NOTES FOR MATH 4510, FALL 2010

DOMINGO TOLEDO

1. Metric Spaces

The following definition introduces the most central concept in the course. Think of the plane with its usual distance function as you read the definition.

Definition 1.1. A *metric space* (X, d) is a non-empty set X and a function $d : X \times X \to \mathbb{R}$ satisfying

(1) For all $x, y \in X$, $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
(2) For all $x, y \in X$, $d(x, y) = d(y, x)$.
(3) For all $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$ (called the *triangle inequality*).

The function d is called the *metric*, it is also called the *distance function.*

1.1. **Examples of metric spaces.**

We now give examples of metric spaces. In most of the examples the conditions (1) and (2) of Definition 1.1 are easy to verify, so we mention these conditions only if there is some difficulty in establishing them. The difficult point is usually to verify the triangle inequality, and this we do in some detail.

Example 1.1. Let $X = \mathbb{R}$ with the usual distance function $d(x, y) = |x - y|$.

The triangle inequality is easy to verify by looking at cases. First, it's clear if two of x, y, z are equal (and both sides of the triangle inequality are equal), so we may assume all are different, and *we keep this assumption in all subsequent examples.* Let's assume $x < z$ (the case $z < x$ will be similar. Then there are 3 possibilities: $y < x < z$, $x < y < z$, $x < z < y$. In the first case $d(x, z) < d(y, z)$ and in the third case $d(x, z) < d(x, y)$, so in both these cases we get the strict inequality $d(x, z) < d(x, y) + d(y, z)$. In the second case we get equality in the triangle inequality: $d(x, z) = d(x, y) + d(y, z)$. This proves the triangle inequality for (X, d). Moreover, it also proves the following: *Equality holds in the triangle inequality if and only if y is between x and z.*

Example 1.2. Let $X = \mathbb{R}^2$ with the usual distance function

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

Date: January 6, 2011.

where $x = (x_1, x_2)$ and $y = (y_1, y_2)$.

To verify the triangle inequality, write, as usual, $u \cdot v$ for the dot product of vectors $u = (u_1, u_2)$ and $v = (v_1, v_2)$ in $\mathbb{R}^2$ (thus $u \cdot v = u_1 v_1 + u_2 v_2$) and $|u|$ for the length $\sqrt{u \cdot u}$. Given 3 points $x, y, z \in \mathbb{R}^2$, let $u = x - y$ and $v = y - z$. Then $u + v = x - z$, so $d(x,z) = |u+v|, d(x,y) = |u|, d(y,z) = |v|$, therefore the triangle inequality is equivalent to

$$|u + v| \leq |u| + |v| \text{ for all } u, v \in \mathbb{R}^2.$$

squaring both sides this is equivalent to

$$|u + v|^2 \leq |u|^2 + 2|u||v| + |v|^2.$$

Using the properties of the dot product, we see that we want

$$|u + v|^2 = (u + v) \cdot (u + v) = u \cdot u + 2u \cdot v + v \cdot \leq u \cdot u + 2|u||v| + v \cdot v,$$

which is equivalent to

$$u \cdot v \leq |u||v|$$

which is half of the familiar Cauchy-Schwarz inequality $|u \cdot v| \leq |u||v|$. Moreover, we have equality in the triangle inequality if and only if $u \cdot v = |u||v|$, which holds (assuming, as we may, that u and v are both non-zero), if and only if u and v are positive multiples of each other. In terms of x, y, z this means that *$d(x,z) = d(x,y) + d(y,z)$ holds if and only if y is in the straight line segment joining x and z.*

Example 1.3. Let $X = \mathbb{R}^n$ with the usual distance function

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

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots y_n)$. The verifications are exactly as for the case $n = 2$ just discussed.

Example 1.4. Let $X = \mathbb{R}^n$ and $d(x,y) = |x_1 - y_1| + \cdots + |x_n - y_n|$. For $n = 2$ this is the usual distance we use when driving in a city laid out in rectangular coordinates like Salt Lake City.

The triangle inequality is easy to verify. We need

$$d(x,z) = \sum_{i=1}^{n} |x_i - z_i| \leq \sum_{i=1}^{n} |x_i - y_i| + \sum_{i=1}^{n} |y_i - z_i|,$$

which follows from the fact that, for each i, from the triangle inequality in $\mathbb{R}$, $|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|$. Moreover,*equality holds in the triangle inequality for d if and only if, for all i, we have $|x_i - z_i| = |x_i - y_i| + |y_i - z_i|$, which happens if and only if y_i lies between x_i and z_i for each $i = 1 \ldots n$.* Thus, given x and z, the set of all y for which $d(x,z) = d(x,y) + d(y,z)$ is a "box" given by these inequalities. See Figure 1.1 for $n = 2$. For any y in the shaded region we have $d(x,y) + d(y,z) = d(x,z)$. Thus there are many more possibilities for equality than in the case of Example 1.2 and Example 1.3 where equality occurs only on a line segment.

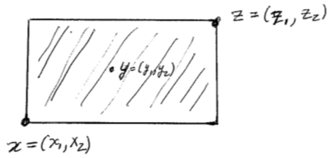

Figure 1.1. Equality Set for Taxicab Metric

Example 1.5. Let $X = \mathbb{R}^n$ and let $d(x,y) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}$.

To prove the triangle inequality $d(x,z) \leq d(x,y) + d(y,z)$, suppose that $d(x,z) = \max\{x_i - z_i|\} = |x_k - z_k|$ for some fixed k, $1 \leq k \leq n$, that is, the maximum is attained at k. Then $|x_k - z_k| \leq |x_k - y_k| + |y_k - z_k|$ and $|x_k - y_k| \leq d(x,y)$ and $|y_k - z_k| \leq d(y,z)$. So $d(x,z) \leq d(x,y) + d(y,z)$ follows. We will not discuss in detail the case of equality, but remark, just as in Example 1.4, there are in general many more possibilities than a line segment.

Example 1.6. Let $X = S^2 = \{x \in \mathbb{R}^3 : |x| = 1\}$, the unit sphere in $\mathbb{R}^3$. Let $d(x,y)$ be the length of the great-circle arc joining x and y. This is the way we measure distances on the surface of the earth. An explicit formula for $d(x,y)$ is easy to find: Let θ be the angle between the unit vectors x and y. The great circle arc connecting x and y is the part of the intersection with S^2 of the plane spanned by x and y lying between these two vectors, and the length of this arc is θ, see Figure 1.2. Thus $\cos\theta = x \cdot y$ (the usual dot product in $\mathbb{R}^3$) so $d(x,y) = \cos^{-1}(x \cdot y)$.

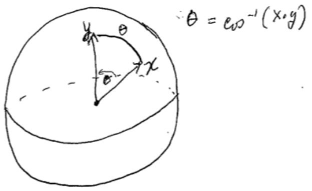

Figure 1.2. Spherical Distance

It is an exercise to verify the triangle inequality assuming another geometric inequality. Let $x_1, \ldots x_m$ be vectors in $\mathbb{R}^n$, and assume $m \leq n$. The *Gram matrix* of $x_1, \ldots, x_m$ is the m by m matrix A whose i,j-entry is $x_i \cdot x_j$. Note that A is a symmetric matrix, since $x_i \cdot x_j = x_j \cdot x_i$.

Theorem 1.1. *If A is the Gram matrix of $x_1, \ldots, x_m$ just defined, then* $\det(A) \geq 0$*, and det*$(A) = 0$ *if and only if the set* $\{x_1, \ldots, x_m\}$ *is linearly dependent.*

Proof. To avoid complicated notation, we only prove the theorem in the case that we need, namely $m = n = 3$, the proof being the same for all m, n. Let

$$A = \begin{pmatrix} x \cdot x & x \cdot y & x \cdot z \\ y \cdot x & y \cdot y & y \cdot z \\ z \cdot x & z \cdot y & z \cdot z \end{pmatrix}$$

be the Gram matrix of 3 vectors $x = (x_1, x_2, x_3), y = (y_1, y_2, y_3), z = (z_1, z_2, z_3) \in \mathbb{R}^3$, and let B be the matrix

$$B = \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix},$$

Clearly we have

$$\begin{pmatrix} x \cdot x & x \cdot y & x \cdot z \\ y \cdot x & y \cdot y & y \cdot z \\ z \cdot x & z \cdot y & z \cdot z \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix} \begin{pmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{pmatrix},$$

in other words, $A = ({}^tB)B$, where ${}^tB$ denotes the transpose matrix. Thus $\det(A) = \det({}^tB)\det(B) = det(B)^2 \geq 0$, and $det(A) = 0$ if and only if $\det(B) = 0$, which, by the definition of B, happens if and only if $\{x, y, z\}$ is linearly dependent.

□

Remark 1.1. Observe that in the case $m = 2$, that is, two vectors, say $x, y \in \mathbb{R}^m$, then Theorem 1.1 says that

$$\det(A) = (x \cdot x)(y \cdot y) - (x \cdot y)^2 \geq 0$$

which is the same as the Cauchy - Schwarz inequality. Recall from Examples 1.2 and 1.3 that this proves the triangle inequality for the ordinary Euclidean metric. In the exercises you will see that the case $m = 3$ proves the triangle inequality for the spherical metric of Example 1.6.

Example 1.7. Let X be any non-empty set and let d be defined by

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

This distance is called a *discrete metric* and (X, d) is called a *discrete metric space.*

It is easy to verify the triangle inequality: only need to consider the case $x \neq z$, in which case at least one of the two inequaities $x \neq y$ and $y \neq z$ must hold. Thus in the triangle inequality the left hand side = 1 and at least one of the two summands on the right hand side = 1, so the right hand side is ≥ 1.

Example 1.8. Let $X = \mathbb{R}^2$ and let d be defined by

$$d(x,y) = \begin{cases} |x-y| \text{ if } x \text{ and } y \text{ are in the same ray from the origin} \\ |x| + |y| \text{ otherwise,} \end{cases}$$

where $|x|$ denotes the usual length of a vector $x \in \mathbb{R}^2$. See Figure 1.3. This metric is called the *French railway metric* and it describes the following hypothetical situation: a country (let's call it France) in which there are railway lines passing through every town but always ending at a fixed city (let's call it Paris). You can travel directly between any two towns that happen to lie on the same railway line to Paris. Otherwise you have to go to Paris and change to another line.

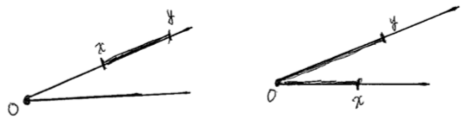

Figure 1.3. The French Railway Metric

There are two ways to verify the triangle inequality. One would be a direct check distinguishing cases, depending on the number of rays in which x, y and z lie and perhaps their relative positions on these rays. We will choose a more roundabout way that illustrates a general reasoning that we will often need in the future. Let us use the following terminology: given two points $x, y \in X$, a *path γ from x to y* is a finite collection $I_1, \ldots I_n$ where

(1) Each I_i is an interval lying in a ray from the origin.
(2) The ending point of I_i is the beginning point of I_{i+1}.
(3) The beginning point of I_1 is x and the ending point of I_n is y.

The *length of a path* is the sum of the lengths of the intervals I_i. We need the following observation: $d(x,y) =$ *the length of the shortest path from x to y.* In fact, the shortest path consists of one interval in case x, y lie on the same ray starting at the origin, and otherwise of two intervals.

The triangle inequality now follows: let γ_1 be a shortest path from x to y and let γ_2 be a shortest path from y to z. Let $\gamma_1\gamma_2$ denote the path formed by γ_1 followed by γ_2, see Figure 1.4. This is a path from x to z of length $d(x,y) + d(y,z)$. Its length cannot be any shorter than that of the shortest path from x to z, thus $d(x,z) \leq d(x,y) + d(y,z)$.

This example illustrates a very useful principle: existence of paths and a reasonable notion of length of paths gives a metric space. We give another example along the same lines. We will see more examples later on.

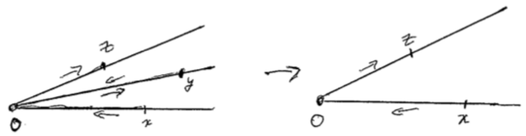

FIGURE 1.4. One Case of the Triangle Inequality

Example 1.9. Let $S \subset \mathbb{R}^3$ be a smooth surface. As a temporary definition of smooth surface, let's say that there is a smooth (meaning infinitely differentiable) function $f : \mathbb{R}^3 \to \mathbb{R}$ so that $S = \{x \in \mathbb{R}^3 : f(x) = 0\}$ and that the gradient $\nabla f \neq 0$ at any point of S. We will study later why this is a reasonable definition. For the moment, keep in mind Example 1.6 where $S^2 \subset \mathbb{R}^3$ is given as the zero set of $f(x) = |x|^2 - 1$. (Check that f is smooth and that ∇f does not vanish on S^2.)

If $x, y \in S$, let us define a *path from x to y* to be a continuous, piecewise differentiable curve γ lying in S, starting at x and ending at y. This means that for some interval $[a, b] \subset \mathbb{R}$, $\gamma : [a, b] \to S \subset \mathbb{R}^3$ is a continuous, piecewise differentiable map. Its *length* $L(\gamma) = \int_a^b |\gamma'(t)|dt$ is defined. *Assume that for all $x, y \in S$ there is a path from x to y.* This assumption is called *connectedness*, a concept that will be discussed in detail later. Define the distance function $d : S \times S \to \mathbb{R}$ by

$$d(x, y) = \inf\{L(\gamma) : \gamma \text{ a path from } x \text{ to } y\}.$$

We use infimum because, in contrast with the last example, it is not clear that a minimum exists (in fact, we will have to give conditions that ensure the existence of a minimum).

To verify that (S, d) is a metric space, we should first check that if $d(x, y) = 0$ then $x = y$. This follows from the fact that, if γ is a path from x to y, then $L(\gamma) \geq |x - y|$, where $|x - y|$ is the usual distance in $\mathbb{R}^3$. This implies that $d(x, y) \geq |x - y|$, so if $d(x, y) = 0$ then $|x - y| = 0$, so $x = y$.

Now to the triangle inequality. This follows the same pattern as the proof in Example 1.8 except that, since we have an infimum rather than a minimum, we have to use some ϵ's. Let x, y, z be fixed, and let $\epsilon > 0$ be given. Then, by the definition of infimum, there exists a path γ_1 from x to y with $L(\gamma_1) < d(x, y) + \frac{\epsilon}{2}$, and there exists a path γ_2 from y to z with $L(\gamma_2) < d(y, z) + \frac{\epsilon}{2}$. Let $\gamma = \gamma_1\gamma_2$ be the piecewise differentiable path γ_1 followed by γ_2 from x to z, see Figure 1.5. Then $d(x, z) \leq L(\gamma) = L(\gamma_1) + L(\gamma_2) < d(x, y) + d(y, z) + \epsilon$. Thus for all $\epsilon > 0$ we have $d(x, z) < d(x, y) + d(y, z) + \epsilon$, thus $d(x, z) \leq d(x, y) + d(y, z)$.

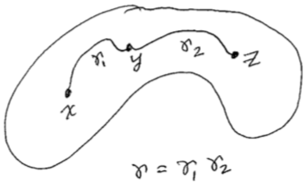

Figure 1.5. Putting Paths Together

Remark 1.2. If we knew that for $x, y \in S^2$ the great circle arc from x to y gives the shortest curve from x to y, then Example 1.6 becomes a special case of Example 1.9. This familiar fact will be proved later. This will give a more conceptual proof of the triangle inequality for the spherical metric than the one suggested in Example 1.6.

Example 1.10. Let $X = \mathbb{Z}$, the integers, and fix a prime number p. For $x, y \in \mathbb{Z}$, $x \neq y$, define $n(x, y)$ to be the exponent of p in the prime factorization of $x - y$, thus $x - y = kp^{n(a,b)}$ where p does not divide k. Define $d : X \times X \to \mathbb{R}$ by

$$d(x, y) = \begin{cases} 0 \text{ if } x = y, \\ p^{-n(x,y)} \text{ if } x \neq y. \end{cases}$$

Thus in this distance, called the *p-adic metric*, closeness means congruence modulo a high power of p. For instance, if $p = 5$, $d(0, 1) = d(0, 2) = d(0, 8) = 1$, while $d(0, 5) = d(0, 15) = \frac{1}{5}$, while $d(0, 25) = d(0, 50) = \frac{1}{25}$, etc.

To check the triangle inequality observe that given $x, y, z \in \mathbb{Z}$, we have

$$n(x, z) \geq \min\{n(x, y), n(y, z)\},$$

because p raised to the exponent on the right hand side divides both $x - y$ and $y - z$, so it certainly divides the sum $x - z$. We therefore have the inequality

$$p^{-n(x,z)} \leq \max\{p^{-n(x,y)}, p^{-n(y,z)}\}$$

which is equivalent to

$$d(x, z) \leq \max\{d(x, y), d(y, z)\}.$$

This inequality is called the *ultrametric inequality* and it immediately implies the triangle inequality because $\max\{d(x, y), d(y, z)\} \leq d(x, y) + d(y, z)$.

Example 1.11. We could modify the last example by taking $X = \mathbb{Q}$, the rational numbers. Each rational number has a prime factorization, where the exponents may now be negative. Fix a prime number p as before, and define $n(x, y)$ in the same way, and use the same formula for the distance. For instance, if $p = 5$ we have, in addition to the examples given above, $d(0, \frac{1}{2}) = 1, d(0, \frac{1}{5}) = d(0, \frac{3}{5}) = d(0, \frac{2}{15}) = 5, d(0, \frac{3}{50}) = 25$, etc. For any prime p we get, as before, a metric space, satisfying the stronger ultrametric inequality.

Definition 1.2. We define notation we will use in referring to some of the metric spaces just introduced.

(1) The metric of Example 1.3 will be called the *Euclidean metric* and, when there is need to distinguish it from other metrics on $\mathbb{R}^n$, will be denoted $d_{(2)}$. Thus
$$d_{(2)}(x,y) = \sqrt{(x_1-y_1)^2+\cdots+(x_n-y_n)^2}.$$
(2) The metric of Example 1.4 will be called the *taxicab metric* or the l^1 *metric* and denoted by $d_{(1)}$. Thus
$$d_{(1)}(x,y) = |x_1-y_1|+\cdots+|x_n-y_n|.$$
(3) The metric of Example 1.5 will be called the *supremum metric* or *sup metric* or l^∞*-metric* and denoted $d_{(\infty)}$. Thus
$$d_{(\infty)}(x,y) = \max\{|x_1-y_1|,\ldots,|x_n-y_n|\}.$$
(4) The metric of Example 1.6 will be called the *spherical metric* and denoted d_{S^2}.
(5) The metric of Example 1.9 will be called the *intrinsic metric* on $S \subset \mathbb{R}^3$.

Remark 1.3. We will see that the spherical metric is the same as the intrinsic metric on $S^2 \subset \mathbb{R}^3$.

1.2. **Constructions of Metric Spaces.**

There are some standard constructions of new metric spaces from given ones. The most common one is that of subspaces:

1.2.1. *Subspaces.*

Let (X,d) be a metric space and let $Y \subset X$. let $d' = d|_{Y\times Y}$ (the restriction of d to $Y \times Y$. Then (Y,d') is a metric space, called a *subspace* of (X,d). We usually write simply d for the restricted distance d'.

Examples of Subspaces

(1) $\mathbb{Q}$ is a subspace of $\mathbb{R}$.
(2) Any interval is a subspace of $\mathbb{R}$, for instance $(0,\infty)$ is a subspace of $\mathbb{R}$.
(3) S^2 is a subspace of $\mathbb{R}^3$. But the subspace metric is *not the same* as the spherical metric of Example 1.6. If d' is the restriction to $S^2 \times S^2$ of the Euclidean metric $d_{(2)}$ on $\mathbb{R}^3$ and d_{S^2} is the spherical metric on S^2, then clearly $d'(x,y) \le d_{S^2}(x,y)$ for all $x,y \in S^2$, and equality holds iff $x = y$.
(4) More generally, if $S \subset \mathbb{R}^3$ is a surface as in Example 1.9, then we get two distance functions on S: the subspace distance d' (restriction of the Euclidean distance) and the *intrinsic distance* d as defined in

> Example 1.9. We have again that $d'(x,y) \leq d(x,y)$ (in fact, this is how the fact that $d(x,y) = 0 \Rightarrow x = y$ was proved in Example 1.9). The case of equality is more subtle, it certainly holds if the straight line segment joining x and y lies in S.

1.2.2. *Product Spaces.* If (X_1, d_1) and (X_2, d_2) are metric spaces, their *product* is the space $(X_1 \times X_2, d)$ where

$$d((x_1, x_2), (y_1, y_2)) = \max\{d_1(x_1, y_1), d_2(x_2, y_2)\}$$

for all $(x_1, x_2), (y_1, y_2) \in X_1 \times X_2$. A similar definition can be made for the product of more than two factors. Note the analogy with the definition ((3) of Definition 1.2)of the supremum metric. Other definitions of the metric on the product are possible, but this is a convenient choice.

1.2.3. *Functions of the distance.* Suppose (X, d) is a metric space, and suppose that $f : [0, \infty) \to \mathbb{R}$ is a strictly increasing function with $f(0) = 0$ which is *sub-linear*: $f(a + b) \leq f(a) + f(b)$ holds for all $a, b \in [0, \infty)$. Then it is not hard to see that $f \circ d : X \times X \to \mathbb{R}$ is also a metric on X, that is, $(X, f \circ d)$ is a metric space. Details are in a homework problem.

1.3. **Limits.** One of the virtues of Definition 1.1 is that it allows the formulation of many familiar concepts from real analysis, with essentially the same definitions and proofs. We give some examples.

By a *sequence* in a metric space (X, d) we mean, as usual, a function $\mathbb{N} \to X$, written $\{x_n\}$.

Definition 1.3. Let $\{x_n\}$ be a sequence in (X, d).

(1) Let $x \in X$. We say $\lim\{x_n\} = x$ iff for all $\epsilon > 0$ there is an $N(= N(\epsilon) \in \mathbb{N}$ so that $d(x, x_n) < \epsilon$ for all $n > N$.
(2) We say that $\{x_n\}$ *converges* iff there exists $x \in X$ so that $\lim\{x_n\} = x$.
(3) We say that $\{x_n\}$ is a *Cauchy sequence* iff for all $\epsilon > 0$ there exists $N \in \mathbb{N}$ so that $d(x_m, x_n) < \epsilon$ for all $m, n > N$.

Theorem 1.2. *If $\{x_n\}$ converges, then $\{x_n\}$ is a Cauchy sequence.*

Proof. Suppose $\lim\{x_n\} = x$ and let $\epsilon > 0$. Then by (1) of Definition 1.3 there exists $N \in \mathbb{N}$ so that $d(x_n, x) < \frac{\epsilon}{2}$ for all $n > N$. If $m, n > N$, by the triangle inequality we have

$$d(x_m, x_n) \leq d(x_m, x) + d(x, x_n) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

hence $\{x_n\}$ is a Cauchy sequence. □

Observe how this proof uses the defining properties of metric spaces. The use of the triangle inequality is clear, also the symmetry of the distance ((2) of Definition 1.1 is used. As another example, we give a proof that also uses part (1) of Definition 1.1. In fact, this should be proved before the notation of that definition is introduced, so that the notation makes sense.

Theorem 1.3. *If $\{x_n\}$ converges, then its limit is unique.*

Proof. Suppose $\lim\{x_n\} = x$ and $\lim\{x_n\} = y$. Given $\epsilon > 0$ there exits N_1 so that $d(x_n, x) < \frac{\epsilon}{2}$ for all $n > N_1$ and there exists N_2 so that $d(x_n, y) < \frac{\epsilon}{2}$. Then, if $n > \max\{N_1, N_2\}$, we have

$$d(x, y) \leq d(x, x_n) + d(x_n, y) < \frac{\epsilon}{2} + \frac{\epsilon}{2} < \epsilon.$$

Since $d(x, y) < \epsilon$ for all $\epsilon > 0$, $d(x, y) = 0$, therefore (by (1) of Definition 1.1 we have $x = y$. Thus the limit is unique. □

Example 1.12. If we use the p-adic metric of Example 1.10 the convergent sequences we get may be unexpected. For example, the sequence $\{p^n\}$ converges to 0 since $d(p^n, 0) = p^{-n}$, thus, given $\epsilon > 0$, $d(p^n, 0) < \epsilon$ when $n > -\log_p(\epsilon)$.

Going back to Theorem 1.2, a familiar fact from analysis is that the converse holds for $X = \mathbb{R}$ (with usual distance) and $X = \mathbb{R}^n$ (with Euclidean metric). But it need not hold for all metric spaces (X, d). For example, we know that it does not hold for $X = \mathbb{Q}$, the set of rational numbers, with the usual distance $d(x, y) = |x - y|$. In fact, the validity of the converse is made into a definition:

Definition 1.4. A metric space (X, d) is called *complete* if every Cauchy sequence converges.

Thus $\mathbb{R}$ and $\mathbb{R}^n$ are complete, while $\mathbb{Q}$ is not complete (all with their usual distances).

1.4. **Maps Between Metric Spaces.** Let (X, d) and (Y, d') be metric spaces, and let $f : X \to Y$.

Definition 1.5. (1) Let $x \in X$. The map f is *continuous* at x iff for all $\epsilon > 0$ there exists a $\delta > 0$ so that for all $y \in X$, if $d(x, y) < \delta$, then $d'(f(x), f(y)) < \epsilon$.

(2) The map f is *continuous* iff it is continuous at all $x \in X$. Explicitly, f is continuous iff for all $x \in X$ and $\epsilon > 0$ there exists a $\delta(= \delta(x, \epsilon))$ so that $d'(f(x), f(y)) < \epsilon$ for all $y \in X$ with $d(x, y) < \delta$.

(3) The map f is *uniformly continuous* iff for all $\epsilon > 0$ there exists a $\delta(= \delta(\epsilon))$ so that $d'(f(x), f(y)) < \epsilon$ for all $x, y \in X$ with $d(x, y) < \delta$

(4) The map f is called *Lipschitz* iff there exists a constant $C > 0$ so that $d'(f(x), f(y)) \leq Cd(x, y)$ holds for all $x, y \in X$. The constant C is called *a Lipschitz constant* for f. If a smallest Lipschitz constant exists, then it is called *the* Lipschitz constant for f.

(5) The map f is *bi-Lipschitz* iff there exist constants $C_1, C_2 > 0$ so that $C_1 d(x, y) \leq d'(f(x), f(y)) \leq C_2 d(x, y)$ holds for all $x, y \in X$.

(6) The map f is an *isometry* iff $d'(f(x), f(y)) = d(x, y)$ for all $x, y \in X$.

Familiarity with the difference between continuity and uniform continuity for real functions is assumed. For example, the continuous function $f(x) = \frac{1}{x}$ on $(0, \infty)$ is uniformly continuous on $[1, \infty)$ but not on $(0, 1]$.

Remark 1.4. If (X, d) and (Y, d') are metric spaces, we often use the notation $f : (X, d) \to (Y, d')$ to mean:

(1) $f : X \to Y$,
(2) In the whole discussion we are using the metric d on the domain X and the metric d' on the target Y.

The notation does not imply that there is any relation among the three functions f, d, d'. It is just a reminder of which metrics are being used in the domain and the target. It is particularly important in the case that $X = Y$ but $d \neq d'$, we need to keep straight which metric we are using in domain and target.

Theorem 1.4. *If $f : (X, d) \to (Y, d')$ is Lipschitz, then f is uniformly continuos.*

Proof. Suppose $d'(f(x), f(y)) \leq Cd(x, y)$ and let $\epsilon > 0$. Let $\delta = \frac{\epsilon}{C}$. Then for all $x, y \in X$, $d(x, y) < \delta \Rightarrow d'(f(x), f(y)) < C\delta = \epsilon$, thus f is uniformly continuous. (Thus Lipschitz means that in the definition of uniform continuity, δ can be chosen as a linear function of ϵ). □

Here is a simple way to get Lipschitz functions. We state it for $\mathbb{R}$, but similar theorems can be formulated and proved in $\mathbb{R}^n$.

Theorem 1.5. *Suppose $I \subset \mathbb{R}$ is an interval, suppose $f : I \to \mathbb{R}$ is differentiable, and suppose that $|f'|$ is bounded on I: there exists $C > 0$ so that $|f'(x)| \leq C$ for all $x \in I$. Then f is Lipschitz on I with Lipschitz constant C.*

Proof. Given x and y in I, by the Mean Value Theorem there exists ξ between x and y so that $f(x) - f(y) = f'(\xi)(x - y)$. Then $|f(x) - f(y)| = |f'(\xi)||x - y| \leq C|x - y|$. □

For example, we see readily that $f(x) = \frac{1}{x}$ is Lipschitz on $[1, \infty)$ with Lipschitz constant 1 since $|f'(x)| = |\frac{-1}{x^2}| \leq 1$ on $[1, \infty)$. In particular f is uniformly continuous on $[1, \infty)$ as asserted earlier.

1.5. **Equivalences Between Metric Spaces.**

We will define various equivalences between metric spaces by assuming that the maps defined in the last section are bijective, with suitable additional requirements when needed.

Definition 1.6. Let (X, d) and (Y, d') be metric spaces, and let $f : X \to Y$ be a map. We say that:

(1) The map f is a *homeomorphism* iff f is continuous, $f^{-1} : Y \to X$ exists, and f^{-1} is continuous. If a homeomorphism f exists, we say that (X, d) and (Y, d') are *homeomorphic.*
(2) The map f is a *bi-Lipschitz equivalence* iff f is surjective and bi-Lipschitz. If a bi-Lipschitz equivalence exists we say that (X, d) and (Y, d') are *bi-Lipschitz equivalent.*
(3) The spaces (X, d) and (Y, d') are *isometric* iff there exists a surjective isometry $f : (X, d) \to (Y, d')$.

These equivalence relations go from very loose to very strict. More precisely, they are related as follows:

Theorem 1.6. *Let (X, d) and (Y, d') be metric spaces.*

(1) *If (X, d) and (Y, d') are isometric, then they are bi-Lipschitz equivalent.*
(2) *If (X, d) and (Y, d') are bi-Lipschitz equivalent, then they are homeomorphic.*

Proof. For the first part, observe that if the two spaces are isometric, this is the same thing as saying that they are bi-Lipschitz equivalent with constants $C_1 = C_2 = 1$.

For the second part, first observe that if f is a bi-Lipschitz equivalence, then f is injective: If $f(x) = f(y)$, then $d'(f(x), f(y)) = 0$, so $C_1 d(x, y) = 0$, so $d(x, y) = 0$, so $x = y$. Since f is surjective, then f^{-1} exists. Moreover, for all $x, y \in Y$, $C_1 d(f^{-1}(x), f^{-1}(y)) \leq d'(f(f^{-1}(x)), f(f^{-1}(y))) = d'(x, y)$. This is the same as $d(f^{-1}(x), f^{-1}(y)) \leq \frac{1}{C_1} d'(x, y)$, in other words, f^{-1} is Lipschitz (with Lipschitz constant $\frac{1}{C_1}$), thus f^{-1} is continuous, thus f is a homeomorphism. □

Example 1.13. Recall the distances $d_{(1)}, d_{(2)}, d_{(\infty)}$ on $\mathbb{R}^n$ of Definition 1.2. They are related by the following inequalities (the first in a homework problem, the remaining two are similar but easier).

(1) $d_{(2)}(x, y) \leq d_{(1)}(x, y) \leq \sqrt{n}\ d_{(2)}(x, y)$.
(2) $d_{(\infty)}(x, y) \leq d_{(2)}(x, y) \leq \sqrt{n}\ d_{(\infty)}(x, y)$.
(3) $d_{(\infty)}(x, y) \leq d_{(1)}(x, y) \leq n\ d_{(\infty)}(x, y)$.

These inequalities mean that the identity map is a bi-Lipschitz equivalence between any pair of these metrics, and the constants displayed turn out to be optimal. Moreover, it is easy to see that the identity map is *not* an isometry between any pair. For instance, for each $n > 1$, the distance from the origin to the point $(\frac{1}{n}, \dots, \frac{1}{n})$ is different in all three metrics on $\mathbb{R}^n$.

Example 1.14. A more delicate question is: can there be any isometry between two of these metrics? To see that to give a negative answer is not as obvious as it may seem at first sight, and to see a non-trivial example of an isometry, check the following: *The map $f : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $f(x_1, x_2) = (x_1 + x_2, x_1 - x_2)$ is an isometry from $(\mathbb{R}^2, d_{(1)})$ to $(\mathbb{R}^2, d_{(\infty)})$.*

Example 1.15. The last example indicates that it may not be so easy to prove that two spaces are not isometric, in other words, to prove that no f satisfying (6) of Definition 1.5 can exist. This usually requires some invariants that distinguish two metrics. For instance, it seems very clear to the eye that $(\mathbb{R}^n, d_{(2)})$ and $(\mathbb{R}^n, d_{(1)})$ are not isometric. Here's an possible way to distinguish them: Given a metric space (X, d) and two points $x, z \in X$, define the *equality set of the triangle inequality*, $E_d(x, z)$, by

$$E_d(x, z) = \{y \in X : d(x, z) = d(x, y) + d(y, z)\}.$$

It is not hard to prove that if $f : (X, d) \to (Y, d')$ is an isometry, then $E_{d'}(f(x), f(z)) = f(E_d(x, z))$. We know from Examples 1.3 and 1.4 that these equality sets are different for $d_{(2)}$ and $d_{(1)}$. This can be used to prove that they are *not* isometric. More details in the homework.

Example 1.16. Here's a more challenging question. Are $(\mathbb{R}^3, d_{(1)})$ and $(\mathbb{R}^3, d_{(\infty)})$ isometric? Note that the trick of Example 1.14 doesn't work. Make a conjecture and see if you can prove it..

2. Groups of Isometries

Let (X, d) be a metric space and let f, g be isometries of (X, d) onto itself. Then the composition $f \circ g$ is an isometry, since $d(f \circ g(x), f \circ g(y)) = d(f(g(x)), f(g(y))) = d(g(x), g(y)) = d(x, y)$. Also the inverse f^{-1} is defined and is also an isometry, since $d(f^{-1}(x), f^{-1}(y)) = d(f(f^{-1}(x), f(f^{-1}(y))) = d(x, y)$. This means that the set of all isometries is a *group* under composition.

Definition 2.1. Let $Isom(X, d) = \{f : X \to X : f$ is an isometry of (X, d) onto itself$\}$ denote the set of all isometries of (X, d). If $x \in X$, let $Isom(X, d)_x = \{f \in Isom(X, d) : f(x) = x\}$, the set of isometries of X that fix the point x. (This is often called the *stabilizer* of x, or the *isotropy group* of x.)

Theorem 2.1. *The set $Isom(X, d)$ is a group under composition. The subset $Isom(X, d)_x$ is a subgroup of $Isom(X, d)$ (under composition).*

Proof. We have just verified that the composition of two isometries is an isometry, and that the inverse of an isometry is an isometry. We thus have a binary operation $Isom(X, d) \times Isom(X, d) \to Isom(X, d)$ that assigns to $f, g \in Isom(X, d)$ their composition $f \circ g$. It is easy to verify the group axioms:

(1) The associative law $f \circ (g \circ h) = (f \circ g) \circ h$ holds for all $f, g, h \in Isom(X, d)$. This is always true for composition of maps.

(2) There exists $e \in Isom(X,d)$ such that $e \circ f = f \circ e = f$ for all $f \in Isom(X,d)$. Take $e = id$, the identity map $id : X \to X$.
(3) For all $f \in Isom(X,d)$ there exists $f^{-1} \in Isom(X,d)$ such that $f^{-1} \circ f = f \circ f^{-1} = e$. Take f^{-1} to be the usual inverse map.
Finally, if $x \in X$ and $f, g \in Isom(X,d)$ are such that $f(x) = x$ and $g(x) = x$, then $f \circ g(x) = f(g(x)) = f(x) = x$, and $f^{-1}(x) = x$ since $f(x) = x$. So $f \circ g$ and $f^{-1} \in Isom(X,d)_x$, so this subset is a subgroup.

□

The group of isometries of a metric space may be very small, in fact it may consist just of the identity. We next study a case where the group is big.

2.1. **Isometries of Euclidean Space.** We study first the group of isometries of $\mathbb{R}^2$ with the Euclidean metric $d_{(2)}$. In this section we'll write simply d for $d_{(2)}$, since this is the only metric we consider. The goal is to find all isometries of $(\mathbb{R}^2, d)$ and to describe the group structure.

2.1.1. *Affine transformations.* We first recall some facts from linear algebra. A transformation $L : \mathbb{R}^2 \to \mathbb{R}^2$ is called a *linear transformation* iff for all $r \in \mathbb{R}$ and for all $x, y \in \mathbb{R}^2$ we have $L(rx) = rL(x)$ and $L(x+y) = L(x) + L(y)$. This is equivalent to saying that for all $r, s \in \mathbb{R}$ and for all $x, y \in \mathbb{R}^2$, we have $L(rx + sy) = rL(x) + sL(y)$. Such a transformation is determined by $L((1,0))$ and $L((0,1))$ since $L((x_1, x_2)) = x_1 L((1,0)) + x_2 L((0,1))$. This is usually encoded in form of a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

where $(a,c) = L((1,0))$ and $(b,d) = L((0,1))$. If $(y_1, y_2) = L((x_1, x_2))$, then

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Note that column vectors, rather than row vectors, are used in this correspondence between matrices and linear transformations. The reason is that composition then corresponds to matrix mutliplication: If $L_1(x) = A_1 x$ and $L_2(x) = A_2 x$, then $L_1 \circ L_2(x) = L_1(L_2(x)) = A_1 A_2 x$. So we should always write points in $\mathbb{R}^n$ as column vectors rather than row vectors. But this is typographically very clumsy, so the usual convention is to write points as row vectors, keeping in mind that whenever matrix formulas such as $y = Ax$ are used, we temporarily write x, y, etc as column vectors.

The same correspondence and the same conventions are used for linear maps $L : \mathbb{R}^n \to \mathbb{R}^m$: $L(x) = Ax$ for some $n \times m$ matrix A.

Definition 2.2. A map $f : \mathbb{R}^n \to \mathbb{R}^n$ is called an *affine-linear transformation* iff there exist an $n \times n$ matrix A and a vector $b \in \mathbb{R}^n$ such that $f(x) = Ax + b$.

Thus an affine linear transformation is the composition of a translation and a linear transformation. Such a transformation sends lines to lines, planes to planes, etc.

2.1.2. *Some isometries of the Euclidean plane.* Some familiar isometries of $\mathbb{R}^2$ are the translations, the rotations, the reflections. These are all affine linear transformations of $\mathbb{R}^2$, namely, they are of the form $f(x) = Ax + b$ where b is a vector and A is a 2×2 matrix. We use the following terminology:

Definition 2.3. An afiine linear transformation $f(x) = Ax + b$ of $\mathbb{R}^2$ is called

(1) *translation by b and denoted t_b* if $A = I$, the unit matrix. Then $f(x) = t_b(x) = x + b$.

(2) *A rotation about the origin counterclockwise by an angle θ, denoted by R_θ,* if $b = 0$ and

$$A = R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{2.1}$$

(3) *a reflection about the line $\{t(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})\}$, denoted by S_θ* if $b = 0$ and

$$A = S_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \tag{2.2}$$

Remark 2.1. The terminology in the first two parts of the definition are clear and familiar. The fact that S_θ is a reflection in the asserted line will be shown in the homework. It is also clear and famliar that these transformations are isometries of $\mathbb{R}^2$. This is clear for translations. For rotations and reflections we use the fact that $|Ax| = |x|$ for $A = R_\theta$ or S_θ. (If you are not familiar with this fact, check it by using the fomulas for distance and the identity $\cos^2\theta + \sin^2\theta = 1$. For rotations, check that $(\cos\theta\, x_1 - \sin\theta\, x_2)^2 + (\sin\theta\, x_1 + \cos\theta\, x_2)^2 = x_1^2 + x_2^2$, with a similar check for reflections.) Then $d(Ax, Ay) = |Ax - Ay| = |A(x - y)| = |x - y| = d(x, y)$. Thus A is an isometry, as is any composition $f(x) = Ax + b$ of isomertries. So all the affine transformations we are considering are isometries.

Remark 2.2. The formulas for the above transformations are sometimes more convenient by identifying $\mathbb{R}^2$ with $\mathbb{C}$ and using complex operations. In terms of the complex variable z *real* affine linear transformation is of the form $f(z) = az + b$ or $f(z) = a\bar{z} + b$ where $a, b \in \mathbb{C}$. The formula for translation t_b is the same: $t_b(z) = z + b$. The formula for (2.1) becomes

$$R_\theta(z) = e^{i\theta} z \tag{2.3}$$

while (2.2) becomes

$$S_\theta(z) = e^{i\theta}\bar{z} \tag{2.4}$$

2.1.3. *The main theorem.* We want to prove that the examples $f(x) = Ax+b$ just discussed give *all* isometries of $\mathbb{R}^2$. The only difficulty is proving that every isometry of $\mathbb{R}^2$ is an affine linear transformation. Once we know this fact, then it is not hard to classify these transformations and obtain a complete list.

Theorem 2.2. *Let $f : (\mathbb{R}^2, d) \to (\mathbb{R}^2, d)$ be an isometry. Then f is affine-linear: there exists a vector $b \in \mathbb{R}^2$ and a 2×2 matrix A so that $f(x) = Ax+b$ for all $x \in \mathbb{R}^2$.*

Proof. First we make the reduction to the case $f(0) = 0$. Namely, let $f : \mathbb{R}^2 \to \mathbb{R}^2$ be an isometry, define a new isometry by $g(x) = f(x) - f(0)$, in other words, $g(x) = t_{-f(0)} \circ f$. Then g is an isometry with $g(0) = 0$, so if we can prove that there exists a matrix A so that $g(x) = Ax$, then $f(x) = Ax + b$, where $b = f(0)$.

To prove that if g is an isometry with $g(0) = 0$ then $g(x) = Ax$ for some matrix A is the same as proving that g is a linear transformation: $g(x+y) = g(x) + g(y)$ and $g(rx) = rg(x)$ for all $x, y \in \mathbb{R}^2$ and for all $r \in \mathbb{R}$. We will give two different proofs of this fact. The first proof is based on the following lemma concerning certain equality sets for the triangle inequality in $\mathbb{R}^2$ (straight line segments)

Lemma 2.1. *Let a, b be positive real numbers. Define a subset $E(a,b)$ of $(\mathbb{R}^2)^3$ (the set of triples of points of $\mathbb{R}^2$) by*

$$E(a,b) = \{(x,y,z) : x,y,z \in \mathbb{R}^2,\ d(x,y) = a,\ d(y,z) = b \textit{ and } d(x,z) = a+b\}.$$

Suppose that $(x_1,y_1,z_1),(x_2,y_2,z_2) \in E(a,b)$, and suppose that two of the three equalities $x_1 = x_2, y_1 = y_2, z_1 = z_2$ holds. Then the third equality also holds.

Proof. It two of the equalities hold, then both triples (x_1,y_1,z_1) and (x_2,y_2,z_2) lie on the same straight line segment of length $a+b$ and the three points are situated on the segment in the same order and with the same distances, so all three must coincide. To give a more detailed proof, we could divide it into three cases:

(1) $x_1 = x_2 = x,\ z_1 = z_2 = z$: Then y_1 and y_2 lie on the straight line through x and z between x and z and at distance a from x (also distance b from z), thus $y_1 = y_2$.

(2) $x_1 = x_2 = x,\ y_1 = y_2 = y$: Then z_1 and z_2 lie on the straight line through x and y on the opposite side of y from x and at distance b from y, thus $z_1 = z_2$.

(3) $y_1 = y_2 = y,\ z_1 = z_2 = z$: Then x_1 and x_2 lie on the straight line through y and z on the opposite side of y from z and at distance a from y, thus $x_1 = x_2$.

□

With this lemma we can now prove the theorem in two more steps:

(1) *Proof that $g(rx) = rg(x)$ for all $r \in \mathbb{R}$ and $x \in \mathbb{R}^2$*: It is clear for $r = 0, 1$, so let's assume $r \neq 1$ and consider three cases, corresponding, in the same order, to the three cases in the proof of the lemma:
   (a) $0 < r < 1$: Let $a = rd(0,x)$ and $b = (1-r)d(0,x)$. Then $(0, rx, x) \in E(a,b)$. Since g is an isometry with $g(0) = 0$, $(0, g(rx), g(x)) \in E(a,b)$. But since g is an isometry, $(0, rg(x), g(x)) \in E(a,b)$. By the lemma, $g(rx) = rg(x)$,
   (b) $r > 1$: Let $a = d(0,x)$ and $b = (r-1)d(0,x)$. Then $(0, x, rx) \in E(a,b)$. Again, since g is an isometry with $g(0) = 0$ we get that both $(0, g(x), g(rx)), (0, g(x), rg(x) \in E(a,b)$, so by the lemma we have $g(rx) = rg(x)$.
   (c) $r < 0$: Let $a = |r|d(0,x)$ and $b = d(0,x)$. Then $(rx, 0, x) \in E(a,b)$ and again, since g is an isometry with $g(0) = 0$ we get both $(g(rx), 0, g(x))$ and $(rg(x), 0, g(x)) \in E(a,b)$, so, by the lemma, $g(rx) = g(x)$.

   Thus $g(rx) = rg(x)$ for all $r \in \mathbb{R}$ and for all $x \in \mathbb{R}^2$.

(2) *Proof that $g(x+y) = g(x) + g(y)$ for all $x, y \in \mathbb{R}^2$*: Let $x, y \in \mathbb{R}^2$ Then $(x, \frac{x+y}{2}, y)) \in E(a,a)$, where $a = d(x,y)/2 > 0$ (since $\frac{x+y}{2}$ is the midpoint of the segment from x to y). Since g is an isometry $(g(x), g(\frac{x+y}{2}), g(y))$, and $(g(x), \frac{g(x)+g(y)}{2}, g(y)) \in E(a,a)$. By the lemma, $\frac{g(x+y)}{2} = \frac{g(x)+g(y)}{2}$, and, since $g(0) = 0$, by the first part (with $r = \frac{1}{2}$) we have $g(\frac{x+y}{2}) = \frac{g(x+y)}{2}$. Thus $\frac{g(x+y)}{2} = \frac{g(x)+g(y)}{2}$, and cancelling the denominators we get $g(x+y) = g(x) + g(y)$ as desired. This finishes the first proof of the theorem.

Another proof of the theorem is based on the following lemma, and on the fact that we already know many isometries of $\mathbb{R}^2$. This is much as the proof in section 1.4 of [10].

Lemma 2.2. *Suppose $f : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry with $f((0,0) = (0,0)$, $f((1,0)) = (1,0)$ and $f((0,1)) = (0,1)$. Then $f = id$.*

Proof. We need to prove that $f(x_1, x_2) = (x_1, x_2)$ for all $(x_1, x_2) \in \mathbb{R}^2$. Let

$$a = d((0,0), (x_1, x_2),\ b = d((1,0), (x_1, x_2)),\ c = d((0,1), (x_1, x_2)).$$

Since $f((x_1, x_2))$ is at the same distances a, b, c from the three points $(0,0)$, $(1,0)$, $(0,1)$ it is enough to show that a, b, c, determine (x_1, x_2)

uniquely. But this is clear by squaring the distances and solving:

$$\begin{aligned}
&a^2 = x_1^2 + x_2^2,\\
&b^2 = (x_1 - 1)^2 + x_2^2 = x_1^2 - 2x_1 + 1 + x_2^2 = a^2 - 2x_1 + 1,\\
&c^2 = x_1^2 + (x_2 - 1)^2 = x_1^2 + x_2^2 - 2x_2 + 1 = a^2 - 2x_2 + 1,
\end{aligned}$$

where, in the second and third equations we substituted the first. This clearly can be solved as:

$$\begin{aligned}
x_1 &= \tfrac{a^2-b^2+1}{2},\\
x_2 &= \tfrac{a^2-c^2+1}{2}.
\end{aligned}$$

Thus we can find x_1, x_2 from the three distances a, b, c, so $f((x_1, x_2)) = (x_1, x_2)$. □

We can now prove the theorem: Suppose $f : \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry, and let $g(x) = f(x) - f(0)$ as before, so that g is an isometry of $\mathbb{R}^2$ with $g(0) = 0$. Then $g((1, 0))$ is at distance one from $(0, 0)$, so $g((1, 0)) = (\cos\theta, \sin\theta)$ for some θ. Then $g((0, 1)) = (u, v)$ must be at distance one from $(0, 0)$ and at distance $\sqrt{2}$ from $(1, 0)$, which gives the equations

$$\begin{aligned}
&u^2 + v^2 = 1\\
&(u - \cos\theta)^2 + (v - \sin\theta)^2 + 2.
\end{aligned}$$

Expanding the second equation and substituting the first we get

$$-2u\cos\theta - 2v\sin\theta = 0, \text{ equivalently, } u\cos\theta + v\sin\theta = 0,$$

which clearly has only two solutions with $u^2 + v^2 = 1$, namely

$$\begin{aligned}
&(u, v) = (-\sin\theta, \cos\theta) \text{ or}\\
&(u, v) = (\sin\theta, -\cos\theta).
\end{aligned}$$

In the first case we have that g agrees *on the three points* $(0, 0), (1, 0), (0, 1)$ with the linear transformation with matrix

$$A = R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

of Equation 2.1 while, in the second case, g agrees *on the same three points* with the linear transformation with matrix

$$A = S_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}.$$

of Equation 2.2 In either case we have that the isometry $A^{-1}g$ sends the three points $(0, 0)$, $(1, 0)$, $(0, 1)$ to themselves, thus, by the lemma, $A^{-1}g = id$ or $g = A$. Thus $f(x) = Ax + b$, where $b = f(0)$ and this concludes the second proof of the theorem.

□

Remark 2.3. Observe that the first proof of Theorem 2.2 does not use the assumption $n = 2$, so it is valid for any n. (The second proof could be modified to work for every n, but we'll not do so). So the theorem is true in this generality, and we state it as such:

Theorem 2.3. *Let $f : (\mathbb{R}^n, d) \to (\mathbb{R}^n, d)$ be an isometry. Then f is affine-linear: there exists a vector $b \in \mathbb{R}^n$ and a $n \times n$ matrix A so that $f(x) = Ax + b$ for all $x \in \mathbb{R}^n$.*

2.1.4. *Orthogonal Matrices.* Once we know Theorem 2.3 we can get more detailed information. The matrix A is not arbitrary, it gives a linear transformation of $\mathbb{R}^n$ that is an isometry (since $Ax = f(x) - b$ and both f and t_b are isometries). In particular A preservs distance from the origin, which means $Ax \cdot Ax = x \cdot x$ for all x, equivalently, $(Ax)^t Ax = (x^t)x$, equivalently, $x^t A^t A x = x^t x$ which happens for all x if and only if $A^t A = I$. Here A^t means the transpose matrix. A matrix A that satisfies $A^t A = I$ is called an *orthogonal matrix.* Then $AA^t = I$ also holds (a left inverse is also a right inverse for $n \times n$ matrices) If the equation $A^t A = I$ is written explicitly, it says that the columns of A have length one and are pairwise orthogonal. (If you are not familiar with this, write everything out explicitly for $n = 2$: if $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, write down the equation $(ax_1 + bx_2)^2 + (cx_1 + dx_2)^2 = x_1^2 + x_2^2$, compare coefficients to get $a^2 + c^2 = 1$, $b^2 + d^2 = 1$, $ab + cd = 0$, and check that this is the same as $A^t A = I$.)

2.2. **The Euclidean and Orthogonal Groups.**

The equation $A^t A = I$ is equivalent to $A^t = A^{-1}$, so orthogonal matrices are invertible, and A^{-1} is also orthogonal, since it equals A^t and $(A^t)^t = A$ and $A^t A = I$ is equivalent to $AA^t = I$. The product AB of two orthogonal matrices A, B is orthogonal since $(AB)^t(AB) = B^t A^t A B = B^t B = I$. This means that the set of orthogonal matrices forms a group under matrix multiplication. Also, if A is orthogonal, then $\det(I) = \det(A^t A) = \det(A^t)\det(A) = \det(A)^2$, so $\det(A) = \pm 1$. Moreover, since $\det(AB) = \det(A)\det(B)$, we have that det is a homomorphism. Thus the following definition makes sense:

Definition 2.4. We denote by $O(n)$ the set of orthogonal matrices, by $SO(n)$ the set of orthogonal matrices with determinant one, by $E(n)$ the set of isometries of $\mathbb{R}^n$ and by $SE(n)$ the set of isometries $Ax + b$ of $\mathbb{R}^n$ with $\det(A) = 1$. The elements of $SE(n)$ are called the *proper isometries* (or the *orientation preserving isometries*) of $\mathbb{R}^n$. The elements of $E(n)$ which are *not* in $SE(n)$ are called the *improper isometries* (or the *orientation reversing isometries*) of $\mathbb{R}^n$.

Remark 2.4. The notation $O(n)$, $SO(n)$ is standard. Unfortunately there does not seem to be a standard notation for what we call $E(n)$, $SE(n)$.

We will use the notation $f_{A,b}$ for the isometry $f_{A,b}(x) = Ax + b$ of $\mathbb{R}^n$.

Definition 2.5. Define a map $l : E(n) \to O(n)$ by $l(f_{A,b}) = A$. The matrix $l(f)$ is called the *linear part* of f.

Theorem 2.4. (1) *The sets $E(n)$, $SE(n)$, $O(n)$, $SO(n)$ are groups (under composition or matrix multiplication as the case may be).*

(2) *The map $l : E(n) \to O(n)$ is a group homomorphism with kernel the group of translations of R^n, which is a group isomorphic to the group $\mathbb{R}^n$ (vector addition).*
(3) *The map* $\det : O(n) \to \{1, -1\}$ *is a group homomorphism with kernel $SO(n)$.*
(4) *The map* $\det \circ l : E(n) \to \{1, -1\}$ *is a group homomorphism with kernel $SE(n)$*

Proof. We know, by Theorem 2.1, that the group of isometries of any metric space is a group, thus $E(n)$ is a group. We have just verified directly that $O(n)$ is a group. This also follows from the second part of Theorem 2.1 since $O(n)$ is the subgroup of $E(n)$ that fixes the origin. That $SO(n)$ and $SE(n)$ are groups will follow from (3) and (4).

To prove (2), note that $f_{A_1,b_1} \circ f_{A_2,b_2}(x) = A_1(A_2x + b_2) + b_1 = A_1A_2x + b_1 + A_1b_2$, thus

$$f_{A_1,b_1} \circ f_{A_2,b_2} = f_{A_1A_2,b_1+A_1b_2} \tag{2.5}$$

From this we see that $l(f_{A_1,b_1} \circ f_{A_2,b_2}) = A_1A_2 = l(A_1)l(A_2)$, thus l is a homomorphism. Its kernel is $\{f_{A,b} : A = I\} = \{f_{I,b} : b \in R^n\}$ which is a sub-group of $E(n)$, the group of translations $\{t_b : b \in \mathbb{R}^n\}$, which is isomorphic to $\mathbb{R}^n$ since $t_{b_1} \circ t_{b_2} = t_{b_1+b_2}$. This proves (2). Then (3) and (4) are clear since det is a homomorphism and kernels of homomorphisms are subgroups. □

Remark 2.5. Note that the fourth part of this Theorem says that $SE(n)$ is a subgroup of index 2 of $E(n)$, so its complement in $E(n)$, the set of improper isometries, is a coset. This also means : the composition of two proper or two improper isometries is proper, while the composition of a proper and an improper isometry (in either order) is improper.

Remark 2.6. Observe that the set $E(n)$ is in one-to-one correspondence with the set $O(n) \times \mathbb{R}^n$, namely $f_{A,b} \in E(n) \leftrightarrow (A, b) \in O(n) \times \mathbb{R}^n$. This one-to-one correspondence takes the product of Equation 2.5 to the product

$$(A_1, b_1)(A_2, b_2) = (A_1A_2, b_1 + A_1b_2). \tag{2.6}$$

This is a group structure on the product $O(n) \times \mathbb{R}^n$, but *it is not isomorphic to the product group structure*

$$(A_1, b_1)(A_2, b_2) = (A_1A_2, b_1 + b_2). \tag{2.7}$$

The group structure of Equations 2.5 and 2.6 is called *a semi-direct product* of $O(n)$ and $\mathbb{R}^n$.

Remark 2.7. The subgroup $\mathbb{R}^n$ of translations is a *normal subgroup* of $E(n)$ since it is the kernel of a homomorphism. The group $E(n)$ contains many subgroups isomorphic to $O(n)$, but none of these are normal subgroups. For instance the subgroup $O(n)$ itself, namely $O(n) = \{f_{A,0}\} \subset E(n)$ of isometries that preserve the origin 0 is not a normal subgroup, because, for any fixed $b \neq 0$, we see that $t_b \circ f_{A,0} \circ t_B^{-1}(x) = A(x-b)+b = f_{A,b-Ab} \notin O(n)$. We will shortly discuss this in more detail for the case $n = 2$.

2.3. **The Euclidean Group in 2 Dimensions.**

Next we classify the isometries of $\mathbb{R}^2$ by dividing them into 4 classes according to properness and fixed points.

2.3.1. *Classification of Proper Isometries.* Let $f \in SE(2)$ be a proper isometry of $\mathbb{R}^2$, and assume $f \neq id$. A point $x \in \mathbb{R}^2$ is called a *fixed point* of f iff $f(x) = x$. To find fixed points it will be convenient to use the complex notation of Equation 2.3 To solve $f(z) = e^{i\theta}z + b = z$ is the same as solving $z - e^{i\theta}z = (1 - e^{i\theta})z = b$, which can be solved iff $e^{i\theta} \neq 1$. So there are two cases:

(1) *f has no fixed points.* This happens if and only if $e^{i\theta} = 1$, which is the same as $f(z) = z + b$, which is the same as f being a translation. Thus $f \in SE(2)$ *has no fixed points if and only if f is a translation.*

(2) *f has a fixed point.* This happens if and only if $e^{i\theta} \neq 1$, in which case the fixed point, which we denote by c, is given by $c = \frac{b}{1-e^{i\theta}}$. Thus we see that in this case the fixed point c is unique. The interpretation of this fixed point is that f *is a rotation with center c.* This can be seen as follows: A rotation by angle θ with center c is obtained from the rotation R_θ about origin by first translating the whole plane by t_{-c} so that c moves to the origin, then applying R_θ, then translating the whole plane back by t_c so that the origin goes back to c. In formulas, $f(z) = e^{i\theta}(z - c) + c = e^{i\theta}z + (c - e^{i\theta}c) = e^{i\theta} + b$, thus our solution for the fixed point found the center of rotation. In summary: *$f \in SE(2)$, $f \neq id$, has a fixed point if and only if f is a rotation (by a non-trivial angle) about a center $c \in \mathbb{R}^2$, and c is the unique fixed point of f.*

Remark 2.8. The interpretation of conjugation of a rotation by a translation as translating the center of rotation gives a very clear picture of why the subgroup $SO(2) \subset SE(2)$ cannot be a normal subgroup: if it were, then rotations about any point c would be the same collection of transformations as the rotations about any other point c', which we know by experience not to be true. This explains why the group structure of $SE(2)$ must follow the pattern of Equation (2.6) rather than that of (2.7). Note also that $SO(2)$ and $\mathbb{R}^2$ are both abelian groups, so if $SE(2)$ had the group law of (2.7), then it would be abelian, which is not the case.

Remark 2.9. This example illustrates what conjugacy of isometries means. Roughly speaking, two isometries are conjugate if they act in the same way but maybe in different locations (as rotations by the same angle but with two different centers) or in reference to different objects, like reflections in different mirrors that we will see below.

Remark 2.10. Here's a familiar consequence of the group law of Equations (2.5) and (2.6). Take rotations about different centers, but with opposite angles, say $f_1(x) = R_{-\theta}(x)$ and $f_2(x) = R_\theta(x) + b$. Then $f_1 \circ f_2(x) = x + R_{-\theta}b$ which is a translation. Thus composing rotations with different centers (but angles adding to zero) produces a translation. This is familiar to anybody who has had to parallel-park a car.

2.3.2. *Classification of Improper Isometries.* We now study the fixed points of improper isometries $f \in E(2) \setminus SE(2)$. These isometries are of the form $f(x) = S_\theta(x) + b$ in real notation (2.2), or $f(z) = e^{i\theta}\bar{z}$ in complex notation (2.4). We need first to understand the linear part S_θ. It is easy to check that for all θ we have $S_\theta^2 = id$, thus its eigenvalues are ± 1, and since their product is the determinant, which is -1, we must have that one eigenvalue is 1, the other is -1. This means that there is an orthonormal basis $\{w_1, w_2\}$ for $\mathbb{R}^2$ so that $S_\theta w_1 = w_1$ and $S_\theta w_2 = -w_2$.

There is a very standard and convenient way of writing this. Let's call v the eigenvector w_2, thus $S_\theta(v) = -v$ and $|v| = 1$ (this determines v uniquely up to sign). Then it is easy to check that

$$S_\theta(x) = x - 2(x \cdot v)v, \tag{2.8}$$

because $S_\theta(v) = v - 2(v \cdot v)v = v - 2v = -v$ and if $w \perp v$, then $S_\theta(w) = w - 2(w \cdot v)v = w - 0 = w$, so this linear transformation fixed everything perpendicular to v and maps v to $-v$, as required. We need of course to relate v to θ. It is a homework problem to work out that $v = (-\sin\frac{\theta}{2}, \cos\frac{\theta}{2})$, which is equivalent to saying that S_θ is reflection in $v^\perp$, the line through the origin perpendicular to this vector, namely the line

$$v^\perp = \{(x_1, x_2) : -\sin\frac{\theta}{2}\, x_1 + \cos\frac{\theta}{2}\, x_2 = 0\} = \{t(\cos\frac{\theta}{2}, \sin\frac{\theta}{2}) : t \in \mathbb{R}\}, \tag{2.9}$$

as asserted when the notation S_θ was introduced in Equation (2.2). This line $v^\perp$ is called the *mirror of the reflection* S_θ, or S_θ is called the *reflection in the line* $v^\perp$. Another way to visualize S_θ is as a conjugate of S_0, the reflection in the x_1-axis (in complex notation $S_0(z) = \bar{z}$) by the rotation that takes the x_1-axis to the mirror $v^\perp$: $S_\theta(X) = R_{\frac{\theta}{2}} S_0 R_{\frac{\theta}{2}}^{-1}$.

The equation $f(x) = S_\theta x + b = x$ for fixed points is equivalent to $x - 2(x \cdot v)v + b = x$, in other words

$$2(x \cdot v)v = b, \tag{2.10}$$

which is one equation for the two unknowns (x_1, x_2). There are two cases

(1) *b is a multiple of v.* Then this multiple must be the number $b \cdot v$, and (2.10) has infinitely many solutions, namely the line $x \cdot v = \frac{1}{2}b \cdot v$ or $(x - \frac{b}{2}).v = 0$, equivalently, $x \in \frac{b}{2} + v^\perp$, the translate of the mirror $v^\perp$ of S_θ by the vector $\frac{b}{2}$. Geometrically, this means that *f is a reflection on the mirror* $\frac{b}{2} + v^\perp$, which is parallel to the mirror

$v^\perp$ of S_θ. Another way to visualize f is as a *conjugate* of S_θ, in a similar way as we just did with rotations about different centers: $f = t_{\frac{b}{2}} S_\theta t_{\frac{b}{2}}^{-1}$, namely, translate x by $-\frac{b}{2}$, then reflect on $v^\perp$ and then translate by $\frac{b}{2}$ has the effect of reflecting in the mirror $\frac{b}{2} + v^\perp$. This can be easily checked by formulas: the conjugate is $S_\theta(x - \frac{b}{2}) + \frac{b}{2} = S_\theta x + \frac{b}{2} + \frac{b}{2} = S_\theta x + b = f(x)$ where the first equality uses $S_\theta b = -b$.

(2) *b is not a multiple of v.* Then the equation (2.10) has no solutions, and f has no fixed points. But there's one distinguished line L which is *invariant* under f in the sense that $f(L) \subset L$. This is the line $x \cdot v = \frac{1}{2} b \cdot v$, the translate of the mirror $v^\perp$ of the linear part S_θ by half the normal component of the vector b, and along this line f is translation by the vector $b - (b \cdot v)v \neq 0$ which is the component of b in the direction of $v^\perp$. This can be easily seen from the formula (2.8), since $x \cdot v = \frac{1}{2} b \cdot v$ yields $f(x) = x - 2(x \cdot v)v + b = x + (b - (b \cdot v)v)$. In this case f is called a *glide reflection with axis L.* Any glide reflection is conjugate to the standard model $g_a(x_1, x_2) = (x_1 + a, -x_2)$ (or $g_a(z) = \bar{z} + a,\ a \in \mathbb{R}, a \neq 0$ in complex notation). This is a *glide reflection along the x_1-axis by a distance a.*

We can summarize the classification of isometries (different from the identity) in the following table:

	Proper	*Improper*
With fixed points	Rotations	Reflections
Without fixed points	Translations	Glide Reflections

The reflections are in some sense the most basic isometries of $\mathbb{R}^2$ in the sense that all isometries may be obtained by composing reflections. More precesily:

Theorem 2.5. *The composition of two reflections is:*

(1) *A translation if the mirrors of both reflections are parallel. Precisely, if b is a vector perpendicular two both mirrors and of length the distance between them, then their composition is $t_{\pm b}$ (sign depending on the order).*
(2) *A rotation by angle $\pm 2\alpha$ and centered at the intersection of their mirrors if they meet at an angle α (the sign depending on the order)*

The composition of three reflections is either a reflection or a glide-reflection. Every glide reflection can be obtained by composing three reflections, two of the mirrors being parallel and the third perpendicular to both.

Proof. The two statements about composition of two reflections are easy to verify. Since the composition of three reflections must be an improper motion, the next statement follows from the classification. The last statement follows by taking one mirror to be the invariant line (axis) of the glide reflection and the other two mirrors perpendicular to the axis placed so as to obtain the necessary translation. □

Corollary 2.1. *Every isometry of $\mathbb{R}^2$ can be obtained by composing one, two or three reflections. In particular, the group $E(2)$ is generated by reflections.*

See Section 1.4 of [10] for another proof of this Corollary, and Section 1.5 for another proof of the classification theorem.

3. Topological Spaces

3.1. **Topology of Metric Spaces.**

Let (X, d) be a metric space. We define the following objects, with a terminology motivated by the familiar concepts in the Euclidean plane:

Definition 3.1. Suppose (X, d) is a metric space.

(1) If $x \in X$ and $r > 0$, the set $B(x, r) = \{y \in X : d(x, y) < r\}$ is called the *ball of radius r centered at x*. This set is sometimes called the *open* ball of radius r centered at x.
(2) If $x \in X$ and $r \geq 0$, the set $\bar{B}(x, r) = \{y \in X : d(x, y) \leq r\}$ is called the *closed ball of radius r centered at x.* .
(3) If $x \in X$ and $r \geq 0$, the set $S(x, r) = \{y \in X : d(x, y) = r\}$ is called the *sphere of radius r centered at x.* .

Definition 3.2. A subset $U \subset X$ is called an *open set* if and only if, for every $x \in U$ there exists $r(= r(x)) > 0$ so that $B(x, r) \subset U$.

Theorem 3.1. *For any $x \in X$ and $r > 0$, $B(x, r)$ is an open set.*

Proof. Let $y \in B(x, r)$. We have to find $\rho > 0$ so that $B(y, \rho) \subset B(x, r)$. Guided by the picture in the Euclidean plane, we choose $\rho = r - d(x, y)$. To check $B(y, \rho) \subset B(x, r)$, let $z \in B(y, \rho)$, that is, $d(y, z) < \rho = r - d(x, y)$. Then, by the triangle inequality, $d(x, z) \leq d(x, y) + d(y, z) < d(x, y) + (r - d(x, y)) = r$, thus $d(x, z) < r$, in other words, $B(y, \rho) \subset B(x, r)$ as desired. □

Example 3.1. If $(X, d) = \mathbb{R}^2$ with the usual Euclidean metric $d_{(2)}$, then the balls and spheres are the usual balls and spheres, the open sets are the usual open sets. Same holds for $\mathbb{R}^n$, any n.

Example 3.2. If $(X, d) = \mathbb{R}^2$ with the taxicab metric $d_{(1)}$, then the balls and spheres are not the usual Euclidean balls and spheres, but they give the same open sets. One general principle at work here is: *bi-Lipschitz metrics give the same open sets.* By this we mean (see Definition 1.5): Suppose d, d' are metrics on X and that there exist constants $C_1, C_2 > 0$ so that $C_1 d'(x, y) \leq d(x, y) \leq C_2 d'(x, y)$. Then using B for d-balls and B' for d'-balls, we get $B'(x, r) \subset B(x, C_2 r)$ and $B(x, r) \subset B'(x, r/C_1)$. Then, if $U \subset X$ is d-open, $x \in U$ and $r > 0$ is such that $B(x, r) \subset U$, then

$B'(x, r/C_1) \subset U$, so U is d'-open, similarly in the other direction. We will see shortly that the necessary and sufficient condition for two metrics to give the same open sets is that they be *homeomorphic* (see Def 1.5).

Example 3.3. Suppose X is any non-empty set and let $d : X \times X \to \mathbb{R}$ be the discrete metric of Example 1.7. Then

$$B(x,r) = \begin{cases} \{x\} & \text{if } 0 < r \leq 1 \\ X & \text{if } r > 1. \end{cases}$$

Thus every subset of X is open: if $S \subset X$ is any subset and $x \in S$, then, say, $B(x, \frac{1}{2}) = \{x\} \subset S$, so S is open.

Example 3.4. If (X, d) is any metric space, the empty set is open. This is a "vacuously true" statement, namely, the negation of Definition 3.2 would begin: there exists $x \in U$ so that ... which could never be true for $U = \emptyset$.

Definition 3.3. A subset $F \subset X$ is called a *closed set* if and only if its complement $X \setminus F$ is an open set.

Theorem 3.2. *For all $x \in X$ and for all $r \geq 0$, the closed ball $\bar{B}(x, r)$ is a closed set.*

Proof. Just as with the proof of Theorem 3.1, we guide ourselves by the Euclidean picture. Let $x \in X$ and $r \geq 0$. We have to prove that the complement $X \setminus \bar{B}(x,r) = \{y \in X : d(x,y) > r\}$ is an open set. Given $y \in X \setminus \bar{B}(x,r)$ we need to find $\rho > 0$ so that $B(y,\rho) \subset X \setminus \bar{B}(x,r)$. Drawing the picture in $\mathbb{R}^2$ suggests trying $\rho = d(x,y) - r$. So suppose $z \in B(y,\rho)$, that is, $d(z,y) < d(x,y) - r$. Then the triangle inequality gives $d(x,y) \leq d(x,z) + d(z,y)$, equivalently, $d(x,z) \geq d(x,y) - d(z,y) > d(x,y) - (d(x,y) - r) = r$, as desired (where the last inequality uses the assumption $d(y,z) < d(x,y) - r$, and the inequality gets reversed when subtracting).

□

Remark 3.1. This proof would be slightly shorter if we use an equivalent form of the triangle inequality:

$$|d(x,z) - d(y,z)| \leq d(x,y).$$

Geometrically, in any triangle the difference of the lengths of two sides is at most the length of the third side. This inequality is easily derived from the usual triangle inequality: Start from $d(x,z) \leq d(x,y) + d(y,z)$ and subtract $d(y,z)$ from both sides, getting $d(x,z) - d(y,z) \leq d(x,y)$. Then interchange x, y to get $d(y,z) - d(x,z) \leq d(x,y)$, which together give the above inequality.

3.1.1. *Review of some set theory.* We briefly review some concepts and notations from set theory that we will need often. See the first chapter of [7] for more information.

If X is any set, we write 2^X for the set of all subsets of X, what is often called the *power set* of X. If X and Y are any sets and $f : X \to Y$ is *any* function, we the function $f^{-1} : 2^Y \to 2^X$ is defined by

$$f^{-1}(A) = \{x \in X : f(x) \in A\}. \tag{3.1}$$

The set $f^{-1}(A)$ is called the *inverse image of A* or the *pre-image of A*. Observe that it is defined for *any* function, it is by no means implied nor needed that the original function $f : X \to Y$ be invertible. There is another function associated to f, denoted by the same letter, namely $f : 2^X \to 2^Y$, defined by

$$f(A) = \{f(x) : x \in A\} \tag{3.2}$$

The pre-image function behaves very nicely with respect to all the set operations, for example:

Theorem 3.3. *If $f : X \to Y$, then the following hold for all $A, B \subset Y$:*

(1) $f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$,
(2) $f^{-1}(A \cap B) = f^{-1}(A) \cap f^{-1}(B)$,
(3) *Same for unions and intersections of arbitrary families of subsets.*
(4) $f^{-1}(A \setminus B) = f^{-1}(A) \setminus f^{-1}(B)$,
If, in addition, $g : Y \to Z$, then we also have:
(5) $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$

Proof. The proofs of all these statements are straightforward verifications using the definitions of the objects involved. We verify the last statement: If $A \subset Z$, then $x \in (g \circ f)^{-1}(A) \Leftrightarrow (g \circ f)(x) \in A \Leftrightarrow g(f(x)) \in A \Leftrightarrow f(x) \in g^{-1}(A) \Leftrightarrow x \in f^{-1}(g^{-1}(A)) \Leftrightarrow x \in f^{-1} \circ g^{-1}(A)$. □

We do not give corresponding statements for the image of sets $f : 2^X \to 2^Y$ because they are less useful, more complicated, and harder to remember. They usually involve inclusions rather than equalities.

3.1.2. *Continuous maps.* Let (X, d) and (Y, d') be metric spaces. Recall from Definition 1.5(1) what it means for a map $f : X \to Y$ to be *continuous*. The following theorem gives a very useful characterization of continuous maps:

Theorem 3.4. *A map $f : (X, d) \to (Y, d')$ is continuous if and only if the following holds: for each open set $U \subset Y$, its pre-image $f^{-1}(U) \subset X$ is also open.*

Proof. One implication: Suppose f is continuous in the sense of Definition 1.5, suppose $U \subset Y$ is open, and let $x \in f^{-1}(U)$. Since $f(x) \in U$ and U is open, there exists $\epsilon > 0$ so that $B'(f(x), \epsilon) \subset U$, where B' denotes a d'-ball. Since f is continuos, there exists $\delta > 0$ so that if $y \in X$ and $d(x, y) < \delta$, then $d'(f(x(, f(y)) < \epsilon$, in other words, $B(x, \delta) \subset f^{-1}(B'(f(x), \epsilon) \subset f^{-1}(U)$, so $f^{-1}(U)$ is open.

The opposite implication: Suppose that for all open subsets $U \subset Y$, $f^{-1}(U) \subset X$ is open. Given $x \in X$ and $\epsilon > 0$, since $B'(f(x), \epsilon) \subset Y$ is open, thus $f^{-1}(B'(f(x), \epsilon) \subset X$ is open. Since $x \in f^{-1}(B'(f(x), \epsilon)$, there exists $\delta > 0$ so that $B(x, \delta) \subset f^{=1}(B'(f(x), \epsilon)$. But this says exactly that for all $y \in X$, if $d(x, y) < \delta$, then $d'(f(x), f(y)) < \epsilon$. Therefore f is continuous.

□

Here are some immediate and useful cosequences:

Corollary 3.1. *A map $f : (X, d) \to (Y, d')$ is continuous if and only if the following holds: for each closed set $F \subset Y$, its pre-image $f^{-1}(F) \subset X$ is also closed.*

Proof. By definition, $F \subset Y$ is closed if and only if $X \setminus F$ is open and by Theorem 3.3, $f^{-1}(Y \setminus F) = f^{-1}(Y) \setminus f^{-1}(F) = X \setminus f^{-1}(F)$ is open, which happens if and only if $f^{-1}(F)$ is closed. Hence all pre-images of closed sets are closed if and only if all pre-images of open sets are open, as asserted. □

Corollary 3.2. *The composition of continuos maps is continuous. Precisely, suppose $f : (X, d) \to (Y, d')$ and $g : (Y, d') \to (Z, d'')$ are continuous. Then the composition $g \circ f : (X, d) \to (Z, d'')$ is continuous.*

Proof. Using the last part of Theorem 3.3, if $U \subset Z$ is open, then $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$ which is open because $g^{-1}(U)$ is open (continuity of g) and thus $f^{-1}(g^{-1}(U))$ is open (continuity of f). □

Corollary 3.3. *Let $f : (X, d) \to (Y, d')$ be a continuos map. Then f is a homeomorphism if and only if f is bijective, and for all open subsets $U \subset X$, $f(U) \subset Y$ is open. The last condition can be replaced by: for all closed subsets $F \subset X$, $f(F) \subset Y$ is closed.*

Proof. If f is bijective, then $f^{-1} : Y \to X$ is defined, and if $U \subset X$ is open, then $(f^{-1})^{-1}(U) = f(U)$ is open, thus f^{-1} is also continuous and f is a homeomorphism. Same reasoning with closed sets. □

Another variation of the same reasoning is:

Corollary 3.4. *Let $f : (X, d) \to (Y, d')$ be a bijective map (not assumed continuous). Then f is a homeomorphism if and only if the following holds: A subset $A \subset X$ is open if and only if $f(A) \subset Y$ is open. Equivalently: a subset $A \subset X$ is closed if and only if $f(A) \subset Y$ is closed.*

The following examples show some immediate applications of the theorems and corollaries just proved.

Example 3.5. One familiar example of how these characterizations of continuity are used is the following. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function. Then the following sets are open: $\{x : f(x) \neq 0\}$, $\{x : f(x) > 0\}$, $\{x : 1 < f(x) < 3\}$, etc, since they are pre-images of open sets in $\mathbb{R}$, namely they are $f^{-1}((-\infty, 0) \cup (0, \infty))$, $f^{-1}((0, \infty))$, $f^{-1}((1, 3))$, etc. Similarly, the following sets are closed: $\{x : f(x) = 0\}$, $\{x : 0 \leq f(x) \leq 1\}$, etc, since they are the pre-images of closed subsets of $\mathbb{R}$, namely $f^{-1}(\{0\})$, $f^{-1}([0, 1])$, etc.

Example 3.6. Let (X, d) be a discrete metric space as in Example 1.7 and let (Y, d') be any metric space. Then any map $f : (X, d) \to (Y, d')$ is continuous, because, as we saw in Example 3.3, every subset of X is open. If (Y, d') is also discrete, then $f : (X, d) \to (Y, d')$ is a homeomorphism if and only if it is bijective.

Example 3.7. Let d, d' be two metrics on X. Then they have the same open sets if and only if the identity map is a homeomorphism, as mentioned at the end of Example 3.2.

3.1.3. *The Collection of Open Sets.*

Theorem 3.5. *Let (X, d) be a metric space.*

1. *Let $\{U_\alpha\}_{\alpha \in A}$ be a collection of open subsets of X indexed by a set A. Then the union $\cup_{\alpha \in A} U_\alpha$ is an open set.*
2. *Let $U_1, \cdots, U_n$ be a* finite *collection of open subsets of X. Then their intersection $U_1 \cap \cdots \cap U_n$ is an open set.*

Proof. For (1), suppose $x \in \cup_{\alpha \in A} U_\alpha$. By definition of union, there exists $\alpha_0 \in A$ so that $x \in U_{\alpha_0}$. Since U_{α_0} is open, there exists an $r > 0$ so that $B(x, r) \subset U_{\alpha - 0}$. Then $B(x, r) \subset \cup_{\alpha \in A} U_\alpha$, so this last set is open.

For (2), suppose $x \in U_1 \cap \cdots \cap U_n$. Then, by definition of intersection, $x \in U_i$ for $i = 1, \cdots, n$. Since each U_i is open, there exists $r_i > 0$ so that $B(x, r_i) \subset U_I$ for $i = 1, \cdots, n$. Let $r = \min\{r_1, \cdots, r_n\}$. Then $B(x, r) \subset U_1 \cap \cdots \cap U_n$, so this last set is open. □

There is of course a corresponding theorem for closed sets:

Theorem 3.6. *Let (X, d) be a metric space.*

1. *Let $\{F_\alpha\}_{\alpha \in A}$ be a collection of closed subsets of X indexed by a set A. Then the intersection $\cap_{\alpha \in A} F_\alpha$ is a closed set.*
2. *Let $F_1, \cdots, F_n$ be a* finite *collection of closed subsets of X. Then their untion $F_1 \cup \cdots \cup F_n$ is a closed set.*

Proof. This follows directly from the last theorem and the properties of complements of unions or intersections as intersections or unions of complements. For example, to prove $\cap_{\alpha\in A}F_\alpha$ is closed if each F_α is closed, need to show that $X \setminus \cap F_\alpha$ is open. But $X \setminus \cap F_\alpha = \cup(X \setminus F_\alpha)$ which is a union of open sets (since each F_α is closed), hence open by the last theorem. □

3.2. **Topologies and Continuity.**

It turns out that a very good way of discussing continuity is to turn the last theorems into definitions.

Definition 3.4. Let X be any non-empty set. A subset $\mathcal{T} \subset 2^X$ is called a *topology* on X if and only if the following hold:

(1) $\emptyset \in \mathcal{T}$ and $X \in \mathcal{T}$.
(2) If A is any index set and for each $\alpha \in A$, $U_\alpha \in \mathcal{T}$, then $\cup_{\alpha\in A}U_\alpha \in \mathcal{T}$.
(3) If $U_1, \cdots, U_n \in \mathcal{T}$, then $U_1 \cap \cdots \cap U_n \in \mathcal{T}$.

Briefly, a topology on X is a collection of subsets of X that contains $\emptyset$ and X, and which is closed under the operations of arbitrary union and finite intersection.

Definition 3.5. A *topological space* is a pair $(X, \mathcal{T})$ where X is a non-empty set and $\mathcal{T}$ is a topology on X.

Definition 3.6. Let $(X, \mathcal{T})$ be a topological space. A subset $U \subset X$ is called an *open set* (or, if more than one topology is being discussed, a $\mathcal{T}$*-open set*) if and only if $U \in \mathcal{T}$. A subset $F \subset X$ is called a *closed set* (or a $\mathcal{T}$*-closed set* if needed) if and only if $X \setminus F \in \mathcal{T}$.

In other words, the elements of $\mathcal{T} \subset 2^X$ are the subsets of X that we decide to call open sets. Their complements in X are the subsets that we decide to call closed sets.

Remark 3.2. An equivalent way of defining a topology on X would be to give the collection of its closed sets. Namely, suppose we have a collection $\mathcal{C} \subset 2^X$ with the properties:

(1) $\emptyset \in \mathcal{C}$ and $X \in \mathcal{C}$.
(2) If A is any index set and for each $\alpha \in A$, $F_\alpha \in \mathcal{C}$, then $\cap_{\alpha\in A}F_\alpha \in \mathcal{C}$.
(3) If $F_1, \cdots, F_n \in \mathcal{C}$, then $F_1\cup, \cdots, F_n \in \mathcal{C}$.

(briefly, $\mathcal{C}$ contains $\emptyset$, X, and is closed under arbitrary intersections and finite unions), then $\mathcal{C}$ is the collection of open sets of a unique topology $\mathcal{T}$ on X, namely

$$\mathcal{T} = \{X \setminus F : F \in \mathcal{C}\}.$$

Sometimes it is more convenient to define a topology on X by defining the collection of closed sets rather than the collection of open sets.

Definition 3.7. Let $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$ be topological spaces. A map $f : X \to Y$ is called *continuous* if and only if, for all $U \in \mathcal{T}$, we have that $f^{-1}(U) \in \mathcal{T}$. A map $f : X \to Y$ is called a *homeomorphism* if and only if it is continuous, f^{-1} exists, and f^{-1} is continuous.

Thus a map $f : X \to Y$ is called continuous if and only if the pre-image of each $\mathcal{T}'$-open set in Y is a $\mathcal{T}$-open set in X.

Remark 3.3. Just as in the notation we explained in Remark 1.4, we use the notation $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ to mean:

(1) $f : X \to Y$,
(2) In the whole discussion, the topology $\mathcal{T}$ is being used in the domain X and the topology $\mathcal{T}'$ is being used in the target Y.

Just as in the case of metric spaces, this notation is particularly important when $X = Y$ but $\mathcal{T} \neq \mathcal{T}'$.

Just as with Corollary 3.1, we have the following characterization of continuity (with the same proof):

Theorem 3.7. *A map $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ is continuous if and only if the preimage $f^{-1}(F)$ of each $\mathcal{T}'$-closed set $F \subset Y$ is a $\mathcal{T}$-closed subset of X.*

Just as with Corollary 3.2. we have that the composition of continuous maps is continuous (again with the same proof):

Theorem 3.8. *Let $(X, \mathcal{T})$, $(Y, \mathcal{T}')$ and $(Z, \mathcal{T}'')$ be topological spaces. Let $f : (X, \mathcal{T}) \to (Y, \mathcal{T}')$ and $g : (Y, \mathcal{T}') \to (Z, \mathcal{T}'')$ be continuous maps. Then the composition $g \circ f : (X, \mathcal{T}) \to (Z, \mathcal{T}'')$ is continuous.*

3.2.1. *Examples of Topological Spaces.*

Example 3.8. Let (X, d) be any metric space, and let $\mathcal{T}_d$ be the collection of open sets as defined in Definition 3.2. Then, by Theorem 3.5, the collection $\mathcal{T}_d \subset 2^X$ is a topology on X.

Example 3.9. In the special case that $X = \mathbb{R}^n$ and d is the Euclidean metric of Example 1.3 we call the resulting metric topology $\mathcal{T}_d$ the *Euclidean topology* and denote it by $\mathcal{T}_E$.

Example 3.10. Let X be any non-empty set and let $\mathcal{T}_{disc} = 2^X$. This is called the *discrete topology* on X. Every subset of X is open. Note that this is a special case of the last example, namely $\mathcal{T}_{disc}$ is the same as the metric topology of the discrete metric, see Examples 1.7 and 3.3.

Example 3.11. Let X be any non-empty set and let $\mathcal{T}_{ind} = \{X, \emptyset\}$. This example is at the opposite extreme of the last one: it is the *smallest* collection in 2^X that satisfies Definition 3.4, while the last example gave the *largest* one. This is often called the *indiscrete topology.*

Example 3.12. Let $X = \{a, b\}$ be a two element set. Then besides the discrete and indiscrete topologies on X there are precisely two other topologies: $\{\emptyset, \{a\}, X\}$ and $\{\emptyset, \{b\}, X\}$, see Example 4 in p. 72 of [7].

Example 3.13. Let X be any infinite set and let $\mathcal{T}_{CF} \subset 2^X$ be defined by

$$U \in \mathcal{T}_{CF} \text{ if and only if } \begin{cases} U = \emptyset & \text{or} \\ X \setminus U \text{ is a finite set.} \end{cases}$$

The subscript CF stands for "complement of finite sets". This topology is perhaps more natural to define in terms of it closed sets, namely $F \subset X$ is $\mathcal{T}_{CF}$-closed if and only if either $F = X$ or F is a *finite* subset of X.

It is instructive to check that $\mathcal{T}_{CF}$ is a topology. It is more natural to check that the collection of $\mathcal{T}_{CF}$-closed sets satisfies the properties of Remark 3.2. In this paragraph, let "closed" always mean $\mathcal{T}_{CF}$-closed. Clearly X and $\emptyset$ are closed. Suppose $\{F_\alpha\}_{\alpha \in A}$ is a collection of closed sets. If there exists $\alpha_0 \in A$ so that $F_{\alpha_0} \neq X$, then F_{α_0} is a finite set, and hence $\cap_\alpha F_\alpha \subset F_{\alpha_0}$ is finite, hence closed. Otherwise, $\cap_\alpha F_\alpha = X$, which is also closed. Similarly, if $F_1, \cdots, F_n$ is a finite collection of closed sets, then its union is either X (if one of the $F_i = X$) or a finite set (otherwise), hence also closed.

Example 3.14. In the special case $X = \mathbb{R}$ we will call the topology $\mathcal{T}_{CF}$ the *Zariski topology* and denote it $\mathcal{T}_Z$. This is a special case of the Zariski topology widely used in algebraic geometry, in which closed sets are common zeros of polynomials.

3.2.2. *Examples of Continuous Maps.* Let $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$ be topological spaces. It should be reasonable from the definition of continuity that, for a map $f : X \to Y$, having many open sets in $\mathcal{T}$ or few open sets in $\mathcal{T}'$ should make it easy for f to be continuous, while having few open sets in $\mathcal{T}$ or many in $\mathcal{T}'$ should make continuity hard. Let's see some examples.

Example 3.15. Let $\mathcal{T}$ be the discrete topology $\mathcal{T}_{disc}$. Then for *any* $\mathcal{T}'$ and for *any map* $f : X \to Y$, we have that $f : (X, \mathcal{T}_{disc}) \to (Y, \mathcal{T}')$ is continuous. For, given any $U \in \mathcal{T}'$, we have that $f^{-1}(U) \subset X$, hence $f^{-1}(U) \in \mathcal{T}_{disc}$, and f is continuous.

Example 3.16. Let $\mathcal{T}'$ be the indiscrete topology $\mathcal{T}_{ind}$. Then for *any topology* $\mathcal{T}$ and for *any map* $f : X \to Y$, we have that $f : (X, \mathcal{T}) \to (Y, \mathcal{T}_{ind})$ is continuous. For, if $U \in \mathcal{T}_{ind}$, then either $U = \emptyset$ or $U = Y$, so $f^{-1}(U) = \emptyset$ or X, in both cases elements of $\mathcal{T}$, so f is continuous.

Example 3.17. Let $(X, \mathcal{T})$ and $(Y, \mathcal{T}')$ be arbitrary, and let $f : X \to Y$ be a constant map: $f(x) = y_0$ for all $x \in X$. Then f is continuous: If $u \in \mathcal{T}'$, then

$$f^{-1}(U) = \begin{cases} X \text{ if } y_0 \in U, \\ \emptyset \text{ otherwise.} \end{cases}$$

In either case $f^{-1}(U) \in \mathcal{T}$ and f is continuous.

Example 3.18. Sometimes the only continuous maps are constant. For example, let X be any set but $\mathcal{T} = \mathcal{T}_{ind}$, and let $(Y, \mathcal{T}') = (\mathbb{R}, \mathcal{T}_E)$. If $f : (X, \mathcal{T}) \to (\mathbb{R}, \mathcal{T}_E)$ is continuous, then, for any $y \in \mathbb{R}$, $f^{-1}(\{y\})$ is either $\emptyset$ or X. Since f is a function, this means that for some $y_0 \in \mathbb{R}$, $f^{-1}(\{y_0\}) = X$, in other words, $f(x) = y_0$ for all $x \in X$ and f is a constant function. We will later see (after the discussion of connectedness) that if $\mathcal{T}' = \mathcal{T}_{disc}$, then any continuous map $f : (\mathbb{R}, \mathcal{T}_E) \to (Y, \mathcal{T}_{disc})$ is constant.

Example 3.19. Suppose $X = Y$. Then $id : (X, \mathcal{T}) \to (X, \mathcal{T}')$ is continuous if and only if $\mathcal{T}' \subset \mathcal{T}$. For example, $id : (\mathbb{R}, \mathcal{T}_E) \to (\mathbb{R}, \mathcal{T}_Z)$ is continuous (since finite sets are closed in the Euclidean topology), while $id : (\mathbb{R}, \mathcal{T}_Z) \to (\mathbb{R}, \mathcal{T}_E)$ is not continuous (since there are Euclidean closed sets that are neither finite nor all of $\mathbb{R}$).

Example 3.20. Let $f, g : \mathbb{R} \to \mathbb{R}$ be defined by $f(x) = x^2$ and $g(x) = \sin(x)$. Both are continuous functions $(\mathbb{R}, \mathcal{T}_E) \to (\mathbb{R}, \mathcal{T}_E)$. Check the following: both are continuous functions $(\mathbb{R}, \mathcal{T}_E) \to (\mathbb{R}, \mathcal{T}_Z)$; $f : (\mathbb{R}, \mathcal{T}_Z) \to (\mathbb{R}, \mathcal{T}_Z)$ is continuous, while $g : (\mathbb{R}, \mathcal{T}_Z) \to (\mathbb{R} : \mathcal{T}_Z)$ is *not* continuous.

3.3. **Limits.**

3.3.1. *Neighborhoods and Limits.*

Let $(X, \mathcal{T})$ be a topological space.

Definition 3.8. Let $x \in X$. A subset $U \subset X$ is called a *neighborhood of* x if and only if U is open and $x \in U$.

Remark 3.4. Many authors use the terminology *open neighborhood* for what we have called a neighborhood, and use the word neighborhood of x to mean a set which contains an open set containing x.

Neighborhoods can be used much as balls to extend the definitions of various familiar concepts of metric spaces. But some care is needed. For example, we could be tempted to make the following definition:

Definition 3.9. Let $\{x_n\}$ be a sequence in $(X, \mathcal{T})$. (Recall that this means that we have a function $\mathbb{N} \to X$ that to $n \in \mathbb{N}$ assigns $x_n \in X$.) If $x \in X$, we say that $\{x_n\}$ *converges to* x if and only if for every neighborhood U of x there exists $N \in \mathbb{N}$ so that $x_n \in U$ whenever $n > N$.

Then we are tempted to write $\lim\{x_n\} = x$. We have to be careful with this notation, since this definition need not give us what we think it does. If we write $\lim\{x_n\} = x$, we are tacitly assuming that limits are unique, that is, if $\{x_n\}$ converges to x and converges to y, then $x = y$, as we know to be true for metric spaces, see Theorem 1.3. Now consider the following example:

Example 3.21. Consider $(\mathbb{R}, \mathcal{T}_Z)$ as in Example 3.14, and let $x_n = n$. Pick any $x \in \mathbb{R}$, say pick $x = 7$. Then $\{n\}$ converges to 7: if U is a neighborhood of 7 and $U \neq \mathbb{R}$, then $U = \mathbb{R} \setminus F$ for some finite set $F \subset \mathbb{R}$, and $7 \neq F$. Let

M be the largest element of F. Then, if $n > M$, then $n \notin F$, thus $n \in U$. So $\{n\}$ converges to $x = 7$. The same argument holds for any $x \in \mathbb{R}$. So for any $x \in \mathbb{R}$, $\{n\}$ converges to x. Thus limits are not unique, and the notation $\lim\{n\} = x$ does not make sense.

3.3.2. *Hausdorff Spaces, Metrizable Spaces.* The proof of Theorem 1.3 could be rephrased so that it depends on the following: if (X, d) is a metric space, $x, y \in X$ and $x \neq y$, and $c = \frac{d(x,y)}{2}$, then $B(x, c) \cap B(y, c) = \emptyset$. This suggests the following condition for the uniqueness of limits:

Definition 3.10. A topological space $(X.\mathcal{T})$ is called a *Hausdorff space* if and only if given any two points $x, y \in X$, $x \neq y$, there exists a neighborhood U_x of x and a neighborhood U_y of y so that $U_x \cap U_y = \emptyset$.

Theorem 3.9. *Let $(X, \mathcal{T})$ be a Hausdorff space, and let $\{x_n\}$ be a sequence in X. If $\{x_n\}$ converges to x and converges to y, then $x = y$.*

Proof. Suppose $\{x_n\}$ converges both to x and y and $x \neq y$. Then there exist neighborhoods U_x, U_y of x, y respectively so that $U_x \cap U_y = \emptyset$. Since $\{x_n\}$ converges to x and y there exist $N_1, N_2 \in \mathbb{N}$ so that $x_n \in U_x$ for all $n > N_1$ and $x_n \in U_y$ for all $n > N_2$. Thus for all $n > \max(N_1, N_2)$ we have $x_n \in U_x \cap U_y$, contradicting $U_x \cap U_y = \emptyset$ □

Finally, the following terminology is useful and standard:

Definition 3.11. A topological space $(X, \mathcal{T})$ is *metrizable* if and only if there exists a metric d on X so that $\mathcal{T} = \mathcal{T}_d$, the metric topology.

Theorem 3.10. *Suppose $(X, \mathcal{T})$ is a metrizable topological space. Then it is Hausdorff.*

Proof. Let d be a metric on X so that $\mathcal{T}_d = \mathcal{T}$. If $x, y \in X$ and $x \neq y$, then $d(x, y) > 0$ and if $2c = d(x, y)$, then, by the triangle inequality, $B(x, c)$ and $B(y, c)$ are disjoint neighborhoods of x and y. □

Example 3.22. The discussion of Example 3.21 shows that $(\mathbb{R}, \mathcal{T}_Z)$ is not a Hausdorff space. In fact, if U and V are any two non-empty open sets, then $U \cap V \neq \emptyset$ since it is the complement of a finite set.

3.3.3. *Interior, Closure, Boundary.*

Definition 3.12. Let $(X, \mathcal{T})$ be a topological space and let $A \subset X$.

(1) The *interior* of A, denoted by A° is defined by

$$A^\circ = \cup\{U \subset X : U \text{ is open in } X \text{ and } U \subset A\}.$$

Equivalently, A° is the largest open set contained in A.

(2) The *closure* of A, denoted by $\bar{A}$ is defined by

$$\bar{A} = \cap\{F \subset X : F \text{ is closed and } A \subset F\}.$$

Equivalently, $\bar{A}$ is the smallest closed set containing A.

(3) The *boundary* (also called *frontier* of A), denoted by ∂A, is defined by $\partial A = \bar{A} \setminus A^\circ$.

These sets have the following alternative characterizations:

Theorem 3.11. *Let $A \subset X$.*

(1) *$x \in A^\circ$ if and only if there exists a neighborhood U of x with $U \subset A$.*
(2) *$x \in \bar{A}$ if and only if for every neighborhood U of x, $U \cap A \neq \emptyset$.*
(3) *$x \in \partial A$ if and only if for every neighborhood U of x, $U \cap A \neq \emptyset$ and $U \cap (X \setminus A) \neq \emptyset$.*
(4) *A is open if and only if $A = A^\circ$ and A is closed if and only if $A = \bar{A}$.*

Proof. For the first part, the definition $x \in A^\circ \Leftrightarrow x \in U$ for some U open, $U \subset A$, which is equivalent to U being a neighborhood of x contained in A. For the second part, from the definition we see that $x \notin \bar{A} \Leftrightarrow x \in X \setminus F$ for some F closed so that $A \subset F \Leftrightarrow x$ has a neighborhood U (namely, $X \setminus F$) so that $U \cap A = \emptyset$, which is the negation of the second statement, thus proving this statement. The third statement is equivalent, by the first two statements, to $x \in \bar{A} \setminus A^\circ$, thus $x \in \partial A$. The fourth statement is clear from the definitions. □

Definition 3.13. If $A \subset X$ and $x \in X$, then x is called a *limit point of* A if and only if it satisfies condition (2) of Theorem 3.11: for every neighborhood U of x, $U \cap A \neq \emptyset$.

Thus a set is closed if and only if it contains all its limit points. In a metric space, if x is a limit point of A, for every $n \in \mathbb{N}$ we could take $U = B(x, \frac{1}{n}$ and obtain that for each $n \in \mathbb{N}$ there exists $x_n \in A$ with $d(x, x_n) < \frac{1}{n}$. Thus x is the limit of the sequence $\{x_n\}$.

3.4. **Basis for a Topology.** Let $(X, \mathcal{T})$ be a topological space.

Definition 3.14. A subset $\mathcal{B} \subset 2^X$ is called a *basis for* $\mathcal{T}$ if and only if every element of $\mathcal{T}$ is a union of elements of $\mathcal{B}$. More explicitly, $\mathcal{B}$ is a basis if and only if, for each open set $U \in \mathcal{T}$ and for every $x \in U$ there exists $B \in \mathcal{B}$ such that $x \in B$ and $B \subset U$.

Example 3.23. Suppose (X, d) is a metric space. Then

$$\mathcal{B} = \{B(x, r) : x \in X,\ r > 0\}$$

is a basis for $\mathcal{T}_d$, the metric topology, and so is

$$\mathcal{B}' = \{B(x, \frac{1}{k}) : x \in X,\ k \in \mathbb{N}\}.$$

The fact that $\mathcal{B}$ is a basis is immediate from the definition of open sets in (X, d). To show that $\mathcal{B}'$ is a basis, it is enough to show that for each U open in (X, d) and for each $x \in U$ there exists $k \in \mathbb{N}$ so that $B(x, \frac{1}{k}) \subset U$. This is easy to do: by the definition of open set, there exists $r > 0$ so that $B(x, r) \subset U$. Choose $k \in \mathbb{N}$ so that $\frac{1}{k} < r$. Then $B(x, \frac{1}{k}) \subset B(x, r) \subset U$, so we are done. This shows that $\mathcal{B}'$ is also a basis for $\mathcal{T}_d$.

Example 3.24. Specializing the above example to $\mathbb{R}^n$ with d each of the metrics $d_{(1)}$, $d_{(2)}$, $d_{(\infty)}$ we obtain a basis $\mathcal{B}_d$ for the topology of $\mathbb{R}^n$ by balls of different shapes and all possible radii, and the corresponding balls $\mathcal{B}'_d$ of radii reciprocals of natural numbers. Moreover, for each of these metrics d could also use the collection

$$\mathcal{B}^*_d = \{B_d(x, \frac{1}{k}) : x \in \mathbb{Q}^n,\ \ k \in \mathbb{N}\}.$$

Note that the centers of the balls have all their coordinates rational. The interest of these collections is that it each is a *countable* collection that generates the uncountable collection of open sets in $\mathbb{R}^n$.

We now prove that $\mathcal{B}^*_d$ is a basis for the Euclidean topology $\mathcal{T}_E$ on $\mathbb{R}^n$. To prove this it is enough to prove that any ball $B(x, r)$ in the metric d is a union of elements of $\mathcal{B}^*_d$, in other words, given any $y \in B(x, r)$ there exists $x \in \mathbb{Q}^n$ and $k \in \mathbb{N}$ so that $y \in B(z, \frac{1}{k}) \subset B(x, r)$. Since there exists an r' so that $B(y, r') \subset B(x, r)$ (can take $r' = r - d(x, y)$, see the proof of Theorem 3.1), it enough to find z, k so that $y \in B(z, \frac{1}{k}) \subset B(y, r')$, in other words, just need to check the statement for $y = x$ the center of the ball. To reiterate, it suffices to prove that *for all $x \in \mathbb{R}^n$ and for all $r > 0$ there exists $z \in \mathbb{Q}^n$ and $k \in \mathbb{N}$ so that $x \in B(z, \frac{1}{k}) \subset B(x, r)$.*

Suppose we know the density of $\mathbb{Q}^n$ in $\mathbb{R}^n$: *for all $x \in \mathbb{R}^n$ and for all $\epsilon > 0$ there exists $z \in \mathbb{Q}^n$ so that $d(x, z) < \epsilon$.* Then the above statement is easy to prove: Given x and r, there exists $z \in \mathbb{Q}^n$ such that $d(x, z) < \frac{r}{2}$ and there exists $k \in \mathbb{N}$ so that $\frac{1}{k} < \frac{r}{2}$. Then, if $d(y, z) < \frac{1}{k}$, then

$$d(y, x) \leq d(y, z) + d(z, x) < \frac{r}{2} + \frac{r}{2} = r$$

Thus $x \in B(z, \frac{1}{k}) \subset B(x, r)$, as desired, so $\mathcal{B}^*_d$ is a basis for $\mathbb{R}^n$.

We assume that the density statement is known for $\mathbb{R}$: for all x in $\mathbb{R}$ and all $\epsilon > 0$ there exists $z \in \mathbb{Q}$ so that $|x - z| < \epsilon$. The statement immediately follows for $\mathbb{R}^n$ and the metric $d_{(\infty)}$ by applying the statement for $\mathbb{R}$ in each coordinate: for any $x = (x_1, \cdots, x_n) \in \mathbb{R}^n$ and $\epsilon > 0$, for each i there exists $z_i \in \mathbb{Q}$ so that $|x_i - z_i| < \epsilon$, thus, letting $z = (z_1, \cdots, z_n)$, $d_{(\infty)}(x, z) = \max\{|x_i - z_i|\} < \epsilon$. Finally, it d is $d_{(1)}$ or $d_{(2)}$, then use the comparisons of Example 1.13. For example, given x and ϵ, to find $z \in \mathbb{Q}^n$ with $d_{(2)}(x, z) < \epsilon$, find $z \in \mathbb{Q}^n$ with $d_{(\infty)}(x, z) < \frac{\epsilon}{\sqrt{n}}$, then by (2) of Example 1.13, $d_{(2)}(x, z) < \epsilon$.

Remark 3.5. One use of a basis is that many statements have only to be checked for elements of the basis. For example, if we are given a basis $\mathcal{B}_Y$ for $\mathcal{T}_y$, to check that a map $f : (X, \mathcal{T}_X) \to (Y, \mathcal{T}_y)$ is continuous it is enough to check that $f^{-1}(B)$ is open for all $B \in \mathcal{B}_Y$. Namely, if U is open in Y, then $U = \cup_\alpha B_\alpha$ for some collection $\{B_\alpha\}$ of elements of $\mathcal{B}$, thus $f^{-1}(U) = f^{-1}(\cup_\alpha U_\alpha) = \cup_\alpha f^{-1}(B_\alpha)$ is open in X since it is a union of open sets.

Another example of the same principle: x is a limit point of A if and only if $B \cap A \neq \emptyset$ for all $B \in \mathcal{B}$ so that $x \in B$.

3.4.1. *Defining a Topology from a Basis.* It is important to be able to reverse the above procedure. In other words: take a non-empty set X and a collection $\mathcal{B} \subset 2^X$, and try to *define a topology on* X by declaring $\mathcal{B}$ to be a basis. More precisely, given $\mathcal{B}$, define $\mathcal{T} \subset 2^X$ to be the set of all unions of elements of $\mathcal{B}$, that is, define $U \subset X$ to be an element of $\mathcal{T}$ if and only if for all $x \in U$ there exists $B \in \mathcal{B}$ so that $x \in B$ and $B \subset U$. We need to know that this is a topology, namely that it satisfies the three properties of Definition 3.4. It is clear that half of (1) and (2) are satisfied: $\emptyset \in \mathcal{T}$ and $\mathcal{T}$ is closed under arbitrary unions. But it need not be true that $X \in \mathcal{T}$ or that (3) is satisfied: $\mathcal{T}$ need not be closed under finite intersections. But if we add these as an assumption, then $\mathcal{T}$ is a topology with basis $\mathcal{B}$:

Theorem 3.12. *Let X be a non-empty set and let $\mathcal{B} \subset 2^X$ be a collection of subsets that satisfies:*

1. *For every $x \in X$ there exists $B \in \mathcal{B}$ such that $x \in B$.*
2. *For every $B_1, B_2 \in \mathcal{B}$ and for every $x \in B_1 \cap B_2$ there exists $B \in \mathcal{B}$ such that $x \in B$ and $B \subset B_1 \cap B_2$.*

Let $\mathcal{T} = \{U \subset X : \text{for all } x \in U \text{ there exists } B \in \mathcal{B} \text{ with } x \in B \text{ and } B \subset U\}$. Then $\mathcal{T}$ is a topology on X and $\mathcal{B}$ is a basis for $\mathcal{T}$.

Proof. The first condition says that $X \in \mathcal{T}$ and the second condition implies that $\mathcal{T}$ is closed under intersections of two sets: if $U_1, U_2 \in \mathcal{T}$ and $x \in U_1 \cap U_2$, then there exist $B_1, B_2 \in \mathcal{B}$ so that $x \in B_1 \subset U_1$ and $x \in B_2 \subset U_2$. Since $x \in B \subset B_1 \cap B_2 \subset U_1 \cap U_2$, we have that $U_1 \cap U_2 \in \mathcal{T}$ whenever $U_1, U_2 \in \mathcal{T}$. A straightforward induction argument then implies that (3) of Definition 3.4 holds. It is clear that $\emptyset \in \mathcal{T}$, and, if for all $\alpha \in A$ we have $U_\alpha \in \mathcal{T}$, and if $x \in \cup_{\alpha \in A} U_\alpha$, then $x \in U_{\alpha_0}$ for some $\alpha_0 \in A$, so there exists $B \in \mathcal{B}$ so that $x \in B \subset U_{\alpha_0} \subset \cup_{\alpha \in A} U_\alpha$, thus $\cup_{\alpha \in A} U_\alpha \in \mathcal{T}$ and (2) of Definition 3.4 is also satisfied. Thus $\mathcal{T}$ is a topology on X. By the definition of $\mathcal{T}$ it is clear that $\mathcal{B}$ is a basis for $\mathcal{T}$. □

3.4.2. *The Product Topology.* Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces. There is a natural way to topologize the product $X \times Y$, but this natural way requires the concept of basis. Let

$$\mathcal{B} = \{U \times V : U \in \mathcal{T}_X \text{ and } V \in \mathcal{T}_Y\} \tag{3.3}$$

It is easy to check that $\mathcal{B}$ satisfies the conditions of Theorem 3.12. Namely, if $(x, y) \in X \times Y$, since $X \times Y \in \mathcal{B}$, (1) is clearly satisfied. If $B_1 = U_1 \times V_1$ and $B_2 = U_2 \times V_2$ and $(x, y) \in B_1 \cap B_2$, then $x \in U_1 \cap U_2$ and $y \in V_1 \cap V_2$, so letting $B = (U_1 \cap U_2) \times (V_1 \cap V_2)$, we have that $(x, y) \in B \subset B_1 \cap B_2$, thus (2) is also satisfied, and $\mathcal{B}$ is the basis for a unique topology $\mathcal{T}_{X\times Y}$ on $X \times Y$. This topology is called the *product topology* on $X \times Y$.

Note that this collection $\mathcal{B}$ is actually *closed under finite intersections*, because of the identity (which holds for arbitrary subsets of X and Y, not just open sets):

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2) \tag{3.4}$$

that we used above to prove (2). But $\mathcal{B}$ is *not closed under unions*. This is easily visualized in $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. The elements of $\mathcal{B}$ are "rectangles " but unions of rectangles need not be rectangles.

Remark 3.6. We could modify the definition of $\mathcal{B}$ in Equation 3.3 by letting $\mathcal{B}_X$ be a basis of $\mathcal{T}_X$ and $\mathcal{T}_Y$ be a basis for $\mathcal{T}_Y$ and defining $\mathcal{B}' \subset \mathcal{B}$ by

$$\mathcal{B}' = \{U \times V : U \in \mathcal{B}_X \text{ and } V \in \mathcal{B}_Y\}$$

It is easy to check that $\mathcal{B}'$ also satisfies the conditions of Theorem 3.12 and that $\mathcal{B}'$ is also a basis for the product topology $\mathcal{T}_{X\times Y}$. These verifications are left as an exercise. They depend, of course, on the above identity (3.4) for intersections of products.

Remark 3.7. In Subsection 1.2.2 we defined the cartesian product of metric spaces. The metric topology resulting from that definition and the product topology just defined are the same topology. It would be an instructive exercise to verify this. Keep in mind the basic example of $\mathbb{R} \times \mathbb{R}$ and $(\mathbb{R}^2, d_{(\infty)})$.

Here are two useful properties of the product topology. We use the notation p_X and p_Y for the projection maps $p_X : X \times Y \to X$ and $p_Y : X \times Y \to Y$ defined by $p_X(x, y) = x$ and $p_Y(x, y) = y$.

Theorem 3.13. *Let $(X, \mathcal{T}_X)$, $(Y, \mathcal{T}_Y)$ and $(Z, \mathcal{T}_Z)$ be topological spaces.*

(1) *The product topology $\mathcal{T}_{X\times Y}$ is the smallest topology that makes both projections p_X and p_Y continuous.*
(2) *A map $f : Z \to X \times Y$ is continuous with respect to the product topology if and only if both compositions $p_X \circ f$ and $p_Y \circ f$ are continuous.*

Proof. For the first part we note that $p_X : X \times Y \to X$ is continuous if and only if for all open $U \subset X$, $U \times Y$ is open in $X \times Y$. Since this is open in the product topology, p_X is continuous. Similarly, p_Y is continuous if and only if for all open $V \subset Y$, $X \times V$ is continuous, so p_Y is also continuous in the product topology. Moreover, if $\mathcal{T}$ is any topology which makes p_X and p_Y continuous, then it must contain all the sets $\{U \times Y : U \in \mathcal{T}_X\}$ and $\{X \times V : V \in \mathcal{T}_Y\}$, therefore $\mathcal{T}$ must contain all their two-fold intersections $\{U \times V : U \in \mathcal{T}_X \text{ and } V \in \mathcal{T}_Y\}$. Since this is a basis for $\mathcal{T}_{X \times Y}$ we must have $\mathcal{T}_{X \times Y} \subset \mathcal{T}$, thus proving the first statement.

For the second part, first note that f continuous certainly implies that $p_X \circ f$ and $p_y \circ Y$ are continuous, since compositions of continuous maps are continuous. For the converse, if $p_X \circ f$ is continuous, then $(p_X \circ f)^{-1}(U) = f^{-1} \circ p_X^{-1}(U) = f^{-1}(U \times Y$ is open for each $U \in \mathcal{T}_X$ and similarly $f^{-1}(X \times V)$ is open for all $V \in \mathcal{T}_Y$, thus the same is true for their intersections: all $f^{-1}(U \times V)$ are open. Since these sets form a basis for $\mathcal{T}_{X \times Y}$, f is continuous.

□

4. Subspaces and Quotient Spaces

Let X and Y be sets and let $f : X \to Y$ be a map. We know from Examples 3.15 and 3.16 that f is continuous if either the discrete topology is given to X (the largest possible topology) or if the indiscrete topology is given to Y (the smallest possible topology). We want to find optimal intermediate topologies that make f continuous under the assumption of a given topology on the domain or target.

Theorem 4.1. *Let X and Y be sets and let $f : X \to Y$.*

(1) *Given a topology $\mathcal{T}_Y$ on Y there is a smallest topology $\mathcal{T}_X$ on X that makes f continuous, namely $\mathcal{T}_X = \{f^{-1}(U) : U \in \mathcal{T}_Y\}$.*
(2) *Given a topology $\mathcal{T}_X$ on X there is a largest topology $\mathcal{T}_Y$ that makes f continuous., namely $\mathcal{T}_Y = \{U \subset Y : f^{-1}(U) \in \mathcal{T}_X\}$. (In this case we usually only consider the case where f is surjective.)*

Proof. To prove (1), note that if we let $\mathcal{T}_X$ be as in the statement, then $X = f^{-1}(Y) \in \mathcal{T}_X$ and $\emptyset = f^{-1}(\emptyset) \in \mathcal{T}_X$. Since $f^{-1}(\cup U_\alpha) = \cup f^{-1}(U_\alpha)$ and $f^{-1}(U) \cap f^{-1}(V) = f^{-1}(U \cap V)$ it follows that $\mathcal{T}_X$ is closed under arbitrary unions and finite intersections (since $\mathcal{T}_Y$ is), thus $\mathcal{T}_X$ is a topology on X. If $\mathcal{T}$ is any topology on X so that f is continuous, then for all $U \in \mathcal{T}_Y$ we have that $f^{-1}(U) \in \mathcal{T}$. Therefore $\mathcal{T}_Y \subset Y$, in other words, $\mathcal{T}_Y$ is the smallest topology making f continuous.

To prove (2), we check, using the same ingredients as in the first part, that $\mathcal{T}_Y$ is a topology on Y. If $\mathcal{T}$ is any topology on Y so that f is continuous, then, given $U \in \mathcal{T}$, we must have that $f^{-1}(U) \in \mathcal{T}_X$, in other words, $U \in \mathcal{T}_Y$, therefore $\mathcal{T} \subset \mathcal{T}_Y$ and $\mathcal{T}_Y$ is the largest topology that makes f continuous.

Note that if $U \subset Y \setminus f(X)$ is *any* subset, then $f^{-1}(U) = \emptyset \in \mathcal{T}_X$, thus $U \in \mathcal{T}_Y$. Thus $\mathcal{T}_Y$ gives $Y \setminus f(X)$ the discrete topology. Since this has nothing to do with the map f, it is only reasonable to consider the case where f is *surjective* in part (2).

□

Remark 4.1. By taking complements, we could equally well have defined the topologies of Theorem 4.1 in terms of *closed* sets. In other words, for part (1), we could have defined $\mathcal{T}_X$ as the topology whose closed sets are $\{f^{-1}(F) : F \subset Y \text{ is closed in } \mathcal{T}_Y\}$. Recall that this means that $\mathcal{T}_X = \{X \setminus f^{-1}(F) : F \subset Y \text{ is closed in } \mathcal{T}_Y\}$. Then $\mathcal{T}_X$ is a topology on X and it is the smallest topology that makes f continuous.

Similarly, for part (2) of Theorem 4.1, we could define $\mathcal{T}_Y$ as the topology whose closed sets are $\{F : f^{-1}(F) \text{ is closed in } \mathcal{T}_X\}$. The equivalence of the two definitions in both parts follows, as usual, from the identity $f^{-1}(Y \setminus F) = X \setminus f^{-1}(F)$.

4.1. **The Subspace Topology.** We specialize the first part of Theorem 4.1 to the case that $X \subset Y$ and f is the inclusion. The resulting topology of X is called the *subspace topology*. More explicitly, observing that in this case, for $U \subset Y$, $f^{-1}(U) = U \cap X$, we get the following description of the topology:

Definition 4.1. Let $(Y, \mathcal{T}_Y)$ be a topological space, and let $X \subset Y$. The *subspace topology* $\mathcal{T}_X$ on X is defined to be $\mathcal{T}_X = \{U \cap X : U \in \mathcal{T}_Y\}$.

The subspace topology can be hard to picture. We give a couple of situations where it is a familiar topology.

Recall that in (1.2.1) we defined a subspace of a metric space. in the present context, suppose $\mathcal{T}_Y$ is the metric topology of a metric d on Y and let $d' = d|_{X \times X}$ be the subspace metric on X.

Theorem 4.2. *Let (Y, d) be a metric space, let $X \subset Y$ and $\mathcal{T}_Y = \mathcal{T}_d$ the metric topology. Then the subspace topology $\mathcal{T}_X$ agrees with the metric topology $\mathcal{T}_{d'}$ of the subspace metric $d' = d|_{X \times X}$.*

Proof. Observe that if $x \in X$ and $r > 0$, then $B_X(x, r) = \{y \in X : d(x, y) < r\} = \{y \in Y : d(x, y) < r\} \cap X = B_Y(x, r) \cap X$, thus $B_X(x, r)$ is open in the subspace topology, thus any open set in the metric topology is open in the subspace topology. Conversely, if $U \subset X$ is open in the subspace topology and $x \in U$, then there exists an open set $V \subset Y$ so that $U = V \cap X$. Since V is open, there exists $r > 0$ so that $B_Y(x, r) \subset V$. Then $B_X(x, r) = B_Y(x, r) \cap X \subset U = V \cap X$, thus U is open in the metric topology of X. □

Another situation where it is simple to see the subspace topology is the following:

Theorem 4.3. *Suppose X is open in Y. Then a subset $U \subset X$ is open in X if and only if it is open in Y. Similarly, if X is closed in Y, a subset $F \subset X$ is closed in X if and only if it is closed in Y.*

Proof. Suppose X is open in Y and $U \subset X$ is open in X. Then there exists an open set $V \subset Y$ such that $U = X \cap V$. Since X is open in Y, so is $X \cap V$, so U is open in Y. Conversely, if $U \subset X$ is open in Y, then $U = X \cap U$ is open in X. The proof for closed subsets is similar. □

4.2. **The Quotient Topology.** We now turn to the second part of Theorem 4.1. This theorem justifies making the following definition:

Definition 4.2. Let $(X, \mathcal{T}_X)$ be a topological space and let $f : X \to Y$.

(1) The *quotient topology*, also called the *identification topology* on Y is the topology $T_y = \{U \subset Y : f^{-1}(U) \in \mathcal{T}_X\}$.
(2) A surjective continuous map $f : X \to Y$ between topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ is called an *identification* if $\mathcal{T}_Y$ is the quotient (or identification) topology just defined.

Let us keep some concrete examples in mind as we develop this concept.

Example 4.1. Let $S^1 \subset \mathbb{R}^2$ be the unit circle. Define $f : \mathbb{R} \to S^1$ by $f(t) = (\cos t, \sin t)$. Let $f_1 = f|_{[0,2\pi]} : [0, 2\pi] :\to S^1$ and let $f_2 = f|_{[0,2\pi)} : [0, 2\pi) \to S^1$. All three of f , f_1 and f_2 are continuous surjections. Let's prove that f and f_1 are identifications, but f_2 is not. To show that a continuous map f is an identification is the same as showing that for all subsets A of the target, $f^{-1}(A)$ open implies that A is open. For f this is true, because f has the property that for any open $V \subset \mathbb{R}$, $f(V)$ is open in S^1. This is clear because it is clear that small (meaning, say, of length less than π) open intervals in $\mathbb{R}$ have open image in S^1, and all open sets are unions of small intervals. So, if $A \subset S^1$ has the property that $f^{-1}(A)$ is open in $\mathbb{R}$, then $f(f^{-1}(A))$ is open in S^1. But for a *surjective map* we have that $f(f^{-1})(A) = A$, thus A is open.

To prove that f_1 is an identification, let $A \subset S^1$ and suppose $f_1^{-1}(A)$ is open in $[0, 2\pi]$. If $t \in f_1^{-1}(A)$, we consider two cases:

(1) $t \in (0, 2\pi)$. Then there exists $\epsilon > 0$ so that $(t - \epsilon, t + \epsilon) \subset (0, 2\pi)$ and $f_1((t - \epsilon, t + \epsilon))$ is a neighborhood of $f(t)$ contained in A.
(2) $t = 0$ or $t = 2\pi$. Then we must have that the other endpoint 2π or 0 is also in $f_1^{-1}(A)$, since $f_1(0) = f_1(2\pi) = (1, 0) \in S^1$. Then there is an $\epsilon > 0$ so that $[0, \epsilon) \cup (2\pi - \epsilon] \subset f_1^{-1}(A)$, therefore $f_1([0, \epsilon) \cup (2\pi - \epsilon, 2\pi]) = f((-\epsilon, \epsilon))$ is a neighborhood of $f_1(t)$ which is contained in A.

Therefore, in both cases we found a neighborhood of each point of A which is contained in A, so A is an open set and f_1 is an identification.

But for f_2 the situation is different: If $A = \{(\cos t, \sin t) : 0 \le t < \pi\}$, then A is not open in S^1 but $f_2^{-1}(A) = [0, \pi)$ which is open in $[0, 2\pi)$. Therefore f_2 is not an identification.

Let us formalize the proof just given that f is an identification:

Definition 4.3. A map $f : X \to Y$ of topological spaces is called an *open map* if and only if, for all open $U \subset X$, $f(U) \subset Y$ is open. Similarly, f is called a *closed map* if and only if, for all closed $F \subset X$, $f(F) \subset Y$ is a closed set.

Example 4.2. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x, y) = x$ (projection to the first factor). Then f is an open map (because $f(B(x, y), r) = (x - r, x + r)$ is open in $\mathbb{R}$ and the collection $\{B((x, y), r) : (x, y) \in \mathbb{R}^2, r > 0\}$ is a basis for the topology of $\mathbb{R}^2$). But f is not a closed map: let $F = \{xy = 1\}$ (a hyperbola). As the zero set of a continuous function it is a closed set, but $f(F) = \{x \neq 0\}$ which is not closed in $\mathbb{R}$.

The argument given for f in Example 4.1 shows the following:

Theorem 4.4. *Let X and Y be topological spaces and let $f : X \to Y$ be a continuous surjection and an open map. Then f is an identification. Similarly, if $f : X \to Y$ is a continuous surjection and a closed map, then f is an identification.*

Proof. We have to prove that $U \subset Y$ is open if and only if $f^{-1}(U)$ is open. Since f is continuous, U open implies that $f^{-1}(U)$ is open. Since f is an open map, $f^{-1}(U)$ open implies that $f(f^{-1}(U))$, and since f is surjective, $f(f^{-1}(U)) = U$, thus U is open. Similarly, if f is a continuous surjection and a closed map, we prove in the same way that $F \subset Y$ is closed if and only if $f^{-1}(F)$ is closed, hence, by Remark 4.1, Y has the quotient topology and f is an identification.

□

Remark 4.2. Theorem 4.4 gives sufficient conditions for f to be an identification. But these are not necessary conditions. For example, the map f_1 of Example 4.1 is not an open map: $[0, \pi)$ is open in $[0, 2\pi]$ but $f_1([0, \pi))$ is not open in S^1. Also the map f of the same example is open but not closed: Let $F = \{\frac{1}{n} + 2\pi n : n \in \mathbb{N}\}$. Then F is a discrete subset of $\mathbb{R}$, hence closed, but $f(F) = \{(cos(\frac{1}{n}), \sin(\frac{1}{n}))\}$ is not closed since it does not contain its limit point $(1, 0)$.

The reason that the terms "quotient" or "identification" topology are used is that we often apply this to quotients by equivalence relations. We could also think of quotients as making suitable identifications. We could say the following:

Remark 4.3. Let X and Y be two sets. Then the following are equivalent:

(1) A surjective map $f : X \to Y$.
(2) A partition of X into disjoint sets indexed by Y, that is, a collection $\{X_y\}_{y\in Y}$ where, for each y, $X_y \subset X$, $X = \cup_{y\in Y} X_y$, and $X_{y_1} \cap X_{y_2} = \emptyset$ whenever $y_1 \neq y_2$.
(3) An equivalence relation on X with equivalence classes in one to one correspondence with Y.

The equivalences are easy to see: Given (1), define the partition in (2) by $X_y = f^{-1}(y)$, and given the partition (2), define $f : X \to Y$ by $f(x) = y$ if and only if $x \in X_y$. Thus (1) is equivalent to (2). Similalrly, given a partition (2), define an equivalence relation on X by $x_1 \sim x_2$ if and only if there is a $y \in Y$ so that $x_1 \in X_y$ and $x_2 \in X_y$. This is easily checked to be an equivalence relation, and its equivalence classes are in one to one correspondence with Y, thus we have (3). Finally, given (3), define the partition of X to be the equivalence classes. Since these are in one to one correspondence with Y, we can label them as $\{X_y\}_{y\in Y}$, and this gives (2).

The following theorem gives a useful characterization of the quotient topology.

Theorem 4.5. *Let X, Y and Z be topological spaces. Suppose that maps f and g are given as in the following diagram, and that f is an identification.*

$$\begin{tikzcd}
X \arrow[d, "f"] \arrow[dr, "g \circ f"] & \\
Y \arrow[r, "g"'] & Z
\end{tikzcd}$$

Then g is continuous if and only if $g \circ f$ is continuous.

.

Proof. If g is continuous then certainly $g \circ f$ is continuous by Corollary 3.2. What is specific to the identification topology is the converse, which is proved as follows: if $g \circ f$ is continuous, then for each open $U \subset Z$, $(g \circ f)^{-1}(U)$ is open in X. But $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$, thus, since f is an identification, $g^{-1}(U)$ is open in Y, so g is continuous. □

.

This theorem is usually applied in the following equivalent form. Suppose that $f : X \to Y$ is an identification as in the theorem, and suppose we are given a continuous map $h : X \to Z$ with the property that h is constant on the fibers of f (the sets $f^{-1}(y)$, $y \in Y$. In other words, suppose that $h(x) = h(x')$ whenever $f(x) = f(x')$. Then we can define a map $g : Y \to Z$ by as follows: given $y \in Y$, choose $x \in X$ so that $f(x) = y$, and define

$g(x) = h(x)$. The above condition implies that this is well=defined: Given $y \in Y$, if we choose x' so that $f(x') = y$, then $f(x) = f(x')$, so, by the assumption on h, $h(x) = h(x')$, so the point $g(y)$ depends just on y, and not on the representative x chosen to define $g(y)$. We then have the following theorem:

Theorem 4.6. *In the following diagram, suppose that X, Y and Z are topological spaces, f is an identification and h is constant on the fibers of f, so that the map g as in the above discussion is well-defined.*

$$\begin{tikzcd}
X \arrow[r, "h"] \arrow[d, "f"] & Z \\
Y \arrow[ur, "g"] &
\end{tikzcd}$$

Then g is continuous if and only if h is continuous.

Proof. Since, by the definition of g, $h = g \circ f$, this is the same as Theorem 4.5. □

Example 4.3. We can apply this Theorem to the identification $f : \mathbb{R} \to S^1$ of Example 4.1. Say we take $Z = \mathbb{R}$, then we obtain the familiar fact that there is a one-to-one correspondence between continuous periodic functions on $\mathbb{R}$, with period 2π, and continuous functions on the circle S^1.

4.3. **Surfaces as Identification Spaces.** We now apply Theorem 4.6 to define various surfaces. The procedure is in some cases similar to what we saw in Example 4.1 when we saw the circle could be described either as a quotient of $\mathbb{R}$ or as a quotient of $[0, 2\pi]$. See Chapter 4 of [5] for more discussion (and pictures) of this procedure.

Example 4.4. We can picture the torus (= surface of a doughnut) as a surface of revolution in $\mathbb{R}^2$, obtained by rotating a circle of radius one centered at $(2, 0, 0)$ about the z-axis. As such it has parametric equations $(x, y, z) = ((2+\cos\phi)\cos\theta, (2+\cos\phi)\sin\theta, \sin\phi)$, $0 \le \theta, \phi \le 2\pi$. In the same way that we showed in Example 4.1 that S^1 is an identification space of $\mathbb{R}$, we can show that the torus is an identification space of $\mathbb{R}^2$, where the equivalence relation on $\mathbb{R}^2$ is $(x, y) \sim (x+2\pi m, y+2\pi n)$ for all $m, n \in \mathbb{Z}$. We can picture the equivalence classes as translating (x, y) by any element of $2\pi\mathbb{Z}^2$, where $\mathbb{Z}^2 \subset \mathbb{R}^2$ is the integral lattice. From now on it would be convenient to reparametrize to get rid of the factors of 2π, and let's agree that by *torus* we mean the quotient of $\mathbb{R}^2$ by the equivalence relation $(x, y) \sim (x + m, y + n)$ for all $m, n \in \mathbb{Z}$. This quotient space is denoted $\mathbb{R}^2/\mathbb{Z}^2$, and we write $p_1 : \mathbb{R}^2 \to \mathbb{R}^2/\mathbb{Z}^2$ for the natural map ("projection") that to (x, y) assigns its equivalence class $(x, y) + \mathbb{Z}^2$.

Now, there is a more economical way to represent the torus, just as we did with S^1 in Example 4.1. Namely, let $S = [0,1] \times [0,1]$ be the unit square. Then the composition of the inclusion of S in $\mathbb{R}^2$ with the projection of $\mathbb{R}^2$ to $\mathbb{R}^2/\mathbb{Z}^2$ is surjective, and identifies certain points on the boundary of S: let $\sim$ be the equivalence relation $(x,0) \sim (x,1)$ and $(0,y) \sim (1,y)$ on S (meaning that these are the equivalence classes with more than one element, the points (x,y) with $0 < x, y < 1$ are equivalent just to themselves). Note also that $(0,0) \sim (0,1) \sim (1,0) \sim (1,1)$, thus this one equivalence class has 4 elements, while the equivalence classes $(x,0) \sim (x,1)$ for $0 < x < 1$ and $(0,y) \sim (1,y)$, for $0 < y < 1$ have two elements. We write $p : S \to S/\sim$ for the natural map that to (x,y) assigns its equivalence class.

The conventional way of describing this identification space is to draw a square and indicate by arrows which sides are identified and how. Sides with similar arrows are identified, imagining that we travel at the same speed on both sides in direction of the arrow, and identify corresponding points. The identfiication space $T = S/\sim$ just defined would be indicated as follows:

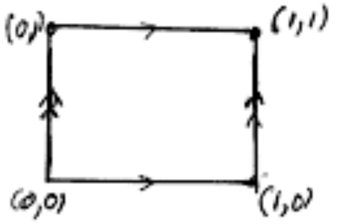

Figure 4.1. Torus

We will see more examples below of how these conventions are used to define identification spaces.

The above convention describes the set $S/\sim$. The topology on this set is the identification topology resulting from the topology on S. This just follows from the definitions, but, if we want to picture the topology explicitly, we picture the sets $p^{-1}(U)$. It is enough to give a basis. If $(x,y) \in S^o$, the interior of S, then we can take balls $B((x,y),\epsilon) \subset S^o$ for small enough ϵ. If we take a point $(x,0)$ with $0 < x < 1$, then any set $p^{-1}(U)$ that contains $(x,0)$ must also contain the equivalent point $(x,1)$ and a neighborhood of that point. So in our basis we could choose neighborhoods of $p((x,0))$ to have pre-image $B_S((x,0),\epsilon) \cup B_S((x,1),\epsilon)$ for $\epsilon(x)$ sufficiently small. By B_S we mean a ball in the metric space S as a subspace of $\mathbb{R}^2$. Similarly for $(0,y)$ we could choose $B_S((0,y),\epsilon) \cup B_S((1,y),\epsilon)$. The picture is:

Finally for a corner we get $B_S((0,0)\epsilon) \cup B_S((1,0),\epsilon) \cup B_S((0,1),\epsilon) \cup B_S((1,1),\epsilon)$:

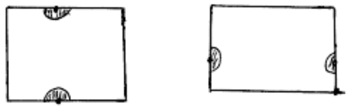

Figure 4.2. Neighborhoods in Identification Space

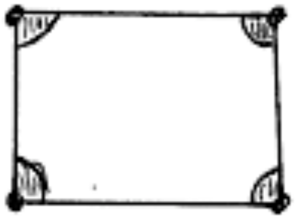

Figure 4.3. Neighborhood of the Corner

The two description we have given of the torus T can be summarized in the following commutative diagram

$$\begin{tikzcd}
S \subset \mathbb{R}^2 \arrow[r, "p_1"] \arrow[d, "p"'] & \mathbb{R}^2/\mathbb{Z}^2 \\
T = S/\sim \arrow[ur, "g"'] &
\end{tikzcd}$$

By Theorem 4.6 we see that g is continuous. Moreover, from the very definition of $S/\sim$, we see that g is a bijection: each point of S contains at least one member of each equivalence class in $\mathbb{R}^2/\mathbb{Z}^2$, thus g is surjective. And two points in S are equivalent under $\sim$ if and only if they are equivalent in $\mathbb{R}^2$ under translation by the integral lattice $\mathbb{Z}^2$, so g is injective. From the definitions of the topologies we see that g is an open map: The images of the basic open sets just described for T are the sets whose pre-image under p_1 are the sets $\cup\{B((x+m, y+n)\epsilon) : m, n \in \mathbb{Z}\}$, which are open in $\mathbb{R}^2$. Since an open continuous bijection is a homeomorphism, we see that g is a homeomorphism.

Since we have these two descriptions of the torus, we choose the more economical one as the official definition:

Definition 4.4. The *torus* T is the identification space $T = S/\sim$ of the unit square S as just defined in the previous example.

We can use the same pattern to define other surfaces. For example:

Definition 4.5. The *Klein Bottle* K is the identification space $K = S/\sim$ of the unit square S where $(x, 0) \sim (x, 1)$ and $(0, y) \sim (1, 1 - y)$, with the quotient topology.

Thus using the convention we explained above when describing the torus, K can be described by the diagram

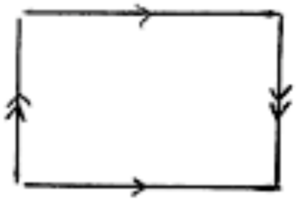

Figure 4.4. The Klein Bottle

Note that the horizontal arrows go in the same direction indicating $(x, 0) \sim (x, 1)$ while the vertical arrows go in the opposite direction indicating $(0, y) \sim (1, 1 - y)$. The quotient topology can be explicitly defined and illustrated in a fashion analogous to the discussion of the torus in Example 4.4.

While the identification of the torus T with a surface in $\mathbb{R}^3$ is easy to visualize (see, for example, p. 300 of [6]), the Klein bottle can only be realized as a surface with self-intersections. See p. 308 of [6] for pictures and explanation.

Here's a more familiar surface. Make sure you make a paper model to make the definition concrete.

Definition 4.6. The *Möbius Band* M is the identification space $M = [0, 1] \times [-1, 1]/(0, y) \sim (1, -y)$, with the quotient topology. (This is also called the *closed Möbius band*. A variation of the definition would be the *open Möbius band*, the quotient $[0, 1] \times (-1, 1)/(0, y) \sim (1, -y)$)

Thus the identification picture for M would be

Follow the identifications to verify that the top and bottom line combine to give a closed curve (homeomorphic to a circle). In fact, the horizontal line in the middle, $\{(x, 0) : 0 \leq x \leq 1\}$, is a circle, and every pair of horizontal lines equidistant from this central line also gives a circle (twice as long as the middle one). Verify this in the identification picture, and also in a paper model.

Finally, as a more challenging exercise in visualization, we could define the *surface of genus two* as the quotient of an octagon in the plane by the identifications in the boundary indicated by: See the pictures in pp. 300–301 of [6] to see in more detail how the identifications on the boundary of

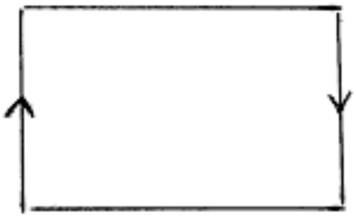

FIGURE 4.5. The Möbius Band

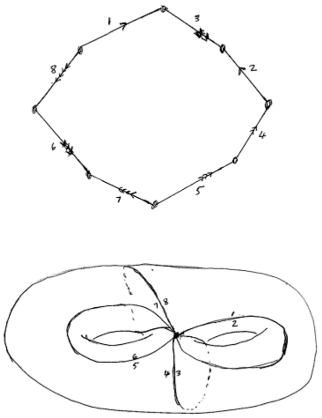

FIGURE 4.6. A Surface of Genus 2

the octagon indicated on the top picture leads to the surface in the bottom picture.

5. Connected Spaces

A topological space X is said to be *disconnected* if there exist open sets $U, V \subset X$, both non-empty, so that $U \cap V = \emptyset$ and $X = U \cup V$. If such open sets exist, we say that U, V *disconnect* X. A topological space is said to be *connected* if it is not disconnected, in other words:

Definition 5.1. A topological space X is *connected* if and only if, whenever $U, V \subset X$ are disjoint open sets such that $X = U \cup V$, then either $U = \emptyset$ or $V = \emptyset$.

Theorem 5.1. *The following conditions on a topological space X are equivalent:*

(1) *X is connected.*
(2) *If $E, F \subset X$ are disjoint closed subsets so that $X = E \cup F$, then either $E = \emptyset$ or $F = \emptyset$.*

(3) *The only subsets of X that are both open and closed are X and $\emptyset$.*
(4) *Every continuous map $f : X \to \{0, 1\}$, (where $\{0, 1\}$ has the discrete topology) is constant.*

Proof. By taking complements it is clear that (1) and (2) are equivalent. A subset $A \subset X$ is both open and closed if and only if both A and $X \setminus A$ are open, and these two sets are disjoint and their union is X, so (1) and (3) are equivalent. If $f : X \to \{0, 1\}$ is a continuous function, then $U = f^{-1}(\{0\})$ and $V = f^{-1}(\{1\})$ are disjoint open sets whose union is X, and if U, V are disjoint open sets whose union is X, then the function which is 0 on U and 1 on V is a continuous function from X to $\{0, 1\}$, so (1) and (4) are equivalent. □

One reason for the choice of definition of connectedness is to make the following theorem clear:

Theorem 5.2. *Let X, Y be topological spaces and let $f : X \to Y$ be a surjective continuous map. If X is connected, then Y is connected.*

Proof. Suppose Y is not connected and suppose U, V disconnect Y. Then $f^{-1}(U)$ and $f^{-1}(V)$ disconnect X, since they are disjoint open sets whose union is X, and the surjectivity of f guarantees that they are both non-empty. □

Corollary 5.1. *Suppose $f : X \to Y$ is a homeomorphism. Then X is connected if and only if Y is connected.*

Example 5.1. It is easy to give examples of disconnected spaces: A discrete space with more than one point, $\mathbb{R} \setminus \{0\} = (-\infty, 0) \cup (0, \infty)$, etc, are disconnected spaces. It is harder to give examples of connected spaces. One non-trivial example of a connected space would be the space $(\mathbb{R}, \mathcal{T}_Z)$ of Example 3.14, because, as we saw in Example 3.22, any two non-empty open sets in $(\mathbb{R}, \mathcal{T}_Z)$ have non-empty intersection, so we cannot possibly disconnect this space.

The main non-trivial example of a connected space is the unit interval. Note that the proof of connectedness has to use the completeness of $\mathbb{R}$, which we do in the form of the existence of the infimum of a non-empty set which is bounded below.

Theorem 5.3. *The interval $[0, 1] \subset \mathbb{R}$ is connected.*

Proof. Suppose $[0, 1] = U \cup V$ where U, V are disjoint open sets with union $[0, 1]$, and label them so that $0 \in U$. If $V \neq \emptyset$, then $a = \inf(V) \in \mathbb{R}$ exists. Moreover, a must be a limit point of V (if this is not a familiar fact, prove it as an exercise in the definitions). In particular, since $[0, 1]$ is closed in $\mathbb{R}$, $a \in [0, 1]$ We cannot have $a \in U$ because U would be a neighborhood of a disjoint from V, contradicting that a is a limit point of V. We cannot have

$a \in V$ because, if so, we would first have $a > 0$ because $0 \in U$, and then, since V is open, there would be an $\epsilon > 0$ so that $(a-\epsilon, a] \subset V$, contradicting that a is a lower bound for V. Thus $V = \emptyset$ and $[0,1]$ is connected. □

Once we have this example of a connected space we can derive many others. In order to do this, it is useful to use the following concept:

Definition 5.2. A topological space X is said to be *path connected* if and only if, for all $x, y \in X$ there exists a continuous map $\phi : [0,1] \to X$ with $\phi(0) = x$ and $\phi(1) = y$. We call such a map ϕ a *path from x to y*.

Theorem 5.4. *Suppose X is path connected. Then X is connected.*

Proof. Suppose X is not connected, and let U, V be disjoint, non-empty open sets whose union is X. Pick $x \in U$ and $y \in V$. If X were path connected there would be a continuous map $\phi : [0,1] \to X$ with $\phi(0) = x \in U$ and $\phi(1) = y \in V$, thus $\phi^{-1}(U)$ and $\phi^{-1}(V)$ would be non-empty, disjoint open sets with union $[0,1]$, contradicting the connectedness of $[0,1]$. Thus X is not path connected, proving the theorem. □

Examples of path connected spaces are plentiful, so we get many examples of connected spaces.

Definition 5.3. A subset $C \subset \mathbb{R}^n$ is called *convex* if and only if, for all $x, y \in C$, the straight line segment $\overline{xy} \subset C$.

Theorem 5.5. *Let $C \subset \mathbb{R}^n$ be convex. Then C is path connected, in particular, C is connected.*

Proof. Let $x, y \in C$. Since $\overline{xy} \subset C$, the map $\phi : [0,1] \to \mathbb{R}^n$ defined by $\phi(t) = (1-t)x + ty$ has image contained in C and is therefore a path from x to y in C. □

This gives many examples of connected subspaces of $\mathbb{R}^n$:

Example 5.2. The following spaces are convex, hence connected:

(1) $\mathbb{R}^n$ for any n
(2) Any interval in $\mathbb{R}$.
(3) Any half-space in $\mathbb{R}^n$: let $l : \mathbb{R}^n \to \mathbb{R}$ be any linear function and $c \in \mathbb{R}$, then $\{x : l(x) > c\}$ as well as $\{x : l(x) \geq c\}$.
(4) Any ball (open or closed) in any of the metrics $d_{(1)}, d_{(2)}, d_{(\infty)}$ of Definition 1.2.

The class of convex sets is relatively small, we can visualize many other path connected spaces. In order to systematically do this, it is useful to have a concept of concatenation of paths. There are many ways to do this, for instance, for many purposes we do not need the domain of our paths to be $[0,1]$, any interval would do. For other purposes we will see later, it is useful to always use the domain $[0,1]$. Let us make the following definition:

Definition 5.4. Let $\phi, \psi : [0,1] \to X$ be continuous maps, and assume that $\phi(1) = \psi(0)$. We define the *concatenation of* ϕ *and* ψ, (also called the *composition of* ϕ *and* ψ), denoted $\phi \cdot \psi$, to be the map $[0,1] \to X$ defined by

$$\phi \cdot \psi(t) = \begin{cases} \phi(2t) & \text{if} \quad 0 \leq t \leq \frac{1}{2}, \\ \psi(2t-1) & \text{if} \quad \frac{1}{2} \leq t \leq 1. \end{cases}$$

Also, let the *inverse path* of ϕ to be the map $\phi^{-1} : [0,1] \to X$ defined by

$$\phi^{-1}(t) = \phi(1-t).$$

In particular, ϕ^{-1} is a path from $\phi(1)$ to $\phi(0)$.

Warning: The meaning of inverse path is different from the meaning of inverse function, even though the same notation is used. It should be clear from the context what is meant.

This definition is easy to visualize. Say $\phi(0) = x$, $\phi(1) = \psi(0) = y$ and $\psi(1) = z$. Then we are saying that a path from x to y can be followed by a path from y to z to form a path from x to z. Note that this is the same construction that we used in Example 1.9 to define a distance function of a surface in $\mathbb{R}^3$, except that now we are making the construction more precise. Since we choose to parametrize the paths by $[0,1]$, in order to concatenate the two paths, we re-parametrize ϕ to have domain $[0, \frac{1}{2}]$ and ψ to have domain $[\frac{1}{2}, 1]$ and then literally put the re-paremetrized paths next to each other. The inverse path means running along the same path in the opposite direction. Clearly the inverse path is continuous, and for the continuity of the concatenation we just need to check the following:

Lemma 5.1. *If* $\phi, \psi : [0,1] \to X$ *are continuous, and* $\phi(1) = \psi(0)$*, then* $\phi \cdot \psi : [0,1] \to X$ *is also continuous.*

The proof follows immediately from the following useful general principle, that we state explicitly for future use:

Lemma 5.2. *Suppose* X, Y *are topological spaces,* $X = A \cup B$*, where* A *and* B *are closed subsets. Suppose we are given maps* $f : A \to Y$ *and* $g : B \to Y$ *such that* $f|_{A \cap B} = g|_{A \cap B}$*. Define a map* $F : X \to Y$ *by*

$$F(x) = \begin{cases} f(x) & \textit{if} \quad x \in A, \\ g(x) & \textit{if} \quad x \in B. \end{cases}$$

Then F *is well defined, and it is continuous if and only if* f *and* g *are both continuous (in the subspace topology). The same statement holds if* A *and* B *are both open sets.*

Proof. It is clear that F is well-defined, since f and g agree on $A \cap B$. Since $F|A = f$ and $F|B = g$, the continuity of F implies that of f and g. Conversely, if f and g are both continuous and $C \subset Y$ is a closed set, then $F^{-1}(C) = (F^{-1}(C)\cap A)\cup(F^{-1}(C)\cap B) = f^{-1}(C)\cup g^{-1}(C)$. By Theorem 4.3 we have that $f^{-1}(C)$ and $g^{-1}(C)$, which by hypothesis of continuity are closed in A, B respectively, are also closed in X. Thus $F^{-1}(C)$ is closed in X, so F is continuous. The proof for the case in which A and B are open sets is similar.

□

Lemma 5.1 follows immediately by taking $X = [0,1] = [0,\frac{1}{2}] \cup [\frac{1}{2}.1] = A\cup B$ and f, g the restrictions of the definition of $\phi\cdot\psi$ to the two subintervals.

Example 5.3. The space $(\mathbb{R}^2, d_{FR})$ of Example 1.8 (the French railway metric) is clearly path connected: given $x, y \in \mathbb{R}^2$, if they are in the same ray from the origin the straight line segment joining them gives a path between them, otherwise we concatenate the path from x to 0 with the path from 0 to y to join them by a path.

Example 5.4. A simple application of Lemma 5.1 is to show that for $n \geq 2$, $\mathbb{R}^n \setminus \{0\}$ is path connected. Let $x, y \in \mathbb{R}^n \setminus \{0\}$. If $0 \notin \overline{xy}$, then $\phi(t) = (1-t)x + ty$ is a path from x to y. If $0 \in \overline{xy}$, then y is a (negative) multiple of x. Since $n \geq 2$, we can choose a vector z linearly independent from x, hence also linearly independent from y. Let $\phi(t) = (1-t)x + tz$ and let $\psi(t) = (1-t)z + ty$. Then $\phi(t)$ and $\psi(t)$, being linear non-trivial combinations of x and z, are never 0, so these are paths in $\mathbb{R}^n \setminus \{0\}$ from x to z and from z to y respectively, so by Lemma 5.1, $\phi \cdot \psi$ is a path in $\mathbb{R}^n \setminus \{0\}$ from x to y, thus this space is path connected.

Question: Why doesn't the above argument work for $n = 1$?

We finally have a way to distinguish some topological spaces that should "obviously" not be homeomorphic :

Theorem 5.6. *There is no homeomorphism between* $\mathbb{R}$ *and* $\mathbb{R}^n$ *for* $n \geq 2$.

Proof. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ were a homeomorphism, $n \geq 2$. Then $f|\mathbb{R}^n \setminus \{0\} : \mathbb{R}^n \setminus \{0\} \to \mathbb{R} \setminus \{f(0)\}$ would be a homeomorphism. But $\mathbb{R}^n \setminus \{0\}$ is connected for $n \geq 2$ while $\mathbb{R} \setminus \{f(0)\} = (-\infty, f(0)) \cup (f(0), \infty)$ is disconnected, contrary to Corollary 5.1. □

Remark 5.1. It is more difficult to prove that $\mathbb{R}^n$ and $\mathbb{R}^m$ are not homeomorphic for $m \neq n$, $m, n \geq 2$. More subtle topological invariants are needed to distinguish these spaces.

Remark 5.2. Using the same ideas as in the proof of Theorem 5.6 it is not hard to prove that $[0,1]$ and $[0,1] \times [0,1]$ are not homeomorphic. This of course means that a segment and a rectangle are not homeomorphic. This can be used, together with the calculations of the equality sets in the triangle inequality for the Euclidean and Taxicab distances in $\mathbb{R}^2$ (see Examples 1.2 and 1.4) to complete a proof in the homework problems that these two metric spaces are not isometric (since the equality sets $E_d(x,z)$ are not homeomorphic, see the discussion in Example 1.15).

5.1. **Connected Components.** Let X be a topological space. Define a relation on X by $x \sim y$ if and only if there is a connected subset $C \subset X$ so that $x \in C$ and $y \in C$. This is an equivalence relation: It is clearly reflexive ($x \sim x$ since $\{x\}$ is connected), it is clearly symmetric ($x \sim y$ if and only if $y \sim x$). It requires a proof to show that it is transitive. To show that $x \sim y$ and $y \sim z$ implies $x \sim z$, it would be natural to take connected subsets $C_1, C_2 \subset X$ so that $x, y \in C_1$ and $y, z \in C_2$ and argue that $C_1 \cup C_2$ is connected. The first part of the following lemma (for a collection of two connected sets) shows that this is indeed the case, proving this is an equivalence relation:

Lemma 5.3. (1) *Let $\{C_\alpha\}_{\alpha \in A}$ be a collection of connected subsets of X, and assume that $\cap C_\alpha \neq \emptyset$. Then $\cup C_\alpha$ is connected.*
(2) *Let $C \subset X$ be connected. Then its closure $\bar{C}$ is connected.*

Proof. We use the fourth characterization of connectedness from Theorem 5.1. L For the first part, let $\cup C_\alpha \to \{0,1\}$ be continuous, and $x_0 \in \cap C_\alpha$. Then $f|_{C_\alpha}$ is a constant, which must be $f(x_0)$. Thus $f(x) = f(x_0)$ for all $x \in \cup C_\alpha$, thus f is constant and $\cup C_\alpha$ is connected.

For the second part, suppose $f : \bar{C} \to \{0,1\}$ is a continuous function, let $x \in \bar{C}$, and let $a = f(x)$. Then $f^{-1}(\{a\})$ is an open set containing x, thus, by part (2) of Theorem 3.11, $f^{-1}(\{a\}) \cap C \neq \emptyset$. Let $y \in C \cap f^{-1}(\{a\})$. Then $f(y) = a$. Since C is connected, $f|_C$ is constant, so this constant must be a, so $f(x) = a$ for any $x \in \bar{C}$, thus $\bar{C}$ is connected. □

We are therefore justified in making the following definition:

Definition 5.5. Let X be a topological space. Define two equivalence relations on X:

(1) Let x be equivalent to y if and only if there is a connected subset $C \subset X$ containing x and y. The equivalence classes are called the *connected components* of X.
(2) Let x be equivalent to y if and only if there exists a path in X from x to y. The equivalence classes are called the *path components* of X.

For the second part of the definition, note that the relation in question is clearly reflexive. The inverse path shows that it is symmetric, and concatenation of paths shows that it is transitive. Thus it also is an equivalence relation. It is clear that path components are contained in connected components, and in many, but not all, situations they coincide. See Chapter 4, Section 6 of [7] for an example where the two notions differ.

Example 5.5. Connected components (and path components) can be used to distinguish topological spaces. It is clear that homeomorphic spaces have the same number of connected components, and the same is true for path components. This can be used, for example, to prove that the subsets of $\mathbb{R}^2$ in the shape of the letter X and the shape of the letter Y are not homeomorphic. There is a point $p \in X$ with the property that $X \setminus \{p\}$ has 4 connected components, while for every $q \in Y$, $Y \setminus \{q\}$ has at most 3 connected components. So there could be no homeomorphism between X and Y. It is a standard exercise to use similar reasoning to classify the letters of the Roman alphabet up to homeomorphism.

Here are some general properties of connected components:

Theorem 5.7. *Let X be a topological space and let $x \in X$, and let C_x denote the connected component of X containing x.*

1. *C_x is the largest connected subset of X containing x: If $A \subset X$ is connected and $x \in A$, then $A \subset C_x$.*
2. *C_x is closed in X.*

Proof. By definition, $C_x = \{y \in X : \text{there exists a connected set } B \text{ such that } x, y \in B\} = \cup\{B \subset X : B \text{ is connected and } x \in B\}$ is a union of connected sets with non-empty intersection. By Lemma 5.3, C_x is connected. Moreover, if A is any connected set containing x, then A is an element of this collection, so A is contained in its union, in other words, $A \subset C_x$, as asserted. To prove the second part, use the second part of Lemma 5.3: $\bar{C}_x$ is connected, hence $\bar{C}_x \subset C_x$, hence C_x is closed.

□

5.2. **Locally Path Connected Spaces.**

Definition 5.6. A topological space is called *locally path connected* if it has a basis consisting of path connected open sets.

Remark 5.3. We could state the condition more explicitly as follows: X is locally path connected if and only if for every $x \in X$ and every open subset $U \subset X$ with $x \in U$, there exists a path connected open set V such that $x \in V \subset U$.

Remark 5.4. In general, given any property $\mathcal{P}$ of open sets, a space X is said to be *locally* $\mathcal{P}$ if and only if it has a basis of open sets with property $\mathcal{P}$. For example, a space is *locally connected* if it has a basis of connected open sets.

Example 5.6.

(1) If $X \subset \mathbb{R}^n$ is an open set, then it is locally path connected since the balls $B(x,r)$ contained in X form a basis, are convex, hence path connected.

(2) Let $A = \{(x, \sin(\frac{1}{x}) : 0 < x < \frac{1}{2\pi}\} \subset \mathbb{R}^2$, and let $X = \bar{A}$. Then $X = A \cup B$ where $B = \{(0,y) : -1 \leq y \leq 1\}$. X is connected since A is connected, but it is not locally connected, hence not locally path connected. Small neighborhoods in X of points in B are not connected. See Chapter 4, Section 6 of [7] for more details.

Theorem 5.8. *Suppose X is connected and locally path connected. Then X is path connected.*

Proof. Let $x \in X$. Let $U = \{y \in X :$ there exists a path $\phi : [0,1] \to X$ from x to $y\}$. We will show:

(1) *U is open*: For any $y \in U$ there exists an open, path connected set $V \subset X$ so that $y \in V$. If $z \in V$, then there exists a path $\psi : [0,1] \to V$ with $\psi(0) = y$ and $\psi(1) = z$, Then $\phi \cdot \psi : [0,1] \to X$ is a path from x to z, thus $z \in U$, thus given any $y \in U$ there exists an open set $V \subset X$ so that $y \in V \subset U$, therefore U is open, as claimed.

(2) *$X \setminus U$ is open*: Suppose $y \in X \setminus U$. There exists a path connected open set $V \subset X$ so that $y \in V$. Let $z \in V$. Then there exists a path $\psi : [0,1] \to V$ from y to z. If there were a path $\phi : [0,1] \to X$ from x to z, then $\phi \cdot \psi^{-1}$ would be a path from x to y, contradicting the choice of y. Thus $z \in X \setminus U$, so by the same reasoning as above $X \setminus U$ is open.

Finally, since $x \in U$ we know that $U \neq \emptyset$. Since X is connected we must have $X \setminus U = \emptyset$, in other words, $X = U$, thus X is path connected. □

Remark 5.5. The proof of Theorem 5.8 can be applied to connectedness by other classes of paths, not necessarily the same as the class of continuous paths. All that is needed is that the class of paths be closed under concatenation and inverse. If $X \subset \mathbb{R}^n$ two such classes of paths are the *piecewise linear paths*, meaning continuous paths $\phi : [0,1] \to X \subset \mathbb{R}^n$ so that there exists a subdivision of $[0,1]$ into subintervals so that the restriction of ϕ to each subinterval is a linear map to $\mathbb{R}^n$. The class of *piecewise differentiable paths* is defined in exactly the same way. Then we can make the following definitions:

Definition 5.7. Let $X \subset \mathbb{R}^n$. We say that X is

(1) *piecewise linearly connected* if given any $x, y \in X$ there exists a piecewise linear path $\phi : [0,1] \to X$ from x to y. It is *locally piecewise linearly connected* if it has a basis of piecewise linearly connected open sets.

(2) *piecewise differentiably connected* and *locally piecewise differentiably connected* are defined in exactly the same way.

Theorem 5.9. *Let $X \subset \mathbb{R}^n$*

1. *Suppose X is connected and locally piecewise linearly connected. Then X is piecewise linearly connected.*
2. *Suppose X is connected and locally piecewise differentiably connected. Then X is piecewise differentiably connected.*

Proof. Same as the proof of Theorem 5.8. □

Corollary 5.2. *Let $U \subset \mathbb{R}^n$ be open and connected. Then U is piecewise linearly connected.*

Proof. Since balls in $\mathbb{R}^n$ are convex, hence piecewise linearly connected, U is locally piecewise linearly connected. Apply the theorem. □

5.3. **Existence Theorems.** One application of connectedness is to prove existence theorems for solutions of equations. One familiar theorem from real analysis is the intermediate value theorem, that we can formulate in more generality:

Theorem 5.10. *Let X be a connected space and let $f : X \to \mathbb{R}$ be continuous. Suppose for some $x, y \in X$ we have that $f(x) = a < f(y) = b$. Then, given any number $c \in (a, b)$, there exists $z \in X$ with $f(z) = c$.*

Proof. Suppose not: there is $c \in (a, b)$ so that $c \notin f(X)$. Then $f(X) = (f(X) \cap (-\infty, c)) \cup (f(X) \cap (c, \infty))$ is the disjoint union of two non-empty open sets, contradicting the fact that $f(X)$, as the continuous image of a connected space, must be connected (Theorem 5.2). □

As application of the intermediate value theorem we will prove the version of the implicit function theorem that we need. We note that the same proof would work for the zero set of any smooth function from an open set $U \subset \mathbb{R}^{n+1}$ to $\mathbb{R}$, but we will be mainly using the case $n = 2$, so we will just state this case. The theorem is easier to visualize when $n = 1$, and it would be useful to do this when looking at the theorems, proofs, and examples. There is also a version of the theorem for functions with target $\mathbb{R}^m$ for $m > 1$, but the proof is more involved in this case; it would require the inverse function theorem where we use the intermediate value theorem.

By a *smooth function* we mean a C^∞-function, although C^1 would be enough in this theorem. Using C^∞ is often an expedient way of avoiding counting how many derivatives are used in a proof.

Theorem 5.11. *Let $U \subset \mathbb{R}^3$ be open and let $f : \mathbb{R}^3 \to \mathbb{R}$ be a smooth function. Let $S = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) = 0\}$ be the zero set of f. Suppose $(x_0, y_0, z_0) \in S$ and suppose that $\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$. Then there exist $\epsilon, \delta > 0$ and a smooth function $g : B((x_0, y_0), \delta) \to (z_0 - \epsilon, z_0 + \epsilon) \subset \mathbb{R}$ so that $S \cap (B(x_0, y_0), \delta) \times (z_0 - \epsilon, z_0 + \epsilon)) = \{(x, y, g(x, y)) : (x, y) \in B((x_0, y_0), \delta)\}$.*

The theorem says that, under the hypothesis of the non-vanishing of $\frac{\partial f}{\partial z}$ at (x_0, y_0, z_0), there is a neighborhood of the form $B_1 \times B_2$, where B_1 and B_2 are balls in $\mathbb{R}^2$, $\mathbb{R}$ respectively, so that $S \cap (B_1 \times B_2)$ is the graph of a function $g : B_1 \to B_2$. In other words, the relation $f(x, y, z) = 0$ defines z "implicitly" as a function of x and y for (x, y) close enough to (x_0, y_0).

Proof. We may assume $\frac{\partial f}{\partial z}(x_0, y_0, z_0) > 0$ (otherwise change f to $-f$). Let $c = \frac{1}{2}\frac{\partial f}{\partial z}(x_0, y_0, z_0) > 0$. By continuity of $\frac{\partial f}{\partial z}$, there exists a neighborhood of (x_0, y_0, z_0) on which $\frac{\partial f}{\partial z}(x, y, z) > c$, and we may take this neighborhood to be of the form $B((x_0, y_0), \delta_0) \times (z_0 - \epsilon, z_0 + \epsilon)$ for some $\delta_0, \epsilon > 0$. In particular for each $(x, y) \in B((x_0, y_0), \delta_1)$ we have that $f(x, y, z)$ is a strictly increasing function of z for $z_0 - \epsilon \leq z \leq z_0 + \epsilon$. It follows that $f(x_0, y_0, z_0 + \epsilon) > 0$, $f(x_0, y_0, z_0 - \epsilon) < 0$, and, by continuity of f, there exists $\delta > 0$ so that $f(x, y, z_0 + \epsilon) > 0$ and $f(x, y, z_0 - \epsilon)$ (choose a δ_1 the works for $f(x, y, z_0 + \epsilon)$, a δ_2 that works for $f(x, y, z_0 - \epsilon)$, both smaller than δ_0, and let δ be the smaller of δ_1, δ_2).

By the intermediate value theorem (Theorem 5.10), for each $(x, y) \in B((x_0, y_0), \delta)$ there exists a $z \in (z_0 - \epsilon, z_0 + \epsilon)$ so that $f(x, y, z) = 0$. Since f is a strictly increasing function of z, this value of z is unique, call it $g(x, y)$. This gives us the desired function $g : B((x_0, y_0), \delta) \to (z_0 - \epsilon, z_0 + \epsilon)$, since, by construction of g, we have that $S \cap (B((x_0, y_0), \delta) \times (z_0 - \epsilon, z_0 + \epsilon)) = \{(x, y, g(x, y) : (x, y) \in B((x_0, y_0), \delta)\}$.

It remains to prove that g is a smooth function. It is easy to see that g is continuous. This is an easy consequence of the uniqueness: given $(x_1, y_1, z_1) \in B((x_0, y_0) \times (z_0 - \epsilon, z_0 + \epsilon)$ and given $\epsilon' > 0$ sufficiently small, repeat the same construction to find a $\delta' > 0$ and a function, say h, so that $S \cap (B((x_1, y_1), \delta') \times (z_1 - \epsilon', z_1 + \epsilon')) = \{(x, y, h(x, y) : (x, y) \in B((x_1, y_1), \delta')\}$. By the uniqueness of the solution, we must have $g = h$ on $B((x_1, y_1), \delta')$, hence $g((B((x_1, y_1), \delta') \subset (z_1 - \epsilon', z_1 + \epsilon')$. Since $z_1 = g(x_1, y_1)$ and $\epsilon' > 0$ is arbitrary, this is exactly the statement of continuity of g at (x_1, y_1).

We will next check that g is differentiable. By definition of differentiability of f, we have that

$$f(x+\Delta x, y+\Delta y, z+\Delta z) - f(x, y, z) = (f_x+\epsilon_1)\Delta x + (f_y+\epsilon_2)\Delta y + (f_z+\epsilon_3)\Delta z$$

where f_x, f_y, f_z denote the partial derivatives of f with respect to the indicated variable, each evaluated at the point (x, y, z), and $\epsilon_i, i = 1, 2, 3$, are functions of $x, \Delta x, y, \Delta y, z, \Delta z$ so that $\epsilon_i \to 0$ as $(\Delta x, \Delta y, \Delta z) \to (0, 0, 0)$.

Suppose we evaluate this on the graph $z = g(x, y)$. We have that both terms in the left hand side vanish, thus

$$0 = (f_x + \epsilon_1')\Delta x + (f_y + \epsilon_2')\Delta y + (f_z + \epsilon_3')\Delta g$$

where $\epsilon_i'(x, \Delta x, y, \Delta y) = \epsilon_i(x, \Delta x, y, \Delta y, g(x,y), \Delta g)$. Since g is continuous, $\Delta g \to 0$ as $(\Delta x, \Delta y) \to (0,0)$, thus $\epsilon_i' \to 0$ as $(\Delta x, \Delta y) \to (0,0)$.

Solving the above equation for Δg we get

$$\Delta g = -\frac{f_x + \epsilon_1'}{f_z + \epsilon_3'}\Delta x - \frac{f_y + \epsilon_2'}{f_z + \epsilon_3'}\Delta y$$

which makes sense since $f_z > c > 0$, so there is no problem in dividing by $f_z + \epsilon_3'$. Moreover this can be re-written as

$$\Delta g = -(\frac{f_x}{f_z} + \epsilon_1'')\Delta x - (\frac{f_y}{f_z} + \epsilon_2'')\Delta y$$

where $\epsilon_i'' \to 0$ as $(\Delta x, \Delta y) \to (0,0)$, thus g is differentiable and

$$g_x = -\frac{f_x}{f_z}(x, y, g(x,y)) \text{ and } g_y = -\frac{f_y}{f_z}(x, y, g(x,y)) \tag{5.1}$$

from which it is clear that the partial derivatives g_x, g_y are continuous, thus g is of class C^1. This procedure can be continued to show that g is C^∞. □

Example 5.7. Let $f(x,y,z) = x^2 + y^2 + z^2 - 1$. Then the set $\{(x,y,z) : f(x,y,z) = 0\}$ is the unit sphere $S^2 \subset \mathbb{R}^3$. Let $(x_0, y_0, z_0) = (0,0,1)$, the north pole. Then $\frac{\partial f}{\partial z}(0,0,1) = 2 \neq 0$, and we can see visually that we can choose $\delta = \epsilon = 1$ in the statement of the implicit function theorem (although our proof requires a smaller ϵ) , and $g(x,y) = \sqrt{1 - x^2 - y^2}$. If (x_0, y_0, z_0) is any other point of the upper hemisphere, that is, if $z_0 > 0$, then $g(x,y) = \sqrt{1 - x^2 - y^2}$ also works, but the largest δ we can take is $1 - \sqrt{x_0^2 + y_0^2}$ (and we could choose $\epsilon = z_0$). If (x_0, y_0, z_0) is in the lower hemisphere, that is, $z_0 < 0$, then we must choose $g(x,y) = -\sqrt{1 - x^2 - y^2}$ and the largest size of the δ would be $1 - \sqrt{x_0^2 = y_0^2}$ (and we could take $\epsilon = |z_0|$). Finally, if (x_0, y_0, z_0) is on the equator, that is, if $z_0 = 0$, then for all $\delta > 0$ and $\epsilon > 0$, whenever $(x,y,z) \in S^2 \cap (B((x_0, y_0), \delta) \times (-\epsilon, \epsilon))$, so is $(x, y, -z)$, so this intersection cannot be a graph $z = g(x,y)$. This does not contradict the implicit function theorem, because at these points $\frac{\partial f}{\partial z}(x_0, y_0, 0) = 0$, so the implicit function theorem does not apply. This also shows the necessity of the condition $\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$ in the statement of the theorem.

6. Smooth Surfaces

We now define what is meant by a topological surface and a differentiable surface. The same concepts can be defined in any dimension, they are called *topological manifold* and *differentiable manifold* or *smooth manifold.*

Definition 6.1. A topological space S is called:

(1) A *topological surface* if it is a Hausdorff space with a countable basis and it has the property that every $x \in S$ has a neighborhood U

which is homeomorphic to an open set in $\mathbb{R}^2$, in other words, there exists a covering $\{U_\alpha\}_{\alpha\in A}$ for some index set A, and for each $\alpha \in A$ there exists a homeomorphism $\phi_\alpha : U_\alpha \to V_\alpha$, where $V_\alpha \subset \mathbb{R}^2$ is open. These homeomorphisms are called *coordinate charts.*

(2) A *differentiable surface* or a *smooth surface* if it is a topological surface and the above homeomorphisms (or coordinate charts) can be chosen to have the following property: whenever $U_\alpha \cap U_\beta \neq \emptyset$, the homeomorphism $\phi_\alpha \circ \phi_\beta^{-1} : \phi_\beta(U_\alpha \cap U_\beta) \to \phi_\alpha(U_\alpha \cap U_\beta)$ is smooth. The maps $\phi_\alpha \circ \phi_\beta^{-1}$ are called the *transition maps* between charts.

Remark 6.1. Many clarifications are in order concerning these definitions:

(1) In a first reading ignore the conditions that S be Hausdorff and have a countable basis. These conditions are needed for correctness of the definition, but will be automatic in all the examples we will see.

(2) The important condition for us is that S look locally like the plane $\mathbb{R}^2$. It is clear how to state this topologically. The terminology of charts comes from the usual picture we have of maps of the earth, where we take small pieces of the surface of the earth and consider them as part of a plane. The collection of charts is usually called an *atlas.*

(3) In $\mathbb{R}^2$ there is a notion of differentiable function. This uses more than the topology of $\mathbb{R}^2$, it also uses the linear structure. It is not clear how to transfer this concept to a more general space. The point of the definition of differentiable surface is to make sense of differentiable function. A function $f : S \to \mathbb{R}$ will be defined to be differentiable if it is differentiable in every chart. In other words, if and only if, for all $\alpha \in A$, $f \circ \phi_\alpha^{-1} : V_\alpha \to \mathbb{R}$ is differentiable. The latter statement makes sense because $V_\alpha \subset \mathbb{R}^2$, but, for the definition to be independent of the chosen chart, we need the condition stated in the second part of the definition.

(4) Observe that $(\phi_\alpha \circ \phi_\beta^{-1})^{-1} = \phi_\beta \circ \phi_\alpha^{-1}$. Thus the condition that the maps $\phi_\alpha \circ \phi_\beta^{-1}$ be smooth for all α and β for which they are defined implies that their inverses are also smooth. A smooth map between open sets in $\mathbb{R}^n$ that is invertible and whose inverse is also smooth is called a *diffeomeorphism.* Thus all the transition maps are diffeomorphisms.

And examples are also in order:

Example 6.1. Let $U \subset \mathbb{R}^2$ be an open set. Then it is a smooth surface: we know it is Hausdorff, has countable basis, and it can be covered by one chart $id : U \to U$, so the conditions of the second part are automatic.

Example 6.2. The implicit function theorem gives us many examples of smooth surfaces. Let $U \subset \mathbb{R}^3$ be open and let $f : U \to \mathbb{R}$ be smooth. Let $S = \{(x, y, z) \in \mathbb{R}^3 : f(x, y, z) = 0\}$ be the zero set of f and finally make the most important assumption: *the gradient* $\nabla f \neq 0$ *at any point of* S. Recall that $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z})$, so the the assumption is that at each point of S at least one of the partial derivatives of f does not vanish. We have the following theorem:

Theorem 6.1. *Under the above hypothesis, the space* S *is a smooth surface.*

Proof. Let $p^z : \mathbb{R}^3 \to \mathbb{R}^2$ be the projection with kernel the z-axis, that is, $p^z(x, y, z) = (x, y)$, and define p^x, p^y similarly. Given any $(x_0, y_0, z_0) \in S$ at least one partial derivative does not vanish at this point. Say $\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$. Then the implicit function theorem gives a neighborhood U of $(x_0 m y_0, z_0)$, $U = B((x_0, y_0), \delta) \times (z_0 - \epsilon, z_0 + \epsilon)$ and a smooth function $g : B((x_0, y_0), \delta) \to (z_0 - \epsilon, z_0 + \epsilon)$ so that $S \cap U = \{(x, y, g(x, y)) : (x, y) \in B((x_0, y_0), \delta)\}$. In particular we see that $p^z|_{S\cap U} : S \cap U \to B((x_0, y_0), \delta)$ is a chart, with inverse map $G(x, y) = (x, y, g(x, y))$.

If $(x_1, y_1, z_1) \in S$ is another point, we can use the same reasoning. If $\frac{\partial f}{\partial z}(x_1, y_1, z_1) \neq 0$, we obtain a similar chart $p^z|_{S\cap V}$, and, if they intersect, that is, $U \cap V \cap S \neq \emptyset$, then on this intersection the inverse of p^z is also G, so the transition function is $p^z \circ G = id$ is smooth.

If $\frac{\partial f}{\partial z}(x_1, y_1, z_1) = 0$, then some other partial does not vanish. Suppose, say $\frac{\partial f}{\partial y}(x_1, y_1, z_1) \neq 0$. Then the implicit function theorem gives us a neighborhood $V = B((x_1, z_1), \delta') \times (y_1 - \epsilon', y_1 + \epsilon')$ and a smooth function $h : B((x_1, z_1), \delta') \to (y_1 - \epsilon', y_1 + \epsilon')$ so that $S \cap V = \{(x, h(x, z), z)) : (x, z) \in B((x_1, z_1), \delta')\}$. Therefore $p^y|_{S\cap V} : S \cap V \to B((x_1, z_1), \delta')$ is a chart, with inverse $H(x, z) = (x, h(x, z), z)$. If this neighborhood $S \cap V$ of (x_1, y_1, z_1) intersects the neighborhood $S \cap U$ of (x_0, y_0, z_0) considered above, then the transition maps associated to this intersection are $p^y \circ G(x.y) = p^y(x, y, g(x, y)) = (x, g(x, y))$ and its inverse map $p^z \circ H(x, z) = p^z(x, h(x, z), y) = (x, h(x.z))$ which are smooth maps. Since S can be covered by charts of these forms and we have checked that the transition functions that can occur are smooth, we see that S is indeed a smooth surface.

□

Example 6.3. We now specialize the general principle of Theorem 6.1 and its proof to the case of the unit sphere $S^2 \subset \mathbb{R}^3$ defined by $f(x, y, z) = x^2 + y^2 + z^2 - 1 = 0$. Let us cover S^2 by the six sets $U_z^\pm$, $U_y^\pm$, $U_x^\pm$ defined by $U_z^+ = S^2 \cap \{z > 0\}$, $U_z^- = S^2 \cap \{z < 0\}$, $U_y^+ = S^2 \cap \{y > 0\}$, and so on. Write p^x, p^y, p^z for the restrictions to S^2 of the orthogonal projections of $\mathbb{R}^3 \to \mathbb{R}^2$ with kernel the corresponding axis, so $p^x(x, y, z) = (y, z)$, $p^y(x, y, z) = (x, z)$ and $p^z(x, y, z) = (x, y)$. Let D be the open unit disk in $\mathbb{R}^2$ and define charts $\phi_z^\pm : U_z^\pm \to D$ by $\phi_z^\pm = p^z|_{U_z^\pm}$, and define $\phi_y^\pm : U_y^\pm \to D$ and $\phi_x^\pm : U_x^\pm \to D$ in the similar way using the projections p^y, p^x respectively. These maps are indeed charts, because there inverses are given, as in the proof of

Theorem 6.1, by the graph of the implicit function. Thus $(\phi_z^+)^{-1}(x,y) = (x, y, \sqrt{1-x^2-y^2})$, $(\phi_z^-)^{-1}(x,y) = (x,y,-\sqrt{1-x^2-y^2})$, $(\phi_y^+)^{-1}(x,z) = (x, \sqrt{1-x^2-z^2}, z)$, etc. From this it is easy to compute the transition maps. For example, for $U_z^+ \cap U_y^+$ we have

$$\phi_z^+ \circ (\phi_y^+)^{-1}(x,z) = p^z((x, \sqrt{1-x^2-z^2}, z) = (x, \sqrt{1-x^2-z^2}),$$

which is smooth. In fact, we know that it must be a diffeomorphism of $\phi_y(U_z^+ \cap U_y^+) = \{(x,z) \in D : z > 0\}$ onto $\phi_z(U_z^+ \cap U_y^+) = \{(x,y) \in D : y > 0\}$. To check this directly, observe that, since this map is

$$(x,z) \to (x, \sqrt{1-x^2-z^2}),$$

which is the identity on the first coordinate, it is a diffeomorphism if and only if the map $z \to \sqrt{1-x^2-z^2}$ of the second coordinate is a diffeomorphism of the interval $(0, \sqrt{1-x^2})$ to itself. This is indeed the case by the restriction imposed on the interval. Without this restriction the map on the second coordinate would not be injective, for instance, it would fail on $(-\sqrt{1-x^2}, \sqrt{1-x^2})$ where it is two-to-one from this interval onto half the interval.

In the same way we can check that all other transition maps are diffeomorphisms. Thus S^2 is a smooth surface.

Example 6.4. To show that the condition $\nabla F \neq 0$ at every point of S is needed, let's look at a few examples:

(1) $f(x,y,z) = xyz$. Then $\nabla f = (yz, xz, xy) = (0,0,0)$ when at leat two of x, y, z vanish. The zero set S is the union of the coordinate planes, is not locally homeomorpic to the plane along any of the coordinate axes. The origin is also a singular point, looking more complicated than the others.
(2) $f(x,y,z) = x^2+y^2-z^2$. Then $\nabla f = (2x, 2y, -2z) = (0,0,0)$ exactly at the origin, which is lies on S and is a singular point. S is a cone with vertex at the origin.
(3) $f(x,y,z) = x^2 - y^2 z$. Then $\nabla f = (2x, -2yz, y^2) = (0,0,0)$ precisely on the z-axis $x = y = 0$. The zero set S is the union of the z-axis and the surface shown below, called the Whitney umbrella, because the negative z-axis (not shown) would be its handle.

6.1. **Smooth surfaces as metric spaces.** In Example 1.9 we started the discussion of how a smooth surface $S = \{f = 0\} \subset \mathbb{R}^3$, where ∇f never vanishes on S, can be made into a metric space with its *intrinsic distance*. We can now finish the discussion that this metric is defined for all *connected* smooth surfaces.

Theorem 6.2. *Let S be the zero set of a function f as in Example 6.2. Then S is locally piecewise differentiably path connected. In particular, if S is connected, then S is piecewise differentiably path connected.*

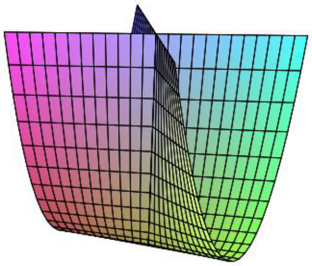

FIGURE 6.1. The Whitney Umbrella

Proof. The proof of Theorem 6.1 gives us that S can be covered by open sets $U \subset \mathbb{R}^3$ which are graphs of functions $g : B \to I$ where $B \subset \mathbb{R}^2$ is a ball and $I \subset \mathbb{R}$ is an open interval: $U \cap S = \{(x, y, g(x, y) : (x, y) \in B\}$ or $U \cap S = \{x, g(x, z), z) : (x, z) \in B\}$ or $U \cap S = \{(g(y, z), y, z) : (y, z) \in B\}$ as the case may be. Since B is convex, in particular piecewise differentiably path connected, and g is smooth, given any two points in $U \cap S$, they can be connected by a smooth path: project the two points to B, connect their projections by a straight line segment, and map this segment back to $U \cap S$ by the smooth map $(x, y) \to (x, y, g(x, y))$ or one of its two variants permuting the coordinates as the case may be. Thus S is locally differentiably path connected. Then Theorem 5.9 does the rest. □

Therefore, if S is a connected smooth surface in $\mathbb{R}^3$, we can define the *intrinsic distance* $d : S \times S \to \mathbb{R}$ as in Example 1.9:

$$d(x, y) = \inf\{L(\gamma) : \gamma \text{ a piecewise differentiable path from } x \text{ to } y\}. \tag{6.1}$$

This infimum is defined, since x and y can always be connected by a piecewise differentiable path. *But this infimum need not be attained.* For example, if $S = \mathbb{R}^2 \setminus \{0\}$, then the infimum defining $d(x, -x) = 2|x|$ is not attained by a path in S. But in many situations it is attained. We will show some examples in section 6.1.2 below.

6.1.1. *Arc Length.* Let $S \subset \mathbb{R}^3$ be a smooth surface and let $\gamma : [0, 1] \to S$ be a piecewise differentiable path. Recall this means that the composition $\gamma : [0, 1] \to S \subset \mathbb{R}^3$ is piecewise differentiable. Write $\gamma(t) = (x(t), y(t), z(t))$. Then the length of γ, $L(\gamma)$ is defined to be

$$L(\gamma) = \int_0^1 |\gamma'(t)|\, dt = \int_0^1 \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\, dt, \tag{6.2}$$

which we could also write as

$$L(\gamma) = \int_\gamma \sqrt{dx^2 + dy^2 + dz^2} \; = \int_\gamma ds, \tag{6.3}$$

where traditionally we write $ds^2 = dx^2 + dy^2 + dz^2$. If γ is piecewise differentiable then this integrals are always defined.

It will be important for calculations to be able to change coordinates. If our curve lies in the domain of some coordinate chart, as in Definition 6.1, then the inverse of this chart gives a differentiable map from an open set $U \subset \mathbb{R}^2$ to S, in other words, we can express (x, y, z) in this chart in S as functions of two variables, say $(u, v) \in U$. Then γ corresponds to a curve $(u(t), v(t))$, $0 \le t \le 1$, and we can work out the length of γ by the chain rule. It would be convenient to write $\mathbf{x} = (x, y, z)$. Then $\gamma(t) = \mathbf{x}(u(t), v(t))$, $\gamma'(t) = \mathbf{x}_u u'(t) + \mathbf{x}_v v'(t)$ (where the subscripts denote partial derivatives) and $\gamma'(t)\cdot\gamma'(t) = (\mathbf{x}_u u' + \mathbf{x}_v v')\cdot(\mathbf{x}_u u' + \mathbf{x}_v v') = (\mathbf{x}_u\cdot\mathbf{x}_v)u'^2 + 2(\mathbf{x}_u\cdot\mathbf{x}_v)u''v' + (\mathbf{x}_v\cdot\mathbf{x}_v v)v'^2$. This last equation is usually written symbolically in differential form as

$$ds^2 = (\mathbf{x}_u\cdot\mathbf{x}_v)du^2 + 2(\mathbf{x}_u\cdot\mathbf{x}_v)dudv + (\mathbf{x}_v\cdot\mathbf{x}_v)dv^2, \tag{6.4}$$

or as

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2, \tag{6.5}$$

where the g_{ij} are the coeffients of Equation 6.4: $g_{11} = \mathbf{x}_u\cdot\mathbf{x}_u$, $g_{12} = \mathbf{x}_u\cdot\mathbf{x}_v$ and $g_{22} = \mathbf{x}_v\cdot\mathbf{x}_v$. They are smooth functions of u, v and geometrically, $g_{11}(u, v) = \mathbf{x}_u\cdot\mathbf{x}_v$ is the magnitude squared of the tangent vector at (u, v) of the curve obtained by varying u and holding v constant, $g_{22}(u, v)$ has the same interpretation with u and v interchanged, while g_{12} is the dot product between the tangent vectors of the two curves just considered.

Example 6.5. One familiar example is the use of polar coordinates in $\mathbb{R}^2$: $x = r\cos\theta$, $y = r\sin\theta$. Then $dx = \cos\theta\ dr - r\sin\theta\ d\theta$, $dy = \sin\theta\ dr + r\cos\theta\ d\theta$ and $ds^2 = dx^2 + dy^2 = (\cos\theta\ dr - r\sin\theta\ d\theta)^2 + (\sin\theta\ dr + r\cos\theta\ d\theta)^2$ which simplifies to

$$ds^2 = dr^2 + r^2\ d\theta^2, \tag{6.6}$$

which is the familiar formula for arclength in polar coordinates. We should be a bit careful when using polar coordinates, since they do not follow the assumption we made above that it be the inverse of a chart. The transformation $(r, \theta) \to (r\cos\theta, r\sin\theta)$ is not invertible unless we carefully restrict its domain, and its image does not cover all of $\mathbb{R}^2$ in an invertible way. But, with our knowledge of the identification topology, we can say, for instance, that polar coordinates give a map $[0, \infty] \times [0, 2\pi] \to \mathbb{R}^2$ which identifies $0 \times [0, 2\pi]$ to a point, and identifies $(r, 0)$ with $(r, 2\pi)$ for each $r \in [0\infty)$. It is an instructive exercise that polar coordinates give a homeomorphism of this identification space to $\mathbb{R}^2$. There are other convenient identifications we could use to explain polar coordinates, for example, we could take $[0, \infty)\times\mathbb{R}$

and identify $0 \times \mathbb{R}$ to a point, and identify (r, θ) with $(r, \theta + 2\pi n)$ for each integer n. Or we could use $[0, \infty) \times I$ where $I \subset \mathbb{R}$ is any interval of length 2π and use the appropriate identifications on the boundary.

Example 6.6. Another example is the use of spherical coordinates on the unit sphere $S^2 \subset \mathbb{R}^3$. If $\mathbf{x} = (x, y, z) \in S^2$, let ϕ denote the angle between $\mathbf{x}$ and the positive z-axis, and let θ be the angle between the projection of $\mathbf{x}$ to the xy-plane and the positive x-axis. Then we have $x = \sin\phi\cos\theta$, $y = \sin\phi\sin\theta$, and $z = \cos\phi$. Consequently $dx = \cos\phi\cos\theta\ d\phi - \sin\phi\sin\theta\ d\theta$, $dy = \cos\phi\sin\theta\ d\phi + \sin\phi\cos\theta\ d\theta$, and $dz = -\sin\phi\ d\phi$. A short computation gives

$$ds^2 = d\phi^2 + \sin^2\phi\ d\theta^2. \tag{6.7}$$

6.1.2. *Absolute Minimizers.* We now give some examples of curves of minimum length joining two points. The simplest example is of course a line segment in the plane, say the segment $(x, 0)$, $0 \le x \le a$ for some fixed $a > 0$. Suppose γ is a piecewise smooth curve joining $(0,0)$ and $(a, 0)$. Then $\gamma(t) = (x(t), y(t))$, $0 \le t \le 1$, and $x(0) = 0$, $x(1) = a$. So

$$\begin{aligned} L(\gamma) = &\ \textstyle\int_0^1 \sqrt{x'(t)^2 + y'(t)^2}dt \ge \int_0^1 \sqrt{x'(t)^2}dt \ge \\ &\textstyle\int_0^1 x'(t)dt = x(1) - x(0) = a, \end{aligned} \tag{6.8}$$

which shows that any curve from $(0,0)$ to $(a,0)$ has length at least a. Since the line segment has length a, this shows that the line segment is gives the absolute minimum of the length of connecting curves.

Note that the calculation in 6.8 actually gave more: the length of any curve connecting the y axis with the line $x = a$ is at least a.

This calculation can also be done in polar coordinates (taking, perhaps, some care with the identifications explained in Example 6.5 in the case where the curve crosses the boundary of the chosen domain of the coordinate system). If γ is a curve from the origin to a point on the circle $r = a$, in other words, $\gamma(t) = (r(t), \theta(t))$, where $r(0) = 0$ and $r(1) = a$, then

$$\begin{aligned} L(\gamma) \ = &\ \textstyle\int_0^1 \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2}\ dt \ge \int_0^1 \sqrt{r'(t)^2}\ dt \ge \\ &\textstyle\int_0^1 r'(t)\ dt = r(1) - r(0) = a, \end{aligned} \tag{6.9}$$

which shows that the length of any curve from the origin to the circle $r = a$ is at least a. Since a line segment from the origin to this circle has length a, this shows again that line segments minimize length between their endpoints.

Finally, let's discuss the case of the unit sphere $S^2 \subset \mathbb{R}^3$. Let's take curves γ from the north pole $N = (0, 0, 1)$ to a point other than the south pole, in other words, to a point with $\phi = \alpha$, where $0 < \alpha < \pi$, say the point $(0, \sin\alpha, \cos\alpha)$ corresponding to $\theta = \frac{\pi}{2}$ and $\phi = \alpha$ in spherical coordinates of Example 6.6. As before, we take a curve $\gamma(t) = (\phi(t), \theta(t))$ with $\phi(0) = 0$

and $\phi(1) = \alpha$ and compute:

$$(6.10)\ L(\gamma) \ = \ \int_0^1 \sqrt{\phi'(t)^2 + \sin^2\phi(t)\ \theta'(t)^2}\ dt \geq \int_0^1 \sqrt{\phi'(t)^2}\ dt \geq \int_0^1 \phi(t)\ dt = \phi(1) = \phi(0) = \alpha,$$

showing that any curve from the north pole N (corresponding to $\phi = 0$) to a point on the parallel $z = \cos\alpha$ (corresponding to $\phi = \alpha$) has length at least alpha. Since the great circle arc from the north pole to this point has length α, this shows the following theorem. In the theorem, the *shortest great circle arc joining* $\mathbf{x}$ *and* $\mathbf{y}$*,* $\mathbf{y} \neq \pm\mathbf{x}$, we mean the following. First, the *great circle* determined by $\mathbf{x}$ and $\mathbf{y}$ we mean the intersection with S^2 of the plane $< \mathbf{x}, \mathbf{y} >$ determined by $\mathbf{x}$ and $\mathbf{y}$ (the *span* of $\mathbf{x}$ and $\mathbf{y}$ on the language of liner algebra). They determine a plane because $\mathbf{x} \neq \pm\mathbf{y}$. This intersection is a circle of radius 1 containing $\mathbf{x}$ and $\mathbf{y}$, and by the *shortest great circle arc* we mean the shorter of the two arcs in this circle joining $\mathbf{x}$ and $\mathbf{y}$. There is a shorter one again because $\mathbf{x} \neq \pm\mathbf{y}$.

Theorem 6.3. *Let* $\mathbf{x}, \mathbf{y} \in S^2$*,* $\mathbf{y} \neq \pm\mathbf{x}$*, and let* γ *be the shortest great circle arc joining* $\mathbf{x}$ *and* $\mathbf{y}$*. Then* γ *is the shortest curve on* S^2 *joining* $\mathbf{x}$ *and* $\mathbf{y}$*.*

Proof. If $\mathbf{x} = N$ the north pole, then $\mathbf{y}$ would be different form the south pole, and we have just proved that the shortest great circle arc minimizes length. If $\mathbf{x}$ is any other point on S^2, then there is a rotation R of $\mathbb{R}^3$ that takes $\mathbf{x}$ to N: $R(\mathbf{x}) = N$. Then $R(\mathbf{y})$ is different from the south pole, thus the shortest great circle arc from $R(\mathbf{x})$ to $R(\mathbf{y})$ minimizes, so does R^{-1} of this arc, which is the shortest great circle arc joining $\mathbf{x}$ and $\mathbf{y}$. □

Remark 6.2. If $\mathbf{y} = -\mathbf{x}$, say if we take N and the south pole $\mathbb{N}$, then the computation of Equation 6.10 with $\alpha = \pi$ shows that any great circle arc passing through N and $-N$ is still length minimizing, its length is π. But there are infinitely many such arcs, one for each value of θ. So minimizers exist, but are not unique. But this is good enough to give us the following theorem:

Theorem 6.4. *The spherical metric of Example 1.6 is the same metric on* S^2 *as the intrinsic metric of Example 1.9.*

Proof. We have seen that for any $\mathbf{x}, \mathbf{y} \in S^2$,

$$\cos^{-1}(\mathbf{x} \cdot \mathbf{y}) = \inf\{L(\gamma) : \gamma \text{ a piecewise differentiable path from } \mathbf{x} \text{ to } \mathbf{y}\},$$

since, for $\mathbf{x}$ the north pole, the left hand side is ϕ, where (ϕ, θ) are the spherical coordinates of $\mathbf{y}$, and so is the right hand side. □

6.2. **Geodesics.** We now study length minimizing curves in the general smooth surface, generalizing the discussion in the plane and sphere. A look at the sphere shows that the concept of length minimizing is more subtle than in the plane. Experience has shown that it is easier to look at these curves from the point of view of differential equations. We begin by deriving this equation as a necessary condition for minimizing length.

6.2.1. *The First Variation Formula for Arclength.* Let $S \subset \mathbb{R}^3$ be a smooth surface and let $\gamma : [0, L_0] \to S$ be a smooth curve, parametrized by arclength, of length L_0. To derive a necessary condition for γ to be the shortest curve joining its endpoints $P = \gamma(0)$ and $Q = \gamma(1)$, it is natural to consider *variations of* γ, meaning *smooth maps*

$$\tilde{\gamma} : [0, L_0] \times (-\epsilon, \epsilon) \to S \text{ with } \tilde{\gamma}(s, 0) = \gamma(s) \text{ for all } s \in [0, L_0].$$

If, in addition, we have that

$$\tilde{\gamma}(0, t) = P, \; \tilde{\gamma}(L_0, t) = Q \text{ for all } t \in (-\epsilon, \epsilon),$$

we say that $\tilde{\gamma}$ is a *variation of* γ *preserving the endpoints*. Thus a variation of γ is a one parameter family of curves, depending on a $t \in (-\epsilon, \epsilon)$, where the curve $t = 0$ is γ. A variation of γ preserves endpoints if all these curves join P and Q. Moreover, it is assumed that this family is a smooth map of the rectangle $[0, L_0] \times (-\epsilon, \epsilon)$ to S. Note that s is arclength on γ but not on the other curves. Let

$$L(t) = \int_0^{L_0} \left(\tilde{\gamma}_s(s,t) \cdot \tilde{\gamma}_s(s,t)\right)^{1/2} \, ds$$

be the length of the t-th curve of the variation $s \to \tilde{\gamma}(s, t)$. A necessary condition for γ to be length minimizing is that for *every* variation of γ preserving the endpoints, $\frac{dL}{dt}(0) = 0$.

Let's compute this derivative for an arbitrary variation (not necessarily preserving endpoints), and then specialize to endpoint preserving. First, differentiating under the integral sign we get

$$\frac{dL}{dt} = \int_0^{L_0} \frac{1}{2}(\tilde{\gamma}_s(s,t) \cdot \tilde{\gamma}_s(s,t))^{-1/2}(2 \; \tilde{\gamma}_{st}(s,t) \cdot \tilde{\gamma}_s(s,t)) \; ds.$$

Evaluating at $t = 0$ and using the fact that $\tilde{\gamma}(s, 0) = \gamma(s)$ is parametrized by arclength, equivalently, $\tilde{\gamma}_s(s,0) \cdot \tilde{\gamma}_s(s,0) = 1$, we get

$$\frac{dL}{dt}(0) = \int_0^{L_0} \tilde{\gamma}_{st}(s,0) \cdot \tilde{\gamma}_s(s,0) \; ds.$$

Next, integrate by parts, using the formula

$$(\tilde{\gamma}_t(s,0) \cdot \tilde{\gamma}_s(s,0))_s = \tilde{\gamma}_{ts}(s,0) \cdot \tilde{\gamma}_s(s,0) \; + \; \tilde{\gamma}_t(s,0) \cdot \tilde{\gamma}_{ss}(s,0)$$

and noting that, by the smoothness of $\tilde{\gamma}$, we have equality of mixed partials: $\tilde{\gamma}_{st} = \tilde{\gamma}_{ts}$:

$$\frac{dL}{dt}(0) = (\tilde{\gamma}_t(s,0) \cdot \tilde{\gamma}_s(s,0))|_0^{L_0} - \int_0^{L_0} \tilde{\gamma}(s,0) \cdot \tilde{\gamma}_{ss}(s,0) \; ds$$

Let us simplify this formula. First recall that $\tilde{\gamma}(s,0) = \gamma(s)$, thus $\tilde{\gamma}_s(s,0) = \gamma'(s)$ and $\tilde{\gamma}_{ss}(s,0) = \gamma''(s)$. Next, define a vector field $V(s)$ along γ by

$$V(s) = \tilde{\gamma}_t(s,0).$$

This is called the *variation vector field.* It tells us how we are moving away from γ at $t = 0$. More precisely, $V(s)$ is a vector based at $\gamma(s)$ and is the velocity vector of the curve $t \to \tilde{\gamma}(s,t)$ at $t = 0$, so it is telling us the velocity at which $\gamma(s)$ initially moves under the variation. Observe that if the variation preserves endpoints, then $V(0) = 0$ and $V(L_0) = 0$, since these point do not move at all.

Using this notation, we can rewrite the above formula as

$$\frac{dL}{dt}(0) = V(s) \cdot \gamma'(s)|_0^{L_0} \;-\; \int_0^{L_0} V(s) \cdot \gamma''(s)\; ds.$$

Noting that $V(s)$ is a vector tangent to S at the point $\gamma(s)$, the inner product under the integral sign is the same as $V(s) \cdot \gamma''(s)^T$, where $\gamma''(s)^T$ denotes the *tangential component* of $\gamma''(s)$. So we can finally rewrite the formula as

$$\frac{dL}{dt}(0) = V(s) \cdot \gamma'(s)|_0^{L_0} \;-\; \int_0^{L_0} V(s) \cdot \gamma''(s)^T\; ds. \tag{6.11}$$

This is called the *first variation formula for arclength.*

6.2.2. *The Geodesic Equation.* Let us now see what the first variation formula implies for our original problem, where the variation preserves endpoints. Then the first term of formula 6.11 vanishes, we get just the second term, which must vanish for all possible variation vector fields V.

Theorem 6.5. *$\int_0^{L_0} V(s) \cdot \gamma''(s)^T\, ds = 0$ for all possible variations of γ with fixed endpoints if and only if the tangential component γ''^T of γ'' vanishes: $\gamma''(s)^T = 0$ for all $s \in [0, L_0]$.*

Example 6.7. Before proceeding to the proof of the theorem, let us look at the example of the sphere S^2. Using spherical coordinates, let γ be the curve, depending on ϕ, given by holding ϕ constant, in other words, for fixed ϕ, $0 < \phi < \pi$, the curve

$$\gamma(\theta) = (\sin\phi\cos\theta, \sin\phi\sin\theta, \cos\phi),$$

which is a "parallel" on the sphere. It is parametrized proportional to arclength s, thus $\gamma''(\theta)$ is a constant multiple of $\gamma''(s)$. Then

$$\gamma''(\theta) = (-\sin\phi\cos\theta, -\sin\phi\cos\theta, 0),$$

which is perpendicular to the sphere if and only if it is a multiple of the vector $\gamma(\theta)$, which happens if and only if $\cos\phi = 0$, that is, $\phi = \pi/2$, in other words, γ is the equator. Thus the only parallel that satisfies the equation $\gamma''(s)^T = 0$ is the equator, which is the only parallel that is a great circle.

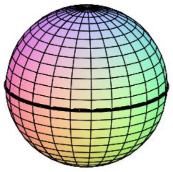

Figure 6.2. Equator and Meridians are Geodesics

This confirms that our differential equation does characterize great circles, which our intuition tells us are the geodesics on the sphere. (Note also that in this variation of the equator, the equator is a local maximum, rather than a local minimum. In fact, as a function of the parameter ϕ in the variation, $L(\phi) = 2\pi \sin \phi$ which has a maximum at $\phi = \pi/2$.) In the same spirit we can readily verify that all the meridians $\theta = const$ are geodesics. Note that ϕ is arclength along the meridians.

Proof of the Theorem. Let us begin with two observations:

1. Since $\gamma''^T \cdot \gamma' = 0$ (because $\gamma' \cdot \gamma'$ is a constant), if we let $N(s)$ be a unit vector field along γ perpendicular to $\gamma'(s)$, then $\gamma''^T(s)$ is a multiple of $N(s)$. This multiple is traditionally written $\kappa_g(s)$ and (up to sign) is called the *geodesic curvature* of γ. This terminology will be discussed later, for the moment let's just write $\gamma''(s)^T = \kappa_g(s)N(s)$ for some smooth function κ_g.
2. It suffices to take $V(s)$ to be a multiple of $N(s)$: $V(s) = f(s)N(s)$ for some smooth function f on $[0, L_0]$. Then the integral in question becomes $\int_0^{L_0} f(s)\kappa_g(s)\, ds$ and we need to prove that if this integral is 0 for all f arising from variations of γ, then $\kappa_g(s) = 0$ for all $s \in [0, L_0]$.

We need the following lemma:

Lemma 6.1. *Let $(a, b) \subset \mathbb{R}$ be an interval. Then there is a smooth function $\phi : \mathbb{R} \to \mathbb{R}$ that is positive on (a, b) and vanishes on $\mathbb{R} \setminus (a, b)$.*

Proof. First check that the function defined by

$$f(x) = \begin{cases} e^{-1/x} \text{ if } x > 0, \\ 0 \text{ if } x \leq 0. \end{cases}$$

is smooth (of class C^∞). In fact, all its derivatives are defined and vanish at 0. Then, if $a < b$, the function $\phi(x) = f(x - a)f(b - x)$ satisfies the requirements of the lemma. This is the picture for $(a, b) = (0, 1)$:

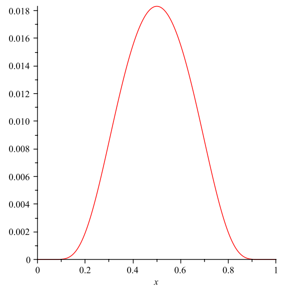

FIGURE 6.3. Smooth "Bump" Function

□

Now we can prove the theorem. Following observation (2) above, suppose $\kappa_g(s_0) \neq 0$, say $\kappa_g(s_0) > 0$ for some $s_0 \in (0, L_0)$. Then there exists an interval $(a,b) \subset (0, L_0)$ containing s_0 on which $\kappa_g > 0$. We may also assume that $\gamma|_{(a,b)} : (a,b) \to U \subset S$ where one of the projections p^x, p^y, p^z, let's call it p, maps U diffeomorphically to its image V. Let $g : V \to U$ be the inverse of $p|_U$, as in the proof of Theorem 6.1. Let ϕ be as in the Lemma and let $f = \phi|_{[0,L_0]}$. Then $\int_0^{L_0} f(s)\kappa_g(s)\, ds > 0$, contradicting the assumption, provided that the field $f(s)N(s)$ is a variation field, that is, provided that there exists a variation $\tilde{\gamma} : [0, L_0] \times (-\epsilon, \epsilon) \to S$, written $\tilde{\gamma}(s,t)$, of γ with $\tilde{\gamma}_t(s, 0) = f(s)N(s)$. But this is indeed the case. For example, we can define $\tilde{\gamma}$ by

$$\tilde{\gamma}(s,t) = \begin{cases} g(p(\gamma(s) + tf(s)N(s))) \text{ if } s \in (a,b), \\ \gamma(s) \text{ otherwise.} \end{cases}$$

in other words, form the variation $\gamma(s) + tf(s)N(s)$ by curves in $\mathbb{R}^3$ that give the desired variation vector field $f(s)N(s)$ but need not lie on S, and force them to lie on S by projecting to S in the indicated manner. This uses that p is defined on all of $\mathbb{R}^3$, so that $p(\gamma(s) + tf(s)N(s))$ makes sense. To get the correct derivative with respect to t at $t = 0$ we use that for $t = 0$ we are on S. Since g and $p|_U$ are inverse to each other, one gets from the chain rule that

$$\begin{aligned} \tilde{\gamma}_t(s,0) &= (d_{p(\gamma(s))}g \circ d_{\gamma(s)}p)(\frac{\partial}{\partial t}(\gamma(s) + tf(s)N(s))|_{t=0}) \\ &= (d_{p(\gamma(s))}g \circ d_{\gamma(s)}p)(f(s)N(s)) \\ &= f(s)N(s), \end{aligned}$$

because $d_{p(x)}g \circ d_x p = id$ on the tangent space to S at any $x \in U \subset S$, and dg, dP denote, as usual, the differentials of g and p. □

Definition 6.2. A smooth curve $\gamma : (a,b) \to S$ is called a *geodesic* in S if it satisfies $\gamma''(s)^T = 0$ for all $s \in (a,b)$.

Note that if γ is a geodesic, then $|\gamma'(s)|$ is constant, that is, γ is a constant speed curve, equivalently, parametrized proportional to arclength. The reason is that $(\gamma' \cdot \gamma')' = 2\gamma'' \cdot \gamma' = 2\gamma''^T \cdot \gamma' = 0$ if γ is a geodesic.

There are several notations used for γ''^T. For instance,

Definition 6.3. Let $\gamma : (a, b) \to S$ be a smooth curve and $V : (a, b) \to \mathbb{R}^3$ a smooth vector field along γ, meaning that V is a smooth map and for all $s \in (a, b)$, $V(s) \in T_{\gamma(s)}S$, the tangent plane to S at $\gamma(s)$.

(1) The tangential component $V'(s)^T$ is called the *covariant derivative* of V and is denoted DV/DS.
(2) γ is a *geodesic* if and only if $D\gamma'/Ds = 0$ for all $s \in (a, b)$.

Other notations for $D\gamma'/Ds = 0$ are commonly used, for example $D^2\gamma/Ds^2 = 0$, $D_{\gamma'}\gamma' = 0$, and some others.

6.2.3. *The Geodesic Equation in Local Coordinates.* To study the geodesic equation $\gamma''(s)^T = 0$ in more detail, we restrict γ to an interval that lies in the domain of some chart, and we use the notation of the second paragraph of Subsection 6.1.2, namely we have the smooth map $\mathbf{x} : U \to S$ which is the inverse of the chart, where $U \subset \mathbb{R}^2$ is open and we use u, v for the coordinates on U.

For each point $P \in S$ we write T_PS for the tangent plane of to S at P. This is the two-dimensional subspace of $\mathbb{R}^3$ of vectors which are tangent to S at P. For each $(u, v) \in U$, the vectors $\mathbf{x}_u(u, v)$ and $\mathbf{x}_v(u, v)$ form a basis for $T_{\mathbf{x}(u,v)}S$. The curve $\gamma(s) = \mathbf{x}(u(s), v(s))$ for some curve $(u(s), v(s))$ in U. We compute γ''. First, by the chain rule, $\gamma' = \mathbf{x}_u u' + \mathbf{x}_v v'$, differentiating once more using the product rule and chain rule, and combining some terms, we get

$$\gamma'' = \mathbf{x}_u u'' + \mathbf{x}_v v'' + \mathbf{x}_{uu}(u')^2 + 2\mathbf{x}_{uv}u'v' + \mathbf{x}_{vv}(v')^2.$$

Notice that the first two terms are tangential. So, to find γ''^T we need to find the tangential component of the sum of the last three terms. We do not need to do this explicitly at this moment (more will be said later), all we need is the general shape of the formula. The tangential component of the sum of the last three terms is of the form

$$q_1(u, v, u', v')\mathbf{x}_u + q_2(u, v, u', v')\mathbf{x}_v,$$

where q_1 and q_2 are quadratic functions of u', v' with coefficients smooth functions of u, v, written more explicitly below. Putting this together we see that the equation $\gamma''^T = 0$ is equivalent to a system of second order ODE's

$$\begin{aligned} u'' + \Gamma^1_{11}u'^2 + 2\Gamma^1_{12}u'v' + \Gamma^1_{22}v'^2 = 0 \\ v'' + \Gamma^2_{11}u'^2 + 2\Gamma^2_{12}u'v' + \Gamma^2_{22}v'^2 = 0 \end{aligned} \tag{6.12}$$

where the six coefficients $\Gamma^i_{jk} = \Gamma^i_{jk}(u, v)$ are smooth functions on U.

We will later discuss how to obtain formulas for the coefficients. For the time being all we need is that it is a system of second order ODE's where the coefficients of u'', v'' are 1. There is a standard existence and uniqueness theorem for the initial value problem, together with a theorem on the smooth dependence of the solution on the initial conditions. Let us write $\mathbf{u} = \mathbf{u}(s) = ((u(s), v(s))$ for a solution of the system 6.12. Let us write p for a point in U and $\mathbf{v}$ for a vector in $\mathbb{R}^2$, which we think of as a tangent vector to U at p.

Theorem 6.6. *Given any $p_0 \in U$ and any $\mathbf{v}_0 \in \mathbb{R}^2$ there exists a neighborhood W of $(p_0, \mathbf{v}_0)$ in $U \times \mathbb{R}^2$ and an interval $(-a, a) \subset \mathbb{R}$ so that for any $(p, \mathbf{v}) \in W$ there exists a unique solution $\mathbf{u}(s) = (u(s), v(s))$ of the system 6.12 satisfying the initial conditions $\mathbf{u}(0) = p$ and $\mathbf{u}'(0) = \mathbf{v}$. Let $\mathbf{u}(s, p, \mathbf{v})$ denote this solution. It depends smoothly on the initial conditions $p, \mathbf{v}$ in the sense that the map $\mathbf{u} : (-a, a) \times W \to U$ given by $(s, p, \mathbf{v}) \mapsto \mathbf{u}(s, p, \mathbf{v})$ is smooth.*

A proof of this theorem, stated for a system $\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t))$ of first order equations, where $U \subset \mathbb{R}^n$ is an open set, $I \subset \mathbb{R}$ is an open interval, and $\mathbf{f} : I \times U \to \mathbb{R}^n$ is a smooth map, can be found in any rigorous text on ODE's, for example, in Chapter 2 of [2] or Chapter 4 of [1]. (See also Chapter 4 of [3], paticularly sections 4.6 and 4.7 for a discussion of the geodesic equation.) A second order system in n unknown functions is equivalent to a first order system in $2n$ unknown functions. Note that our system is equivalent to a first order system of a more special form, $\mathbf{x}'(t) = \mathbf{f}(\mathbf{x})$, an autonomous system ($\mathbf{f}$ does not depend on t).

Our solution $\mathbf{u}(s, p, \mathbf{v})$ satisfies the identity

$$\mathbf{u}(rs, p, \mathbf{v}) = \mathbf{u}(s, p, r\mathbf{v}) \quad \text{for any } r \in \mathbb{R} \tag{6.13}$$

because both sides are solutions of the ODE with value p and first derivative $r\mathbf{v}$ at $s = 0$.

Fix $p \in U$. To simplify the calculations, we may make a linear change of coordinates (u, v) so that $p = (0, 0) = 0$ (by translating the coordinates) and so that, at 0, the differential of our parametrization $\mathbf{x}$ of S , $d_0\mathbf{x} : \mathbb{R}^2 = T_0\mathbb{R}^2 \to T_{\mathbf{x}(0)}S$, is an isometry. This last requirement is achieved as follows. The set $\{\mathbf{v} \in \mathbb{R}^2 : |d_0\mathbf{x}(\mathbf{v})| = 1\}$ is an ellipse. If it is a circle, multiply by a factor to make the circle of radius one. If it is not a circle, apply the linear transformation with eigenvectors pointing in the direction of the axes and eigenvalues the inverses of the semi-axes, to take this ellipse into a circle of radius one. Another way of saying this is that, at 0, $dx^2 + dy^2 + dz^2 = du^2 + dv^2$, equivalently, that the coefficients g_{ij} of Equation 6.5 satisfy $g_{11}(0) = g_{22}(0) = 1$ and $g_{12}(0) = 0$.

By Theorem 6.6, for any $\mathbf{v}_0$ so that $|\mathbf{v}_0| = 1$, there exists a neighborhood V of $\mathbf{v}_0$ and an $a > 0$ so that the solution $\mathbf{u}(s, 0.\mathbf{v})$ exists for all $(s, \mathbf{v}) \in (-a, a) \times V$. By the *compactness* of the circle $S^1 = \{|\mathbf{v}| = 1\}$, it can be covered by finitely many such V, and taking b to be the smallest of the corresponding a's, we get the following lemma:

Lemma 6.2. *There exists $b \in (0, \infty]$ so that the solution $\mathbf{u}(s, 0, \mathbf{v})$ of the geodesic equation 6.12 is defined for all $(s, \mathbf{v}) \in (-b, b) \times S^1$.*

In other words, *for any fixed length $c < b$ all geodesics through 0 in all directions $\mathbf{v} \in S^1$ are defined up to c.* Note that $b = \infty$ is possible, in fact, it is the ideal situation.

The reason for requiring that $d_0\mathbf{x}$ be an isometry is to insure that s is arclength along these solutions $\mathbf{u}(s, 0, \mathbf{v})$ with $|\mathbf{v}| = 1$, where $|\mathbf{v}|$ is the Euclidean length in $\mathbb{R}^2$. Otherwise we would have to use the length measurement $|d_0\mathbf{x}(\mathbf{v})| = \sqrt{g_{11}(0)v_1^2 + 2g_{12}(0)v_1v_2 + g_{22}(0)v_2^2}$ where g_{ij} are as in Equation 6.5 and $\mathbf{v} = (v_1, v_2)$.

Using the formula 6.13, for any $\mathbf{v} \in \mathbb{R}^2$, $\mathbf{v} \neq 0$, we have $\mathbf{u}(1, 0, \mathbf{v}) = \mathbf{u}(|\mathbf{v}|, 0, \mathbf{v}/|\mathbf{v}|)$ is defined provided $|\mathbf{v}| < b$, with b as in Lemma 6.2. In other words, the map $\mathbf{v} \mapsto \mathbf{u}(1, 0, \mathbf{v})$ is defined and smooth on the ball $\{|\mathbf{v}| < b\}$. Let us call this map $f : B(0, b) \to U$, and let's compute its differential at 0, $d_0f(v) = \lim_{t\to 0}(f(t\mathbf{v}) - f(0))/t = \lim_{t\to 0} f(t\mathbf{x})/t = \lim_{t\to 0} \mathbf{u}(1, 0, t\mathbf{v})/t = \lim_{t\to 0} \mathbf{u}(t, 0, \mathbf{v})/t = \mathbf{u}'(0, 0, \mathbf{v}) = \mathbf{v}$, where the second to last equality is Equation 6.13 and the last equality is the definition of $\mathbf{u}(s, p, \mathbf{v})$ in terms of initial conditions. Thus we get $d_0f = id$. By the *inverse function theorem* we get that *there exists an $\epsilon > 0$ so that $f|_{B(0,\epsilon)}$ is a diffeomorphism of $B(0, \epsilon)$ onto its image.*

6.2.4. *Exponential Map and Geodesic Polar Coordinates.* We transfer the information just obtained in local coordinates back to the surface S. Recall that $0 \in U \subset \mathbb{R}^2$, that $\mathbf{x} : U \to S$ is a diffeomorphism onto its image, $P = \mathbf{x}(0)$ and that $d_0\mathbf{x} : T_0U = \mathbb{R}^2 \to T_PS$ is an isometry.

For $V \in T_PS$, let $\gamma(s, P, V)$ be the solution of $\gamma''(s)^T = 0$ satisfying $\gamma(0) = P$ and $\gamma'(0) = V$. Our discussion of the geodesic equation in the the local coordinates $(u, v) \in U$ proves the following theorem:

Theorem 6.7. (1) *There is $b \in (0, \infty]$ so that $\gamma(1, P, V)$ is defined for all $v \in B(0, b) \subset T_PS$.*

(2) *Define a map $\exp_P : B(0, b) \to S$ by $\exp_P(V) = \gamma(1, P, V)$. Then the differential $d_P \exp_P : T_PS \to T_PS$ is the identity.*

(3) *There exists $\epsilon > 0$ so that $\exp_P|_{B(0,\epsilon)}$ is a diffeomorphism of $B(0, \epsilon)$ onto its image.*

Proof. For any $r \leq b$, where b is as in Lemma 6.2, we have the following diagram

$$\begin{tikzcd}
B(0,r) \subset \mathbb{R}^2 \arrow[r, "d_0\mathbf{x}"] \arrow[d, "f"'] & B(0,r) \subset T_PS \arrow[d, "\exp_P"] \\
U \arrow[r, "\mathbf{x}"] & S
\end{tikzcd}$$

where the left half is the discussion in local coordinates just finished in subsection 6.2.3, and the right half is the map just defined. We have just proved the three parts of this theorem for the left half of the diagram, the diffeomeorphism $\mathbf{x}$ transfers the theorem to the right half. For part (1) take $r = b$, for part (3) take $r = \epsilon$ as in the last sentence of subsection 6.2.3.

□

The traditional notation and terminology for this map comes from the fact that in some examples the matrix exponential could be seen as a special case of this map:

Definition 6.4. The map $\exp_P : B(0,b) \to S$ defined in (2) of Theorem 6.7 is called the *exponential map at* P.

To make matters concrete, let's keep in mind the example $S = S^2$ and $P = N$ the north pole. The rays through the origin in T_NS^2 are mapped to the meridians (great circles through N). Note that $\exp_N$ is defined on the whole tangent space ($b = \infty$ in Theorem 6.7), but its restriction to the ball of radius r is a diffeomorphism only for $r < \pi$.

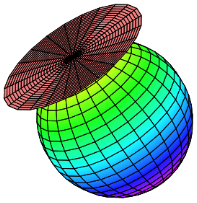

FIGURE 6.4. Exponential Map for the Sphere

The parametrization of a neighborhood of $P \in S$ by the ball $B(0,\epsilon) \subset T_pS$ turns out to be a very natural one. We will change notation, forget the arbitrary parametrization $\mathbf{x}(u,v)$ of subsection 6.2.3 and for the rest of this section we will use the convenient letters u, v for rectangular coordinates in T_PS with respect to some orthonormal basis $\mathbf{e}_1, \mathbf{e}_2$, and the convenient notation $\mathbf{x}$ for the map $\exp_P : B(0,\epsilon) \to S$, that is, $\mathbf{x}(u,v) = \exp_P(u\mathbf{e}_1 + v\mathbf{e}_2)$ for $(u,v) \in B(0,\epsilon) \subset \mathbb{R}^2$. A glance at Figure 6.4 suggests that we should also use the associated polar coordinates (r,θ) so that $u = r\cos\theta, v = r\sin\theta)$. When doing so, we will use the usual abbreviated, if somewhat inaccurate notation $\mathbf{x}(r,\theta)$ for $\mathbf{x}(r\cos\theta, r\sin\theta)$.

Definition 6.5. The parametrization $\mathbf{x} : B(0,\epsilon) \to S$ just defined will be called a *normal coordinates centered at* P. When this parametrization is expressed in polar coordinates, the coordinates r, θ will be called *geodesic polar coordinates centered at* P.

Figure 6.4 also suggests that the curves $r = const$ and $\theta = const$ should be perpendicular to each other. This is indeed the case:

Theorem 6.8. *(Gauss's Lemma): In a geodesic polar coordinate system* $\mathbf{x} : B(0,\epsilon) \to S$, $\mathbf{x}_r \cdot \mathbf{x}_\theta = 0$. *Equivalently, in this coordinate system,* $ds^2 = dr^2 + g(r,\theta)^2\, d\theta^2$ *for some positive smooth function* g.

Proof. This follows immediately from the first variation formula 6.11. For fixed $r_0 < \epsilon$, and any $\theta \in [0, 2\pi]$, the curve $\gamma(\cdot,\theta) : [0, r_0] \to S$ given by $\gamma(r,\theta) = \mathbf{x}(r,\theta)$ is a geodesic of length r_0, so its length $L(\theta)$ is independent of θ. For any fixed θ_0, $\gamma(r,\theta)$ is then a variation of $\gamma(\cdot,\theta_0)$ by geodesics of constant length, keeping $\gamma(0,\theta) = P$ fixed, and variation vector field $V(r) = \mathbf{x}_\theta(r,\theta)$ Thus formula 6.11 reads

$$0 = L'(\theta_0) = \mathbf{x}_\theta \cdot \mathbf{x}_r|_{r=0}^{r=r_0} = \mathbf{x}_\theta(r_0,\theta_0) \cdot \mathbf{x}_r(r_0,\theta_0).$$

Since r_0, θ_0 are arbitrary, this means that $\mathbf{x}_\theta \cdot \mathbf{x}_r = 0$ everywhere, as asserted. Recalling formulas 6.4 and 6.5, we see that $\mathbf{x}_r \cdot \mathbf{x}_r = 1$ (since, for each θ, $\mathbf{x}(r,\theta)$ is a unit speed geodesic), $\mathbf{x}_r \cdot \mathbf{x}_\theta = 0$ (as just proved) and $\mathbf{x}_\theta \cdot \mathbf{x}_\theta = g_{22}$. Since g_{22} is a positive smooth function, we can write $g_{22} = g^2$ for some positive smooth function g. □

Now that we have geodesic polar coordinates, we can repeat the reasoning we used in Equations 6.9 and 6.10 in any surface. First, Gauss's Lemma justifies the following terminology:

Definition 6.6. In a geodesic polar coordinate system centered at P, the curves $r \mapsto \mathbf{x}(r,\theta)$, $0 \le \theta \le 2\pi$, are called the *geodesic rays through* P. The curves $\theta \mapsto \mathbf{x}(r,\theta)$ are called the *geodesic circles centered at* P.

Theorem 6.9. *Let* $\mathbf{x} : B_T(0,\epsilon) \to S$ *be a geodesic polar coordinate system centered at* P, *where* B_T *denotes the ball in the Euclidean metric of the tangent plane* T_PS.

(1) *For any* $0 \le r_0 < r_1 < \epsilon$ *and any fixed* θ, *the geodesic segments* $\mathbf{x}(r,\theta)$, $r_0 \le r \le r_1$ *are the shortest piecewise differentiable curves in* S *joining a point in the geodesic circle* $r = r_0$ *to a point in the geodesic circle* $r = r_1$.

(2) *In particular, the geodesic rays through P are the shortest piecewise differentiable curves in S joining P to any other point Q in $\mathbf{x}(B_T(0,\epsilon))$. This length is $d_S(P,Q)$, where d_S is the intrinsic distance on S as defined in Example 1.9 or Definition 1.2(5).*

(3) *Let $B_S(P,r)$ denote the ball of given center and radius in the intrinsic distance d_S. Then, for any $r<\epsilon$, $B_S(P,r)=\mathbf{x}(B_T(0,\epsilon))$.*

Proof. We argue as we did in polar or spherical coordinates in Equations 6.9 or 6.10. Consider first a smooth curve $\gamma(t)=\mathbf{x}(r(t),\theta(t))$, $0\le t\le 1$ lying in the image of the geodesic polar coordinate system, and suppose that $r(0)=r_0$ and $r(1)=r_1$. Then we have

$$\begin{aligned}(6.14)\qquad L(\gamma) \;=\; &\int_0^1 \sqrt{r'(t)^2+g(r(t),\theta(t))^2\;\theta'(t)^2}\;dt \ge \\ &\int_0^1 \sqrt{r'(t)^2}\;dt \ge \int_0^1 r'(t)\;dt \;=r(1)-r(0)=r_1-r_0.\end{aligned}$$

Observe that the first inequality is strict unless $\theta'=0$, that is, θ is constant, that is, γ lies on a geodesic ray. The second inequality is strict unless $r'\ge 0$, that is, r is an increasing function of t, that is, we are covering a segment $\mathbf{x}(r,\theta)$, $r_1\le r\le r_2$, monotonically. Thus $L(\gamma)>r_2-r_1$ unless γ covers a segment monotonically. Since the length of the segment is r_2-r_1, it is an absolute minimizer among the curves considered: smooth curves lying in $\mathbf{x}(B_T(0,\epsilon))$.

If γ is just piecewise smooth, but still lies in the coordinate system, divide $[0,1]$ into subintervals by taking $0=t_0<t_1<\cdots<t_n=1$, where $\gamma_i|_{[t_{i-1},t_i]}$ is smooth. Let $\rho_i=r(t_i)$. The same reasoning as in Equation 6.14, but refined to take into account the possibility that $\rho_i<\rho_{i-1}$, gives the more useful inequality

$$L(\gamma_i)\ge\int_{t_{i-1}}^{t_i}\sqrt{r'(t)^2}\;dt\ge\begin{cases}\int_{t_{i-1}}^{t_i} r'(t)\;dt=\rho_i-\rho_{i-1} & \text{if } \rho_{i-1}<\rho_i,\\ \int_{t_{i-1}}^{t_i} -r'(t)\;dt=\rho_{i-1}-\rho_i & \text{if } \rho_i<\rho_{i-1}.\end{cases}$$

By more useful inequality we mean that the first inequality is useless in the second case, because it gives a negative lower bound, while the second inequality is equally useless in the first case.

In either case we get the inequality $L(\gamma_i)\ge|\rho_i-\rho_{i-1}|$, with equality if and only if γ_i travels monotonically along a segment $\mathbf{x}(r,\theta_i)$, for some fixed θ_i, and with $\rho_{i-1}\le r\le\rho_i$, or $\rho_i\le r\le\rho_{i-1}$ as the case many be. We thus get

$$L(\gamma)=\sum L(\gamma_i)\ge\sum|\rho_i-\rho_{i-1}|\ge\sum(\rho_i-\rho_{i-1})=\rho_n-\rho_0=r_1-r_0,$$

with equality $L(\gamma)=r_1-r_0$ if and only of all these inequalities are equalities and $\rho_0<\rho_1\cdots<\rho_n$. In particular, each γ_i must be a segment traveled in monotonically increasing fashion. Since γ is a continuous path, all the θ_i must be the same (modulo 2π), hence γ is a segment. Thus segments of geodesic rays absolutely minimize length in the class of piecewise smooth paths in the image of the geodesic coordinate system.

Finally, if $\gamma([0,1])$ does not lie on the image of the geodesic polar coordinate system, then, for some R, $r_1 < R < \epsilon$, γ does not lie in the image of the closed ball $\bar{B}_T(0,R)$. By the continuity of γ there is τ, $0 < \tau < 1$ so that $\gamma(\tau)$ lies on the geodesic circle of radius R and $\gamma([0,\tau])$ lies in the image of $\bar{B}_T(0,R)$. (This can be proved as follows: writing $\gamma(t) = \mathbf{x}(r(t),\theta(t))$, r is a continuous function of t with $r(0) = 0$ and $r(t) > R$ for some t. Thus there exist t_1, $0 < t_1 < t$, so that $r(t_1) = R$. Let $\tau = \inf\{t_1 : r(t_1) = R\}$. It is easily seen that $0 < \tau < 1$ and has the required property.) Then $L(\gamma) \geq L(\gamma|_{[0,\tau]}) \geq R - r_0 > r_1 - r_0$, so it cannot be length minimizing. This proves the first statement of the theorem. See Figure 6.5 for a sketch of what a geodesic coordinate system may look like. The wavy curves represent some of the possibilities we considered in the proof. The geodesic rays realize the distance between geodesic circles.

The remaining two statements in the theorem are easy consequences of the first.

□

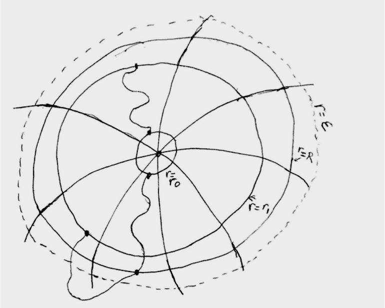

FIGURE 6.5. Geodesic Rays Minimize Length

Remark 6.3. (1) Observe that Theorem 6.9 says, in particular, that sufficiently small balls $B(P,r)$ in the intrinsic distance d_S look roughly like the balls in the Euclidean metric in the plane. In particular, there is a unique minimizing segment from the center P to any other point $Q \in B_S(P,r)$. This is in marked contrast with the Taxicab metric of Example 1.4 or the Supremum distance of Example 1.5 where there are uncountably many shortest curves joining P and Q, no matter how close P and Q are.

(2) It follows easily from Theorem 6.9 that the topology of the intrinsic metric d_S is the subspace topology on $S \subset \mathbb{R}^3$. This fact can also be proved from first principles, without using this detailed theorem.

6.3. **A First Glance at Gaussian Curvature.** We study in more detail the function $g(r,\theta)$ in the expression for ds^2 in geodesic polar coordinates given by Gauss's Lemma (Theorem 6.8)

$$ds^2 = dr^2 + g(r,\theta)^2\ d\theta^2. \tag{6.15}$$

Recall that we can also use rectangular coordinates u, v on $T_P S$ and

$$\begin{aligned} dr^2 + g(r,\theta)^2\ d\theta^2 \ &= \ g_{11}\ du^2 + 2g_{12}\ dudv + g_{22}\ dv^2, \\ \text{were} \ \ g_{11}(0) &= g_{22}(0) = 1 \text{ and } g_{12}(0) = 0, \end{aligned} \tag{6.16}$$

where g_{ij} are smooth functions of u, v, their values at 0 being a consequence of $d_0 \exp_P = id$ (Theorem 6.7 (2)). The two coordinates are related by the usual formulas $u = r\cos\theta$, $v = r\sin\theta$. Let us expand the function $g(r,\theta)$ in powers of r with coefficients that are functions of θ:

$$g(r,\theta) = f_0(\theta) + f_1(\theta)r + f_2(\theta)r^2 + f_3(\theta)r^3 + O(r^4), \tag{6.17}$$

where $O(r^4)$ stands for a function $h(r,\theta)$ so that $|h(r,\theta)|/r^4 \le C$, where C is an absolute constant (independent of θ). Polar coordinates are singular at the origin, but the fact that the functions g_{ij} in right hand side of Equation 6.16 are smooth in u, v imposes strong restrictions on the possible coefficients $f_i(\theta)$ of $g(r,\theta)$. To derive these restrictions, write dr, $d\theta$ in terms of $du.\, dv$:

$$\begin{aligned} dr = (udu + vdv)/r, \quad u/r, v/r \ \text{ homogenous of degree } 0 \\ d\theta = (udv - vdu)/r^2, \quad u/r^2, v/r^2 \ \text{ homogenous of degree } -1 \end{aligned} \tag{6.18}$$

where $r = (u^2{+}v^2)^{\frac{1}{2}}$. Recall that a function $\phi(u,v)$ is said to be *homogeneous of degree* k if $\phi(tu,tv) = t^k\phi(u,v)$ for all $t > 0$. If $k \ge 0$ homogenous polynomials of degree k are homogeneous functions of degree k, but there are homogeneous functions that are not polynomials. For example, u/r and v/r are homogeneous of degree 1, but are not linear functions.

We can expand the left hand side of Equation 6.16 in powers of r using Equation 6.17. If we express dr and $d\theta$ in terms of du and dv by using the formulas 6.18, we must get the expression on the right hand side of Equation 6.16, where the coefficients g_{ij} are smooth functions of u, v, so that the homogenous of degree k in its Taylor expansion have to be homogeneous *polynomials* of degree k in u, v. This polynomial restriction if very strong, and gives the following conclusions:

The coefficients $f_i(\theta)$ in Equation 6.17 satisfy;

(1) $f_0(\theta) = 0$. (Otherwise, there would be a term $f_0(\theta)^2 d\theta^2$ homogeneous with coefficients of du, dv homogeneous of degree -2.)

(2) $f_1(\theta) = 1$. (Because the terms homogeneous of degree 0 are $dr^2 + f_1(\theta)^2 r^2 d\theta^2$, but this must be the same as $du^2 + dv^2 = dr^2 + r^2 d\theta^2$, thus $f_1(\theta) = 1$.)

(3) $f_2(\theta) = 0$. (Because $dr^2 + (r + f_2(\theta)r^2 + \dots)^2 d\theta^2 = dr^2 + (r^2 + 2f_2(\theta)r^3 + \dots)d\theta^2$ and the term homogeneous of degree 1 must be $2f_2(\theta)(udv - vdu)^2/r$ which, in terms of u, v has coefficient of dv^2 equal to $2f_2(\theta)u^2/r$, which is not a polynomial in u unless it is identically zero, thus $f_2 = 0$.)

(4) $f_3(\theta) = c$ where c is a constant, independent of θ. (Because $dr^2 + (r + f_3(\theta)r^3 + \dots)^2 d\theta^2 = dr^2 + (r^2 + 2f_3(\theta)r^4 + \dots)d\theta^2$ and the term homogeneous of degree 2 must be the same as $2f_3(\theta)(udv - vdu)^2$, the coefficient of dv^2 is $2f_3(\theta)u^2$ which is a quadratic polynomial in u, v if and only if $f_3(\theta) = c$, a constant independent of θ

We summarize:

Theorem 6.10. *Let r, θ be a geodesic polar coordinate system centered at $P \in S$. Then $ds^2 = dr^2 + g(r,\theta)^2 d\theta^2$ where $g(r,\theta) = r + cr^3 + O(r^4)$.*

Definition 6.7. The Gaussian curvature of S at P is the number $K(P) = -6c$, with c as in the theorem.

Remark 6.4. This is not the traditional definition of Gaussian curvature, but it is a convenient one for us. Gauss's original definition was extrinsic, and his *Theorema Egregium* was the statement that K is intrinsic. See Subsection 6.4 below for the meaning of intrinsic.

Example 6.8. (1) If $S = \mathbb{R}^2$, then geodesic polar coordinates are the usual polar coordinates, $ds^2 = dr^2 + r^2 d\theta^2$, $g(r,\theta) = r$ and $K(P) = 0$ for all $P \in \mathbb{R}^2$.

(2) If $S = S^2$ and N is the north pole, then we have seen that geodesic polar coordinates are the same as spherical coordinates of Example 6.6, with $\phi = r$ and $ds^2 = dr^2 + \sin^2 r \, d\theta^2$, thus $g(r,\theta) = \sin r = r - r^3/6 + \dots$, thus $K(N) = 1$ (this explains the factor -6). Since there is a rotation of S^2 taking N to any other point P, S^2 , $K(P) = 1$ for all $P \in S^2$.

Remark 6.5. Here is a nice interpretation of the Gaussian curvature $K(P)$. In a geodesic polar coordinate system centered at P, the length of the geodesic circle $C_r = \{(r,\theta) : 0 \le \theta \le 2\pi\}$ is given by

$$(6.19)\ L(C_r) = \int_0^{2\pi} g(r,\theta)\, d\theta = \int_0^{2\pi} (r - (K(P)/6)r^3 + O(r^4))\, d\theta$$
$$= 2\pi r - (K(P)\pi/3)r^3 + O(r^4).$$

Thus $K(P)$ measures the deviation of the formula for circumference from the usual Euclidean formula. For example, for $\mathbb{R}^2$ we get $L(C_r) = 2\pi r$ while for S^2 we get $L(C_r) = 2\pi \sin r = 2\pi r - \pi r^3/3 + \dots$, thus geodesic circles on the sphere are shorter than their counterparts in $\mathbb{R}^2$, as suggested by Figure 6.4.

6.4. **A Quick Glance at Intrinsic Geometry.** Gauss discovered the intrinsic geometry of surfaces, and introduced the geodesic polar coordinates to study it in detail. Intrinsic geometry means the part of the geometry of $S \subset \mathbb{R}^3$ that depends on intrinsic measurements on S, and not on its embedding in $\mathbb{R}^3$. Intrinsic measurements are those that can be reduced to the study of measurements within surface, such as length, angles, area.

We have seen one example in the homework problems. Consider the cylinder $C = \{x^2 + y^2 = 1\} \subset \mathbb{R}^3$, parametrized by $\mathbf{x} : \mathbb{R}^2 \to C \subset \mathbb{R}^3$ where

$$\mathbf{x}(u, v) = (\cos u, \sin u, v). \tag{6.20}$$

Since $\mathbf{x}(u + 2\pi, v) = \mathbf{x}(u, v)$, we can view $\mathbf{x}$ as a map $(\mathbb{R}^2/\sim) \to C$, where $(u, v) \sim (u + 2n\pi, v)$ for all $n \in \mathbb{Z}$. It is easy to check that this map is a homeomorphism. But more is true: we saw in the homework that this map takes geodesics $u'' = 0, v'' = 0$ in $\mathbb{R}^2$ to geodesics in C (spirals and vertical lines) and this map preserves the length of curves. So this map $\mathbf{x}$ is an *isometry* between the intrinsic metrics of the surfaces $\mathbb{R}^2/\sim$ and C.

There is a quick way for checking that a smooth map is an isometry between intrinsic metrics: check that it preserves ds^2, thus it preserves length of curves, thus preserves intrinsic metrics. More formally, to say that a smooth map $f : S_1 \to S_2$ between smooth surfaces S_1, S_2 preserves ds^2 is the same as saying that, for all $P \in S_1$, the differential $d_P f : T_P S_1 \to T_{f(P)} S_2$ is an isometry between the two inner product spaces $T_P S_1, T_{f(P)} S_2 \subset \mathbb{R}^3$. In our example of $\mathbf{x} : (\mathbb{R}^2/\sim) \to C$ this is checked as follows:

$$d\mathbf{x} \cdot d\mathbf{x} = dx^2 + dy^2 + dz^2 = (-\sin u\ du)^2 + (\cos u\ du)^2 + dv^2 = du^2 + dv^2,$$

thus the integrands for arclength correspond, thus lengths of curves correspond, and this map is an isometry in the sense of metric spaces. (This is a *sufficient* condition for isometry of metric spaces. It turns out that it is also a necessary condition, but necessity is harder to prove.)

Another example, let et $\gamma : \mathbb{R} \to \mathbb{R}^2$ be any smooth curve parametrized by arclength, periodic of period 2π, and suppose $\gamma(u_1) \neq \gamma(u_2)$ if $u_1 - u_2$ is not an integral multiple of 2π. Write $\gamma(u) = (x(u), y(u))$. Define a surface $C_\gamma \subset \mathbb{R}^3$ by the formula

$$\mathbf{y}(u, v) = (x(u), y(u), v), \tag{6.21}$$

called the *cylinder on* γ. Then the map $\mathbf{y} : \mathbb{R}^2/\sim \to C_\gamma$ is also an isometry, thus the map $f : C \to C_\gamma$ defined by$f(\mathbf{x}(u, v)) = \mathbf{y}(u, v)$ is an isometry. Since there are infinitely many surfaces isometric to the cylinder C. The isometries, and continuous deformations from one to another.

We have been a bit sloppy since the "surface" $\mathbb{R}^2/\sim$ is a quotient of $\mathbb{R}^2$ rather than a subspace of $\mathbb{R}^3$. But it is a smooth surface in the sense of Definition 6.1. What this means is that we have to enlarge the context in which we consider lengths of curves, we should not restrict ourselves to surfaces in $\mathbb{R}^3$. We will do this next semester.

To finish, let us remark that the Gaussian curvature is invariant under isometries. Since it is 1 for any open subset of S^2 and 0 for any open subset of $\mathbb{R}^2$, we get that *no open subset of S^2 can be isometric to an open subset of $\mathbb{R}^2$.*

References

[1] V. I. Arnold, *Ordinary Differential Equations*, MIT Press, 1973.

[2] E. A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, 1955.

[3] M. A. Do Carmo, *Differential Geometry of Curves and Surfaces*, Prentice-Hall, 1976.

[4] K. F. Gauss, *General Investigations of Curved Surfaces*, Dover, 2005.

[5] A. Hatcher, *Notes on Introductory Point-Set Topology*, available at *http://www.math.cornell.edu/ hatcher/Top/TopNotes.pdf*

[6] D. Hilbert and S. Cohn-Vossen, *Geometry and the Imagination*, Chelsea Pub. Co.

[7] B. Mendelson, *Introduction to Topology*, Dover Publications, 1990.

[8] J. Oprea, *Differential Geometry and its Applicationos*, Prenetice-Hall 1997.

[9] A. Pressley, *Elementary Differential Geometry*, Springer, 2002.

[10] J. Stillwell, *Geometry of Surfaces*, Universitext, Springer-Verlag, 1992.