_{x_1}^{-1}(w(1/N)) = q_{x_1}^{-1}(q(x_1)) = x_1$. We put $x_2 = \tilde{w}_1(2/N)$, and note that $q(x_2) = w(2/N) = y_2$.

Continuing in this way, we get points $x_0, \dots, x_n$ with $q(x_i) = y_i$ and maps $\tilde{w}_i\colon I_i \longrightarrow S^n$ with $\tilde{w}_i((i+1)/N) = x_{i+1} = \tilde{w}_{i+1}((i+1)/N)$. We can thus combine all the maps $\tilde{w}_i$ to get a continuous map $\tilde{w}\colon I \longrightarrow S^n$ with $q \circ \tilde{w} = w\colon I \longrightarrow \mathbb{R}P^n$.

Now let $z\colon I \longrightarrow S^n$ be another map such that $z(0) = x_0$ and $q(z(t)) = w(t)$ for all t. As $q(z(t)) = w(t) = q(\tilde{w}(t))$, we must have $\tilde{w}(t) = \pm z(t)$, so $\tilde{w}(t) = g(t)z(t)$ for some function $g\colon I \longrightarrow \{1, -1\}$. Note that $g(0) = 1$ because $z(0) = \tilde{w}(0) = x_0$. Note also that $\langle z(t), \tilde{w}(t)\rangle = g(t)\|z(t\|^2 = g(t)$. As the functions z and $\tilde{w}$ are continuous, the function $g(t) = \langle z(t), \tilde{w}(t)\rangle$ is also continuous. The only continuous functions from I to $\{1, -1\}$ are constants, and $g(0) = 1$, so $g(t) = 1$ for all t. Thus $\tilde{w}(t) = g(t)z(t) = z(t)$ for all t. This proves that $\tilde{w}$ is unique.

Finally, note that $q(\tilde{w}(1)) = w(1) = y_0 = q(x_0)$, so $\tilde{w}(1) = \pm x_0$. □

Definition 18.7. Let $w\colon I \longrightarrow \mathbb{R}P^n$ be a based loop. Let $\tilde{w}$ be as above, and put

$$\epsilon(w) = \begin{cases} +1 & \text{if } \tilde{w}(1) = x_0 \\ -1 & \text{if } \tilde{w}(1) = -x_0 \end{cases}$$

(so $\tilde{w}(1) = \epsilon(w)x_0$ in all cases.) Equivalently, we have $\epsilon(w) = \langle x_0, \tilde{w}(1)\rangle$.

Remark 18.8. If w is the constant path c_{y_0}, then $\tilde{w}$ is the constant path c_{x_0} and $\epsilon(w) = 1$. For the standard loop v in Definition 18.1, we have $\tilde{v} = u$ and thus $\epsilon(v) = -1$.

Proposition 18.9. *Every element of $\pi_1\mathbb{R}P^n$ is equal to e or to a.*

Proof. If $b \in \pi_1\mathbb{R}P^n$ then we can choose a based loop $w\colon I \longrightarrow \mathbb{R}P^n$ with $b = [w]$, and then we can use the proposition to find a path $\tilde{w}\colon I \longrightarrow S^n$ with $w = q \circ \tilde{w}$ and $\tilde{w}(1) = \pm x_0$. If $\tilde{w}(1) = x_0$ then $\tilde{w}$ is a based loop in S^n. As S^n is simply connected, we deduce that $\tilde{w} \simeq_{\text{re}} c_{x_0}$, so $w = q \circ \tilde{w} \simeq_{\text{re}} q \circ c_{x_0} = c_{y_0}$, so $b = [w] = [c_{y_0}] = e$.

Suppose instead that $\tilde{w}(1) = -x_0$. Then $\tilde{w} * \overline{u}$ is a path from x_0 to x_0, or in other words a based loop in S^n. Again we deduce that $\tilde{w} * \overline{u} \simeq_{\text{re}} c_{x_0}$ and thus $[q \circ \tilde{w}][q \circ \overline{u}] = e$ or equivalently $ba^{-1} = e$, so $b = a$. □

Proposition 18.10. *If z and w are based loops in $\mathbb{R}P^n$ with $d(z(t), w(t)) < \sqrt{2}$ for all t, then $\epsilon(z) = \epsilon(w)$.*

Proof. Let $\tilde{z}, \tilde{w}\colon I \longrightarrow \mathbb{R}P^n$ be as in Proposition 18.6. Put $g(t) = \langle \tilde{z}(t), \tilde{w}(t)\rangle$. Lemma 6.5 tells us that

$$\sqrt{2(1 - g(t)^2)} = d(q(\tilde{z}(t)), q(\tilde{w}(t))) = d(z(t), w(t)) < \sqrt{2},$$

so we cannot have $g(t) = 0$ for any t. On the other hand, g is clearly continuous and $g(0) > 0$, so the Intermediate Value Theorem tells us that $g(1) > 0$ also. We have

$$g(1) = \langle \tilde{z}(1), \tilde{w}(1) \rangle = \langle \epsilon(z)x_0, \epsilon(w)x_0 \rangle = \epsilon(z)\epsilon(w).$$

On the other hand, both $\epsilon(z)$ and $\epsilon(w)$ are ± 1, so

$$\epsilon(z)\epsilon(w) = \begin{cases} +1 & \text{if } \epsilon(z) = \epsilon(w) \\ -1 & \text{if } \epsilon(z) \neq \epsilon(w). \end{cases}$$

As $\epsilon(z)\epsilon(w) = g(1) > 0$, we must have $\epsilon(z) = \epsilon(w)$ as claimed. □

Corollary 18.11. *If z and w are based loops in $\mathbb{R}P^n$ with $z \simeq_{re} w$, then $\epsilon(z) = \epsilon(w)$.*

Proof. Choose a map $h\colon I \times I \longrightarrow \mathbb{R}P^n$ giving a homotopy relative to endpoints between z and w. As $I \times I$ is compact, h must be uniformly continuous, so we can find $\delta > 0$ such that $d(h(s,t), h(s',t')) < \sqrt{2}$ whenever $d((s,t),(s',t')) < \delta$. Choose an integer N such that $N > 1/\delta$, and define maps $z_0, \dots, z_N\colon I \longrightarrow \mathbb{R}P^n$ by $z_i(t) = h(i/N, t)$. As h is a homotpy relative to endpoints, we see that each z_i is a based loop. As h is a homotopy from z to w, we see that $z_0 = z$ and $z_N = w$. As $d((i/N,t),((i+1)/N,t)) = 1/N < \delta$, we see that $d(h(i/N,t), h((i+1)/N,t)) < \sqrt{2}$, or in other words $d(z_i(t), z_{i+1}(t)) < \sqrt{2}$ for all t. Thus, the proposition tells us that $\epsilon(z_i) = \epsilon(z_{i+1})$ for $i = 0, \dots, N-1$, which implies that $\epsilon(z_0) = \epsilon(z_N)$, or in other words $\epsilon(z) = \epsilon(w)$. □

Corollary 18.12. *The standard path v is not homotopic relative to endpoints to c_{y_0}, because $\epsilon(v) = -1$ and $\epsilon(c_{y_0}) = 1$. Thus $a \neq e$ in $\pi_1 \mathbb{R}P^n$.*

19. $\pi_1 SO(3)$

Recall that $SO(3)$ is the set of all 3×3 rotation matrices, or equivalently all matrices A such that $\det(A) = 1$ and AA^T is the identity. We take the identity matrix as the basepoint in $SO(3)$. Define $u\colon I \longrightarrow SO(3)$ by

$$u(t) = \begin{pmatrix} \cos(2\pi t) & -\sin(2\pi t) & 0 \\ \sin(2\pi t) & \cos(2\pi t) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

This is a based loop in $SO(3)$, giving an element $a \in \pi_1 SO(3)$.

Theorem 19.1. *We have $\pi_1 SO(3) = \{e, a\}$, with $a \neq e$ but $a^2 = e$.*

This theorem is in fact a disguised version of Theorem 18.2, as we see from the following remarkable result:

Theorem 19.2. *$SO(3)$ is homeomorphic to $\mathbb{R}P^3$.*

Here will will just give an outline of the proof of the theorem. We define a map $g\colon S^3 \longrightarrow M_3(\mathbb{R})$ by

$$g(w,x,y,z) = \begin{pmatrix} w^2+x^2-y^2-z^2 & 2(xy-wz) & 2(wy+xz) \\ 2(xy+wz) & w^2-x^2+y^2-z^2 & 2(yz-wx) \\ 2(xz-wy) & 2(wx+yz) & w^2-x^2-y^2+z^2 \end{pmatrix}.$$

It is tedious but essentially straightforward to check that $\det(g(w,x,y,z)) = 1$ and $g(w,x,y,z)g(w,x,y,z)^T = I$, so $g(w,x,y,z)$ is a rotation and g gives a map $S^3 \longrightarrow SO(3)$. (Of course one has to use the identity $w^2+x^2+y^2+z^2 = 1$, which holds for $(w,x,y,z) \in S^3$.) Next, one checks that the product of the matrix $g(w,x,y,z)$ with the vector (x,y,z) is (x,y,z) again, so (x,y,z) lies on the axis of rotation. By considering characteristic polynomials one can check that the angle of rotation is $\pm 2\cos^{-1}(w)$. Using a slight extension of this line of argument, one can check that g is surjective, and that $g(a) = g(b)$ iff $b = \pm a$ (which is parallel to the properties of the map $q\colon S^3 \to \mathbb{R}P^3$). It follows from this that there is a bijection $f\colon \mathbb{R}P^3 \longrightarrow SO(3)$ such that $g = f \circ q$, and one can check that f is in fact a homeomorphism.

Of course, the formulae above did not come out of thin air. The "real reason" that the theorem is true comes from the algebra of quaternions, but that is too long a story to explain here.

20. Classification results

We now use our knowledge of π_1 to show that certain pairs of spaces are not homotopy equivalent or homeomorphic to each other, extending the results of Section 9.

Lemma 20.1. *If the group $\mathbb{Z}^n$ is isomorphic to $\mathbb{Z}^m$, then $n = m$.*

Proof. For any abelian groups A and B (written additively), if $A \simeq B$ then $A/2A \simeq B/2B$, and so $|A/2A| = |B/2B|$. If $A = \mathbb{Z}^n$ it is not hard to see that $A/2A = (\mathbb{Z}/2\mathbb{Z})^n$ and so $|A/2A| = 2^n$. Thus, if $\mathbb{Z}^n \simeq \mathbb{Z}^m$ then $2^n = 2^m$ and so $n = m$. □

Proposition 20.2. *The space $(S^1)^n = S^1 \times \ldots \times S^1$ is not homotopy equivalent to $(S^1)^m$ unless $n = m$.*

Proof. Note that $\pi_1 S^1 \simeq \mathbb{Z}$ and $\pi_1(X \times Y) \simeq \pi_1 X \times \pi_1 Y$. It follows that $\pi_1((S^1)^n) \simeq \mathbb{Z}^n$.

Now suppose that $(S^1)^n$ is homotopy equivalent to $(S^1)^m$. Then $\pi_1((S^1)^n) \simeq \pi_1(S^1)^m)$, so $\mathbb{Z}^n \simeq \mathbb{Z}^m$, so $n = m$. □

Proposition 20.3. *No two of the spaces S^1, S^2, E and $\mathbb{R}P^2$ are homotopy equivalent to each other. (Here E is the figure eight, as in Example 10.15).*

Proof. It is enough to show that no two of the groups $\pi_1 S^1$, $\pi_1 S^2$, $\pi_1 E$ and $\pi_1 \mathbb{R}P^2$ are isomorphic. Note that $\pi_1 S^2$ is trivial, but the other groups are not; that $\pi_1 \mathbb{R}P^2$ is finite and nontrivial, but the other three groups are not; that $\pi_1 S^1$ is infinite and abelian, but the others are not; and that $\pi_1 E$ is nonabelian, whereas the other three groups are all abelian. This easily implies the proposition. □

Proposition 20.4. *$\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^3$.*

(Note however that $\mathbb{R}^2$ and $\mathbb{R}^3$ are both homotopy equivalent to a point, and so are homotopy equivalent to each other.)

Proof. If $\mathbb{R}^2$ were homeomorphic to $\mathbb{R}^3$, then $\mathbb{R}^2 \setminus \{0\}$ would be homeomorphic (and thus homotopy equivalent) to $\mathbb{R}^3 \setminus \{a\}$ for some $a \in \mathbb{R}^3$. We know that $\mathbb{R}^2 \setminus \{0\}$ is homotopy equivalent to S^1, and $\mathbb{R}^3 \setminus \{a\}$ is homotopy equivalent to S^2, and S^1 is not homotopy equivalent to S^2, so $\mathbb{R}^2 \setminus \{0\}$ is not homotopy equivalent to $\mathbb{R}^3 \setminus \{a\}$. Thus $\mathbb{R}^2$ cannot be homeomorphic to $\mathbb{R}^3$ after all. □

Corollary 20.5. *S^2 is not homeomorphic to S^3.*

Proof. Let $f\colon S^2 \to S^3$ be a homeomorphism. Choose $a \in S^2$ and put $b = f(a) \in S^3$. Then f gives a homeomorphism $S^2 \setminus \{a\} \simeq S^3 \setminus \{b\}$, but $S^2 \setminus \{a\}$ is homeomorphic to $\mathbb{R}^2$ and $S^3 \setminus \{b\}$ is homeomorphic to $\mathbb{R}^3$, so $\mathbb{R}^2$ is homeomorphic to $\mathbb{R}^3$, contradicting the proposition. □

21. Higher homotopy and homology

The results of the previous section are restricted to spaces of fairly low dimension, essentially because $\pi_1 S^n$ is trivial (and thus independent of n) for $n > 1$. To get similar results for higher-dimensional spaces, we need to define higher homotopy groups $\pi_n X$ for $n > 1$. Here we will just give the basic definitions and mention some important properties without proof.

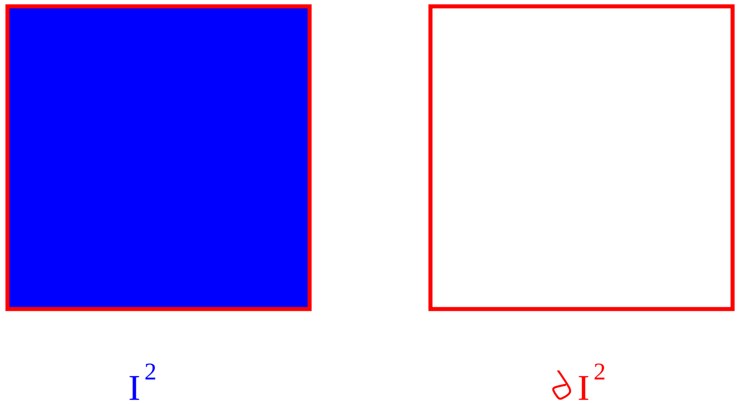
Recall that

$$I^n = [0,1]^n = \{t \in \mathbb{R}^n \ : \ 0 \leq t_i \leq 1 \text{ for all } i\}.$$

We write

$$\partial I^n = \{t \in I^n \ : \ t_i \in \{0,1\} \text{ for some } i\},$$

and call this the *boundary* of I^n.



Definition 21.1. Let X be a based space, and n a natural number. An *n-loop* in X is a map $u\colon I^n \longrightarrow X$ such that $u(t) = x_0$ for all $t \in \partial I^n$. We say that two n-loops u and v are *homotopic relative to the boundary* if there is a map $h\colon I \times I^n \longrightarrow X$ such that

(a) $h(0,t) = u(t)$ for all $t \in I^n$.
(b) $h(1,t) = v(t)$ for all $t \in I^n$.
(c) $h(s,t) = x_0$ for all $s \in I$ and $t \in \partial I^n$.

We write $\pi_n X$ for the set of equivalence classes of n-loops under this equivalence relation. If u and v are n-loops then we define n-loops $\overline{u}$ and $u * v$ by

$$\overline{u}(t_1, \dots, t_n) = u(1 - t_1, t_2, \dots, t_n)$$

$$(u * v)(t_1, \dots, t_n) = \begin{cases} u(2t_1, t_2, \dots, t_n) & \text{if } 0 \leq t_1 \leq 1/2 \\ v(2t_1 - 1, t_2, \dots, t_n) & \text{if } 1/2 \leq t_1 \leq 1 \end{cases}$$

These operations make $\pi_n X$ into a group. A based map $f\colon X \longrightarrow Y$ gives a homomorphism $f_*\colon \pi_n X \longrightarrow \pi_n Y$, and if g is based-homotopic to f then $g_* = f_*$.

Some important properties are as follows:

Proposition 21.2. *$\pi_n X$ is always abelian when $n > 1$.* □

Proposition 21.3. *We have $\pi_n S^n \simeq \mathbb{Z}$, and $\pi_n S^m$ is trivial when $n < m$.*

This is enough to prove that when $n \neq m$, the spheres S^n and S^m are not homotopy equivalent, and the spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ are not homeomorphic.

Unfortunately, it is very difficult to calculate $\pi_n X$ in general. In fact, not all the groups $\pi_n S^m$ are known when $n > m$, although many cases have been calculated. For example

$$\pi_5 S^4 \simeq C_2$$
$$\pi_6 S^4 \simeq C_2$$
$$\pi_7 S^4 \simeq C_{12} \times C_\infty$$
$$\pi_8 S^4 \simeq C_2 \times C_2$$
$$\pi_9 S^4 \simeq C_2 \times C_2$$
$$\pi_{10} S^4 \simeq C_{24} \times C_3.$$

However, using a slightly different approach one can define the *homology groups* $H_n X$ of a space X. The definition is more complicated, but once the framework is in place the calculations are much easier. For example, for $n, m > 0$ we just have

$$H_n S^m = \begin{cases} \mathbb{Z} & \text{if } n = m \\ 0 & \text{otherwise.} \end{cases}$$

These homology groups are in fact the main tool used in the algebraic topology of higher-dimensional spaces.

22. Other applications of the fundamental group

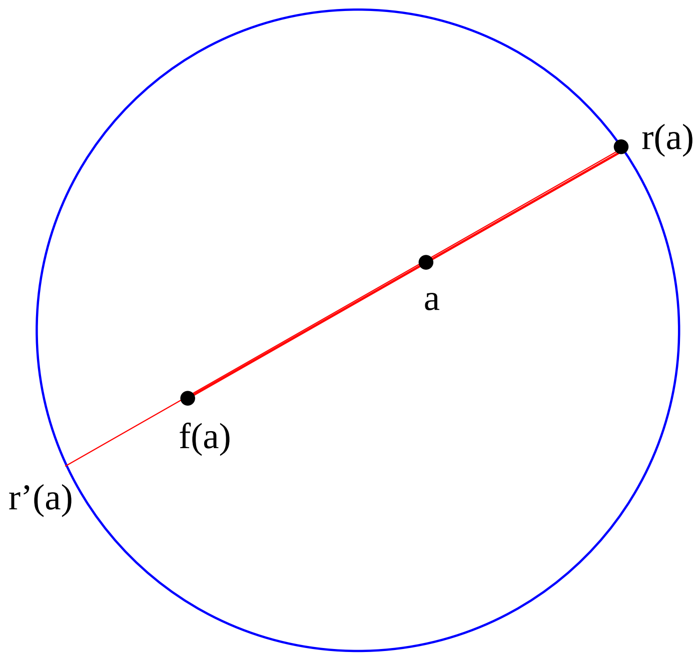
Theorem 22.1. *There is no continuous map $r\colon B^2 \longrightarrow S^1$ such that $r(a) = a$ for all $a \in S^1$.*

Proof. We suppose that r exists and deduce a contradiction. We use 1 as the basepoint in S^1 and in B^2. Let $j\colon S^1 \longrightarrow B^2$ be the inclusion map, defined by $j(a) = a$ for all $a \in S^1$. Thus rj is a map from S^1 to S^1, and by assumption we have $rj(a) = a$ for all $a \in S^1$, in other words rj is just the identity map. It follows that $r_*j_* = (rj)_*\colon \pi_1 S^1 \longrightarrow \pi_1 S^1$ is the identity map. In particular, if a_1 is the usual generator of $\pi_1 S^1$ then $r_*j_*(a_1) = a_1$.

On the other hand, we have $j_*(a_1) \in \pi_1 B^2$ and B_2 is contractible so $\pi_1 B^2$ is trivial so $j_*(a_1) = e$ so $r_*j_*(a_1) = e \neq a_1$. This contradiction completes the proof. □

Theorem 22.2 (The Brouwer Fixed Point Theorem)**.** *Let $f\colon B^2 \longrightarrow B^2$ be a continuous map. Then f has a fixed point, in other words there is a point $a \in B^2$ such that $f(a) = a$.*

Proof. Suppose not. Then for all $a \in B^2$, the point $f(a)$ is different from a, so we can draw a unique line from $f(a)$ through a and on out to infinity. Let $r(a)$ be the point where this line meets S^1.



This gives a function $r\colon B^2 \longrightarrow S^1$. One way to check that it is continuous is to find the formula. The vector from $f(a)$ to a is just $f(a) - a$, so the unit vector in the direction from $f(a)$ to a is $(f(a) - a)/\|f(a) - a\|$. We call this $g(a)$, and note that g gives a continuous map from B^2 to S^1. Thus, the rules $h(a) = \langle f(a), g(a)\rangle$ and $k(a) = \|f(a)\|^2$ define continuous maps $h, k\colon B^2 \longrightarrow \mathbb{R}$.

The point $r(a)$ is obtained by moving from $f(a)$ in the direction $g(a)$ until we reach S^1. Thus $r(a) = f(a) + tg(a)$ for some $t \geq 0$. To find t we note that

$$\begin{aligned} 1 &= \|r(a)\|^2 \\ &= \langle f(a) + tg(a), f(a) + tg(a)\rangle \\ &= \langle f(a), f(a)\rangle + 2t\langle f(a), g(a)\rangle + t^2\langle g(a), g(a)\rangle \\ &= k(a) + 2th(a) + t^2, \end{aligned}$$

so

$$t^2 + 2h(a)t + k(a) - 1 = 0.$$

This quadratic has two roots. We see geometrically that one of them is negative (corresponding to the point $r'(a)$ in the diagram), and one is positive (corresponding to $r(a)$). The positive root is

$$t = -\langle f(a), g(a)\rangle + \sqrt{h(a)^2 + 1 - k(a)}$$

which is again a continuous function of a. Thus, the function

$$r(a) = f(a) + (-\langle f(a), g(a)\rangle + \sqrt{h(a)^2 + 1 - k(a)})g(a)$$

is also continuous, as claimed. (There are also more geometric ways to prove continuity.)

Clearly if $a \in S^1$ then the ray from $f(a)$ to a meets S^1 at a itself, so $r(a) = a$. This contradicts Theorem 22.1. $\square$