NOTES FOR MATH 5510, FALL 2017, V 1

DOMINGO TOLEDO

CONTENTS

Date: December 3, 2017.

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*.

Two notable properties of this definition are:

- Its simplicity.
- Its wide applicability, resulting from the large number and great variety of examples.

The simplicity of the definition is clear from its statement. We proceed to explain the second property by giving examples.

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 1.2. 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 1.3. Let $X = \mathbb{R}^2$ with the usual distance function

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

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

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

squaring both sides this is equivalent to

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

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

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

which is equivalent to

$$u \cdot v \le |u||v|$$

which is half of the familiar Cauchy-Schwarz inequality $|u \cdot v| \le |u||v|$. 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 1.4. 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 1.5. 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 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 1.3 and Example 1.4 where equality occurs only on a line segment.

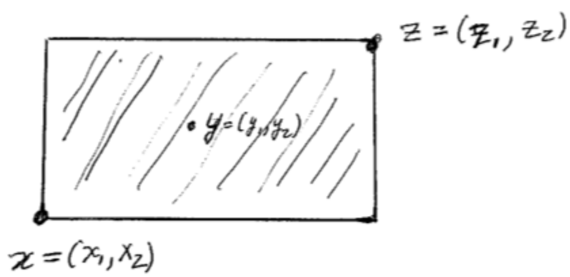

FIGURE 1. Equality Set for Taxicab Metric

Example 1.6. 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 1.5, there are in general many more possibilities than a line segment.

One way to visualize these metrics is by looking at their unit spheres, that is, $\{x \in \mathbb{R}^n : d(0, x) = 1\}$. First, define the terminology:

Definition 1.7. *Let (X, d) be a metric space, $x \in X$ and $r \in \mathbb{R}$, $r > 0$.*

(1) *The* ball of radius r centered at x *is* $B(x, r) = \{y \in X | d(x, y) \leq r\}$.
(2) *The* sphere of radius r centered at x $S(x, r) = \{y \in X | d(x, y) = r\}$.

Figure 2 shows, for $n = 2$ the unit spheres for the the three metrics $d_{(1)}, d_{(2)}$ and $d_{(\infty)}$. The innermost the taxi-cab metric, then the Euclidean, and the outer one is $d_{(\infty)}$.

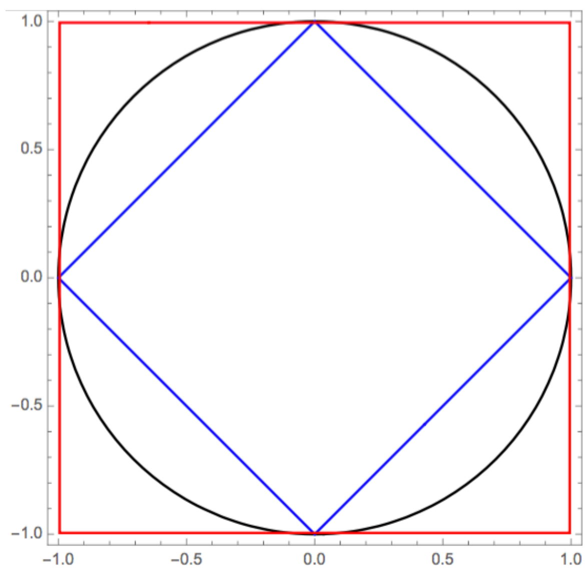

FIGURE 2. Unit Spheres of Examples 1.5, 1.4, 1.6 (ordered from inner to outer).

Example 1.8. 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 ϕ be the angle between the unit vectors x and y. The great circle containing x and y is the intersection with S^2 of the plane through the origin spanned by x and y. (This is well-defined provided $x \neq \pm y$, that is, x and y are nto antipodal points on S^2). The great circle arc connecting x and y is the shorter of the two arcs into which this circle is divided by x and y.

The length of this arc is ϕ, the radian measure of the angle (at the origin) between x and y, see Figure 3. Thus $\cos\phi = x \cdot y$ (the usual dot product in $\mathbb{R}^3$) so the formula for the spherical distance $d(x, y)$ is

$$d(x, y) = \cos^{-1}(x \cdot y). \tag{1}$$

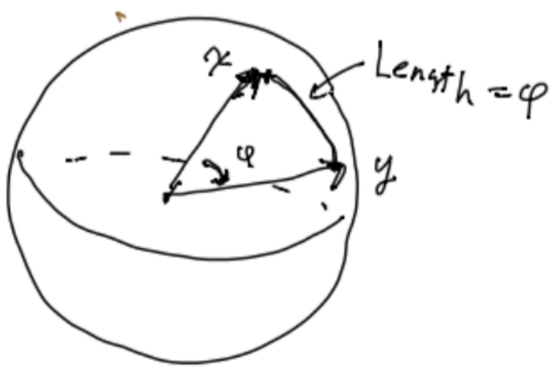

FIGURE 3. Spherical Distance

One way to verify the triangle inequality for the spherical metric is to derive it from 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 $x_i \cdot x_j = x_j \cdot x_i$. The inequality that we want is

Theorem 1.9. *If A is the Gram matrix just defined of m vectors $x_1, \ldots, x_m$ in $\mathbb{R}^n$, 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 1.10. It will be a homework problem for the course to derive the triangle inequality for the spherical distance (1) from the case $m = n = 3$ of Theorem 1.9. Once you do this homework problem it should be clear that we can use the same reasoning for the unit sphere S^n in $\mathbb{R}^{n+1}$, any $n \geq 2$, by defining the distance $d(x, y)$ by the same formula (1) we get the spherical distance in S^n. The triangle inequality can be checked by applying Theorem 1.9 with $m = 3$ in the same way it was applied in the homework to the case $m = n = 3$ to prove the triangle inequality for d_{S^2}.

Remark 1.11. Observe that in the case $m = 2$, that is, two vectors, say $x, y \in \mathbb{R}^m$, then Theorem 1.9 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 1.3 and 1.4 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 1.8.

Example 1.12. 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 inequalities $x \neq y$ 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 1.13. 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 4. This metric is called the *French railway metric* because 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 4. 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 $I_1, \dots I_n$ 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 5. 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 $d(x, z) \leq d(x, y) + d(y, z)$.

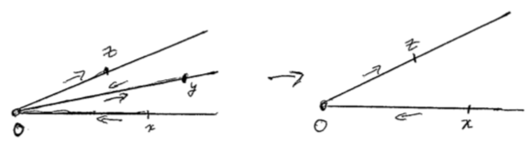

FIGURE 5. One Case of the Triangle Inequality

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.

Example 1.14. Let $S \subset \mathbb{R}^3$ be a *smooth surface*. As a temporary definition of smooth surface, let's say that there is an open subset $U \subset \mathbb{R}^3$ a smooth (meaning infinitely differentiable) function $f : U \to \mathbb{R}$ so that $S = \{x \in U : f(x) = 0\}$ 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 1.8 where $S^2 \subset \mathbb{R}^3$ is given as the zero set of $f(x) = |x|^2 - 1 = x_1^2 + x_2^2 + x_3^3 - 1$. Then $f : \mathbb{R}^3 \to \mathbb{R}$ is smooth and that $\nabla f = (2x_1, 2x_2, 2x_3)$ which vanishes only at the origin. In particular, it does not vanish on S^2. Therefore S^2 is a smooth surface in $\mathbb{R}^3$.

So let $S \subset \mathbb{R}^3$ be a smooth surface. 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* $L(\gamma)$ is defined in the usual way :

$$L(\gamma) = \int_a^b |\gamma'(t)| dt = \int_a^b \sqrt{(x_1')^2 + (x_2')^2 + (x_3')^2} dt \tag{2}$$

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 a distance function $d_S : S \times S \to \mathbb{R}$, called the *intrinsic distance* by

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

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).

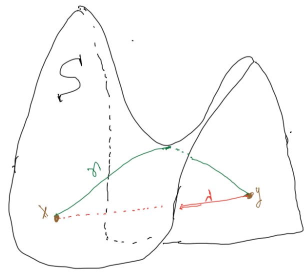

FIGURE 6. Intrinsic Metric on a surface S

To verify that (S, d_S) is a metric space, we should first check that if $d_S(x, y) = 0$ then $x = y$. This follows from the fact that, if γ is a path from x to y, then $L(\gamma) \geq L(\lambda) = |x - y|$, where λ is the straight-line segment in $\mathbb{R}^3$ from x to y, of length $|x-y| = d_{\mathbb{R}^3}(x, y)$ is the usual distance in $\mathbb{R}^3$. This implies that $d_S(x, y) \geq |x-y|$, so if $d_S(x, y) = 0$ then $|x-y| = 0$, so $x = y$. See Figure 6

Now to the triangle inequality. This follows the same pattern as the proof in Example 1.13 except that, since we have an infimum rather than a minimum, we have to use some ϵ's. Let $x, y, z \in S$ 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 $L(\gamma_1) < d_S(x, y) + \frac{\epsilon}{2}$, and there exists a path γ_2 from y to z with $L(\gamma_2) < d_S(y, z) + \frac{\epsilon}{2}$. Let $\gamma = \gamma_1\gamma_2$ be the piecewise differentiable path γ_1 followed by γ_2 from x to z, see Figure 7. Then $d_S(x, z) \leq L(\gamma) = L(\gamma_1) + L(\gamma_2) < d_S(x, y) + d_S(y, z) + \epsilon$. Thus for all $\epsilon > 0$ we have $d_S(x, z) < d_S(x, y) + d_S(y, z) + \epsilon$, thus $d_S(x, z) \leq d_S(x, y) + d_S(y, z)$. This completes the verification that d_S is a metric.

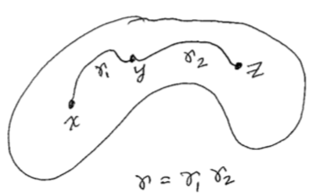

FIGURE 7. Putting Paths Together

Remark 1.15. The metric d_S just defined is called the *intrinsic metric* or *intrinsic distance* because it only allows measurements within the surface S, and does not allow measurement of paths in the surrounding $\mathbb{R}^3$ that are not already in S. In the proof that d_S is a metric we actually proved that for all $x, y \in S$, $d_S(x,y) \geq d_{\mathbb{R}^3}(x,y)$. Moreover, closer examination (and with some reasonable assumptions and some facts we will prove later) shows that for given $x, y \in S$, equality holds in this inequality if and only if the straight line segment in $\mathbb{R}^3$ joining x and y already lies in S.

To make the meaning of intrinsic distance more concrete, imagine that S is a piece of the surface of the earth containing a mountain pass and two towns x and y on opposite sides of the mountain pass, see Figure 6. If you walk along any path γ on S from x to y you have to travel a longer distance than the length of the straight line segment λ in $\mathbb{R}^3$ from x to y. The latter segment can only by realized by digging a tunnel connecting the two sides of the pass.

Remark 1.16. We are all familiar with the statement that for any $x, y \in S^2$ the great circle arc from x to y gives the shortest curve from x to y (this curve is unique if x and y are not antipodal). This statement implies that given $x, y \in S^2$, the infimum of the length of curves in S^2 from x to y is realized by the length of this great-circle arc (the infimum is actually a minimum). It therefore implies the spherical distance of Example 1.8 is the same as the intrinsic distance in S^2. In particular, the more conceptual proof of the triangle inequality for d_S in Example 1.14 applies to d_{S^2}.

The statement that the great-circle arc is the shortest curve connecting its endpoints requires proof. Later we will give general theorems (existence of geodesic and their length-minimmizing properties) that immediately imply the statement about great circles. But it is also worthwhile at this point to prove it directly.

Theorem 1.17. *Let $x, y \in S^2$, let γ_0 be the great-circle arc from x to y, and let γ be any (piecewise differentiable) path from x to y. Then $L(\gamma_0) \leq L(\gamma)$.*

Proof. Given any two points in S^2 we can apply a rigid motion of $\mathbb{R}^3$ fixing the origin that takes x to the north-pole $N = (0,0,1)$, y to a point on $P = (\sin\phi_0, 0, \cos\phi_0)$ on the x_1x_3-plane, and takes S^2 to itself (since it fixes the origin). It is therefore enough to show that the length of any curve from N to P has length at least ϕ_0.

Use spherical coordinates (θ, ϕ) related to the cartesian (x_1, x_2, x_3) by

$$(x_1, x_2, x_3) = (\cos\theta\sin\phi, \sin\theta\sin\phi, \cos\phi),$$

See Figure 8. Our curve $\gamma(t)$ from N to P is determined by two functions $\theta(t)$ and $\phi(t)$, t some interval $[a,b]$ with $\phi(a) = 0$ (the north pole, where θ is indeterminate), $\theta(b) = 0$ and $\phi(b) = \phi_0$ (to end at $P = (\sin\phi_0, 0, \cos\phi_0)$, spherical coordinates $(0, \phi_0)$. Then compute:

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

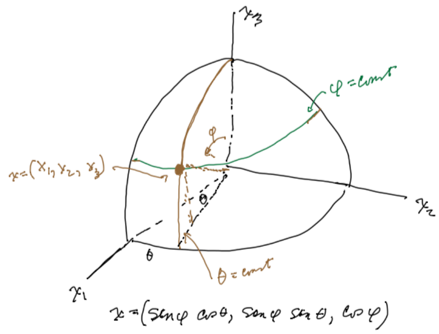

FIGURE 8. Spherical Coordinates

Then $||\gamma'||^2$ is the sum of squares of the three components in the right-hand side. We see right away that the cross terms in the sum of the first two squares cancel out, the sum has to terms with a factor of $(\theta')^2$ that add to $\sin^2\phi\,(\theta')^2$ and three terms with a factor of $(\phi')^2$ that add to $(\phi')^2$. We thus get

$$||\gamma'||^2 = \sin^2\phi\,(\theta')^2 + (\phi')^2 \geq (\phi')^2$$

Therefore

$$L(\gamma) = \int_a^b ||\gamma'(t)||dt \geq \int_a^b |\phi'(t)|dt \geq \int_a^b \phi'(t)dt = \phi(t)|_a^b = \phi_0 = L(\gamma_0)$$

□

Example 1.18. It is hard to resist the temptation of discussing briefly another example, which we will not have time to develop in detail, but which is the most important example for the study of topology and geometry of surfaces. To follow the same pattern as the last two examples, we will consider a surface in a three-dimensional vector space, except this time it will be in *Minkowski space* rather than Euclidean space.

Definition 1.19. Minkowski Space *is* $\mathbb{R}^3$ *with the* Minkowski inner product*, defined as follows, If* $x, y \in \mathbb{R}^3$*, then*

$$x \diamond y = x_1y_1 + x_2y_2 - x_3y_3.$$

The notation $x \diamond y$ is not standard. Note that $x \diamond y$ is a variation of the usual dot product $x \cdot y$ in $\mathbb{R}^3$. It is like $x \cdot y$ in that it is linear in each variable, but unlike $x \cdot y$ in that it is not positive definite: $x \diamond x$ can be positive, negative, or zero; $x \diamond x = 0$ does not imply $x = 0$. The level sets of $x \diamond x$ are visualized as follows: $x \diamond x = 0$ is the cone $x_1^2 + x_2^2 = x_3^2$ (the light cone) which separates the positive and negative vectors. For each non-zero constant c, the set $x \diamond x = c$ is a hyperboloid. Figure 9 shows the levels $c = 1$ (hyperbolid of one sheet), $c = 0$ (cone) and $c = -1$ (hyperboloid of two sheets.)

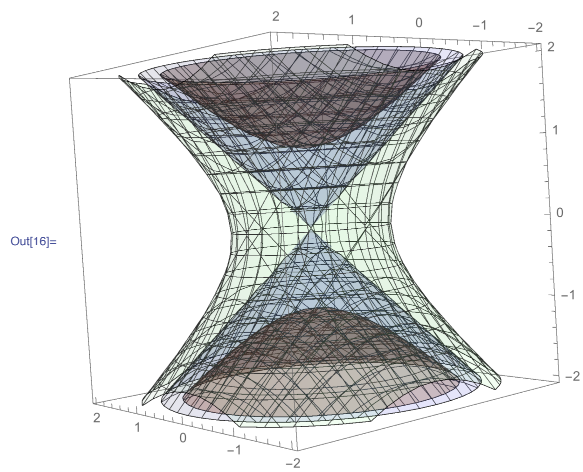

FIGURE 9. Level Sets of Minkowski Squared Norm

Since $x \diamond x$ is not positive definite, it seems like it would be impossible to define an intrinsic metric on a surface S in Minkowski space by using the same procedure as in Example 1.14. The problem is the definition of length of a curve: if $\gamma : [a, b] \to S$ is a piecewise smooth curve, then the integrand of the formula for the Minkowski length of γ (the Minkowski analogue of (33)) would not make sense because $\sqrt{\gamma' \diamond \gamma'}$ would not be real if $\gamma' \diamond \gamma' < 0$.

It is a remarkable fact that there are surfaces S in Minkowski space with the property that $\gamma' \diamond \gamma' > 0$ for all $\gamma : [a, b] \to S$. For such a surface one can repeat almost verbatim the arguments of Example 1.14 and get a metric space.

The most important example with this property is the upper sheet H of the hyperboloid of two sheets $x \diamond x = -1$, see Figure 10.

Here is a quick verification. Suppose $\gamma : [a, b] \to H$ is piecewise differentiable. Then $\gamma(t)$ satisfies $\gamma(t) \diamond \gamma(t) == 1$ for all t. Differentiating this equation we get $\gamma(t) \diamond \gamma'(t) = 0$. The following lemma implies that $\gamma'(t) \diamond \gamma'(t) \geq 0$ for all t:

Lemma 1.20. *Let x and v satisfy $x \diamond x = -1$ and $v \diamond v = 0$. Then $v \diamond v \geq 0$ and $v \diamond v = 0$ if and only if $v = 0$.*

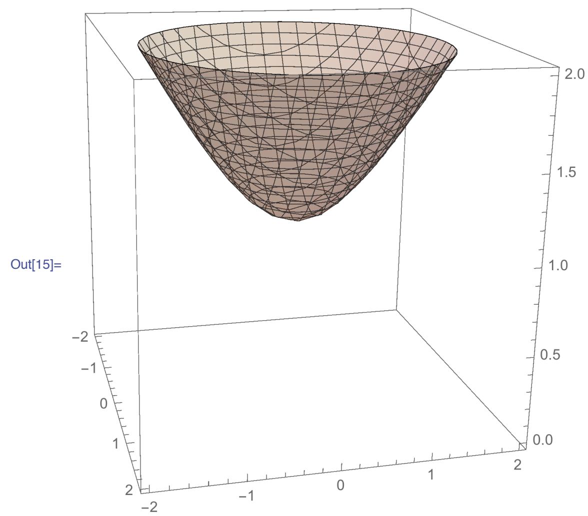

FIGURE 10. Upper Sheet of Hyperboloid $x \diamond x = -1$

Remark 1.21. Geometrically $x \diamond v = 0$ means that v is a vector tangent to H at x. The lemma thus says that all tangent vectors to H have positive Minkowski square-norm.

Proof. This can be done by a simple computation. Suppose $x = (x_1, x_2, x_3)$ and $v = (v_1, v_2, v_3)$. Then the assumptions are that $x_1^2 + x_2^2 - x_3^2 = -1$ and $x_1v_1 + x_2v_2 - x_3v_3 = 0$. We can solve

$$v_3 = \frac{x_1v_1 + x_2v_2}{x_3}$$

because $x_3 > 0$. Then using this expression for v_3 in $v_1^2 + v_2^2 - v_3^2$ and a small computation gives the formula

$$v \diamond v = \frac{v_1^2 + v_2^2 + (x_1v_2 - x_2v_1)^2}{x_3^2}$$

which is positive, and $= 0$ if and only if $v_1 = v_2 = 0$, which in turn implies $v_3 = 0$. $\square$

Definition 1.22. *The* hyperbolic plane *is the surface* $H = \{x : x \diamond x = -1$ *and* $x_3 > 0\}$ *with distance function*

$$d_H(x, y) = \inf_{\{\gamma:[a,b]\to H, \gamma(a)=x, \gamma(b)=y\}} \{\int_a^b \sqrt{\gamma'(t) \diamond \gamma'(t)} dt\}. \tag{3}$$

Remark 1.23. The equation $x_1^2+x_2^2-x_3^2 = -1$ of H has a strange similarity with the equation $x_1^2 + x_2^2 + x_3^2 = 1$ of S^2. There is a lot more to this similarity. For example, it can be proved that an equivalent formula for the distance in H is

$$d_H(x,y) = \cosh^{-1}(x \diamond y) \tag{4}$$

in close analogy with the formula (1) for the spherical distance d_{S^2}.

Example 1.24. 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 1.25. 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 1.26. *We define the terminology and notation that we will use in referring to some of the metric spaces just introduced.*

(1) *The metric of Example 1.4 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 1.5 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 1.6 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 1.8 will be called the* spherical metric *and denoted d_{S^2}.*
(5) *The metric of Example 1.14 will be called the* intrinsic metric *on $S \subset \mathbb{R}^3$, and denoted by d_S.*
(6) *The metric of Example 1.18 will be called the* hyperbolic metric *and denoted d_H.*
(7) *The metrics of Examples 1.24 and 1.25 will be called p-adic metrics. For a given prime p, the p-adic metric will be denoted d_p.*

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 1.8. 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 $d'(x,y) \leq d_{S^2}(x,y)$ 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 1.14, 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 1.14. 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 1.14). 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 1.26)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 1.27. *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 1.28. *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 1.27 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. $\square$

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 1.29. *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 1.30. If we use the p-adic metric of Example 1.24 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 1.28, 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 1.31. *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).

Another fact about metric spaces is that every metric space (X, d) has a *completion*. This is a complete metric space $(\bar{X}, \bar{d})$ so that (X, d) is a *dense* subspace of $(\bar{X}, \bar{d})$. A subspace X of a metric space Y is called *dense* if every $y \in Y$ is the limit of some sequence $\{x_n\}$ in X.

The standard example of a completion is $\mathbb{R}$ as a completion of $\mathbb{Q}$ (both with their usual metrics). The construction of $\mathbb{R}$ from $\mathbb{Q}$ by using Cauchy sequences can be used to construct a completion of any metric space. In other words, given a metric space (X, d), let $Ca(X, d)$ be the collection of Cauchy sequences in X. If $\{x_n\}, \{y_n\} \in Ca(X, d)$, define their distance $d(\{x_n\}, \{y_n\}) = \lim d(x_n, y_n)$. This is not a metric on $Ca(X, d)$ because the distance between two sequences being zero does not imply that the sequences are equal.

To get a metric space, define an equivalence relation on $Ca(X, d)$ by $\{x_n\} \sim \{y_n\}$ if and only if $\lim d(x_n, y_n) = 0$.

Definition 1.32. *The* completion *of* (X, d) *is the set* $\bar{X}$ *of equivalence classes* $[\{x_n\}]$ *of elements of* $Ca(X, d)$. *If* $[\{x_n\}], [\{y_n\}] \in \bar{X}$*, the distance* $\bar{d}([\{x_n\}], [\{y_n\}]$ *is defined to be* $\lim d(x_n, y_n)$.

It has to be checked that $\bar{d}$ is well defined, independent of representatives: if $\{x'_n\} \sim \{x_n\}$ and $\{y'_n\} \sim \{y_n\}$, then $\lim d(x'_n, y'_n) = \lim d(x_n, y_n)$. Then it is clear that $\bar{d}$ is a metric on $\bar{X}$. The set of constant sequences $\{x_n\}$ where $x_n = x$ for all n is a subspace of $(\bar{X}, \bar{d})$ naturally identified with (X, d). It also has to be checked that (X, d) is dense in $(\bar{X}, \bar{d})$.

Observe that it (X, d) is a complete metric space, then the elements of $\bar{X}$ are in one to one correspondence with their limits $x \in X$, so $(\bar{X}, \bar{d}) = (X, d)$. Thus, if (X, d) is complete, then it is its own completion.

Example 1.33. As already mentioned, if $d_{\mathbb{Q}}$ is the usual distance in $\mathbb{Q}$, then $(\bar{\mathbb{Q}}, \bar{d}) = (\mathbb{R}, d_{\mathbb{R}})$, where $d_{\mathbb{R}}$ is the usual distance on $\mathbb{R}$

Example 1.34. Fix a prime number p. Then the completion $(\bar{\mathbb{Z}}, \bar{d_p})$ of $(\mathbb{Z}, d_p)$ is called the ring of *p-adic integers* and the completion $(\bar{\mathbb{Q}}, \bar{d_p})$ of $(\mathbb{Q}, d_p)$ is called the field of *p-adic numbers*.

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

Definition 1.35.
(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 1.36. 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.

Note that the conditions in Definition 1.35 are listed in increasing order of stringency, in the sense that $(6) \implies (5) \implies \ldots \implies (2) \implies (1)$. Moreover, none of these implications can be reversed. Most of these implications are immediate, for example, for $(6) \implies (5)$ just choose $C = C' = 1$. The only implication that may not be immediately familiar is $(4) \implies (3)$. This is the content of the following theorem:

Theorem 1.37. *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 1.38. *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 1.39. *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 loose to strict. More precisely, they are related as follows:

Theorem 1.40. *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 1.41. Recall the distances $d_{(1)}, d_{(2)}, d_{(\infty)}$ on $\mathbb{R}^n$ of Definition 1.26. They are related by the following inequalities (the first in a homework problem, the remaining two are similar but easier).

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

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 $(1, 1, \dots, 1)$ is different in all three metrics on $\mathbb{R}^n$. These innequalities can be visualized by looking at how various spheres are related. See Figure 11, where the figures are presented in the same order as the inequalities above.

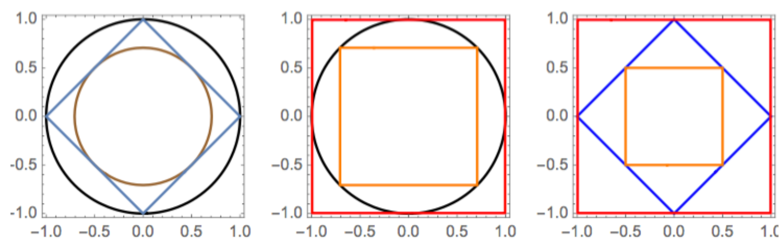

FIGURE 11. Bi-Lipschitz equivalence of the $1, 2, \infty$ metrics

Example 1.42. 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 1.43. 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 1.35 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 1.4 and 1.5 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 1.44. Here's a closely related question. Are $(\mathbb{R}^3, d_{(1)})$ and $(\mathbb{R}^3, d_{(\infty)})$ isometric? Note that the trick of Example 1.42 doesn't work. Make a conjecture about the answer to this question, and then prove it. A look at Figure 12 may help.

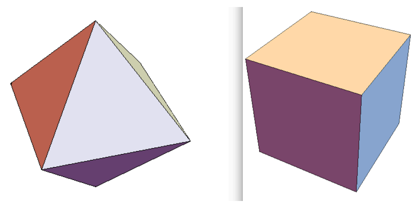

FIGURE 12. Unit Spheres for $d_{(1)}$ and $d_{(\infty)}$ in $\mathbb{R}^3$.

Remark 1.45. Neither of the implications in Theorem 1.40 can be reversed. This can be seen by examples:

(1) Let $f : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ be defined by $f(x) = \arctan(x)$. Then f is a homeomorphism, but cannot be bi - Lipschitz, say, because $(-\frac{\pi}{2}, \frac{\pi}{2})$ is bounded but $\mathbb{R}$ is not. Could also say because $\mathbb{R}$ is complete but $(-\frac{\pi}{2}, \frac{\pi}{2})$ is not. Observe that f is Lipschitz, since $f'(x) = \frac{1}{1+x^2}$ is bounded.
(2) The identitly map $(\mathbb{R}^n, d_{(2)}) \to (\mathbb{R}^n, d_{(1)}$ is bi-Lipschitz, but we have seen in Example 1.43 that these are not isometric (at least for $n = 2$. An analogous argument can be given for any $n > 2$).

Example 1.46. Let $\mathbb{R}^\infty$ denote the space of sequences of real numbers that are eventually zero:

(5) $\mathbb{R}^\infty = \{x = (x_1, x_2, \dots) : x_i \in \mathbb{R}$ and $\exists N$ so that $x_i = 0$ for $i > N\}$.

Note that N depends on the sequence x.

For each n there is a natural inclusion $\mathbb{R}^n \subset \mathbb{R}^\infty$ as

$$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n, 0, 0, \dots)\} \subset \mathbb{R}^\infty$$

In fact, with this convention for the meaning of the inclusion $\mathbb{R}^n \subset \mathbb{R}^\infty$, we have

$$\mathbb{R} \subset \mathbb{R}^2 \subset \cdots \subset \mathbb{R}^n \subset \cdots \subset \mathbb{R}^\infty \text{ and } \mathbb{R}^\infty = \cup_n \mathbb{R}^n.$$

Given $x, y \in \mathbb{R}^\infty$, we can define $d_{(1)}(x, y), d_{(2)}(x, y)$ and $d_{(\infty)}(x, y)$ by the same formulas as in Definition 1.26. More precisely, we can take infinite sums by taking the limit as $n \to \infty$ in those formulas, and observe that they make sense because the entries of x, y are eventually zero, so for each x, ywe get finite sums.

Consider the six inequalities of Example 1.41, illustrated for $n = 2$ in Figure 11. Three of these inequalities, the ones on the left-hand side, are independent of n, so they also hold in $\mathbb{R}^\infty$. These are

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

These inequalities hold for all $x, y \in \mathbb{R}^\infty$. They can be summarized in

$$d_{(\infty)}(x, y) \leq d_{(2)}(x, y) \leq d_{(1)}(x, y) \text{ for all } x, y \in \mathbb{R}^\infty.$$

These are the inequalities that for $n = 2$ are illustrated in Figure 2. We have jsut proved that the same relation holds for all n, therefore for $\mathbb{R}^\infty$; the unit ball of $d_{(1)}$ is contained in the unit ball of $d_{(2)}$ which is in turn contained in the unit ball of $d_{(\infty)}$. Note the order in which the balls appear when talking about containment seems to be opposite to the order in which the inequalities appear. Some thought quickly shows that this is the case, but it can be confusing.

The remaining three inequalities depend on n. In fact for each of these inequalities equality hods at $x_0 = (0, 0, \ldots, 0)$ and $y_0 = (1, 1, \ldots, 1)$ in $\mathbb{R}^n$:

(1) $d_{(1)}((0, \ldots, 0), (1, \ldots, 1)) = 1 + \cdots + 1 = n$.
(2) $d_{(2)}((0, \ldots, 0), (1, \ldots, 1)) = \sqrt{1 + \cdots + 1} = \sqrt{n}$
(3) $d_{(\infty)}((0, \ldots, 0), (1, \ldots, 1)) = \max(1, \ldots, 1) = 1$

So the inequalities in the right hand side of Example 1.41 become equalities:

(1) $d_{(1)}(x_0, y_0) = n = \sqrt{n}\, d_{(2)}(x_0, y_0) = \sqrt{n}\sqrt{n}$
(2) $d_{(2)}(x_0, y_0) = \sqrt{n} = \sqrt{n}\, d_{(\infty)}(x_0, y_0) = \sqrt{n} \cdot 1$
(3) $d_{(1)}(x_0, y_0) = n = n\, d_{(\infty)}(x_0, y_0) = n \cdot 1$

so the constants $\sqrt{n}, \sqrt{n}, n$ in these three are best possible, that is, they are *the* Lipschitz constants, and they go to infinity as $n \to \infty$. In summary, we have proved:

Theorem 1.47. *The identity map of* $\mathbb{R}^\infty$ *is Lipschitz, with Lipschitz constant one, as a map* $(\mathbb{R}^\infty, d_{(1)}) \to (\mathbb{R}^\infty, d_{(2)})$*, as a map* $(\mathbb{R}^\infty, d_{(2)}), \to (\mathbb{R}^\infty, d_{(\infty)})$ *and consequently as a map* $(\mathbb{R}^\infty, d_{(1)}) \to (\mathbb{R}^\infty, d_{(\infty)})$. *None of these three maps is bi-Lipschitz.*

It is easy to see that none of these spaces is *complete*. It is also fairly easy to see what their completions should be:

(1) $\ell_1 = \{(x_1, x_2, \ldots) : x_i \in \mathbb{R} \text{ and } \sum_1^\infty |x_i| < \infty\}$

(2) $\ell_2 = \{(x_1, x_2, \dots) : x_i \in \mathbb{R} \text{ and } \sum_1^\infty |x_i|^2 < \infty\}$
(3) $\ell_\infty = \{(x_1, x_2, \dots) : x_i \in \mathbb{R} \text{ and } \sup\{|x_i|\} < \infty\}$

with metrics defined by $d_{(1)}(x, y) = \sum_1^\infty |x_i - y_i|$, $d_{(2)}(x, y) = \sqrt{\sum_1^\infty |x_i - y_i|^2}$ and $d_{(\infty)}(x, y) = \sup\{|x_i - y_i|\}$ respectively. This is the beginning of another interesting subject that we will not be able to pursue.

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 2.2. *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. But there are some important examples where this group is large. In fact, three if the metric spaces defined in §1, namely Euclidean (Example 1.3), spherical (Example 1.8) and hyperbolic (Definition 1.22 have the largest possible groups of isometries.

2.1. **Isometries of Euclidean Space.** We study first the group of isometries of $\mathbb{R}^n$ 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 $(\mathbb{R}^n, d)$ and to describe the structure of this group of isometies. In the case $n = 2$ we will also give a complete classification of the individual isometries.

2.1.1. **Affine transformations.** We first recall some facts from linear algebra. A transformation $L : \mathbb{R}^n \to \mathbb{R}^m$ is called a *linear transformation* iff for all $r \in \mathbb{R}$ and for all $x, y \in \mathbb{R}^n$ 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 its values on the standard basis vectors e_i, $i = 1, \dots, n$. For this purpose it is best to write the elements of $\mathbb{R}^n$ as column vectors, rather than row vectors. The transformation L is encoded in an m by n matrix A:

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \cdot & & & \\ \cdot & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

with columns the vectors $L(e_i)$. Thus if $y = L(x)$, then

$$\begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \cdot & & & \\ \cdot & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{pmatrix}.$$

in other words, $L(x) = Ax$, where Ax is matrix multiplication. This gives a one-to-one correspondence between linear transformations and matrices. The reason for using column vecors is that composition then corresponds to matrix multiplication: If $L_1(x) = A_1x$ and $L_2(x) = A_2x$, then $L_1 \circ L_2(x) = L_1(L_2(x)) = A_1A_2x$. So we should always write points in $\mathbb{R}^n$ as column vectors rather than row vectors when using matrices to define linear transformations.

.

Definition 2.3. *A map $f : \mathbb{R}^n \to \mathbb{R}^n$ is called an* affine transformation *iff there exist an $n \times n$ matrix A and a vector $b \in \mathbb{R}^n$ such that $f(x) = Ax + b$.*

We by $f_{A,b}$ the affine transformation determined by A and b: $f_{A,b}(x) = Ax + b$

Remark 2.4.
(1) The transformation $f_{A,b}$ uniquely determines A and b: If $A_x + b_1 = A_2 x + b_2$ for all x, then $(A_1 - A_2)x = b_2 - b_1$ for all x, so $A_1 - A_2$ is a constant linear transformation, thus the zero transformation: $A_1 - A_2 = 0$, therefore $b_2 - b_1 = 0$, so $A_1 = A_2$ and $b_1 = b + 2$.
(2) Two special types of affine transformation:
  (a) *Linear*: If $b = 0$: $f_{A,0}(x) = Ax$.
  (b) *Translation*: If $A = I$: $f_{I,b}(x) = x + b$. It is convenient to simplify the notation for translation: $t_b(x) = x + b$.
(3) Thus an affine transformation is the composition of a linear transformation and a translation: $f_{A,b} = f_{I,b} \circ f_{A,0} = t_b A$. Therefore it sends lines to lines, planes to planes, etc. but it need not preserve the origin.
(4) Note the order of composition matters: $f_{A,b} = t_b A = A t_{A^{-1}b}$

2.1.2. **Orthogonal matrices.** It is natural to ask when an affine transformation $= f_{A,b} f : \mathbb{R}^n \to \mathbb{R}^n$ as in Definition 2.3 is an isometry. Using $d(x, y) = |x - y|$, were $|x| = d(0, x)$ denotes the Euclidean norm of x, f is an isometry if and only if $d(f(x), f(y)) = |f(x) - f(y)| = |Ax - Ay| = |A(x - y)| = |x - y|$ for all x, y in $\mathbb{R}^n$. In other words, $f_{A,b}$ is an isometry if and only if $x \to Ax$ is an isometry. This is the same as saying $|Ax| = |x|$ for all x, or $|Ax|^2 = |x|^2$.

Now observe that $x \cdot y = ({}^t x)y$, matrix multiplication of the transpose ${}^t x$ (a row vector) and y (a column vector). So the equation $|Ax|^2 = |x|^2$ becomes ${}^t(Ax)(Ax) =^t xx$. Using ${}^t(Ax) = {}^t x \, {}^t A$ we get that $f_{A,b}$ is an isometry of and only if

$$ {}^t x \, {}^t A \, Ax = {}^t x \, x, \text{ for all } x \in \mathbb{R}^n. \tag{6} $$

We have therefore proved the first statement of the following lemma:

Lemma 2.5. *Let A be an n by n matrix. Then the linear transformation $x \to Ax$ of $\mathbb{R}^n$ is an isometry if and only if A satisfies (6) . This happens if and only if ${}^t A \, A = I$*

Proof. The first statement has already been proved. For the second statement, if $B = {}^t A \, A$, then ${}^t B = {}^t({}^t A \, A) = {}^t A \, {}^t({}^t A)) = {}^t A \, A$ is symmetric, that is, if $B = (b_{ij})$ then $b_{ij} = b_{ji}$. Then (6) reads

$$\sum_{i,j} b_{ij} x_i x_j = \sum_i b_{ii} x_i^2 + 2 \sum_{i<j} b_{ij} x_i x_j = \sum_i x_i^2$$

Comparing the coefficients of these two quadratic functions we see $b_{ii} = 1$ and $b_{ij} = 0$ for $i \neq j$, that is, $B = I$. (6) $\square$

Definition 2.6. *An n by n matrix A is called an* orthogonal matrix *if and only if* ${}^tAA = I$.

Remark 2.7.
1. Note that ${}^tA\,A = I$ is equivalent to $A^{-1} = {}^tA$, in other words, A is invertible and its inverse is simply its transpose.
2. Since $AA^{-1} = I$, we get that if $A\,{}^tA = I$.
3. If the equation ${}^tAA = I$ is written explicitly: $\sum_k a_{ki}a_{kjj} = \delta_{ij}$, it says that the columns of A are orthogonal and of unit length. Since $A\,{}^tA = I$, we get that the rows are also orthogonal and of unit length.

The combination of (6), Lemma 2.5 and Definition 2.6 proves the following theorem:

Theorem 2.8. *An affine transformation $f_{A,b}$ of $\mathbb{R}^n$ is an isometry if and only if A is an orthogonal matrix.*

2.1.3. **Some isometries of the Euclidean plane.** Let us find all the orthogonal 2 by 2-matrices. If

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

then

$${}^tAA = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2+c^2 & ab+cd \\ ab+cd & b^2+d^2 \end{pmatrix}$$

which is the unit matrix if and only if $a^2 + c^2 = 1, b^2 + d^2 = 1$ (thus each column is a vector on the unit circle) and $ab + cd = 0$ (so these vectors are orthogonal). We can choose a number θ so that $a = \cos\theta$ and $c = \sin\theta$. Once we make this choice of θ, there are only two vectors on the unit circle perpendicular to our choice: $b = -\sin\theta, d = \cos\theta$ or $b = \sin\theta, d = -\cos\theta$. So we get two possibilities for A, which we will denote R_θ, S_θ respectively. The matrices, together with their geometric interpretation, are:

1. *A rotation about the origin counterclockwise by an angle θ, denoted by R_θ.*

$$A = R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{7}$$

2. *A reflection about the line $\{t(\cos\frac{\theta}{2}, \sin\frac{\theta}{2})\}$, denoted by S_θ*

$$A = S_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \tag{8}$$

They are distinguished by their determinants:

$$\det(R_\theta) = 1, \ \ \det(S_\theta) = -1, \tag{9}$$

in other words, R_θ is *orientation preserving* while S_θ is *orientation reversing*

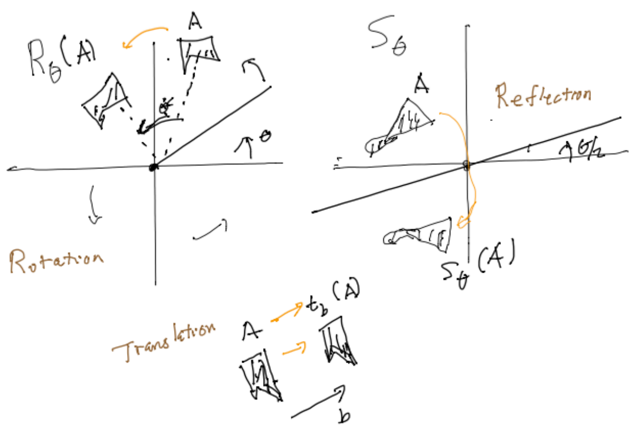

FIGURE 13. Rotations and Reflections

These are the linear transformations that are isometries of the Euclidean plane, in other words, the affine isometries that preserve the origin. The simplest isometry that moves the origin is a t*translation* by a vector $b \in \mathbb{R}^2$:

$$t_b(x) = x + b. \tag{10}$$

See Figure 13 for a description of these transformations. A way to visualize these transformaitons is to choose an (asymmerical) object A and show how it is moved buy each isometry.

Theorem 2.8 implies the following more precise theorem about affine isometries of $\mathbb{R}^2$:

Theorem 2.9. *Let f be an affine isometry of $\mathbb{R}^2$. Then there exists $b \in \mathbb{R}^2$ and $\theta \in \mathbb{R}$ so that either*

(1) $f(x) = R_\theta x + b$
or
(2) $f(x) = S_\theta x + b$.

In particular, all affine isometries of $\mathbb{R}^2$ are obtained by composing the three types of Figure 13

Remark 2.10. 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 (7) becomes

$$R_\theta(z) = e^{i\theta} z \tag{11}$$

while (8) becomes

$$S_\theta(z) = e^{i\theta} \bar{z} \tag{12}$$

2.1.4. **The main theorem.** The reason we have looked at affine isometries of $\mathbb{R}^n$ in so much detail is the remarkable fact that these are *all* the isometries:

Theorem 2.11. *Let $f : (\mathbb{R}^n, d_{(2)}) \to (\mathbb{R}^n, d_{(2)})$ be an isometry. Then f is affine: there exists a vector $b \in \mathbb{R}^n$ and an orthogonal $n \times n$ matrix A so that $f = f_{A,b}$, that is, $f(x) = Ax + b$ for all $x \in \mathbb{R}^n$.*

Proof. First observe that we can reduce to the case $f(0) = 0$. Namely, given any isometry $f : \mathbb{R}^n \to \mathbb{R}^n$ be an isometry, define a new isometry g 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 there exists an orthogonal matrix A so that $g(x) = Ax$, then $f(x) = Ax + b$, where $b = f(0)$.

Suppose g is an isometry with $g(0) = 0$. If we can prove that g is a linear transformation, then $g(x) = Ax$ for some matrix A, and A is necessarily orthogonal by Lemma 2.5. It thus sufficed to prove $g(x+y) = g(x) + g(y)$ and $g(rx) = rg(x)$ holds for all $x, y \in \mathbb{R}^n$ and for all $r \in \mathbb{R}$. This will follow from the nature of the equality sets for the triangle inequality.

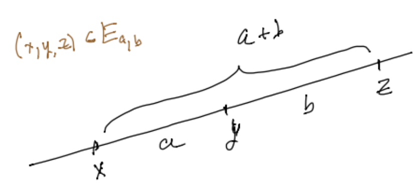

FIGURE 14. Typical Element of $E_{a,b}$

More precisely, given two positive real numbers a, b, define the following set $E_{a,b}$ of triples of points in $\mathbb{R}^n$, that is, $E_{a,b} \subset (\mathbb{R}^n)^3$ (see Figure 14):

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

In other words, these are the triples (x,y,z) for which equality holds in the triangle inequality, and where the distances are given fixed values $a,b,a+b$. Note that, since $a,b>0$, the three elements of any triple are distinct, so they lie in a unique line L, and this line is determined by any two of them. Moreover, *any two of x,y,z determine the third* Suppose $(x,y,z)\in E_{a,b}$ and we know:

Lemma 2.12. *Fix positive real numbers a,b and let $E_{a,b}$ be as just defined.*

(1) *Suppose $x,z\in\mathbb{R}^n$ and $d(x,z)=a+b$. Then there is a unique $y\in\mathbb{R}^n$ iso that $(x,y,z)\in E_{a,b}$.*
(2) *Suppose $x,y\in\mathbb{R}^n$ and $d(x,y)=a$. Then there is a unique $z\in\mathbb{R}^n$ so that $(x,y,z)\in E_{a,b}$.*
(3) *Suppose $y,z\in\mathbb{R}^n$ and $d(y,z)=b$. Then there is a unique $x\in\mathbb{R}^n$ so that $(x,y,z)\in E_{a,b}$.*

Proof. We use the following terminology: if $L\subset\mathbb{R}^n$ is a line and $p,q,r\in L$ are distinct, then we say that

r is between p abd q if r is in the line segment joining p and q.

r lies on the side of q opposite to p if q lies in the line segment joining p and r.

To prove the Lemma, given the twp points in $\mathbb{R}^n$, let L be the line through them. Then the third pont must be:

(1) y is the point on L between x and z at distance a from x.
(2) z is the point on L distance b from y on the side of y opposite to x,
(3) x is the point on L distance a from y on the side of y opposite to z.

□

Lemma 2.13. *Suppose $g:\mathbb{R}^n\to\mathbb{R}^n$ is an isometry and $(x,y,z)\in E_{a,b}$. Then $(g(x),g(y),g(z))\in E_{a,b}$*

Proof. Clear since g preserves all the distances. □

Now go back to the proof that an isometry $g:\mathbb{R}^n\to\mathbb{R}^n$ with $g(0)=0$ must be linear:

(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 us assume $r\neq 1$ and consider three cases

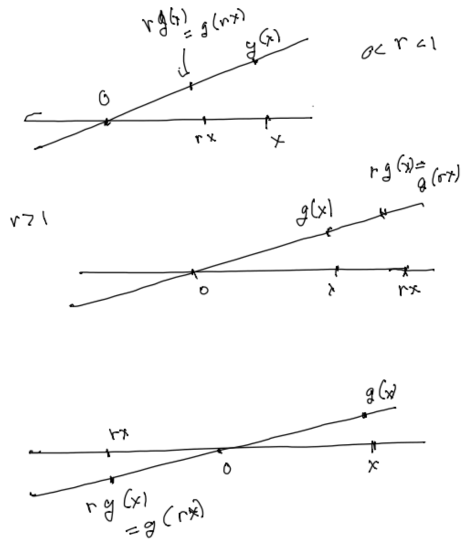

FIGURE 15. Checking Linearity

(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 $g(0) = 0$, by Lemma 2.13 $(0, g(rx), g(x)) \in E_{a,b}$. But also $(0, rg(x), g(x)) \in E_{a,b})$. By the uniqueness part of (1) of Lemma 2.12, get $g(rx) = rg(x)$,

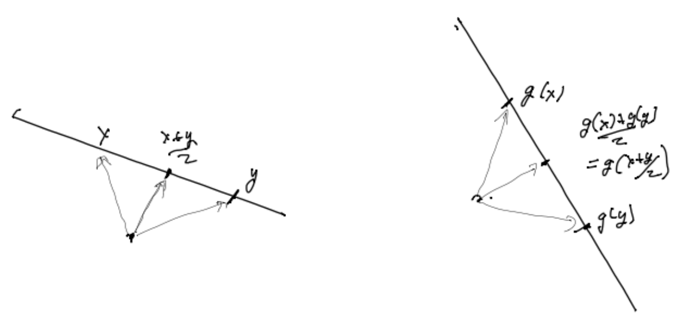

FIGURE 16. Additivity

(b) $r > 1$: Let $a = d(0, x)$ and $b = (r - 1)d(0, x)$. Then get in a similar way that $(0, g(x), g(rx)), (0, g(x), rg(x) \in E_{a,b}$, so by (2) of Lemma 2.12 get $g(rx) = rg(x)$.

(c) $r < 0$: Let $a = |r|d(0, x)$ and $b = d(0, x)$. Then $(g(rx), 0, g(x))$ and $(rg(x), 0, g(x)) \in E_{a,b}$, so, by (3) of the same lemma, $g(rx) = g(x)$.

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

(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 $(g(x), \frac{g(x)+g(y)}{2}, g(y)) \in E(a, a)$ (Lemma 2.13. By Lemma 2.12, $\frac{g(x+y)}{2} = \frac{g(x)+g(y)}{2}$, 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. See Figure 16 This finishes the proof of Theorem 2.11.

□

2.2. **The Euclidean and Orthogonal Groups.** We have seen that orthogonal matrices are invertible, in fact , A being orthogonal is equivalent to. $A^{-1} =^t A$ The product AB of two orthogonal matrices A, B is orthogonal since ${}^t(AB)\ (AB) = {}^tB\ {}^tA\ AB = {}^tB\ B = I$. The unit matrix I is orthogonal. This means that the set of orthogonal matrices forms a group under matrix multiplication. Also, if A is orthogonal, then $\det(I) = \det(A^tA) = \det(A^t)\det(A) = \det(A)^2$, 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 2.14. *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 2.15. The notation $O(n)$, $SO(n)$ is standard. Unfortunately there does not seem to be a standard notation for what we call $E(n)$, $SE(n)$.

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

Definition 2.16. *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 2.17. (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 2.2, that the set 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 2.2 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

$$f_{A_1,b_1} \circ f_{A_2,b_2} = f_{A_1A_2,b_1+A_1b_2} \tag{13}$$

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 2.18. 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 2.19. 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 13 to the product

$$(A_1, b_1)(A_2, b_2) = (A_1A_2, b_1 + A_1b_2). \tag{14}$$

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

$$(A_1, b_1)(A_2, b_2) = (A_1A_2, b_1 + b_2). \tag{15}$$

The group structure of Equations 13 and 14 is called *a semi-direct product* of $O(n)$ and $\mathbb{R}^n$.

Remark 2.20. 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 (= sign of determinant of the linear part) 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 11 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} (t_c^{-1}) 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, see Figure 17 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.*

In summary:

Theorem 2.21. *Let $f \neq id$ be a proper isometry of $\mathbb{R}^2$. Then either*

(1) *f has not fixed points. Then f is a translation.*

(2) *f has a unique fixed point c and f is a rotation by an angle θ (not a multiple of 2π) with center c.*

Remark 2.22. 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, see Figure 17. This explains why the group structure of $SE(2)$ must follow the pattern of Equation (14) rather than that of (15). Note also that $SO(2)$ and $\mathbb{R}^2$ are both abelian groups, so if $SE(2)$ had the group law of (15), then it would be abelian, which is not the case.

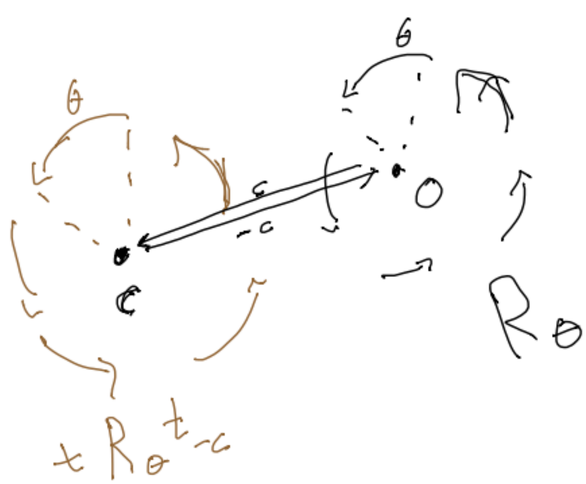

FIGURE 17. Conjugate Rotations

Remark 2.23. 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, as in Figure 17) or in reference to different objects, like reflections in different mirrors that we will see below.

Remark 2.24. Here's a familiar consequence of the group law of Equations (13) and (14). 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 should be familiar to anybody who has parallel-parked 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 (8), or $f(z) = e^{i\theta}\bar{z}$ in complex notation (12). 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 $S_\theta w_2 = -w_2$.

The usual way to find this basis would be to solve linear equations for eigenvectors, which should be familiar. An alternative way would be to use complex notation: For instance, observe that $z = e^{i\frac{\theta}{2}}$ solves $z = e^{i\theta}\bar{z}$. This justifies the claim made in Figure 13 that the fixed line of S_θ makes an angle $\frac{\theta}{2}$ with the x_1 axis. Similarly $z = -e^{i\theta}\bar{z}$ has solution $z = ie^{i\frac{\theta}{2}}$. This would be an orthonormal basis w_1, w_2 as above.

The following terminology for the fixed line of a reflection is very appropriate (see Figure 19:

Definition 2.25. *Let S_θ be as in (8). The fixed line $te^{i\frac{\theta}{2}} = (t\cos\frac{\theta}{2}, t\sin\frac{\theta}{2})$, $t \in \mathbb{R}$, is called the* mirror *of S_θ. (More generally, for any reflection, its fixed line is called its mirror.)*

Remark 2.26. Notice that we just used row vectors to represent points in $\mathbb{R}^2$. We continue to do this For the rest of this chapter.

To classify all improper isometries $f(x) = S_\theta x + b$, it is convenient, and also a good exercise in conjugacy of isometries, to first make the following reduction: it suffices to classify the isometries with linear part S_0 (in other words, with the orthonormal basis w_1, w_2 above bien the standard basis e_1, e_2): Given f, let $g = R_{-\frac{\theta}{2}} f R_{\frac{\theta}{2}}$. Then $g(x) = S_0 x + c$ for $c = R_{\frac{\theta}{2}} b$. In other words,

$$g(x_1, x_2) = (x_1 + c_1, -x_2 + c_2).$$

The equations for (x_1, x_2) to be fixed are:

$$x_1 + c_1 = x_1 \text{ and } 2x_2 = c_2.$$

There are two possibilities:

(1) $c_1 \neq 0$: No soultions, since the first equation is never satisfied. This means that g has no fixed points. We get $g(x_1, \frac{c_2}{2}) = (x_1 + c_1, \frac{c_2}{2})$. In other words *the line $x_2 = \frac{c_2}{2}$ is invariant under g, and their restriction of g to this line is translation by* $c_1 \neq 0$, see Figure 18

Observe that the line $x_2 = \frac{c_2}{2}$ is parallel to the mirror $x_2 = 0$ of S_0 and g interchanges the two sides of this mirror. This is easiest to see if $c_2 = 0$. Then g is simply

$$g(x_1, x_2) = (x_1 + c_1, -x_2) \tag{16}$$

This is called a *glide reflection along the x_1-axis*, see the first picture in Figure 18 If $c_2 \neq 0$, we get the same picture relative to the line $x_2 = \frac{c_2}{2}$, namely we can rewrite the formula for g as

$$g(x_1, (x_2 - \frac{c_2}{2}) + \frac{c_2}{2}) = (x_1 + c_1, -(x_2 - \frac{c_2}{2}) + \frac{c_2}{2})$$

which is the conjugate of (16) by translation by $(0, \frac{c_2}{2})$, as in the second picture in Figure 18

(2) $c_1 = 0$. Then the line the solutions to the fixed-point equations are $(x_1, \frac{c_2}{2}$ with x_1 arbitrary. In other words, the line $x_2 = \frac{c_2}{2}$ is pointwise fixed by g, and g is a reflection in this line., see Figure 19 In other words, *g is a reflection with mirror $x_2 = \frac{c_2}{2}$*, see Definition 2.25.

Finally, the special case of linear part S_0 implies the following theorem:

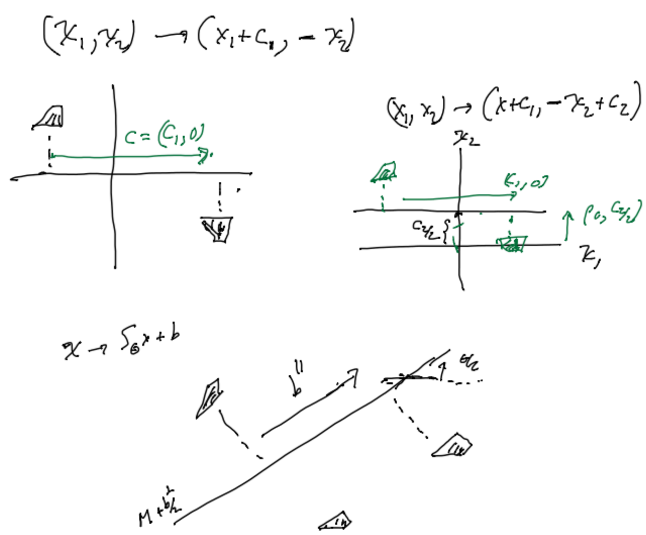

FIGURE 18. Glide Reflections

Theorem 2.27. *Let $f(x) = S_\theta x + b$ be an improper isometry of $\mathbb{R}^2$. Let M denote the mirror of S_θ (see Definition 2.25) and let $b = b^{\|} + b^{\perp}$ denote the components of b parallel, respectively perpendicular, to M.Then either:*

(1) b *is not perpendicular to* M. *Then* f *has no fixed points, it is a glide reflection along the translate* $M + \frac{b^\perp}{2}$ *of* M *by the vector* $b^{\|}$*, see Figure 18*

(2) b *is perpendicular to* M. *Then the translate* $M + \frac{b/}{2}$ *of* M *is fixed by* f*, and* f *is a reflection with mirror* $M + b/2$*, see Figure 19*

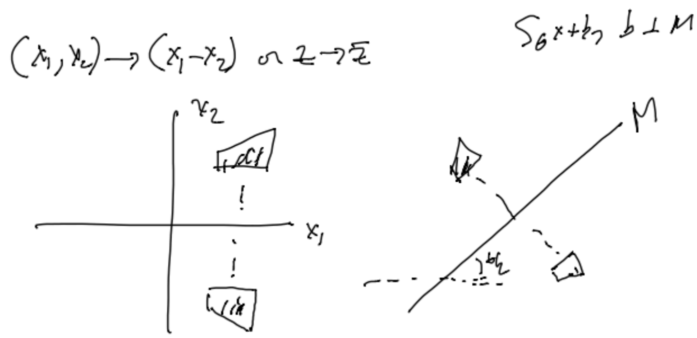

FIGURE 19. Reflections and their Mirrors

We can summarize the classification of isometries (different from the identity) in the following table. In the case of fixed points, the name of the fixed point set is included:

	Proper	*Improper*
With fixed points	Rotations (center)	Reflections (mirror)
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 2.28. *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* $t_{\pm 2b}$ *(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 2.29. *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 [12] for another proof of this Corollary, and Section 1.5 for another proof of the classification theorem.

2.4. **Isometries of the sphere.** Observe that the elements of the orthogonal group $O(n+1)$ are isometries of $\mathbb{R}^{n+1}$ that preserve the origin, so they map the unit sphere $S^n = \{|x| = 1\} \subset \mathbb{R}^{n+1}$ to itself. (Recall $|x| = d(0,x)$.) These are isometries of (S^n, d_{extr}), the *extrinsic* metric on S^n, restriction of the metric $d_{\mathbb{R}^{n+1}}$: of $\mathbb{R}^{n+1}$, that is, $d_{extr}(x,y) = |x-y|$. But they also isometries of the *intrinsic* metric $d_{S^n}(x,y) = \cos^{-1}(x \cdot y)$ of Example 1.8.

One way to see this is the following reasoning. A matrix A is orthogonal if and only if it satisfies (6), which is the same as $Ax \cdot Ax = x \cdot x$ tor all $x \in \mathbb{R}^{n+1}$. Therefore $A(x+y) \cdot A(x+y) = (x+y) \cdot (x+y)$ for all $x, y \in \mathbb{R}^{n+1}$. Expanding both sides of this equation, we get

$$Ax \cdot Ax + 2Ax \cdot Ay + Ay \cdot Ay = x \cdot x + 2x \cdot y + y \cdot y \text{ for all } x, y.$$

Since the corresponding first and last terms on each side of this equality are equal (by (6)), so ate the middle terms, in other words, $Ax \cdot Ay = x \cdot y$ for all x, y, as desired.

For the spherical metric this says $d_{S^n}(Ax, Ay) = \cos^{-1}(Ax \cdot Ay) = \cos^{-1}(x \cdot y) = d_{S^n}(x,y)$. Therefore A is an isometry of (S^n, d_{S^n}), as claimed.

Theorem 2.30. *Let $A \in O(n+1)$. Then the map $x \to Ax$ of $\mathbb{R}^{n+1}$ restricts to an isometry $restr(A)$ of (S^n, d_{S^n}). The map $restr : O(n+1) \to Isom(S^n, d_{s^n})$ so defined is a group isomorphism.*

Proof. We have just checked that the map $restr : O(n+1) \to Isom(S^n)$ is defined, and it is easy to check that it is a group homomorphism. It is clearly injective. It remains to check that it is surjective, this will not be easy to do now. Hope to return to this at the end of the semester.

□

Remark 2.31. It is appropriate to mention, without details, that there is a similar theorem for the isometries of hyperbolic space of Example 1.18. For $n = 2$ (similarly for any $n > 2$) there is the *Lorentz group* of matrices that preserve the Minkowski inner product, and they restrict to isometries of H, giving an analogous isomorphism of groups

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 3.3. *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 3.4. 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 3.5. 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 1.35): 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 1.35).

Example 3.6. Suppose X is any non-empty set and let $d : X \times X \to \mathbb{R}$ be the discrete metric of Example 1.12. 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 3.7. 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 3.8. *A subset $F \subset X$ is called a* closed set *if and only if its complement $X \setminus F$ is an open set.*

Theorem 3.9. *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 3.3, 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 3.10. 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 [8] 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

$$f^{-1}(A) = \{x \in X : f(x) \in A\}. \tag{17}$$

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

$$f(A) = \{f(x) : x \in A\} \tag{18}$$

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

Theorem 3.11. *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 $x \in (g \circ f)^{-1}(A) \Leftrightarrow (g \circ f)(x) \in A \Leftrightarrow g(f(x)) \in A \Leftrightarrow f(x) \in g^{-1}(A) \Leftrightarrow x \in f^{-1}(g^{-1}(A)) \Leftrightarrow x \in f^{-1} \circ g^{-1}(A)$. $\square$

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 1.35(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 3.12. *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 1.35, 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 consequences:

Corollary 3.13. *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 3.11, $f^{-1}(Y \setminus F) = f^{-1}(Y) \setminus f^{-1}(F) = X \setminus f^{-1}(F)$ 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 3.14. *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 3.11, 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 3.15. *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 3.16. *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 3.17. 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 3.18. Let (X, d) be a discrete metric space as in Example 1.12 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 3.6, 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 3.19. 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 3.5.

3.1.3. **The Collection of Open Sets.**

Theorem 3.20. *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 3.21. *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 union $F_1 \cup \cdots \cup F_n$ 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 3.22. *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 3.23. *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 3.24. *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 3.25. 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 3.26. *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 3.27. Just as in the notation we explained in Remark 1.36, 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 3.13, we have the following characterization of continuity (with the same proof):

Theorem 3.28. *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 3.14. we have that the composition of continuous maps is continuous (again with the same proof):

Theorem 3.29. *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 3.30. 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 3.20, the collection $\mathcal{T}_d \subset 2^X$ is a topology on X.

Example 3.31. In the special case that $X = \mathbb{R}^n$ and d is the Euclidean metric of Example 1.4 we call the resulting metric topology $\mathcal{T}_d$ the *Euclidean topology* and denote it by $\mathcal{T}_E$.

Example 3.32. 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 1.12 and 3.6.

Example 3.33. 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 3.22, while the last example gave the *largest* one. This is often called the *indiscrete topology*.

Example 3.34. 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 [8].

Example 3.35. 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 3.25. 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 3.36. 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 3.37. 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 3.38. 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 3.39. 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 3.40. 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 3.41. 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 3.42. 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 3.43. *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 3.44. 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 3.45. *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 1.29. Now consider the following example:

Example 3.46. Consider $(\mathbb{R}, \mathcal{T}_Z)$ as in Example 3.36, 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 1.29 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 3.47. *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 3.48. *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 3.49. *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 3.50. *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 3.51. The discussion of Example 3.46 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 3.52. *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 3.53. *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 3.54. *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 3.53: for every neighborhood U of x, $U \cap A \neq \emptyset$.*

Remark 3.55. Thus a set is closed if and only if it contains all its limit points. This is sometimes taken as the definition of closed set. It has a more immediate appeal than the definition we have chosen. It says that a set is closed if and only if you cannot get outside it by the process of taking limit points. Perhaps the most immediate way to see the equivalence is to say that x is *not* s limit point of A if and only if x has a neighborhood U with $U \cap A = \emptyset$, in other words, x is in the interior of $X \setminus A$. Thus A contains all its limit points if and only if every point of $X \setminus A$ is an interior point. Thus A is closed (in the sense of containing all its limit points) if and only if $X \setminus A$ is open. We chose $X \setminus A$ open as the definition of A being closed mostly for convenience: properties of closed set immediately translate to properties of open sets by usual rules for operations on complements.

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 $\{x_n\}$.

Remark 3.56. Note that in Definition 3.54 we do not require that $U \cap A$ contain a point $y \in A$ with $y \neq x$. So, by this definition, every $x \in A$ is a limit point of A. If every neighborhood U of x contains $y \in U \cap A$ with $y \neq x$ then x is sometimes called an *accumulation point* of A.

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

Definition 3.57. *A subset $\mathcal{B} \subset \mathcal{T}$ 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 3.58. 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 3.59. 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 $z \in \mathbb{Q}^n$ 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 3.3), 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 1.41. 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 1.41, $d_{(2)}(x, z) < \epsilon$.

Remark 3.60. 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 3.22. 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 3.61. *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 $\mathcal{T} = \{U \subset X :$ for all $x \in U$ there exists $B \in \mathcal{B}$ with $x \in B$ and $B \subset U\} \cup \{\emptyset\}$. 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 3.22 holds. By the definition of $\mathcal{T}$, $\emptyset \in \mathcal{T}$. 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 3.22 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 $\mathcal{T}$. $\square$

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

$$\mathcal{B} = \{U \times V : U \in \mathcal{T}_X \text{ and } V \in \mathcal{T}_Y\} \tag{19}$$

It is easy to check that $\mathcal{B}$ satisfies the conditions of Theorem 3.61. 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):

$$(U_1 \times V_1) \cap (U_2 \times V_2) = (U_1 \cap U_2) \times (V_1 \cap V_2) \tag{20}$$

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 3.62. We could modify the definition of $\mathcal{B}$ in Equation 21 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 3.61 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 (20) for intersections of products.

Remark 3.63. 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 3.64. *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 $\{U \times V : U \in \mathcal{T}_X$ 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.

$\square$

3.4.3. **Sub-basis for a topology.** One natural way to phrase the last proof is to use the notion of *sub-basis* for a topology. Briefly, a sub-basis for $\mathcal{T}$ is a collection $\mathcal{B} \subset \mathcal{T}$ with the property that every $U \in \mathcal{T}$ is a union of finite intersections of elements of $\mathcal{T}$:

Example 3.65. The collection of sets

$$\{p_X^{-1}(U) \mid U \in \mathcal{T}_X\} \cup \{p_Y^{-1}(V) \mid V \in \mathcal{T}_Y\}$$

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

To check that a map $f : X \to Y$ is continuous, it is enough to check that $f^{-1}(B) \in \mathcal{T}_X$ for all B in a sub-basis for $\mathcal{T}_Y$, since any open set in Y is a union of intersection $B_1 \cap \cdots \cap B_k$, therefore $f^{-1}(V)$ is a union of sets $f^{-1}(B_1 \cap \cdots \cap B_k) = f^{-1}(B_1 \cap B_k) \cap \cdots \cap f^{-1}(B_k)$ which is open in X. This is the principle we used in the proof of the second statement of Theorem 3.64.

3.5. **Infinite Products.** We now define the product of an arbitrary collection of topological spaces. First, we need to define the product of an arbitrary collection of sets.

Definition 3.66. *Let A be a set and let $\{X_\alpha\}_{\alpha \in A}$ be a collection of sets indexed by A. The* product *of the collection is defined by*

$$\prod_{\alpha \in A} X_\alpha = \{f : A \to \bigcup_{\alpha \in A} X_\alpha \mid \forall \alpha \in A, f(\alpha) \in X_\alpha\}$$

Note that this is a definition purely in set theory. The sets A and X_α are not assumed to have any topology, and the functions $f : A \to \bigcup_\alpha X_\alpha$ are arbitrary functions on sets.

Example 3.67. Suppose $A = \{1, 2\}$. Then a function $f : \{1, 2\} \to X_1 \cup X_2$ with $f(1) \in X_1$ and $f(2) \in X_2$ is completely determined by the pair of values $f(1) = x_1 \in X_1$ and $f(2) = x_2 \in X_2$. thus

$$\prod_{\alpha \in \{1,2\}} X_\alpha = \{(x_1, x_2) \mid x_1 \in X_1, x_2 \in X_2\} = X_1 \times X_2,$$

the usual definition of the Cartesian product $X_1 \times X_2$.

Similarly, for a finite set $A = \{1, 2, \ldots, n\}$, $\prod_{\alpha \in A} X_\alpha$ gives the usual product $\{(x_1, x_2, \ldots, x_n\} | \, x_i \in X_i\} = X_1 \ldots X_n$.

Example 3.68. If $A = \mathbb{N}$, the natural numbers, and $X_i = X$ for all $i \in \mathbb{N}$, then

$$\prod_{i \in \mathbb{N}} X = X^{\mathbb{N}} = \{ \text{ Sequences } \{x_i\}_{i \in \mathbb{N}}\}$$

the set of all sequences in X.

One formulation of the *Axiom of Choice* is the following statement:

If $A \neq 0$ and or all $\alpha \in A$, $X_\alpha \neq \emptyset$, then $\prod_{\alpha \in A} X_\alpha \neq \emptyset$.

In other words, the functions $f : A \to \bigcup_{\alpha \in A} X_\alpha$ with $f(x_\alpha) \in X_\alpha$ "choose", for each $\alpha \in A$, an element of X_α. The axiom of choice is the statement that these choices are always possible. Note that there is no issue here if A is a finite set. The axiom is only needed for arbitrary cardinality.

3.5.1. **Topology on an Infinite Product Space.** Assume now that A is an arbitrary set and that each of the sets X_α has a topology $\mathcal{T}_\alpha$. For each finite subset $F = \{\alpha_1, \ldots \alpha_k\} \subset A$ choose sets $U_{\alpha_i} \in \mathcal{T}_{\alpha_i}$ and let

$$\mathcal{U}_F = U_{\alpha_1} \times \cdots \times U_{\alpha_k} \times \prod_{\alpha \neq \alpha_i} X_\alpha,$$

in other words,

$$\mathcal{U}_F = \{f : A \to \cup_\alpha X_\alpha \mid f(\alpha) \in U_\alpha \text{ for all } \alpha \in F\} \tag{21}$$

Theorem 3.69. *Let $\mathcal{B}_{\prod X_\alpha}$ denote the collection of all the $\mathcal{U}_F$. Then $\mathcal{B}_{\prod X_\alpha}$ satisfies the conditions of Theorem 3.61, hence is the basis for a topology on $\prod_{\alpha \in A} X_\alpha$.*

Proof. The proof is very similar to the proof of the case of two factors in §3.4.2. We replace equation (20) by the following formula: given two finite subsets $F, F' \subset A$ and open sets $\{U_\alpha \,|\alpha \in F\}$ and $\{U'_{\alpha'} \mid \alpha' \in F'\}$, let $F'' = F \cup F'$. Then, for $\beta \in F''$, let

$$V_\beta = \begin{cases} U_\beta & \text{if } \beta \in F, \\ X_\beta & \text{otherwise.} \end{cases} \quad \text{and} \quad V'_\beta = \begin{cases} U'_\beta & \text{if } \beta \in F', \\ X_\beta & \text{otherwise.} \end{cases}$$

Then let $\mathcal{V}_{F''} = \prod_{\beta \in F''} V_\beta \times \prod_{\beta \notin F''} X_\beta$ and $\mathcal{V}'_{F''} = \prod_{\beta \in F''} V'_\beta \times \prod_{\beta \notin F''} X_\beta$. Then

$$\mathcal{U}_F \cap \mathcal{U}'_{F'} = \mathcal{V}_{F''} \cap \mathcal{V}'_{F''} = \prod_{\beta \in F''} (V_\beta \cap \mathcal{V}'_\beta) \times \prod_{\beta \notin F''} X_\beta$$

Denoting the last space $\mathcal{V}''_{F''}$, we get $\mathcal{U}_F \cap \mathcal{U}'_{F'} = V''_{F''}$, so $\mathcal{B}$ is closed under finite intersections and therefore is a basis for a topology.

□

Definition 3.70. *The topology on* $\prod X_\alpha$ *defined by the basis* $\mathcal{B}$ *consisting of all the sets* $\mathcal{U}_F$ *is called the* product topology.

Just as in the case of two factors we have projections, for each $\alpha \in A$

$$p_\alpha : \prod_{\alpha' \in A} X_{\alpha'} \to X_\alpha \text{ defined by } p_\alpha(f) = f(\alpha) \in X_\alpha$$

In this definition we have used the notation of Definition 3.66: an element of $\prod_{\alpha \in A} X_\alpha$ is a function $f : A \to \bigcup X_\alpha$ with the property that, for each $\alpha \in A$, $f(\alpha) \in X_\alpha$.

Theorem 3.71. *Let* $(X_\alpha, \mathcal{T}_\alpha)_{\alpha \in A}$ *be a family of topological spaces, let* $X = \prod_{\alpha \in A} X_\alpha$*, and, for each* $\alpha \in A$*, let* $p_\alpha : \prod X_{\alpha'} \to X_\alpha$ *be the projection*

1. *The product topology is the smallest topology on* X *that makes all projections* $p_\alpha : X \to X_\alpha$ *continuous.*
2. *If* $(Z, \mathcal{T}_Z)$ *is any topological space, a map* $f : Z \to X$ *is continuous if and only if all compositions* $p_\alpha \circ f$ *are continuous.*

Proof. Similar to the proof of Theorem 3.64. Here we use the fact that the collection of all sets $p_\alpha^{-1}(U)$, where U is open in X_α, is a sub-basis (see §3.4.3) for the topology of the product space. □

Remark 3.72. The most important feature of the definition of the product topology is that the open $\mathcal{U}_F$ in the basis restrict only *finitely many factors*. Looking at the formula in Definition 3.66 makes this very clear: $f(\alpha) \in U_\alpha$ for all $\alpha \in F$ puts *no restriction* on the values of f on $A \setminus F$. If A is an infinite set, then the elements of $\mathcal{U}_F$ are arbitrary on most of their domain. One could define other topologies making more restrictions, for instance, we could choose open sets $U\alpha \subset X_\alpha$ for each $\alpha \in A$ and take for a basis the sets $\{f : A \to \cup X_\alpha \,|f(\alpha) \in U_\alpha$ for all $\alpha \in A\}$. This would define a topology (called the "box topology"). It makes all the p_α continuous, but it's larger than the product topology, so (1) of Theorem 3.71 is true. Part (2) of the same theorem is also false. See, for example, [4, 9] for a more detailed discussion of the product topology.

3.6. **The Cantor Set.** A very interesting example of an infinite product space is provided by the Cantor set. We first recall its construction as a subset of the unit interval $[0, 1] \subset \mathbb{R}$:

Start with the unit interval $[0, 1]$, divide it into three equal intervals, and remove the open middle interval. What remains is $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Iterate this construction: to each interval apply the same process: divide it into three equal intervals, remove the open middle interval. For instance, the next step is

$$C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$$

Continue. At the nth stage we get C_n a union of 2^n intervals. Moreover the C_n are nested: $C_1 \supset C_2 \supset \dots$. Consequently

$$C = \bigcap_{n=1}^{\infty} C_n \neq \emptyset$$

This is the Cantor set C. See Figure 20

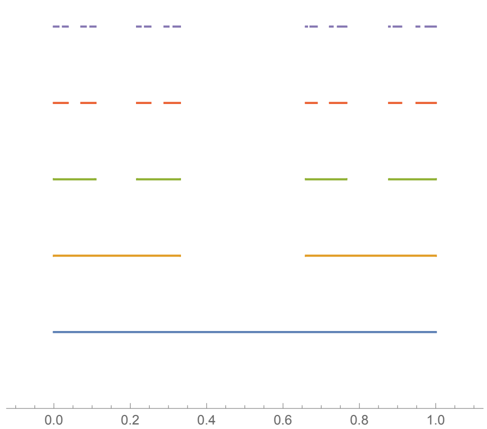

FIGURE 20. Constructing the Cantor Set

Observe that, by definition of the ternary expansion of a real number

$$C = \{\sum_{i=1}^{\infty} \frac{a_i}{3^i} \mid a_i = 0 \text{ or } 2\}$$

Removing the middle thirds removes all points with a one in their ternary expansions, except for right hand endpoints of the intervals, which have two ternary expansions: a finite one ending with a one, or an infinite one with 2 repeated infinitely often. In particular, the elements of C have a *unique* ternary expansion using only 0 and 2. In other words, the map

$$t : \{0,2\}^{\mathbb{N}} \to C \quad \text{defined by} \quad t(\{a_i\}) = \sum_{i=1}^{\infty} \frac{a_i}{3^i} \tag{22}$$

is *bijective*.

Theorem 3.73. *Give* $\{0,2\}$ *the discrete topology, give* $\{0,2\}^{\mathbb{N}}$ *the product topology and give* C *the topology as a subspace of* $[0,1]$ *(equivalently, the metric topology from* $[0,1]$*). Then the map* $t : \{0,2\}^{\mathbb{N}} \to C$ *defined by equation (22) is a homeomorphism.*

Proof. We will prove that t is continuous. The continuity of t^{-1} will be done later. Fix a sequence $a^0 = \{a_i^0\} \in \{0,2\}^{\mathbb{N}}$ and let $\epsilon > 0$. Then choose i_0 so that $\frac{1}{3^{i_0}} < \epsilon$ If $\{a_i\} \in \{0,2\}^{\mathbb{N}}$ and $a_i = a_i^0$ for $i \leq i_0$, then

$$|t(\{a_i\}) - t(\{a_i^0\})| \leq \sum_{i_0+1}^{\infty} \frac{|a_i - a_i^0|}{3^i} \leq \sum_{i_0+1}^{\infty} \frac{2}{3^i} = \frac{1}{3^{i_0}} < \epsilon \tag{23}$$

where we used the geometric series to compute

$$\sum_{i_0+!}^{\infty} \frac{2}{3^i} = \frac{2}{3^{i_0+1}} \sum_{0}^{\infty} \frac{1}{3^i} = \frac{2}{3^{i_0+1}} \frac{3}{2} = \frac{1}{3^{i_0}}$$

Thus if we let

$$U = \{\{a_i\} \in \{0,2\}^{\mathbb{N}} \mid a_i = a_i^0 \text{ for } i \leq i_0\} \tag{24}$$

then U is an open set with $a^0 \in U \subset t^{-1}(B(t(a^0), \epsilon)$. Since a^0 and ϵ are arbitrary, t is continuous. $\square$

Remark 3.74. Observe how the inequality (23) leads to the open set U in the product topology described by equation (24). Restricting the "tail end" of the geometric series $\sum \frac{a_i}{3^i}$ can be achieved by restricting only finitely many a_i.

4. Subspaces and Quotient Spaces

Let X and Y be sets and let $f : X \to Y$ be a map. We know from Examples 3.37 and 3.38 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 4.2. 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 4.3. *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 4.4. *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. $\square$

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

Theorem 4.5. *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. $\square$

4.2. **Compact Spaces.** Let $(X, \mathcal{T})$ be a topological space. An *open cover* $\mathcal{U}$ of X is a collection $\mathcal{U} = \{U_\alpha\}_{\alpha \imath A}$ of open sets (elements of $\mathcal{T}$), so that $\bigcup_{\beta \in B} U_\beta$

$$X = \bigcup_{\alpha \in A} U_\alpha$$

A *finite subcover* of $\mathcal{U}$ means a finite subcollection $U_{\alpha_1}, \ldots, U_{\alpha_n}$ of $\mathcal{U}$ so that

$$X = U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}.$$

Using this language, we have the following definition:

Definition 4.6. *A topological space $(X, \mathcal{T})$ is called* compact *if and only if every open cover of X has a finite subcover. A subset $Y \subset X$ of a topological space $(X, \mathcal{T})$ is called compact if and only if it is a compact topological space when given the subspace topology (Definition 4.3) from $(X, \mathcal{T})$*

Explicitly: $(X, \mathcal{T})$ is compact if and only if, whenever $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ is a collection of open sets such that $X = \bigcup_\alpha U_\alpha$, there is a finite subcollection $U_{\alpha_1}, \ldots, U_{\alpha_n}$ of $\mathcal{U}$ so that $X = U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}$. If $Y \subset X$, then, using the definition (Definition 4.3) of the subspace topology, it is easy to see that Y is compact if and only if whenever $\mathcal{U} = \{U_\alpha\}$ is a collection of open sets in X with

$$Y \subset \bigcup_\alpha U_\alpha,$$

then there exists a finite subcollection $U_{\alpha_1}, \ldots, U_{\alpha_n}$ of $\mathcal{U}$ with

$$Y \subset U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}.$$

Example 4.7. (1) Let X be *finite*. Then X is compact.
(2) Let $(X, \mathcal{T}_{disc})$ be infinite and have the discrete topology. Then X is not compact. In fact, the open cover $\mathcal{U} = \{\{x\} : x \in X\}$ is an open cover of X that has no proper sub-cover.
(3) $\mathbb{R}$ is not compact: The open cover $\mathcal{U} = \{(-n, n) \mid n \in \mathbb{N}\}$ has no finite subcover.
(4) Let $X = \{\frac{1}{n} \, |n \in \mathbb{N}\} \cup \{0\} \subset \mathbb{R}$. then X is compact. Reason:
   (a) Let $\mathcal{U} = \{U_\alpha\}$ be open in $\mathbb{R}$ and $X \subset \bigcup_\alpha U_\alpha$.
   (b) There is α_0 such that $0 \in U_{\alpha_0}$.
   (c) There is $N \in \mathbb{N}$ such that $\frac{1}{n} \in U_{\alpha_0}$ for all $n > N$.
   (d) For $1 \leq n \leq N$, choose U_{α_n} with $\frac{1}{n} \in U_{\alpha_n}$.
   (e) Then $X \subset U_{\alpha_0} \cup \cdots \cup U_{\alpha_N}$.

It is difficult to apply Definition 4.6 directly to prove that a space is compact. The above examples are not typical in this respect. It is easier to prove that a space is *not* compact. It is also easier to derive some properties of compact spaces, and to prove compactness of some spaces given the compactness of some other space. Here are some examples.

Theorem 4.8. (1) *Let X be compact and $C \subset X$ be closed. Then C is compact.*
(2) *Let X ba a Hausdorff space and let $C \subset X$ be compact. Then C is closed.*
(3) *Let X be a compact metric space. Then X is* bounded*: there exists a constant $C > 0$ such that $d(x, y) \leq C$ for all $x, y \in X$.*

Proof. (1) Let $\mathcal{U}$ be a collection of open sets in X with $C \subset \bigcup \mathcal{U}$, and let $\mathcal{V} = \mathcal{U} \cup \{X \setminus C)$. Then $\mathcal{V}$ is an open cover of X, hence it has a finite sub-cover which consists of finitely many elements of $\mathcal{U}$ and possibly $X \setminus C$. Since the latter is disjoint from C, the finitely many elements of $\mathcal{U}$ cover C.
(2) Suppose X is Hausdorff, $C \subset X$ is compact, and $x \notin C$. For each $y \in C$ there exist neighborhoods U_y of y and V_y of x so that $U_y \cap V_y = \emptyset$. Since $c \subset \cup_y U_y$ and C is compact, there exist $y_1, \ldots y_n$ so that

$$C \subset U_{y_1} \cup \cdots \cup U_{y_n}$$

Let

$$V = V_{y_1} \cap \cdots \cap V_{y_n}$$

Then V is a neighborhood of x and $V \cap U_{y_i} = \emptyset$ for $i = 1, \ldots, n$. Thus

$$V \ \cap \ (U_{y_1} \cup \cdots \cup U_{y_n}) \ = \ \emptyset$$

Since $C \subset U_{y_1} \cup \cdots \cup U_{y_n}$, it follows that $V \cap C = \emptyset$. Therefore $X \setminus C$ is open, thus C is closed.
(3) Fix $x_0 \in X$ and let $\mathcal{U} = \{B(x_0, n) \mid n \in \mathbb{N}\}$. Then $\mathcal{U}$ is an open cover of X. Take a finite subcover and let N be the largest radius of a ball in this subcover. Then, for all $x \in X, d(x_0, x) < N$, thus for all $x, y \in X$, $d(x, y) < 2N$.

□

We recall the statement of the Heine-Borel theorem characterizing compact subsets of $\mathbb{R}^n$ (with Euclidean topology). We hope this is a familiar theorem whose proof you have seen in a previous course.

Theorem 4.9. *Let $X \subset \mathbb{R}^n$. Then X is compact if and only if X is closed and bounded.*

4.2.1. **Continuous maps and compactness.** One reason for the definition of compactness is that it makes the following theorem almost obvious:

Theorem 4.10. *Let X, Y be topological spaces, let $f : X \to Y$ be continuous, and let $C \subset X$ be compact. then $f(C)$ is compact.*

Proof. Take an open cover $\mathcal{U} = \{U_\alpha\}$ of $f(C)$, then $f^{-1}(\mathcal{U}) = \{f^{-1}(U_\alpha)\}$ is an open cover of C. Choose a finite subcover $\{f^{-1}(U_{\alpha_i}\}, i = 1, \dots n$ of C. Then $\{U_{\alpha_i}\}, i = 1, \dots, n$ covers $f(C)$. □

This theorem, combined with the Heine-Borel theorem and other facts about compactness (say, as in Theorem 4.8) has many traditional applications. For example, a continuous real valued function on a compact set attains its maximum and its minimum. Here are some other applications:

Theorem 4.11. *Let X be a compact space, let Y be a Hausdorff space, and let $f : X \to Y$ be continuous.*

1. *If $C \subset X$ is closed, then $f(C) \subset Y$ is also closed. (terminology: f is a* closed map*).*
2. *Suppose, in addition, that f is bijective. Then f is a homeomorphism.*

Proof.
1. By (2) of Theorem 4.8 C is compact, by Theorem 4.10 $f(C)$ is compact, by (1) of Theorem 4.8 C is closed.
2. If f is bijective, then $f^{-1} : Y \to X$ exists and $(f^{-1})^{-1}(C) = f(C)$ is closed in Y for all C closed in X, hence f^{-1} is continuous.

□

4.2.2. **Compactness and Products.** In the homework you are asked to prove that if X and Y are compact topological spaces, so is $X \times Y$. This is not too difficult to prove, but it takes a litle bit of work. It is a remarkable feature of the product topology that this remains true of infinite products, with the product topology as defined in §3.5.1. This is known as *Tychonoff's Theorem*:

Theorem 4.12. *Let $A \neq \emptyset$ and let $\{X_\alpha\}_{\alpha \in A}$ be a collection of non-empty spaces indexed by A. Then, if all the X_α are compact, $\prod_{\alpha \in A} X_\alpha$ is also compact.*

Proof. This would take us too far from our path. See [4, 9] for proofs. □

If we accept this fact, then we can finish the proof of Theorem 3.73. Recall we had a map $t : \{0, 2\}^{\mathbb{N}} \to C$ from the infinite product space

$\{0,2\}^{\mathbb{N}}$ to the Cantor set $C \subset [0,1]$, and we proved that t is a continuous bijection. Now we know by Theorem 4.12 that $\{0,2\}^{\mathbb{N}}$ is compact, and from Theorem 4.11 we know that t must be a homeomorphism. This finishes the proof of Theorem 3.73.

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

Definition 4.13. *Let $(X, \mathcal{T}_X)$ be a topological space and let $q : X \to Y$ be surjective.*

(1) *The* quotient topology*, also called the* identification topology *on Y is the topology $T_Y = \{U \subset Y : q^{-1}(U) \in \mathcal{T}_X\}$.*
(2) *A surjective continuous map $q : X \to Y$ 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.*

In other words, a surjective map $q : X \to Y$ is an identification if and only if $U \subset Y$ is open if and only if $q^{-1}(U)$ is open in X. Equivalent formulation: a surjective map $q : X \to Y$ is an identification if and only if $F \subset Y$ is closed in Y iff and only if $q^{-1}(F)$ is closed in X.

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

Example 4.14. 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 $A = \{(\cos t, \sin t) : 0 \leq t < \pi\}$, 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 4.15. *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 4.16. 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 4.14 shows the following:

Theorem 4.17. *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 4.2, Y has the quotient topology and f is an identification.

□

Remark 4.18. Theorem 4.17 gives sufficient conditions for f to be an identification. But these are not necessary conditions. For example, the map f_1 of Example 4.14 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 4.19. Let X and Y be two sets. Then the following are equivalent:

(1) A surjective map $fqX \to Y$.
(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 the elements of Y.

The equivalences are easy to see: Given (1), define the partition in (2) by $X_y = q^{-1}(y)$, and given the partition (2), define $q : X \to Y$ by $q(x) = y$ 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 the elements of 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 4.20. *Let X, Y and Z be topological spaces. Suppose that maps q and g are given as in the following diagram, and that q is an identification.*

$$\begin{array}{ccc} X & \xrightarrow{g \circ q} & Z \\ q \Big\downarrow & \nearrow g & \\ Y & & \end{array} \tag{25}$$

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

.

Proof. If g is continuous then certainly $g \circ q$ is continuous by Corollary 3.14. What is specific to the identification topology is the converse, which is proved as follows: if $g \circ q$ is continuous, then for each open $U \subset Z$, $(g \circ q)^{-1}(U)$ is open in X. But $(g \circ q)^{-1}(U) = q^{-1}(g^{-1}(U))$, thus, since q is an identification, $g^{-1}(U)$ is open in Y, so g is continuous. $\square$

.

This theorem is usually applied in the following equivalent form. Suppose that $q : X \to Y$ 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 q (the sets $q^{-1}(y)$, $y \in Y$). In other words, suppose that $h(x) = h(x')$ whenever $q(x) = q(x')$. Then we can define a map $g : Y \to Z$ as follows: given $y \in Y$, choose $x \in X$ so that $q(x) = y$, 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 $q(x') = y$, then $q(x) = q(x')$, 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 4.21. *In the following diagram, suppose that X, Y and Z are topological spaces, q is an identification and h is constant on the fibers of q, so that the map g as in the above discussion is well-defined.*

$$\begin{array}{ccc} X & \xrightarrow{h} & Z \\ q\big\downarrow & \nearrow g & \\ Y & & \end{array} \tag{26}$$

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

Proof. Since, by the definition of g, $h = g \circ q$, this is the same as Theorem 4.20. $\square$

Example 4.22. We can apply this Theorem to the identification $f : \mathbb{R} \to S^1$ of Example 4.14. 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.

Example 4.23. One word of warning: it can easily happen that $q : X \to Y$ is an identification, X is Hausdorff, yet Y is *not* Hausdorff. Here is a standard example. Let $X = \mathbb{R} \times 0 \cup \mathbb{R} \times 1$, the disjoint union of two copies of $\mathbb{R}$. It can be visualized as the subspace $\{(x, 0) : x \in \mathbb{R}\} \cup \{(x, 1) : x \in \mathbb{R}\} \subset \mathbb{R}^2$. Let

$$x \times 0 \sim x \times 1 \text{ if } x < 0$$

and let $Y = X/\sim$. In other words, the equivalence classes have two elements ($x \times 0$ and $x \times 1$) for $x < 0$, one element (either $x \times 0$ or $x \times 1$)

for $x \geq 0$. One attempt to picture it would be as in Figure 21. The figure suggests that the two points 0×0 and 0×1 are distinct (as they are in the identifcation space, that cannot be drawn in the plane). If U is any neighborhood of 0×0 in Y, then $q^{-1}(U) \subset X$ contains

$$(-\epsilon, \epsilon) \times 0 \cup (-\epsilon, 0) \times 1 \text{ for some } \epsilon > 0$$

Similarly, if V is any neighborhood of 0×1 in Y, then $q^{-1}(V)$ contains

$$(-\epsilon, 0) \times 0 \cup (-\epsilon, \epsilon) \times 1 \text{ for some } \epsilon > 0$$

thus $U \cap V$ contains $q((-\epsilon, 0) \times 0)) \neq \emptyset$. Thus Y is not Hausdorff.

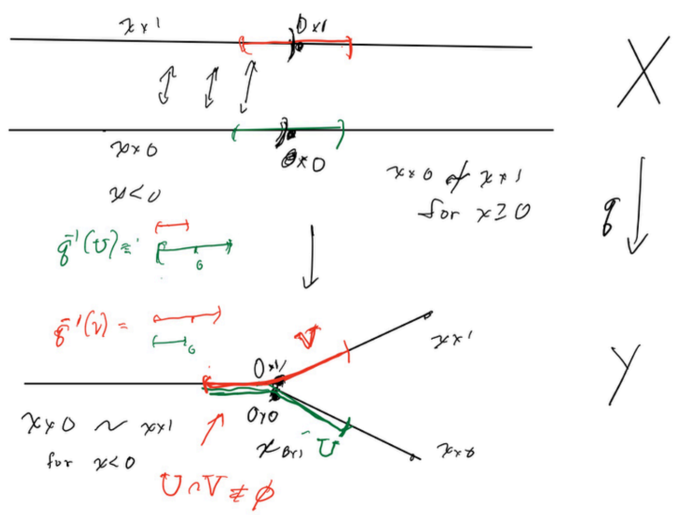

FIGURE 21. A non-Hausdorff Identification

This should be compared with a Hausdorff identification that looks very similar. Start from the same X, call it X_1, but change the equivalence relation to $x \times 0 \sim x \times 1$ for $x \leq 0$. Then the "bad " points 0×0 and 0×1 are no longer distinct in the quotient Y_1, and Y_1 is indeed Hausdorff. See Figure 22

Example 4.24. A more extreme example would be the identification $q : \mathbb{R} \to \mathbb{R}/\mathbb{Q}$. What is the quotient topology on $\mathbb{R}/\mathbb{Q}$ (from the usual topology on $\mathbb{R}$)? Here $\mathbb{R}/\mathbb{Q}$ is the usual quotient group, the set of cosets $\{x + \mathbb{Q} \mid x \in \mathbb{R}\}$. In other words, the equivalence relation on $\mathbb{R}$ is $x \sim y$ if and only if $x - y \in \mathbb{Q}$.

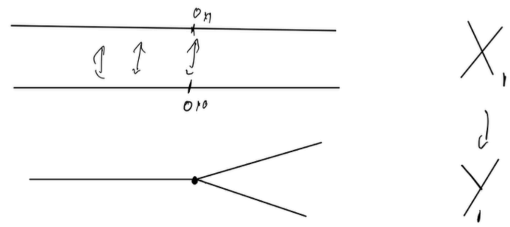

FIGURE 22. A Hausdorff Identification

4.4. **Surfaces as Identification Spaces.** We now apply Theorem 4.21 to define various surfaces. The procedure is in some cases similar to what we saw in Example 4.14 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 [6] for more discussion (and pictures) of this procedure.

Example 4.25. 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 $(x, y, z) = ((2 + \cos\phi)\cos\theta, (2 + \cos\phi)\sin\theta, \sin\phi)$, $0 \leq \theta, \phi \leq 2\pi$. In the same way that we showed in Example 4.14 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 4.14. 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 identfication space $T = S/\sim$ just defined would be indicated as follows:

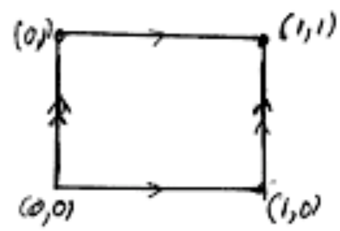

FIGURE 23. Torus

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:

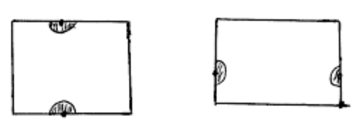

FIGURE 24. Neighborhoods in Identification Space

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

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

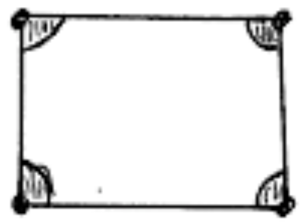

FIGURE 25. Neighborhood of the Corner

$$\begin{array}{rcl} S & \xrightarrow{p_1|_S} & \mathbb{R}^2/\mathbb{Z}^2 \\ p\downarrow & & \downarrow id \\ T = S/\sim & \xrightarrow{g} & \mathbb{R}^2/\mathbb{Z}^2 \end{array}$$

By Theorem 4.21 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 4.26. *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 4.27. *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

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 $(0, y) \sim$

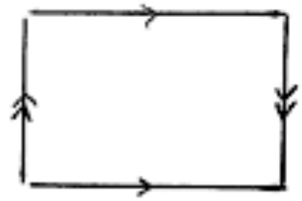

FIGURE 26. The Klein Bottle

$(1, 1 - y)$. The quotient topology can be explicitly defined and illustrated in a fashion analogous to the discussion of the torus in Example 4.25.

While the identification of the torus T with a surface in $\mathbb{R}^3$ is easy to visualize (see, for example, p. 300 of [7]), the Klein bottle can only be realized as a surface with self-intersections. See p. 308 of [7] for pictures and explanation.

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

Definition 4.28. *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

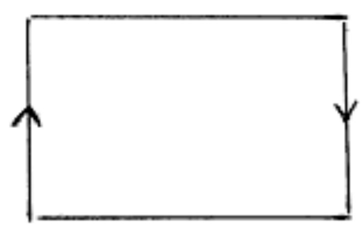

FIGURE 27. The Möbius Band

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 in Figure 28. See the pictures in

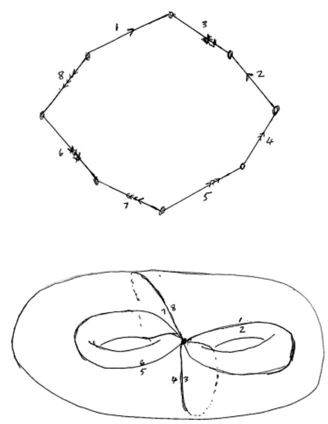

FIGURE 28. A Surface of Genus Two

pp. 300–301 of [7] to see in more detail how the identifications on the boundary of 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 5.2. *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. $\square$

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

Theorem 5.3. *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. $\square$

Corollary 5.4. *Suppose $f : X \to Y$ is a homeomorphism. Then X is connected if and only if Y is connected.*

Example 5.5. 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 3.36, because, as we saw in Example 3.51, 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 5.6. *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. $\square$

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 5.7. *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 5.8. *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. $\square$

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

Definition 5.9. *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 5.10. *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. $\square$

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

Example 5.11. 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 1.26.

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 5.12. *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*

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

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 1.14 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 5.13. *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 5.14. *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*

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

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 4.5 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 5.13 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 5.15. The space $(\mathbb{R}^2, d_{FR})$ of Example 1.13 (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 5.16. A simple application of Lemma 5.13 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 5.13, $\phi \cdot \psi$ 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 5.17. *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 5.4. □

Remark 5.18. 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 5.19. Using the same ideas as in the proof of Theorem 5.17 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 1.3 and 1.5) 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 1.43).

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 5.20. (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 5.2. 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 3.53, $f^{-1}(\{a\}) \cap C \neq \emptyset$. 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 5.21. *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 [8] for an example where the two notions differ.

Example 5.22. 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.

Example 5.23. Connected components can also be used to derive properties of homeomorphisms of spaces. Continuing with the previous example, look again at the subset of the plane in the shape of the letter X. It has 5 distinguished points:

- The point p_0 at the center or the letter X. It is the unique $p \in X$ with the property that $X \setminus \{p\}$ has four connected components.
- The points q_1, q_2, q_3, q_4 the extremities of the four edges emanating from p_0. They are the only $p \in X$ with the property that $X \setminus \{p\}$ is connected.

Thus, if $f : X \to X$ is a homeomorphism, then $f(p_0) = p_0$. Thus p_0 is a *fixed point* of every self-homeomorphism of X. Similarly, the 4 points $q_1, \dots q_4$ must be permuted by any homeomorphism $f : X \to X$. In particular $q_0, \dots q_4$ are *periodic points* of f, meaning $f^4(p) = p$ for $p \in \{q_1, \dots q_4\}$. The *period* of a periodic point p of f is defined to be the smallest $k \in \mathbb{N}$ such that $f^k(p) = p$. The possibilities for the periods of $q_1, \dots, q_4$ are the divisors of 4, namely $1, 2, 4$.

Here are some general properties of connected components:

Theorem 5.24. *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 5.20, 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 5.20: $\bar{C}_x$ is connected, hence $\bar{C}_x \subset C_x$, hence C_x is closed.

□

5.2. **Locally Path Connected Spaces.**

Definition 5.25. *A topological space is called* locally path connected *if it has a basis consisting of path connected open sets.*

Remark 5.26. 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 5.27. 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 5.28.
(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 [8] for more details.

Theorem 5.29. *Suppose X is connected and locally path connected. Then X is path connected.*

Proof. Let $x \in X$. Let $U = \{y \in X : \text{there exists a path } \phi : [0, 1] \to X \text{ from } x \text{ to } y\}$. 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 5.30. The proof of Theorem 5.29 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 5.31. *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 5.32. *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 5.29. □

Corollary 5.33. *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. $\square$

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 5.34. *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 5.3). $\square$

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 5.35. *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 5.34), 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. We need to use the differentiablity of f, in fact, let's use that f is continuously differentiable. We need the following basic lemma on differentiable functions. We state it for $\mathbb{R}^3$, but the same proof works for $\mathbb{R}^n$, any n.

Lemma 5.36. *Let $U \subset \mathbb{R}^3$ be open, let $f : U \to \mathbb{R}$ be of class C^1 (its partial derivatives exist and are continuous on U). Suppose $(x, y, z) \in U$ and let $N = N_{(x,y,z)}$ be a convex nbd of $(0, 0, 0)$ so that $(x, y, z) + N \subset U$. Then there exist functions $\epsilon_x, \epsilon_y, \epsilon_z$ of $(x, y, z, \Delta x, \Delta y, \Delta z)$ defined for all $(\Delta x, \Delta y, \Delta z) \in N$ so that*

$$\begin{aligned} f(x + \Delta x, y + \Delta y, z + \Delta z) - f(x, y, z) &= \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z \\ &+ \epsilon_x \Delta x + \epsilon_y \Delta y + \epsilon_z \Delta z \end{aligned}$$

where the partial derivatives are evaluated at (x, y, z) and $\epsilon_x, \epsilon_y, \epsilon_z \to 0$ as $(\Delta x, \Delta y, \Delta z) \to (0, 0, 0)$.

Proof. Since f is continuously differentiable, the fundamental theorem of calculus gives us

$$f(x+\Delta x, y+\Delta y, z+\Delta z)-f(x,y,z) = \int_0^1 \frac{d}{dt} f(x+t\Delta x, y+t\Delta y, z+t\Delta z)dt$$

Applying the chain rule to the integrand we can rewrite the integral as

$$\int_0^1 \Big(\frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z\Big)dt \tag{27}$$

where the partial derivatives $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$ are all evaluated at the point $(x + t\Delta x, y + t\Delta y, z + t\Delta z)$.

Now each term in this integral can be rewritten as

$$\Big(\int_0^1 \Big(\frac{\partial f}{\partial x}(x,y,z) + \phi(x + t\Delta x, y + t\Delta y, z + t\Delta z)\Big)dt\Big)\Delta x \tag{28}$$

where

$$\phi(x+t\Delta x, y+t\Delta y, z+t\Delta z) = \frac{\partial f}{\partial x}(x+t\Delta x, y+t\Delta y, z+t\Delta z) - \frac{\partial f}{\partial x}(x,y,z).$$

Since $\frac{\partial f}{\partial x}$ is continuous, ϕ is continuous. The continuity of ϕ implies that the functions $\phi(x+t\Delta x, y+t\Delta y, z+t\Delta z)$ of $t \in [0,1]$ converge uniformly to $\phi(x,y,z) = 0$ as $(\Delta x, \Delta y, \Delta z) \to (0,0,0)$. Therefore

$$\int_0^1 \phi(x + t\Delta x, y + t\Delta y, z + t\Delta z)dt \to 0 \text{ as } (\Delta x, \Delta y, \Delta z) \to (0,0,0).$$

Letting $\epsilon_x(x,y,z,\Delta x, \Delta y, \Delta z) = \int_0^1 \phi(x + t\Delta x, y + t\Delta y, z + t\Delta z)dt$, we can rewrite (28) as

$$\frac{\partial f}{\partial x}(x,y,z)\Delta x + \epsilon_x(x,y,z,\Delta x,\Delta y,\Delta z)\Delta x \tag{29}$$

where $\epsilon_x(x,y,z,\Delta x,\Delta y,\Delta z) \to 0$ as $(\Delta x, \Delta y, \Delta z) \to (0,0,0)$.

Reasoning in the same way with the other two terms of (27), we derive the formulas that correspond to (28) and (29). If we rewrite (27) as a sum of terms similar to (29) and go back to the source of (27), we finally arrive at the desired formula

$$\begin{aligned} f(x+\Delta x, y+\Delta y, z+\Delta z) - f(x,y,z) &= \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + \frac{\partial f}{\partial z}\Delta z \\ &+ \epsilon_x \Delta x + \epsilon_y \Delta y + \epsilon_z \Delta z \end{aligned} \tag{30}$$

□

We now resume the proof of Theorem 5.35. Recall we have proved existence and continuity of our function $z = g(x,y)$ satisfying $f(x,y,g(x,y) = 0$. To prove differentiability of g, we evaluate (30) on the graph of g. The left hand side of (30) become

$$f(x+\Delta x, y+\Delta y, g(x+\Delta x, y+\Delta y)) - f(x,y,g(x,y)) = 0$$

since both terms vanish. Therefore (30) becomes

$$0 = (\frac{\partial f}{\partial x} + \epsilon'_x)\Delta x + (\frac{\partial f}{\partial y} + \epsilon'_y)\Delta y + (\frac{\partial f}{\partial z} + \epsilon'_z)\Delta g$$

where $\Delta g = g(x+\Delta x, y+\Delta y) - g(x,y)$ and where the $\epsilon'(x,\Delta x, y, \Delta y) = \epsilon(x, \Delta x, y, \Delta y, g(x,y), \Delta g)$. Since g is continuous, $\Delta g \to 0$ as $(\Delta x, \Delta y) \to (0,0)$, thus the three $\epsilon' \to 0$ as $(\Delta x, \Delta y) \to (0,0)$.

Solving the above equation for Δg we get

$$\Delta g = -\frac{\frac{\partial f}{\partial x} + \epsilon'_x}{\frac{\partial f}{\partial z} + \epsilon'_z}\Delta x - \frac{\frac{\partial f}{\partial y} + \epsilon'_y}{\frac{\partial f}{\partial z} + \epsilon'_z}\Delta y$$

which makes sense since $\frac{\partial f}{\partial z} > c > 0$, so there is no problem in dividing by $\frac{\partial f}{\partial x} + \epsilon'_z$. Moreover this can be re-written as

$$\Delta g = -(\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial z}} + \epsilon''_x)\Delta x - (\frac{\frac{\partial f}{\partial y}}{\frac{\partial f}{\partial z}} + \epsilon''_y)\Delta y$$

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

(31) $$\frac{\partial g}{\partial x} = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial z}}(x,y,g(x,y)) \text{ and } \frac{\partial f}{\partial x} = -\frac{\frac{\partial f}{\partial y}}{\frac{\partial f}{\partial z}}(x,y,g(x,y))$$

from which it is clear that the partial derivatives $\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}$ are continuous, thus g is of class C^1. This procedure can be continued to show that g is C^∞. $\square$

Example 5.37. 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 smooth (or 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 *or simply* charts.
2. *A* smooth surface *(also called* differentiable surface*) if it is a topological surface and the above homeomorphisms 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. Observe that the inverse of $\phi_\alpha \circ \phi_\beta^{-1}$ is $\phi_\beta \circ \phi_\alpha^{-1}$. Thus the requirement that all the transition maps $\phi_\alpha \circ \phi_\beta^{-1}$ be smooth includes, as a consequence, the statement that all the transition maps are smooth and their inverses are also smooth.*

Remark 6.2. Some 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. More generally, given two Euclidean spaces $\mathbb{R}^m, \mathbb{R}^n$, an open set $U \subset \mathbb{R}^m$ and a function $f : U \to \mathbb{R}^n$, we can define what it means for f to be *smooth*. This uses more than just the the topology of $\mathbb{R}^m, \mathbb{R}^n$. It uses the *linear structure* (that is, the vector space structure) in an essential way. It is not clear how to transfer this concept to a more general space. The point of the definition of smooth surface is that it allows us to define a good concept of *smooth function*, as we will see in the Definition 6.3.

(4) Recall that the work *smooth* means *infinitely differentiable*, equivalently, *class C^∞*. We restrict ourselves to this class of functions mostly as a matter of convenience. This way we do not have to count how many derivatives we need in various situations.

Definition 6.3. *Let S be a smooth surface with atlas $\{U_\alpha, \phi_\alpha\}_{\alpha\in A}$, and let $f : S \to \mathbb{R}$ be a function. We say that f is* smooth *if and only if for all $\alpha \in A$, the functions*

$$f \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha) \to \mathbb{R}$$

are smooth.

Remark 6.4. Two observations are in order:

(1) Since $\phi_\alpha(U_\alpha)$ is an open subset of $\mathbb{R}^2$, it makes sense to ask that $f \circ \phi_\alpha$ be smooth.

(2) For this definition of smooth function to make sense, we need to know that it is independent of the chosen charts. In other words, we need to know that when the domains of two charts intersect, they give the same definition of smoothness at points of their intersection.

More precisely, we need to know that if $U_\alpha \cap U_\beta \neq \emptyset$, then $f \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \mathbb{R}$ is smooth if and only if $f \circ \phi_\beta^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \mathbb{R}$ is smooth. This follows from the smoothness of the transition functions:

$$f \circ \phi_\alpha^{-1} = f \circ (\phi_\beta^{-1} \circ \phi_\beta \circ \phi_\alpha^{-1}) = (f \circ \phi_\beta^{-1}) \circ (\phi_\beta \circ \phi_\alpha^{-1})$$

Therefore, since $\phi_\beta \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha \cap U_\beta) \to \phi_\beta(U_\alpha \cap U_\beta)$ is a smooth bijection with smooth inverse $\phi_\alpha \circ \phi_\beta^{-1}$, we see that $f \circ \phi_\alpha^{-1}$ is smooth on $\phi_\alpha(U_\alpha \cap U_\beta)$ if and only if $f \circ \phi_\beta^{-1}$ is smooth on $\phi_\beta(U_\alpha \cap U_\beta)$.

Examples are in order:

Example 6.5. 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 6.6. The implicit function theorem, Theorem 5.35, 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: *at every point of S, the gradient $\nabla f \neq 0$.* Recall that $\nabla f = (\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z})$, so an equivalent formulation of 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 6.7. *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. Suppose, say $\frac{\partial f}{\partial z}(x_0, y_0, z_0) \neq 0$. Then the implicit function theorem gives a neighborhood N_0 of (x_0, y_0, z_0), of the form $N_0 = B((x_0, y_0), \delta) \times (z_0 - \epsilon, z_0 + \epsilon)$, and a smooth function $g : B((x_0, y_0), \delta) \to (z_0 - \epsilon, z_0 + \epsilon)$ so that $S \cap N_0 = \{(x, y, g(x, y)) : (x, y) \in B((x_0, y_0), \delta)\}$. In particular we see that $p^z|_{S\cap N_0} : S \cap N_0 \to B((x_0, y_0), \delta)$ 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 $p^z|_{S\cap N_1}$ where N_1 is a product $N_1 = B((x_1, y_1), \delta_1) \times (z_1 - \epsilon_1, z_1 + \epsilon_1)$. If they intersect, that is, $N_0 \cap N_1 \cap S \neq \emptyset$, then, by the uniqueness of the function g constructed in the proof of Theorem 5.35, $N_0 \cap N_1 \cap S$ is the graph of the same function g over $B((x_0, y_0), \delta) \cap B((x_1, y_1), \delta_1)$. Over this intersection the inverse of p^z is also G, so the transition function is $p^z \circ G = id$ is smooth.

If (x_2, y_2, z_2) is a point where $\frac{\partial f}{\partial z} = 0$, then some other partial derivative does not vanish at this point. Suppose, say $\frac{\partial f}{\partial y}(x_2, y_2, z_2) \neq 0$. Then the implicit function theorem gives us a neighborhood $N_2 = B((x_2, z_2), \delta_2) \times (y_2 - \epsilon_2, y_2 + \epsilon_2)$ and a smooth function $h : B((x_2, z_2), \delta_2) \to (y_2 - \epsilon_2, y_2 + \epsilon_2)$ so that $S \cap N_2 = \{(x, h(x, z), z)) : (x, z) \in B((x_2, z_2), \delta_2)\}$. Therefore $p^y|_{S\cap N_2} : S \cap N_2 \to B((x_2, z_2), \delta_2)$ is a chart, with inverse $H(x, z) = (x, h(x, z), z)$. If this neighborhood $S \cap N_2$ of (x_2, y_2, z_2) intersects the neighborhood $S \cap N_0$ 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 6.8. We now specialize the general principle of Theorem 6.7 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. See Figure 29 for the sets U_z^+, U_y^+ and $U_z^+ \cap U_y^+$.

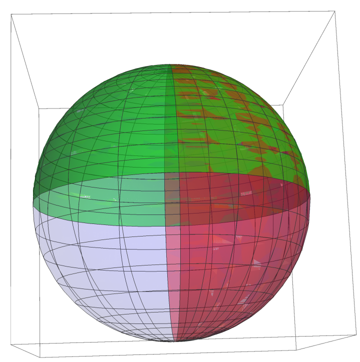

FIGURE 29. The Sets U_z^+, U_y^+ and Their Intersection

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 6.7, 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. Figure 30 shows another view of the sets U_z^+, U_y^+ pulled apart from the sphere. It should be clear that U_z^+ is the graph of a function of (x, y) while U_y^+ is the graph of a function of (x, z).

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 $\phi_z(U_z^+ \cap U_y^+) = \{(x, y) \in D :$

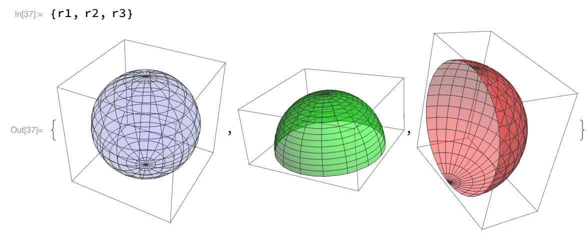

FIGURE 30. Another View of the Sets U_z^+ and U_y^+

$y > 0\}$. 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 6.9. 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 homeomorphic to the plane along any of the coordinate axes. The origin is also a singular point, looking more complicated than the others. See Fiigure 31
(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, see Figure 32
(3) $f(x, y, z) = x^2 + y^2 - z^2 - 1 = 0$. If is instructive to compare this with the last example. The gradient $\nabla f = (2x, 2y, -2z)$ vanishes only at $(0, 0, 0)$ and $f(0, 0, 0) = -1 \neq 0$, so ∇f never vanishes on the set $f(x, y, z) = 0$, which is smooth by the implicit function theorem. $f = 0$ is in fact a hyperboloid of one sheet, see Figure33.
(4) $f(x, y, z) = x^2 - y^2 z$. 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

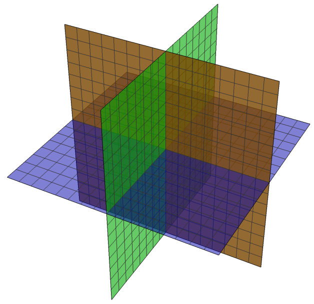

FIGURE 31. The Surface $xyz = 0$

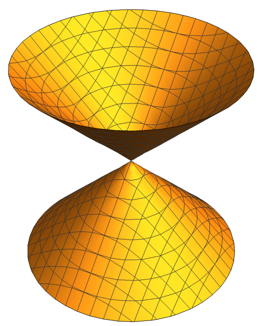

FIGURE 32. The Cone $x^2 + y^2 - z^2 = 0$

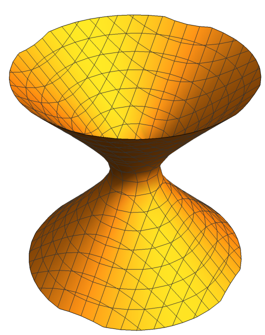

FIGURE 33. The Surface $x^2 + y^2 - z^2 - 1 = 0$

z-axis and the surface shown in Figure 34, called the Whitney umbrella, because the negative z-axis (not shown) would be its handle.

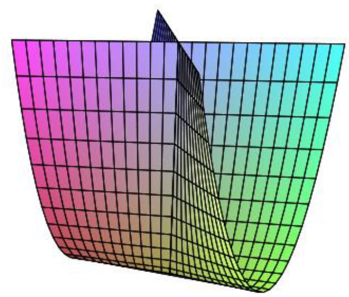

FIGURE 34. The Whitney Umbrella

Example 6.10. So far all the examples (except for Example 6.5) of smooth surfaces that we have considered have been surfaces in $\mathbb{R}^3$ defined by an equation $f(x, y, z) = 0$. But there are many other ways to define smooth surfaces. For instance, the description in Example 4.25 of the torus and Klein bottle can be used to define coordinate charts with smooth transition functions.

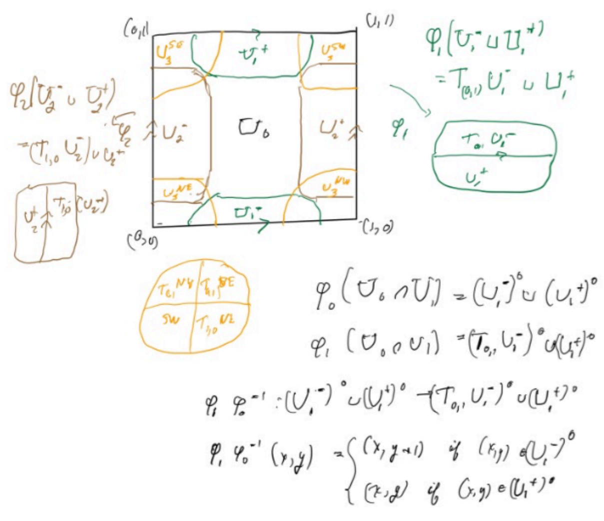

FIGURE 35. Charts for the Torus

For example, for the Torus defined as an identification space of the square $[0, 1] \times [0, 1] \subset \mathbb{R}^2$ by identifications by identifying the sides as in Figure 23 or Figure 35, we can cover the torus by 4 charts as shown in Figure 35, that is

(1) $U_0 =$ interior of the square $[0,1] \times [0,1]$,
(2) $U_1 =$ image in torus of the sets U_1^-, U_1^+
(3) $U_2 =$ image in torus of the sets U_2^-, U_2^+
(4) $U_2 =$ image in torus of a nbd of the corners.

The figure shows in detail the transition function

$$\phi_1 \circ \phi_0^{-1} : \phi_0(U_0 \cap U_1) \to \phi_1(U_0 \cap U_1)$$

This intersection has two connected components, and

$$\phi_1 \circ \phi_0^{-1} : (U_1^-)^0 \cup (U_1^+)^0 \to (T_{(0,1)}(U_1^-)^0) \cup (U_1^+)^0,$$

(where $T_{(0,1)}$ is translation by $(0,1)$, that is, $T_{(0,1)}(x,y) = (x, y+1)$) is given by

$$\phi_1 \circ \phi_0^{-1}(x,y) = \begin{cases} (x, y+1) & \text{if } (x,y) \in (U_1^-)^o, \\ (x,y) & \text{otherwise.} \end{cases}$$

In a similar way one can work out all the other transition functions for this atlas in the torus. Observe one special feature of these transition functions: they are translations on each connected component of their domain. Another way of saying the same thing: the differentials $d(\phi_\alpha \circ \phi_\beta^{-1}$ are always the identity (whenever defined).

In the same way we can work out the transition functions for a similar atlas for the Klein bottle, see figure 36

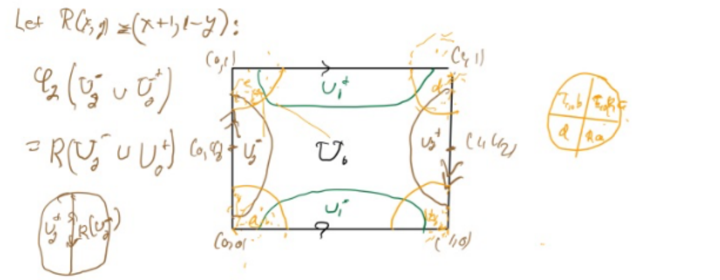

FIGURE 36. Charts for the Klein Bottle

The main difference between the torus and the Klein Bottle is in the chart U_2. While on the torus it is $U_2^+ \cup T_{(1,0)}U_2^-$, for the Klein bottle it is $U_2^+ \cup R(U_2^-)$, where $R(x,y) = (x+1, 1-y)$. This also requires some changes in the chart centered at the corner, see Figure 36. This time the transition functions are either translations of glide reflections on each connected component. The differentials are either the identity or a reflection.

6.1. **Smooth maps involving surfaces.** We have motivated the definition of smooth surface, Definition 6.1, by examining what a smooth function $f : S \to \mathbb{R}$ should be. Since we need a linear structure for the definition of smooth maps, the only reasonable way to define smooth function $f : S \to \mathbb{R}$ is as in Definition 6.3, by requiring that f be smooth in every chart (U_α, ϕ_α) of an atlas. We saw in Remark 6.4 that for this definition to make sense we need the transition functions $\phi_\alpha \circ \phi_\beta^{-1}$ to be smooth.

Now the same principle can be applied to define smoothness of other maps. That is, a map is smooth if and only if it is smooth in all possible way of describing it using local coordinates, in the domain or target or both, as appropriate. Similar arguments to Remark 6.4 show that smoothness of the transition functions guarantees that these definitions make sense.

Definition 6.11. *Let S, T be smooth surfaces with atlas $\{U_\alpha, \phi_\alpha\}_{\alpha\in A}$ and $\{V_\beta, \psi_\beta\}_{\beta\in B}$ respectively.*

(1) *A map $f : S \to \mathbb{R}^n$ is* smooth *if and only if for all $\alpha \in A$, the maps*

$$f \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha) \to \mathbb{R}^n$$

are smooth.

(2) *If $I \subset \mathbb{R}$ is an interval, a continuous map $\gamma : I \to S$ is smooth if and only if I can be covered by a collection $\{I_j\}$ of open subintervals of I (in the subspace topology of $I \subset \mathbb{R}$) so that*
   (a) *For each j there is an $\alpha(j)$ so that $\gamma(I_j) \subset U_{\alpha(j)}$.*
   (b) *For each j, $\phi_{\alpha(j)} \circ \gamma : I_j \to \mathbb{R}^2$ is smooth.*

(3) *If $f : S \to T$ is continuous, and the atlas $\{U_\alpha, \phi_\alpha\}_{\alpha\in A}$ is fine enough so that for each $\alpha \in A$ there is $\beta(\alpha) \in B$ so that $f(U_\alpha) \subset V_{\beta(\alpha)}$, then f is* smooth *if and only if for all $\alpha \in A$ the maps*

$$\psi_{\beta(\alpha)} \circ f \circ \phi_\alpha^{-1} : \phi_\alpha(U_\alpha) \to \mathbb{R}^2$$

are smooth.

Remark 6.12. (1) The first part of the definition is the natural extension of Definition 6.3. It says $f : S \to \mathbb{R}^n$ is smooth if and only if all of its components $f : S \to \mathbb{R}$ are smooth.

(2) For the second part, the continuity of γ imples that all sets $\gamma^{-1}(U_\alpha)$ are open, so each is a union of open intervals (except for the intervals that contain an endpoint of I, if any, but in this case an interval containing and endpoint is open in I)

(3) Similarly, for the third part, by the continuity of f we see that $\{f^{-1}(V_\beta)_{\beta\in B}\}$ is an open cover of S and we can replace the original atlas $\{U_\alpha, \phi_\alpha\}$ by the open sets (the connected components of) the

non-empty intersections $\{U_\alpha \cap f^{-1}(V_\beta)\}$ and the coordinate maps by the restrictions of the ϕ_α to these open sets.

Example 6.13. Here are some examples of smooth maps that we will need:

(1) Let $S \subset \mathbb{R}^3$ be a surface given by an equation $f = 0$ where $\nabla f \neq 0$ at all points of S, as in Theorem 6.7. Let $\iota : S \to \mathbb{R}^3$ be the inclusion. Then ι is smooth.

Proof. As in the proof of Theorem 6.7, S can be covered by charts U_α, ϕ_α where the domain of U_α is the graph G_α of a function $g_\alpha : N_\alpha \to \mathbb{R}$ expressing one variable in terms of the two others: $G_\alpha = (x, y, g_\alpha(x, y)), (x, y) \in N_\alpha$ or $G_\alpha = (x, g_\alpha(x, z), z), (x, z) \in N_\alpha$ or $G_\alpha = (g_\alpha(y, z)), (y, z) \in N_\alpha$ as the case may be. In all cases $\phi_\alpha : G_\alpha \to N_\alpha$ is projection, $\phi_\alpha^{-1} : N_\alpha \to G_\alpha$ is the graph map, and $\iota \circ \phi_\alpha^{-1} = (x, y, g_\alpha(x, y)$ or $(x, g_\alpha(x, z))$ or $(g_\alpha(y, z), y, z)$, as the case may be. In all cases it is smooth.

□

(2) Let $\iota : S \to \mathbb{R}^3$ be as above, let $I \subset \mathbb{R}$ be an interval, and let $\gamma : I \to S$ be continuous. Then $\gamma : I \to S$ is smooth if and only if $\iota \circ \gamma : I \to \mathbb{R}^3$ is smooth.

Proof. Since the composition of smooth maps is smooth, γ smooth $\implies \iota \circ \gamma$ is smooth. Conversely, if $\iota \circ \gamma$ is smooth, following the notation of the last proof, let, for each α, $V_\alpha \subset \mathbb{R}^3$ be an open set, as in the proof of Theorem 6.7 in which $\{f = 0\} \cap V_\alpha = G_\alpha$ and V_α is a product of N_α with an open interval in $\mathbb{R}$. Let $\pi_\alpha : V_\alpha \to N_\alpha$ be the resulting projection. Then, since $\gamma : I \to \mathbb{R}^3$ is continuous, I can be written as a union of subintervals I_j so that $\iota \circ \gamma(I_j) \subset V_{\alpha(j)}$. Since $\gamma(I) \subset S$, $\gamma(I_j) \subset G_{\alpha(j)}$. Therefore $\pi_{\alpha(j)} \circ \iota \circ \gamma = \phi_{\alpha(j)} \circ \gamma$ on I_j. Since the first is smooth, by the assumption that $\iota \circ \gamma$ is smooth, if follows that the second is smooth. Therefore γ is smooth, as desired.

□

6.1.1. **Local connectedness of smooth surfaces**.

Theorem 6.14. *Let S be a smooth surface. Then S is locally piecewise smoothly path connected. In particular, if S is connected, then S is piecewise differentiably path connected.*

Proof. Give S an atlas $\{U_\alpha, \phi_\alpha\}$ in which all sets $\phi_\alpha(U_\alpha)$ are piecewise smoothly path connected, for example, convex. This shows that S is locally piecewise smoothly path connected. Therefore, by the argument of

Theorem 5.29, if S is connected, it is also piecewise smoothly path connected. $\square$

6.2. **Smooth surfaces in $\mathbb{R}^3$ as metric spaces.** In Example 1.14 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 shows that this metric is defined for all *connected* smooth surfaces. First of all, if S is connected, we have just seen in Theorem 6.14, that S is piecewise smoothly path connected

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 1.14:

$$d(x, y) = \inf\{L(\gamma) : \gamma \text{ a piecewise smooth path from } x \text{ to } y\}. \tag{32}$$

This infimum is defined, since x and y can always be connected by a piecewise smooth 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 6.2.2 below.

6.2.1. **Arc Length.** Let $S \subset \mathbb{R}^3$ be a smooth surface and let $\gamma : [0, 1] \to S$ be a piecewise smooth path. Recall (see Example 6.13) this means that the composition $\gamma : [0, 1] \to S \subset \mathbb{R}^3$ is piecewise smooth. Write $\gamma(t) = (x(t), y(t), z(t))$. Then the length of γ, $L(\gamma)$ is defined to be

$$L(\gamma) = \int_0^1 |\gamma'(t)|\, dt = \int_0^1 \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2}\, dt, \tag{33}$$

which we could also write as

$$L(\gamma) = \int_\gamma \sqrt{dx^2 + dy^2 + dz^2} = \int_\gamma ds, \tag{34}$$

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

$$ds^2 = (\mathbf{x}_u \cdot \mathbf{x}_v)du^2 + 2(\mathbf{x}_u \cdot \mathbf{x}_v)dudv + (\mathbf{x}_v \cdot \mathbf{x}_v)dv^2, \tag{35}$$

or as

$$ds^2 = g_{11}du^2 + 2g_{12}dudv + g_{22}dv^2, \tag{36}$$

where the g_{ij} are the coefficients of Equation 35: $g_{11} = \mathbf{x}_u \cdot \mathbf{x}_u$, $g_{12} = \mathbf{x}_u \cdot \mathbf{x}_v$ 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 6.15. 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

$$ds^2 = dr^2 + r^2\ d\theta^2, \tag{37}$$

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 to show 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 6.16. 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

$$ds^2 = d\phi^2 + \sin^2\phi\, d\theta^2. \tag{38}$$

6.2.2. **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

$$\begin{aligned} L(\gamma) = &\int_0^1 \sqrt{x'(t)^2 + y'(t)^2}dt \ge \int_0^1 \sqrt{x'(t)^2}dt \ge \\ &\int_0^1 x'(t)dt = x(1) - x(0) = a, \end{aligned} \tag{39}$$

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 (39) 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 6.15 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

$$\begin{aligned} L(\gamma) = &\int_0^1 \sqrt{r'(t)^2 + r(t)^2\theta'(t)^2}\, dt \ge \int_0^1 \sqrt{r'(t)^2}\, dt \ge \\ &\int_0^1 r'(t)\, dt = r(1) - r(0) = a, \end{aligned} \tag{40}$$

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 6.16. As before, we take a curve $\gamma(t) = (\phi(t), \theta(t))$ with $\phi(0) = 0$ and $\phi(1) = \alpha$ and compute:

$$\begin{aligned} L(\gamma) = &\int_0^1 \sqrt{\phi'(t)^2 + \sin^2\phi(t)\, \theta'(t)^2}\, dt \ge \int_0^1 \sqrt{\phi'(t)^2}\, dt \ge \\ &\int_0^1 \phi'(t)\, dt = \phi(1) - \phi(0) = \alpha, \end{aligned} \tag{41}$$

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 statement of the theorem, by 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 6.17. *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 6.18. If $\mathbf{y} = -\mathbf{x}$, say if we take N and the south pole $\mathbb{N}$, then the computation of Equation 41 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 6.19. *The spherical metric of Example 1.8 is the same metric on* S^2 *as the intrinsic metric of Example 1.14.*

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. □

6.3. **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.

6.3.1. **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

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

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: $\tilde{\gamma}_{st} = \tilde{\gamma}_{ts}$:

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

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

$$\frac{dL}{dt}(0) = V(s) \cdot \gamma'(s)|_0^{L_0} - \int_0^{L_0} V(s) \cdot \gamma''(s)^T\, ds. \tag{42}$$

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

6.3.2. **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 42 vanishes, we get just the second term, which must vanish for all possible variation vector fields V.

Theorem 6.20. $\int_0^{L_0} V(s) \cdot \gamma''(s)^T\, ds = 0$ *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 6.21. 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.

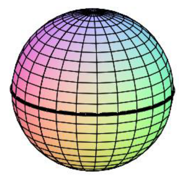

FIGURE 37. Equator and Meridians are Geodesics

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) Let γ be any smooth curve on S parametrized by arc-length. 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 tangent to S and 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 6.22. *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 $(a, b) = (0, 1)$:

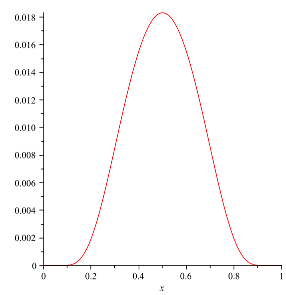

FIGURE 38. Smooth "Bump" Function

□

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 6.7. 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. $\square$

Definition 6.23. *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 6.24. *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 $D\gamma'/Ds$ are commonly used, for example $D^2\gamma/Ds^2$, $D_{\gamma'}\gamma'$, and some others.

6.3.3. **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 6.2.2, 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 $T_P S$ 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

$$\begin{aligned} u'' + \Gamma^1_{11}u'^2 + 2\Gamma^1_{12}u'v' + \Gamma^1_{22}v'^2 = 0 \\ v'' + \Gamma^2_{11}u'^2 + 2\Gamma^2_{12}u'v' + \Gamma^2_{22}v'^2 = 0 \end{aligned} \tag{43}$$

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 43. 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 6.25. *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 43 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], particularly 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

$$\mathbf{u}(rs, p, \mathbf{v}) = \mathbf{u}(s, p, r\mathbf{v}) \text{ for any } r \in \mathbb{R} \tag{44}$$

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 36 satisfy $g_{11}(0) = g_{22}(0) = 1$ and $g_{12}(0) = 0$.

By Theorem 6.25, 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 6.26. *There exists $b \in (0, \infty]$ so that the solution $\mathbf{u}(s, 0, \mathbf{v})$ of the geodesic equation (43) 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 36 and $\mathbf{v} = (v_1, v_2)$.

Using the formula 44, 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 6.26. 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 44 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.*

6.3.4. **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 6.27.
(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 $r \le b$, where b is as in Lemma 6.26, we have the following diagram where the left half is the discussion in local coordinates just finished in subsection 6.3.3, 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 6.3.3.

□

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 6.28. *The map* $\exp_P : B(0, b) \to S$ *defined in (2) of Theorem 6.27 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 6.27), but its restriction to the ball of radius r is a diffeomorphism only for $r < \pi$.

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 6.3.3 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 39 suggests that we should

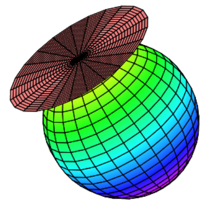

FIGURE 39. Exponential Map for the Sphere

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 6.29. *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 39 also suggests that the curves $r = const$ and $\theta = const$ should be perpendicular to each other. This is indeed the case:

Theorem 6.30. *(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 42. 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 42 reads

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

Since r_0, θ_0 are arbitrary, this means that $\mathbf{x}_\theta \cdot \mathbf{x}_r = 0$ everywhere, as asserted. Recalling formulas 35 and 36, 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. $\square$

Now that we have geodesic polar coordinates, we can repeat the reasoning we used in Equations 40 and 41 in any surface. First, Gauss's Lemma justifies the following terminology:

Definition 6.31. *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 6.32. *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 $T_P S$.*

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 1.14 or Definition 1.26(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 40 or 41. 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

$$\begin{aligned} L(\gamma) &= \int_0^1 \sqrt{r'(t)^2 + g(r(t),\theta(t))^2\,\theta'(t)^2}\,dt \ge \\ &\int_0^1 \sqrt{r'(t)^2}\,dt \ge \int_0^1 r'(t)\,dt \ = r(1) - r(0) = r_1 - r_0. \end{aligned} \tag{45}$$

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 45, but refined to take into account the possibility that $\rho_i < \rho_{i-1}$, gives the more useful inequality

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

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 $L(\gamma_i) \geq |\rho_i - \rho_{i-1}|$, 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 $\rho_{i-1} \leq r \leq \rho_i$, or $\rho_i \leq r \leq \rho_{i-1}$ as the case many be. We thus get

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

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 40 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.

□

Remark 6.33. (1) Observe that Theorem 6.32 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

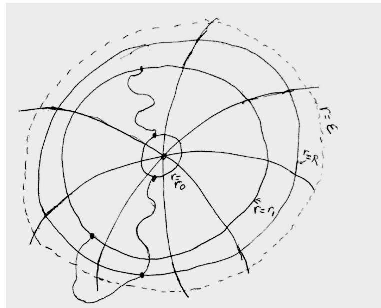

FIGURE 40. Geodesic Rays Minimize Length

any other point $Q \in B_S(P, r)$. This is in marked contrast with the Taxicab metric of Example 1.5 or the Supremum distance of Example 1.6 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 6.32 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.

6.4. **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 6.30). Recall the context: Given a smooth surface $S \subset \mathbb{R}^3$ and a point $P \in S$, there is an $\epsilon > 0$ and a ball$B_T(0,\epsilon)$ about the origin in the tangent space T_PS to S at P so that the map $\mathbf{x} = exp_P$ is a diffeomorphism of $B_T(0,\epsilon)$ to a neighborhood of P in S. If r, θ are polar coordinates in T_PS centered at 0, then the expression of $ds^2 = d\mathbf{x} \cdot d\mathbf{x}$ is

$$ds^2 = dr^2 + g(r,\theta)^2\, d\theta^2 \tag{46}$$

for some function $g(r,\theta)$ defined for $r \geq 0$, positive and smooth for $r > 0$.

First of all, to see that it would be interesting to know more about g, here are two geometric interpretations of this function:

Definition 6.34. *Using the notation just established,*

(1) *The* geodesic circle of radius r_0 centered at P *is, as in Def 6.31, the curve* $C_{r_0} = \{\mathbf{x}(r_0,\theta) \mid 0 \leq \theta \leq 2\pi\}$.

(2) *The* geodesic disk of radius r_0 centered at P *is the surface* $D_{r_0} = \{\mathbf{x}(r,\theta) \mid 0 \leq r \leq r_0, 0 \leq \theta \leq 2\pi\}$

Theorem 6.35. (1) *The length (circumference) of* C_{r_0} *is* $\int_0^{2\pi} g(r_0,\theta)\, d\theta$.

(2) *The area of* D_{r_0} *is* $\int_0^{2\pi}\int_0^{r_0} g(r,\theta)\, dr d\theta$

Proof. Since the length of the tangent vector to C_{r_0} at the point (r_o, θ) is $g(r_0, \theta)$, the first statement is clear. For the second statement, we would have to discuss area, which we are not going to do here. It can be done in a manner totally parallel to the treatment for the plane and the sphere. $\square$

We want to see what restrictions, if any, there are on the function g, and what other information we can extract from this function. Recall that we know two examples, the Euclidean plane, Example 6.15, and the unit sphere, Example 6.16, given by the formulas (37) and (38):

$$dr^2 + r^2 d\theta^2 \quad \text{and} \quad dr^2 + \sin^2 r\, d\theta^2.$$

In the second formula, we changed the notation in (38) from ϕ to r since ϕ is the same as geodesic distance from the north pole in S^2. So we get two examples of a function g:

$$g(r,\theta) = r \quad \text{and} \quad g(r,\theta) = \sin r = r - \frac{r^3}{6} + \dots \tag{47}$$

These formulas clearly illustrate the following restriction on g:

Theorem 6.36. *The function $g(r,\theta)$ of (46) satisfies*

$$g(r,\theta) = r + cr^3 + O(r^4)$$

for some constant c and for $r > 0$, where $O(r^4)$ means a function $f(r,\theta)$ satisfying $|f(r,\theta)| \leq Cr^4$ for some absolute constant C, independent of θ.

Proof. We will see that this restriction on $g(r,\theta)$ is just a consequence of the fact that the formula (46) defines an object that is smooth at the origin. In other words, we could use rectangular coordinates u, v on $T_P S$, related in the usual way: $u = r\cos\theta, v = r\sin\theta$. Under this correspondence we have

$$dr^2 + g(r,\theta)^2\, d\theta^2 \;=\; g_{11}\, du^2 + 2g_{12}\, du dv + g_{22}\, dv^2 \tag{48}$$

where g_{ij} are smooth functions of u, v.

Recall that a function $f(u,v)$ defined on $\mathbb{R}^2 \setminus \{0\}$ is said to be *homogeneous of degree* d if $f(tu,tv) = t^d f(u,v)$ holds for all $t > 0$ and all $(u,v) \in \mathbb{R}^2 \setminus \{0\}$. If $d \geq 0$, an example of a homogeneous function of degree d is a homogeneous polynomial in u, v of degree d. In this case the function extends to $\mathbb{R}^2$.

Since the coefficients $g_{ij}(u,v)$ are smooth functions in a neighborhood of $0 \in \mathbb{R}^2$, the Taylor expansion of g_{ij} at $(0,0)$ gives a sequence $g_{ij,d}$ of homogeneous polynomials of degree d, $d = 0, 1, 2, \dots$ so that for any fixed

d_0,

$$g_{ij}(u,v) = \sum_{d=0}^{d_0} g_{ij,d}(u,v) + O(r^{d_0+1}),$$

where $r = (u^2+v^2)^{1/2}$. Consequently the right-hand side of (48) can be written as a sum

$$\sum_{d=0}^{d_0}(g_{11,d}du^2 + 2g_{12,d}dudv + g_{22,d}dv^2) + O(r^{d_0+1}) \tag{49}$$

To get a similar expansion of the left-hand side of (48), observe first that the Taylor expansion of $g(r,\theta)$ with respect to r, centered at $(0,\theta)$, with coefficients functions of θ, gives a sequence of smooth functions $c_i(\theta)$, $i = 0,1,2,\dots$, periodic of period 2π, so that for any integer $k \geq 0$ there is an expansion

$$g(r,\theta) = c_0(\theta) + c_1(\theta)r + c_2(\theta)r^2 + \cdots + c_k(\theta)r^k + O(r^{k+1}), \tag{50}$$

We will only need the case $k=3$, that is

$$g(r,\theta) = c_0(\theta) + c_1(\theta)r + c_2(\theta)r^2 + c_3(\theta)r^3 + O(r^4), \tag{51}$$

But first we will assume that (50) holds for $k = 4$ (error $O(r^5)$. Then $g(r,\theta)^2$ has expansion, up to $O(r^5)$ as follows:

$$c_0^2+2c_0c_1r+(2c_0c_2+c_1^2)r^2+(2c_0c_3+2c_1c_2)r^3+(2c_0c_4+2c_1c_3+c_2^2)r^4 \tag{52}$$

where we have written simply c_i for the functions $c_i(\theta)$. We will first show that $c_0 = 0$. From this it follows that we only need (51) to get (52) with error $O(r^5)$.

Take the expression (46) and write it as a sum of terms in $r, \theta, dr, d\theta$, that, under the change of variables $u = r\cos\theta, v = r\sin\theta$ correspond to linear combinations of $du^2, dudv$ and dv^2 with coefficients homogeneous functions of a fixed degree d in u, v, as in (49). If all the coefficients in (50) were arbitrary, we would find some homogeneous functions in u, v that are *not* polynomials. This will give the restrictions on the coefficients $c_i(\theta)$.

To see the correspondence, note that $r = \sqrt{u^2+v^2}$ is homogeneous of degree 1. The function $\theta(u,v)$ is not real valued, it has values in $\mathbb{R}/\mathbb{Z}$, but as such satisfies $\theta(tu,tv) = \theta(u,v)$, so formally this is the same as homogeneos of degree 0. Any real valued function of θ is homogeneous of degree 0.

Solving the equations

$$\begin{aligned} du &= \cos\theta \, dr - r\sin\theta \, d\theta \\ dv &= \sin\theta \, dr - r\cos\theta \, d\theta. \end{aligned}$$

for $dr, d\theta$ gives

$$dr = (udu + vdv)/r = \frac{u}{\sqrt{u^2+v^2}}du + \frac{v}{\sqrt{u^2+v^2}}dv \tag{53}$$
$$d\theta = (udv - vdu)/r^2 = \frac{u}{u^2+v^2}dv - \frac{v}{u^2+v^2}$$

Thus dr has coefficients homogeneous of degree 0 and $d\theta$ of degree -1. So neither form is smooth at the origin, since the homogeneous forms are not polynomial. This means that dr^2 and $d\theta^2$ have coefficients homogeneous of degree $0, -2$ respectively. Explicitly:

$$dr^2 = \frac{u^2du_2^2uvdudv + v^2dv^2}{u^2+v^2} \tag{54}$$
$$d\theta^2 = \frac{v^2du^2 - 2uvdudv + u^2dv^2}{(u^2+v^2)^2}$$

Now use all this information in the expansion obtained by using the expansion (52) in the formula (46)

Degree	*terms*	*Conclusion*
-2	$c_0^2 d\theta^2$	$c_0 = 0$
-1	$2c_0c_1rd\theta^2 = 0$	nothing
0	$dr^2 + (c_1^2r^2)d\theta^2$	$c_1 = \pm 1$
1	$c_2r^3d\theta^2$	$c_2 = 0$
2	$c_3r^4d\theta^2$	c_3 constant, independent of θ
3	$c_4r^5d\theta^2$	$c_4(\theta) = a\cos\theta + b\sin\theta$

In more detail:

(1) There can be no term of degree -2, so $c_0^2 = 0$
(2) For degree -1 we get a coefficient $2c_0c_1$, but already know $c_0 = 0$, no new information.
(3) For degree 0 actually get $dr^2 + (2c_0c_2 + c_1^2)r^2d\theta^2$, but already know $c_0 = 0$. The term of degree 0 is $dr^2 + c_1^2r^2d\theta^2$ which is

$$\frac{(u^2 + c_1^2 2v^2)du^2 + 2(1-c_1^2)uvdudv + (v^2 + c_1^2u^2)dv^2}{u^2+v^2}$$

which is a polynomial if and only if $c_1^2 = 1$, in which case it reduces to $du^2 + dv^2$. So $c_1 = \pm 1$, and we should take $c_1 = 1$ since we're assuming $g(r,\theta) > 0$.

(4) For degree 1, get $c_2 r^3 d\theta^2 = c_2(\theta)(udv - vdu)^2/r$. Take any of its coefficients, say $c_2(\theta)u^2/r = c_2(\theta)r\cos^2\theta$. It homogeneous of degree 1. Since it vanishes on the line $\theta = \pi/2$, if it were a linear function, the only possibility would be $au = ar\cos\theta$ or some constant a. Thus $ar\cos\theta = c_2(\theta)r\cos^2\theta$, that is $c_2(\theta) = a/\cos\theta$ which is not a smooth function of θ (blows up at $\theta = \pm\pi/s$). Thus we must have $c_2(\theta) \equiv 0$.

(5) For degree 2 get $c_3 r^4 d\theta = c_3(theta) r^4 d\theta^2 = c_3(\theta)(udv - vdu)^2$. The coefficients can be polynomials if and only if c_3 is a constant, independent of θ.

(6) *Degree* 3 *is beyond our expansion (52)* , but we include it here since the pattern is now clear. We get $c_(\theta)r(udvd - vdu)^2$, homogeneous of degree 1. It is smooth at the origin if and only if it is a linear function of u, v. This happens if and only if $c_4(\theta) = a\cos\theta + b\sin\theta$ for some constants a, b. *This is the first time that this method allows for a coefficient* $c_i(\theta)$ *that is not independent of* θ.

□

Definition 6.37. *The Gaussian curvature of* S *at* P *is the number* $K(P) = -6c$*, with* c *as in the theorem.*

Remark 6.38. 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 6.5 below for the meaning of intrinsic.

Example 6.39. (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 6.16, with $\phi = r$ and $ds^2 = dr^2 + \sin^2 r\, d\theta^2$, thus $g(r, \theta) = \sin r = r - r^3/6 + \ldots$, 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 6.40. Theorem 6.35 gives a nice interpretation of the Gaussian curvature $K(P)$. Recall the meaning of geodesic circle and geodesic disc from Definition 6.34.

$$
\begin{aligned}
L(C_r) &= 2\pi r - \frac{K(P)\pi}{3} r^3 + O(r^4) \\
A(D_r) &= \pi r^2 - \frac{K(P)\pi}{12} r^4 + O(r^5).
\end{aligned}
\tag{55}
$$

Thus $K(P)$ measures the deviation of the formulas for circumference and area of a circle from the usual Euclidean formulas. For example, for $\mathbb{R}^2$ we get $L(C_r) = 2\pi r$ while for S^2 we get $L(C_r) = 2\pi \sin r = 2\pi r - \pi r^3/3 + \ldots$, thus geodesic circles on the sphere are shorter than their counterparts in $\mathbb{R}^2$, as suggested by Figure 39.

6.5. **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

$$\mathbf{x}(u, v) = (\cos u, \sin u, v). \tag{56}$$

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. In other words, γ descends to a map, still denoted γ, defined on the circle: $\gamma : \mathbb{R}/(2\pi\mathbb{Z}) \to \mathbb{R}^2$, and this map is injective.

Write $\gamma(u) = (x(u), y(u))$. Define a surface $C_\gamma \subset \mathbb{R}^3$ by the formula

$$\mathbf{y}(u, v) = (x(u), y(u), v), \tag{57}$$

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 curves $\gamma : \mathbb{R} \to \mathbb{R}^2$ satisfying our requirements of periodicity and injectivity of the map $\mathbb{R}/(2\pi\mathbb{Z}) \to \mathbb{R}^2$, it follows that there are infinitely many surfaces isometric to the cylinder C. Moreover, since the simple closed curves γ can be continuously deformed to each other, the same can be done with the resulting surfaces..

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] J. Dugundji, *Topology, Allyn and Bacon, 1966.*
[5] K. F. Gauss, General Investigations of Curved Surfaces*, Dover, 2005.*
[6] A. Hatcher, Notes on Introductory Point-Set Topology*, available at http://www.math.cornell.edu/ hatcher/Top/TopNotes.pdf*
[7] D. Hilbert and S. Cohn-Vossen, Geometry and the Imagination*, Chelsea Pub. Co.*
[8] B. Mendelson, Introduction to Topology*, Dover Publications, 1990.*
[9] J. R. Munkres, Topology (second edition)*, Prentice Hall, 2000.*
[10] J. Oprea, Differential Geometry and its Applications*, Prentice-Hall 1997.*
[11] A. Pressley, Elementary Differential Geometry*, Springer, 2002.*
[12] J. Stillwell, Geometry of Surfaces*, Universitext, Springer-Verlag, 1992.*
[13] D. J. Struik, Lectures on Classical Differential Geometry*, Second Edition, Dover, 1988.*