Simply Connected Spaces

John M. Lee

1. Doughnuts, Coffee Cups, and Spheres

If you heard anything about topology before you started studying it formally, it is likely that one of the first examples you saw was coffee cups and doughnuts: The surface of a doughnut is topologically equivalent (or homeomorphic) to the surface of a coffee cup with one handle, but not to a sphere. This example is used to illustrate that homeomorphisms can bend, stretch, and shrink but not tear or collapse.

Using the topological tools we have developed so far, you should be able to see how to prove that coffee-cup and doughnut surfaces are homeomorphic. One reasonable model of a "doughnut surface" is the surface of revolution obtained by revolving a circle around the z-axis; another is the product space $S^1 \times S^1$, called the ***torus***. As we have already seen in class, these two spaces are both homeomorphic to a certain quotient space of a square, and thus are homeomorphic to each other. Once you have decided on a specific subset of $\mathbb{R}^3$ to use as a model of a "coffee-cup surface," it is a straightforward (though not necessarily easy!) matter to construct an explicit homeomorphism between it and the doughnut surface. (By virtue of the Closed Map Lemma, or Theorem 26.6 of Munkres, it is sufficient to construct a continuous bijection from one space to the other, because such a map is automatically a homeomorphism.)

But it is a much more difficult problem to show that a torus is *not* homeomorphic to a sphere. Intuitively, we expect that any bijective map from the sphere to the torus would somehow have to "rip a hole" in the sphere, thus violating continuity. But this intuition is far from being a proof. How do we know that a really clever mathematician won't come up with a new idea for a homeomorphism tomorrow? In order to prove that spaces are not homeomorphic to each other, we usually look for topological invariants—topological properties that must be preserved by homeomorphisms. If two spaces have different invariants, they cannot be homeomorphic.

So far in this course, we have seen two powerful topological invariants. The simplest one is *connectedness*: if two spaces are homeomorphic, then they are either both connected or both disconnected. Thus we know that $\mathbb{R} \setminus \{0\}$ is not homeomorphic to $\mathbb{R}$. A more subtle application of connectedness was used in Exercise 1 of §24 of Munkres to show that $\mathbb{R}$ is not homeomorphic to $\mathbb{R}^2$, and that no two of the intervals $(0, 1)$, $(0, 1]$, and $[0, 1]$ are homeomorphic to each other.

Another important topological invariant is *compactness*: a compact space cannot be homeomorphic to a non-compact one. Thus S^2 is not homeomorphic to $\mathbb{R}^2$. But none of these invariants are powerful enough to address the question of whether S^2 is homeomorphic to a torus. Nor are they powerful enough to tell us whether $\mathbb{R}^2$ is homeomorphic to the space $\mathbb{R}^2 \setminus \{0\}$, called the ***punctured plane***. In this note, we introduce a new topologically invariant property called *simple connectedness*, and show that it is powerful enough to distinguish between the sphere and the torus, the plane and the punctured plane, and a good many other pairs of spaces as well.

2. Path Homotopy

The intuition we are trying to capture is that a simply connected space is one that has no "holes," in a certain sense. Roughly speaking, we will detect "holes" by wrapping "strings" around them: a simply connected space is one in which any loop of string can be continuously shrunk to a point while remaining in the space.

Here are the definitions. Suppose X and Y are topological spaces, and let I denote the interval $[0,1] \subset \mathbb{R}$, called the ***unit interval***. If $f, g\colon X \to Y$ are two continuous maps, a ***homotopy from f to g*** is a continuous map $H\colon X \times I \to Y$ satisfying

$$\begin{aligned} H(x,0) &= f(x) \qquad \text{for all } x \in X; \\ H(x,1) &= g(x) \qquad \text{for all } x \in X. \end{aligned}$$

Given such a homotopy, for each $t \in I$ we get a map $H_t\colon X \to Y$ defined by $H_t(x) = H(x,t)$. Each H_t is continuous because it can be written as the composition $H \circ j_t$, where $j_t\colon X \to X \times I$ is the map $j(x) = (x,t)$. Thus we think of H as defining a one-parameter family of maps that vary continuously from $H_0 = f$ to $H_1 = g$ as t goes from 0 to 1.

In this note, we will be working only with a special type of homotopies of paths. Recall that a ***path*** in a topological space X is a continuous map $f\colon [a,b] \to X$, where $[a,b] \subset \mathbb{R}$ is a closed interval. The points $x = f(a)$ and $y = f(b)$ are called the ***starting point*** and ***ending point*** of f, respectively, and we say that f is a ***path from x to y***. As we saw in Chapter 2 of Munkres, a space X is said to be ***path-connected*** if for each $x, y \in X$, there is a path in X from x to y.

The next theorem is an analogue for path connectedness of the main theorem on connectedness.

Theorem 1. *Every continuous image of a path-connected space is path-connected.*

Proof: Suppose X is path-connected, and $G\colon X \to Y$ is a continuous map. Let $Z = G(X)$; we need to show that Z is path-connected. Given $x, y \in Z$, there are points $x_0, y_0 \in X$ such that $x = G(x_0)$ and $y = G(y_0)$. Because X is path-connected, there is a path $f\colon [a,b] \to X$ such that $f(a) = x_0$ and $f(b) = y_0$. Then $G \circ f$ is a path in Z from x to y. □

Suppose $f\colon [a,b] \to X$ and $g\colon [a,b] \to X$ are two paths in X, defined on the same domain $[a,b]$, and both with the same starting point x and ending point y. A ***path homotopy from f to g*** is a continuous map $H\colon [a,b] \times I \to X$ satisfying

$$\begin{aligned} H(s,0) &= f(s) \qquad \text{for all } s \in [a,b]; \\ H(s,1) &= g(s) \qquad \text{for all } s \in [a,b]; \\ H(a,t) &= x \qquad \text{for all } t \in I; \\ H(b,t) &= y \qquad \text{for all } t \in I. \end{aligned}$$

In other words, H is a homotopy from f to g, with the additional requirement that each path H_t in the homotopy starts at x and ends at y. A path homotopy is also sometimes called a ***fixed-endpoint homotopy***. It is useful to think of t as "time," and to visualize a path homotopy as a continuous deformation of the path from

f to g as time runs from $t = 0$ to $t = 1$. In this note, we will consistently use s to denote the parameter along a path (the "space variable"), and reserve t for the "time variable" in a homotopy.

If there exists a path homotopy from f to g, then we say that f and g are ***path-homotopic***.

Lemma 2. *Let X be a topological space, let x, y be points in X, and let $[a, b]$ be a fixed interval in $\mathbb{R}$. "Path-homotopic" is an equivalence relation on the set of all paths in X from x to y with domain $[a, b]$.*

Proof: See Exercise 1. □

The next lemma shows that continuous maps preserve path homotopy.

Lemma 3. *Suppose X and Y are topological spaces, $G\colon X \to Y$ is a continuous map, and $f, g\colon [a, b] \to X$ are paths from x to y. If f and g are path-homotopic, then so are $G \circ f$ and $G \circ g$.*

Proof: Suppose f and g are path-homotopic, and let $H\colon [a, b] \times I \to X$ be a path homotopy from f to g. It follows by composition that $G \circ H$ is continuous, and the hypotheses imply

$$\begin{aligned} G \circ H(s, 0) &= G(x) && \text{for all } s \in [a, b]; \\ G \circ H(s, 1) &= G(y) && \text{for all } s \in [a, b]; \\ G \circ H(a, t) &= G \circ f(t) && \text{for all } t \in I; \\ G \circ H(b, t) &= G \circ g(t) && \text{for all } t \in I. \end{aligned}$$

Thus $G \circ H$ is a path homotopy from $G \circ f$ to $G \circ g$. □

Here is the kind of path we will be mainly concerned with: a ***loop*** is a path whose starting and ending points are the same; that is, a continuous map $f\colon [a, b] \to X$ such that $f(a) = f(b)$. If $x = f(a) = f(b)$ is the common starting and ending point, we say the loop is ***based at x***. A loop is said to be ***null-homotopic*** if it is path-homotopic to the constant loop with the same base point (i.e., the loop $c\colon [a, b] \to X$ such that $c(s) = x$ for all $s \in [a, b]$).

A topological space is said to be ***simply connected*** if it is path-connected and every loop in the space is null-homotopic. A space that is not simply connected is said to be ***multiply connected***. The next theorem shows that simple connectedness (and therefore also multiple connectedness) is a topologically invariant property.

Theorem 4. *Suppose X and Y are homeomorphic topological spaces. Then X is simply connected if and only if Y is simply connected.*

Proof: Let $G\colon X \to Y$ be a homeomorphism, and suppose X is simply connected. It follows from Theorem 1 that Y is path-connected, so we just need to show that every loop in Y is null-homotopic. Let $f\colon [a, b] \to Y$ be a loop. Then $G^{-1} \circ f$ is a loop in X, which is path-homotopic to some constant loop c by hypothesis. It follows from Lemma 3 that $f = G \circ \left(G^{-1} \circ f\right)$ is path-homotopic to the constant loop $G \circ c$. The converse is proved in the same way, with the roles of G and G^{-1} reversed. □

3. Some Simply Connected Spaces

Next we will show that some familiar spaces are simply connected. A subset $K \subset \mathbb{R}^n$ is said to be ***convex*** if whenever x and y are two points in K, the entire line segment from x to y is contained in K. More precisely, this means that for each $t \in [0,1]$, the point $(1-t)x + ty = x + t(y-x)$ lies in K. Clearly $\mathbb{R}^n$ itself is convex. The next lemma gives several more examples of convex sets.

Lemma 5. *The following subsets of $\mathbb{R}^n$ are convex.*

(a) *Every open ball $B(a,\varepsilon)$.*
(b) *Every closed ball $\overline{B}(a,\varepsilon)$.*
(c) *Every product set of the form $K_1 \times \cdots \times K_n$, where $K_1, \ldots, K_n$ are convex subsets of $\mathbb{R}$.*

Proof: To prove that open balls are convex, let $B(a,\varepsilon) \subset \mathbb{R}^n$ be an open ball, and suppose x and y are points in $B(a,\varepsilon)$. Then for each $t \in [0,1]$, the triangle inequality gives

$$\begin{aligned} \|a - ((1-t)x + ty)\| &= \|(1-t)(a-x) + t(a-y)\| \\ &\le \|(1-t)(a-x)\| + \|t(a-y)\| \\ &= (1-t)\|a-x\| + t\|a-y\| \\ &< (1-t)\varepsilon + t\varepsilon = \varepsilon. \end{aligned}$$

This shows that the line segment from x to y is contained in $B(a,\varepsilon)$, and thus $B(a,\varepsilon)$ is convex. The same computation with "$<$" replaced by "$\le$" shows that closed balls are convex.

Finally, suppose $K_1, \ldots, K_n$ are convex subsets of $\mathbb{R}$, and let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be points in $K_1 \times \cdots \times K_n$. Then for each $t \in [0,1]$, the fact that each K_i is convex implies that

$$(1-t)x + ty = \big((1-t)x_1 + ty_1,\ \ldots,\ (1-t)x_n + ty_n\big) \in K_1 \times \cdots \times K_n.$$

Thus $K_1 \times \cdots \times K_n$ is convex. □

The next theorem expresses the most important property of convex sets for our purposes.

Theorem 6. *Suppose K is a convex subset of $\mathbb{R}^n$. If $f, g\colon [a,b] \to X$ are any two paths in K with the same domain and the same starting and ending points, then f and g are path-homotopic to each other.*

Proof: Let $f, g\colon [a,b] \to K$ be two paths in K from x to y. Define a map $H\colon [a,b] \times I \to K$ as follows:

$$H(s,t) = (1-t)f(s) + tg(s).$$

For each s, this map traces out the line segment from $f(s)$ to $g(s)$ as t goes from 0 to 1. The fact that K is convex guarantees that $H(s,t) \in K$ for all $(s,t) \in [a,b] \times I$. The component functions of H are polynomials in t and the component functions of f and g, so H is continuous. It follows immediately from the definition that $H(s,0) = f(s)$ and $H(s,1) = g(s)$ for all $s \in [a,b]$, and $H(a,t) = x$ and $H(b,t) = y$

for all $t \in I$; thus H is a path homotopy from f to g, called a ***straight-line homotopy***. □

Corollary 7. *Every convex subset of $\mathbb{R}^n$ is simply connected.*

Proof: Let K be a convex subset of $\mathbb{R}^n$. We showed earlier in the course that K is path-connected, and Theorem 6 implies that every loop in K is path-homotopic to a constant loop. □

For future use, we note also the following simple corollary of Theorem 6.

Corollary 8. *If X is a topological space that is homeomorphic to a convex subset of $\mathbb{R}^n$, then any two paths in X with the same domain and the same starting and ending points are path-homotopic to each other.*

Proof: Given X as in the hypothesis, let $G\colon X \to K$ be a homeomorphism to a convex subset $K \subset \mathbb{R}^n$. If $f, g\colon [a,b] \to X$ are two paths as in the statement of the lemma, then $G \circ f$ and $G \circ g$ are path-homotopic by the argument in the preceding paragraph, and therefore $f = G^{-1} \circ (G \circ f)$ is path-homotopic to $g = G^{-1} \circ (G \circ f)$ by Lemma 3. □

Next we consider spheres. Recall that for any $n \geq 0$, the ***unit n-sphere*** is the space $S^n \subset \mathbb{R}^{n+1}$ defined by

$$S^n = \big\{ x \in \mathbb{R}^{n+1} \,\big|\, \|x\| = 1 \big\},$$

with the subspace topology. Thus S^0 is the two-element discrete space $\{\pm 1\}$, S^1 is the unit circle in the plane, and S^2 is the familiar spherical surface in $\mathbb{R}^3$, often just called ***the sphere***. The points $N = (0, \ldots, 0, 1)$ and $S = (0, \ldots, 0, -1)$ in S^n are called the ***north pole*** and ***south pole***, respectively.

The following lemma is the key to proving that some spheres are simply connected.

Lemma 9. *For any $n \geq 0$, S^n with the north pole or south pole removed is homeomorphic to $\mathbb{R}^n$.*

Proof: First consider the north pole. Define a continuous map $f\colon S^n \smallsetminus \{N\} \to \mathbb{R}^n$ as follows:

$$f(x_1, \ldots, x_{n+1}) = \frac{1}{1 - x_{n+1}}(x_1, \ldots, x_n).$$

This map is called ***stereographic projection***. It can be described geometrically as follows: Given a point $x = (x_1, \ldots, x_{n+1}) \in S^n \smallsetminus \{N\}$, stereographic projection sets $f(x) = u$, where $(u, 0)$ is the point where the line through N and x meets the subspace $\mathbb{R}^n \times \{0\}$.

Now define another continuous map $g\colon \mathbb{R}^n \to S^n \smallsetminus \{N\}$ by

$$g(u_1, \ldots, u_n) = \frac{1}{\|u\|^2 + 1}\big(2u_1, \ldots, 2u_n, \|u\|^2 - 1\big).$$

A straightforward computation (using the fact that $\|x\|^2 = 1$) shows that $g \circ f(x) = x$ for all $x \in S^n$, and $f \circ g(u) = u$ for all $u \in \mathbb{R}^n$; thus g is a continuous two-sided inverse for f, which shows that f is a homeomorphism.

Finally, note that the reflection map $r\colon S^n \to S^n$ given by $r(x_1, \ldots, x_n, x_{n+1}) = (x_1, \ldots, x_n, -x_{n+1})$ restricts to a homeomorphism from $S^n \smallsetminus \{N\}$ to $S^n \smallsetminus \{S\}$, so $S^n \smallsetminus \{S\}$ is also homeomorphic to $\mathbb{R}^n$. □

We can now begin to address the question of which spheres are simply connected. First, S^0 is the discrete space $\{\pm 1\}$, which is not even connected. The circle S^1 is path-connected, but we will skip over the question of whether it is simply connected for now. It turns out that for $n \geq 2$, each sphere S^n is simply connected. The key is the previous lemma. It implies, for example, that any loop in S^n that does not pass through the north pole can be considered as a loop in $S^n \smallsetminus \{N\}$, which is homeomorphic to $\mathbb{R}^n$ and therefore simply connected. In fact, it can be shown that if x is *any* point in S^n, there is a rotation in $\mathbb{R}^{n+1}$ that takes x to N and restricts to a homeomorphism from $S^n \smallsetminus \{x\}$ to $S^n \smallsetminus \{N\}$, so $S^n \smallsetminus \{x\}$ is also homeomorphic to $\mathbb{R}^n$. Therefore, any loop in S^n that misses at least one point is null-homotopic. Intuitively, one might expect that every loop in the sphere would have to miss at least one point. Unfortunately, this is not the case. In 1890, the Italian mathematician Giuseppe Peano discovered a way to construct a continuous surjective map $f\colon I \to I \times I$, called a ***space-filling curve***. (See Theorem 44.1 in Munkres for an example of such a curve.) Since $I \times I$ is homeomorphic to the closed unit disk $D = \{(x, y) \mid x^2 + y^2 \leq 1\}$ (can you prove it?), and there is a surjective continuous map $p\colon D \to S^2$ (see Example 4 on page 139 of Munkres), we can compose these three maps to obtain a continuous surjective path in S^2:

$$I \xrightarrow{f} I \times I \xrightarrow{\approx} D \xrightarrow{p} S^2.$$

However, as the proof of the next theorem shows, when $n \geq 2$, every loop in S^n is *path-homotopic* to a loop that misses a point, and this suffices to prove that S^n is simply connected.

Theorem 10. *For $n \geq 2$, the unit n-sphere S^n is simply connected.*

Proof: Suppose $n \geq 2$. Example 5 on page 156 of Munkres shows that S^n is path-connected, so we need only show that every loop in S^n is null-homotopic.

Suppose $f\colon [a, b] \to S^n$ is any loop. Let N and S be the north and south poles as above, and define open subsets $U, V \subset S^n$ by $U = S^n \smallsetminus \{N\}$, $V = S^n \smallsetminus \{S\}$. Then both U and V are homeomorphic to $\mathbb{R}^n$. For the time being, let us assume that f is based at a point other than N. We will show that f is path-homotopic to a loop whose image does not contain N.

Because f is continuous, the collection $\{f^{-1}(U), f^{-1}(V)\}$ is an open cover of $[a, b]$. By the Lebesgue number lemma (Lemma 27.5 in Munkres), there is a number $\delta > 0$ such that every subinterval of $[a, b]$ of width less than δ is contained in one of these sets. Choose an integer m such that $(b - a)/m < \delta$, and let $a_i = a + i(b - a)/m$ for $i = 0, \ldots, m$, so that $a = a_0 < a_1 < \cdots < a_m = b$ and $[a_{i-1}, a_i]$ has length $(b - a)/m < \delta$ for each i. It then follows that f maps each subinterval $[a_{i-1}, a_i]$ into U or V (or perhaps both). For each i, denote the restriction of f to the ith subinterval by $f_i = f|_{[a_{i-1}, a_i]}\colon [a_{i-1}, a_i] \to S^n$. By shrinking the codomain, we can consider each f_i as a map into either U or V, which is continuous by the characteristic property of the subspace topology.

Let J be the set of all indices $i \in \{1, \dots, m\}$ such that the image of f_i is contained in U, and let J' be the complementary set $\{1, \dots, m\} \setminus J$. For each $i \in J$, the path f_i does not pass through N, so we leave it alone. On the other hand, for each $i \in J'$, f_i is a path in $V = S^n \setminus \{S\}$. Note that $V \setminus \{N\} = S^n \setminus \{S, N\}$, which is homeomorphic to $\mathbb{R}^n \setminus \{0\}$. Since this space is path-connected (see Example 4 on page 156 of Munkres), there is a path $g_i \colon [a_{i-1}, a_i] \to V \setminus \{N\}$ that has the same starting and ending points as f_i. (This is the step that does not work when $n = 1$, because $\mathbb{R}^1 \setminus \{0\}$ is not connected.) Because V is homeomorphic to $\mathbb{R}^n$, it follows from Corollary 8 that f_i and g_i are path-homotopic in V, and applying Lemma 3 with $G \colon V \to S^n$ equal to the inclusion map shows that f_i and g_i are path-homotopic in S^n. Let $H_i \colon [a_{i-1}, a_i] \times I \to S^n$ be a path homotopy from f_i to g_i.

Now define a new loop $g \colon [a, b] \to S^n$ by

$$g(s) = \begin{cases} f_i(s), & s \in [a_{i-i}, a_i],\ i \in J, \\ g_i(s), & s \in [a_{i-i}, a_i],\ i \in J'. \end{cases}$$

This loop is continuous by the pasting lemma, and it has the same base point as f. By construction, the image of g does not contain N. Then define $H \colon [a, b] \times I \to S^n$ by

$$H(s, t) = \begin{cases} f_i(s), & s \in [a_{i-i}, a_i],\ i \in J, \\ H_i(s, t), & s \in [a_{i-i}, a_i],\ i \in J'. \end{cases}$$

Again, H is continuous by the pasting lemma, and a straightforward verification shows that it is a path homotopy from f to g. Thus f is path-homotopic to a path g whose image lies entirely in U.

Finally, the fact that U is simply connected guarantees that g is path-homotopic in U (and hence also in S^n by Lemma 3) to a constant path. Because path homotopy is transitive, it follows that f is null-homotopic.

The only case we have not dealt with is the case in which f is based at the north pole N. In that case, we just carry out the same argument with the roles of N and S reversed. □

4. The Circle

So far, we have not been able to prove that any particular space is multiply connected. To do so, we have to find a loop for which we can prove that there exists no path homotopy from the given loop to a constant loop. Like many other nonexistence results, this requires some new ideas.

As we noted in the proof of Theorem 10, the proof of that theorem breaks down when $n = 1$. This is not just a flaw in our technique: in fact, as we will see in this section, S^1 is multiply connected.

The basic idea of what we are about to do is simple. To analyze a loop $f \colon [a, b] \to S^1$, we will construct a function that measures the angle from the positive x-axis to the ray through $f(s)$, and compute the net change in angle as s goes from a to b. Because a loop starts and ends at the same point, this net change must be

an integral multiple of 2π. As we will see below, the only loops in S^1 that are null-homotopic are those whose net change in angle is zero.

To begin studying angles in S^1, consider the map $p\colon \mathbb{R} \to S^1$ defined by

$$p(u) = (\cos u, \sin u). \tag{1}$$

We have seen this map earlier in the course; it is a quotient map that "winds" the real line infinitely many times around the circle. For each $u \in \mathbb{R}$, the point $p(u)$ is a point on the circle, and the number u is one of the possible values of the angle measured counterclockwise from the positive x-axis to the ray that starts at the origin and passes through $p(u) \in S^1$. It follows from standard properties of trigonometric functions that $p(u) = p(u')$ if and only if $u - u'$ is an integral multiple of 2π.

If W is a subset of S^1, an ***angle function for W*** is a continuous map $\theta\colon W \to \mathbb{R}$ such that $p \circ \theta = i_W$, or equivalently $(x, y) = (\cos\theta(x, y), \sin\theta(x, y))$ for all $(x, y) \in W$. The next theorem shows that it is impossible to find an angle function defined on all of S^1.

Theorem 11. *There does not exist a continuous function $\theta\colon S^1 \to \mathbb{R}$ such that $p \circ \theta = i_{S^1}$.*

Proof: See Exercise 2. □

However, for some proper subsets of S^1, angle functions are easy to construct. For example, define four open subsets $X_\pm, Y_\pm \subset S^1$ by

$$\begin{aligned} X_+ &= \{(x, y) \in S^1 \mid x > 0\}, \quad Y_+ = \{(x, y) \in S^1 \mid y > 0\}, \\ X_- &= \{(x, y) \in S^1 \mid x < 0\}, \quad Y_- = \{(x, y) \in S^1 \mid y < 0\}. \end{aligned} \tag{2}$$

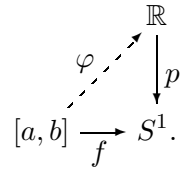
On each of these subsets, we can define a continuous angle function as follows:

$$\begin{aligned} \theta(x, y) &= \arctan\frac{y}{x}, && (x, y) \in X_+; \\ \theta(x, y) &= \operatorname{arccot}\frac{x}{y}, && (x, y) \in Y_+; \\ \theta(x, y) &= \pi + \arctan\frac{y}{x}, && (x, y) \in X_-; \\ \theta(x, y) &= \pi + \operatorname{arccot}\frac{x}{y}, && (x, y) \in Y_-. \end{aligned} \tag{3}$$

(Recall that the arctangent and arccotangent functions are defined as follows: for any real number r, $\arctan r$ is the unique angle $\theta \in (\pi/2, \pi/2)$ such that $\tan\theta = r$, and $\operatorname{arccot} r$ is the unique angle $\theta \in (0, \pi)$ such that $\cot\theta = r$. It is proved in advanced calculus courses that both functions are continuous.)

Now suppose $f\colon [a, b] \to S^1$ is a path. An ***angle function for f*** is a continuous function $\varphi\colon [a, b] \to \mathbb{R}$ such that $p \circ \varphi = f$, or equivalently $f(s) = (\cos\varphi(s), \sin\varphi(s))$ for all $s \in [a, b]$. (Note the distinction between an angle function for a *path* and an angle function for a *subset of* S^1, defined earlier.) An angle function for f is also

often called a ***lift of f***, because it makes the following diagram commute:



(A diagram like this one is said to ***commute***, or to be ***commutative***, if the maps between two spaces obtained by following arrows around either side of the diagram are equal. In this case, to say the diagram commutes is to say that $p \circ \varphi = f$.)

If there were an angle function for all of S^1, say $\theta\colon S^1 \to \mathbb{R}$, then our task would be easy: for any path $f\colon [a,b] \to S^1$ we could just define $\varphi = \theta \circ f$; then the fact that $p \circ \theta = i_{S^1}$ would guarantee that $p \circ \varphi = p \circ \theta \circ f = i_{S^1} \circ f = f$, so φ would be an angle function for f. But, as we saw above, this is impossible. Nonetheless, as the next theorem shows, it is always possible to construct an angle function for a path.

Theorem 12. *Every path in S^1 admits an angle function.*

Proof: Let $f\colon [a,b] \to S^1$ be a path. Define $X_\pm, Y_\pm \subset S^1$ as in (2) above. Then the four subsets $\{f^{-1}(X_+), f^{-1}(Y_+), f^{-1}(X_-), f^{-1}(Y_-)\}$ form an open cover of $[a,b]$. As in the proof of Theorem 10, let δ be a Lebesgue number for this cover, let m be an integer large enough that $(b-a)/m < \delta$, and let $a_i = a + (b-a)/m$ for $i = 0, \ldots, m$. Then for each i, f maps $[a_{i-1}, a_i]$ into one of the four sets $X_\pm, Y_\pm \subset S^1$.

We will prove by induction on k that for each $k = 1, \ldots, m$, there exists an angle function for f on the subinterval $[a_0, a_k]$. To begin, we can define an angle function φ_1 for f on the subinterval $[a_0, a_1]$ by setting $\varphi_1 = \theta \circ f$, where θ is given by one of the four formulas in (3), depending on which of the four subsets contains the image $f\big([a_0, a_1]\big)$. Suppose now that $k \geq 1$, and assume that we have defined an angle function $\varphi_k\colon [a_0, a_k] \to \mathbb{R}$ for the restriction of f to $[a_0, a_k]$. If $k = m$, we are done. If not, there is an angle function $\widetilde{\varphi}\colon [a_k, a_{k+1}] \to \mathbb{R}$, again defined by $\widetilde{\varphi} = \theta \circ f$ where θ is given by one of the formulas in (3). The numbers $\varphi_k(a_k)$ and $\widetilde{\varphi}(a_k)$ satisfy $p(\varphi_k(a_k)) = f(a_k) = p(\widetilde{\varphi}(a_k))$, so there is an integer N such that $\varphi_k(a_k) = \widetilde{\varphi}(a_k) + 2\pi N$. Define a new function $\varphi_{k+1}\colon [a_0, a_{k+1}] \to \mathbb{R}$ by

$$\varphi_{k+1}(s) = \begin{cases} \varphi_k(s), & s \in [a_0, a_k] \\ \widetilde{\varphi}(s) + 2\pi N, & s \in [a_k, a_{k+1}]. \end{cases}$$

These two definitions agree at the point a_k where they overlap, so φ_{k+1} is continuous by the pasting lemma. It is an angle function for f on $[a_0, a_{k+1}]$ because $p(\varphi_k(s)) = f(s)$ for $s \in [a_0, a_k]$ and $p(\widetilde{\varphi}(s) + 2\pi N) = p(\widetilde{\varphi}(s)) = f(s)$ for $s \in [a_k, a_{k+1}]$. This completes the induction. □

An angle function for a path is not unique, because we can always add an integral multiple of 2π and get a different angle function for the same path. But the next theorem shows that this is the only leeway.

Theorem 13. *Let $f\colon [a,b] \to S^1$ be a path. Any two angle functions for f differ by a constant integral multiple of 2π, and if two such angle functions agree for some $s_0 \in [a,b]$, then they agree for all $s \in [a,b]$.*

Proof: To prove the first statement, suppose φ_1 and φ_2 are both angle functions for the same path f. Define a map $g\colon [a,b] \to \mathbb{R}$ by $g(s) = (\varphi_1(s) - \varphi_2(s))/(2\pi)$. Since $p(\varphi_1(s)) = f(s) = p(\varphi_2(s))$ for each s, it follows that g is integer-valued. We proved in class that any integer-valued continuous function on a connected domain is constant, so g is a constant function, which implies that φ_1 and φ_2 differ by a constant integral multiple of 2π. The second statement then follows from the first, because if two angle functions agree at one point, then their constant difference must be zero. □

Now suppose $f\colon [a,b] \to S^1$ is a loop based at a point $(x_0, y_0) \in S^1$, and let $\varphi\colon [a,b] \to \mathbb{R}$ be an angle function for it. Because $p(\varphi(a)) = f(a) = f(b) = p(\varphi(b))$, it follows that $\varphi(a)$ and $\varphi(b)$ differ by an integral multiple of 2π. We define the ***winding number of f*** (also called the ***degree of f***) to be the integer

$$N(f) = \frac{1}{2\pi}(\varphi(b) - \varphi(a)).$$

Intuitively, $N(f)$ represents the net number of times $f(s)$ "winds around the circle," with a positive winding number indicating that it winds counterclockwise, and a negative one indicating clockwise.

Lemma 14. *The winding number of a loop is well defined, independently of the choice of angle function used to compute it.*

Proof: This is an immediate consequence of the fact that two different angle functions for the same loop must differ by a constant. □

Computing the winding number of a specific loop is usually just a matter of writing down an appropriate angle function, and subtracting its ending value from its starting value. Here are some examples.

Example 15. [Winding Numbers]

(a) Every constant loop in S^1 has winding number zero, because every angle function for it is constant.

(b) Let n be any integer, and define a loop $f_n\colon [0, 2\pi] \to S^1$ by

$$f_n(s) = (\cos ns, \sin ns).$$

Then $\varphi(s) = ns$ is an angle function for f_n, so its winding number is $N(f_n) = (\varphi(2\pi) - \varphi(0))/(2\pi) = n$. This shows that for every integer, there is a loop in S^1 with that winding number.

The next lemma shows that any two loops that stay sufficiently close to each other have the same winding number.

Lemma 16. *Suppose $f, g\colon [a,b] \to S^1$ are two loops with the same base point. If $|g(s) - f(s)| < 2$ for all $s \in [a,b]$, then $N(f) = N(g)$.*

Proof: Let $\varphi_1\colon [a,b] \to \mathbb{R}$ and $\varphi_2\colon [a,b] \to \mathbb{R}$ be angle functions for f and g, respectively. Since $p(\varphi_1(a)) = f(a) = g(a) = p(\varphi_2(a))$, the two numbers $\varphi_1(a)$ and $\varphi_2(a)$ differ by $2\pi N$ for some $N \in \mathbb{Z}$. After replacing φ_2 by $\varphi_2 + 2\pi N$, we can assume that $\varphi_1(a) = \varphi_2(a)$.

Let $n = N(f)$ and $m = N(g)$, and assume for the sake of contradiction that $n \neq m$. Without loss of generality, we can assume $m > n$. This means that $\varphi_1(b) = \varphi_1(a) + 2\pi n$ and $\varphi_2(b) = \varphi_2(a) + 2\pi m$. Because $\varphi_1(a) = \varphi_2(a)$, we conclude that $\varphi_2(b) - \varphi_1(b) = 2\pi(m-n)$. Since $m-n$ is a positive integer, we have $m-n \geq 1$, and thus

$$\varphi_2(b) - \varphi_1(b) \geq 2\pi.$$

Now the function $g\colon [a,b] \to \mathbb{R}$ defined by $g(s) = \varphi_2(s) - \varphi_1(s)$ is continuous, and satisfies $g(a) = 0$ and $g(b) \geq 2\pi$. Thus by the intermediate value theorem, there is some $s_0 \in [a,b]$ such that $g(s_0) = \pi$. But this means $\varphi_2(s_0) = \varphi_1(s_0) + \pi$, and so

$$\begin{aligned} g(s_0) &= p(\varphi_2(s_0)) = (\cos\varphi_2(s_0), \sin\varphi_2(s_0)) \\ &= (\cos(\varphi_1(s_0)+\pi), \sin(\varphi_1(s_0)+\pi)) = (-\cos\varphi_1(s_0), -\sin\varphi_1(s_0)) = -f(s_0). \end{aligned}$$

From this it follows that

$$|g(s_0) - f(s_0)| = |2g(s_0)| = 2,$$

contradicting the hypothesis. □

Here is our main theorem about loops in the circle.

Theorem 17. *Let $f, g\colon [a,b] \to S^1$ be loops with the same base point. Then f and g are path-homotopic if and only if they have the same winding number.*

Proof: Suppose first that f and g have the same winding number. As in the preceding proof, there are angle functions $\varphi_1, \varphi_2\colon [a,b] \to S^1$ for f and g, respectively, such that $\varphi_1(a) = \varphi_2(a)$. The assumption that f and g have the same winding number then implies that $\varphi_1(b) = \varphi_2(b)$. By Theorem 6, the two paths φ_1 and φ_2 are path-homotopic in $\mathbb{R}$. Since $f = p \circ \varphi_1$ and $g = p \circ \varphi_2$, it follows from Lemma 3 that f and g are path-homotopic.

Conversely, suppose f and g are path-homotopic, and let $H\colon [a,b] \times I \to S^1$ be a path homotopy from f to g. Because H is a continuous function from a compact metric space to another metric space, it is uniformly continuous by Theorem 27.6 in Munkres. Applying the definition of uniform continuity with $\varepsilon = 2$, we conclude that there exists $\delta > 0$ such that $\|H(s,t) - H(s',t')\| < 2$ whenever $\|(s,t) - (s',t')\| < \delta$. Let m be an integer large enough that $1/m < \delta$, and for $i = 0, \ldots, m$, let $H_i\colon [a,b] \to S^1$ be the path $H_i(s) = H(s, i/m)$. For any $i \in \{1, \ldots, m\}$ and any $s \in [a,b]$, we have $\|(s, i/m) - (s, (i-1)/m)\| = 1/m < \delta$, and thus $\|H_i(s) - H_{i-1}(s)\| < 2$. It follows from Lemma 16 that H_{i-1} and H_i have the same winding number. Therefore, $N(f) = N(H_0) = N(H_1) = \cdots = N(H_m) = N(g)$, so f and g have the same winding number. □

Theorem 18. *S^1 is not simply connected.*

Proof: The loop $f_1\colon [0, 2\pi] \to S^1$ given by $f_1(s) = (\cos s, \sin s)$ has winding number 1 (see Example 15), while any constant loop has winding number zero. Thus f_1 is not path-homotopic to a constant loop. □

5. Retractions

We now have exactly one space that we know to be multiply connected: the unit circle. Can we find others? Of course, spaces that are homeomorphic to the unit circle—such as ellipses, squares, or other circles—are also multiply connected. But can we find more interesting examples?

Unlike connectedness and compactness, neither simple connectedness nor multiple connectedness is preserved by arbitrary continuous maps (see Exercise 3). However, if we restrict our attention to a particular kind of map, then we can say something useful. Recall that a ***retraction*** is a continuous map $r\colon X \to A$, where X is a topological space and A is a subspace of X, such that $r(a) = a$ for all $a \in A$ (see Exercise 2 on page 144 of Munkres). If $A \subset X$ and there exists a retraction from X to A, then we say A is a ***retract*** of X.

Here is the tool that will enable us to identify other multiply connected spaces.

Theorem 19. *A retract of a simply connected space is simply connected.*

Proof: Suppose X is a simply connected space, $A \subset X$ is a subspace, and $r\colon A \to X$ is a retraction. Let $j\colon A \to X$ denote the inclusion map. It follows immediately from the definition of a retraction that $r \circ j = i_A$, the identity map of A.

Suppose $f\colon [a,b] \to A$ is a loop in A based at $x \in A$. Then $j \circ f\colon [a,b] \to X$ is a loop in X, also based at x. Because X is simply connected, $j \circ f$ is path-homotopic in X to the constant loop $c\colon [a,b] \to X$ given by $c(s) \equiv x$.

It follows from Lemma 3 that $r \circ (j \circ f)$ is path-homotopic in A to the constant loop $r \circ c$. However, the fact that r is a retraction implies that

$$r \circ (j \circ f) = (r \circ j) \circ f = i_A \circ f = f,$$

and $r \circ c$ is a constant loop because $r \circ c(s) = r(x) = x$ for all $s \in [a,b]$. Thus f is null-homotopic, which shows that A is simply connected. □

This theorem says that under the assumption that A is a retract of X, if X is simply connected then A is simply connected. But it is actually the contrapositive of this last statement that is most useful, as a way to prove a space is *not* simply connected. For ease of reference, we state this as a corollary.

Corollary 20. *If X is a topological space that retracts onto a multiply connected subspace, then X is multiply connected.* □

Here are some typical applications of this theorem; additional applications are described in the exercises. Recall the two homeomorphism problems that we introduced in the introduction:

- Is the torus homeomorphic to the sphere?
- Is the punctured plane homeomorphic to the plane?

Now we have the tools to answer both questions.

Theorem 21. *The punctured plane is multiply connected.*

Proof: Consider the map $r\colon \mathbb{R}^2 \setminus \{0\} \to S^1$ defined by

$$r(x,y) = \left(\frac{x}{\sqrt{x^2+y^2}},\ \frac{y}{\sqrt{x^2+y^2}} \right).$$

This is continuous because its component functions are continuous, and it satisfies $r(x,y) = (x,y)$ for $(x,y) \in S^1$; thus r is a retraction. Since S^1 is multiply connected, so is $\mathbb{R}^2 \setminus \{0\}$. □

Corollary 22. *The punctured plane is not homeomorphic to* $\mathbb{R}^2$.

Proof: The punctured plane is multiply connected, but $\mathbb{R}^2$ is not. □

Theorem 23. *The torus is multiply connected.*

Proof: Define a map $r\colon S^1 \times S^1 \to S^1 \times \{(1,0)\}$ by $r((x_1,y_1),(x_2,y_2)) = ((x_1,y_1),(1,0))$. It is easy to check that r is a retraction onto the subspace $S^1 \times \{(1,0)\}$, which is homeomorphic to S^1. Since S^1 is multiply connected, so is $S^1 \times S^1$. □

Corollary 24. *The torus is not homeomorphic to the sphere.* □

We end this note by pointing out that the idea of simple connectedness can be carried much further than we have done here. For example, we still do not have the tools to show that the punctured plane is not homeomorphic to a twice-punctured plane, or that a torus is not homeomorphic to a two-holed doughnut surface. The problem is that simple connectedness is a qualitative property; in order to answer questions like these, one must develop a way of measuring *quantitatively* how badly a space fails to be simply connected.

This can be done by moving beyond topological invariants that are simple yes/no questions, and developing ones that are more quantitative. The simplest such invariant is called the *fundamental group of a space*. It is a *group* in the algebraic sense (i.e., a set with an associative multiplication operator, in which there is a multiplicative identity and each element has a multiplicative inverse) attached to each topological space. The elements of the group are equivalence classes of loops with a fixed base point modulo path homotopy, and two loops (or, more precisely, two equivalence classes) are multiplied by first following one loop and then the other.

With some work, one can show that set of equivalence classes satisfies all of the required properties to form a group, and that homeomorphic spaces have fundamental groups that are *isomorphic* (meaning there is a bijective correspondence between them that preserves multiplication, identities, and inverses). This is the starting point for the vast branch of topology called *algebraic topology*. Chapters 9–14 of Munkres give a good introduction to the fundamental group; for more advanced topics in algebraic topology, I recommend Munkres's *Elements of Algebraic Topology* or Allen Hatcher's *Algebraic Topology*, or you might try my book *Introduction to Topological Manifolds*.

Exercises

1. Prove Lemma 2 (path homotopy is an equivalence relation). [Hint: for transitivity, given a homotopy H from f to g and another homotopy H' from g to h, define a new homotopy from f to h by first following H at double the normal speed, and then following H' at double the normal speed. Be sure to verify that your new homotopy is continuous.]

2. Prove Theorem 11 (there is no angle function on all of S^1). [Hint: Use the result of Exercise 2 in §24 of Munkres.]

3. Give explicit examples to show that continuous images of simply connected spaces need not be simply connected, and continuous images of multiply connected spaces need not be multiply connected. (No proofs necessary; just give a precise description of the spaces and the maps.)

4. Suppose A is any subset of $\mathbb{R}^2$ that contains some circle but not its center. Prove that A is multiply connected.

5. Suppose U is an open subset of $\mathbb{R}^2$, and V is obtained from U be deleting a nonempty finite set of points. Prove that U is multiply connected.

6. Suppose U is an open subset of $\mathbb{R}^2$, and W is obtained from U by deleting a countably infinite set of points. Is W necessarily path-connected? Is it necessarily multiply connected? [Hint: How many lines are there through a point in $\mathbb{R}^2$? How many circles are there centered at a point?]

7. Suppose $X_1, \ldots, X_n$ are nonempty topological spaces. Prove that the product space $X_1 \times \ldots \times X_n$ is simply connected if and only if all of the spaces $X_1, \ldots, X_n$ are simply connected.

8. Prove that for each $n \geq 1$, $\mathbb{R}^n \setminus \{0\}$ is homeomorphic to $S^{n-1} \times \mathbb{R}^+$, and conclude that $\mathbb{R}^n \setminus \{0\}$ is simply connected when $n \geq 3$.

9. Prove that $\mathbb{R}^2$ is not homeomorphic to $\mathbb{R}^n$ if $n \geq 3$.

10. Which of the following spaces are homeomorphic to each other? Prove your answers correct. You may use any theorems or previous exercises from this note or from Chapters 1–3 of Munkres. [Hint: Begin by making a table showing which spaces are connected, which are compact, and which are simply connected. There are exactly seven different spaces here, up to homeomorphism.]

(a) $\mathbb{R}$
(b) $\mathbb{R} \setminus \{0\}$
(c) $\mathbb{R}^2$
(d) $\mathbb{R}^2 \setminus \{0\}$
(e) $\mathbb{R}^3$
(f) $S^1 \times \mathbb{R}$
(g) $S^1 \times S^1$
(h) $S^1 \times S^2$
(i) $S^2 \times S^2$
(j) $S^1 \setminus \{(0,1)\}$
(k) $S^1 \setminus \{(0,1),(0,-1)\}$
(l) $S^2 \setminus \{(0,0,1)\}$
(m) $S^2 \setminus \{(0,0,1),(0,0,-1)\}$