GEOMETRY AND TOPOLOGY MATH2049
2023-2024

DANIEL KASPROWSKI

Abstract. These notes are based on the MATH2049 notes from 2022-2023 by Jacek Brodzki with some minor edits.

Contents

Date: January 2024.

1. Euclidean geometry

1.1. **Metrics and isometries.**

Understanding distance is central to our intuition of the physical universe in which we live. When we say that something is on a 'human scale' it suggests that the object we are trying to describe is of a size comparable to that of a human being. The way we measure distances determines our perception of shape, and if we were not able to judge accurately distances between objects, it would be difficult to survive a single day.

There are many different possible ways to measure distances, and they all share the simple properties given in the following definition. This is one of the central notions of this course.

Definition 1.1. Let X be a non-empty set. A *metric* (or a distance function) on X is a map $d : X \times X \to \mathbb{R}$ which satisfies the following properties:

(1) d is *positive definite*: for all $x, y \in X$ $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.
(2) d is *symmetric*: for every $x, y \in X$, $d(x, y) = d(y, x)$.
(3) d satisfies the *triangle inequality*: for every $x, y, z \in X$

$$d(x, z) \leq d(x, y) + d(y, z).$$

Definition 1.2. A *metric space* is a set X equipped with a metric d. We will write (X, d), or (X, d_X) if there is more than one metric space, to stress which metric provides the distance on the set X.

In our study of geometry, we will be interested in two main themes: finding interesting examples of metric spaces, and the study of maps that preserve distances. The following definition makes this precise.

Definition 1.3. Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f : X \to Y$ is said to be an *isometry* (or *distance-preserving*) if $d_Y(f(x), f(y)) = d_X(x, y)$ for all $x, y \in X$.

Exercise 1.4. Prove that any isometry is injective.

Exercise 1.5. Prove that the composition of any two isometries is an isometry.

1.2. **Euclidean space.**

We will discuss many examples of metric spaces in this course, and we will begin with the Euclidean space. Many of the facts presented in this section are well-known to you, and we include them to introduce the necessary notation.

We recall that for every $n \geq 1$ the space $\mathbb{R}^n$ consists of points $\boldsymbol{x}$ which are n-tuples of real numbers. Hence $\boldsymbol{x} \in \mathbb{R}^n$ if and only if

$$\boldsymbol{x} = (x_1, \ldots, x_n),$$

where $x_i \in \mathbb{R}$ for all $i = 1, \ldots, n$.

This space is equipped with a norm (called the *Euclidean norm*) which for every $\boldsymbol{x} \in \mathbb{R}^n$ is defined by

$$||\boldsymbol{x}|| = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}. \tag{1.6}$$

We can interpret $||\boldsymbol{x}||$ as the length of the position vector of the point $\boldsymbol{x} \in \mathbb{R}^n$.

The space $\mathbb{R}^n$ is also equipped with the *scalar product* (also called the *dot product* or the *inner product*), which, for any two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, is given by

$$\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i.$$

This inner product is symmetric:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}.$$

We say that two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are *orthogonal* if and only if $\boldsymbol{x} \cdot \boldsymbol{y} = 0$.

Notation 1.7. Please note that the notation $\boldsymbol{x} \cdot \boldsymbol{y}$ is used to denote the scalar product in spaces like $\mathbb{R}^n$. We will also use the notation $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ as this is convenient in certain contexts. Note also that the scalar product of two vectors in $\mathbb{R}^n$ is a real number.

There is another convenient way to represent the dot product via matrix multiplication. If we treat a vector $\boldsymbol{x} \in \mathbb{R}^n$ as a column vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix},$$

that is an $n \times 1$ matrix, then its transpose is

$$\boldsymbol{x}^T = (x_1, x_2, \ldots, x_n)_{1 \times n}.$$

Then the scalar product of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ can be written as the product of two matrices:

$$\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{y}.$$

This notation is useful for manipulating formulae involving matrices. For example, if A is an $n \times n$ matrix, and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ then

$$(A\boldsymbol{x}) \cdot \boldsymbol{y} = (A\boldsymbol{x})^T \boldsymbol{y} = \boldsymbol{x}^T A^T \boldsymbol{y}.$$

By the associativity of matrix multiplication, the expression on the right can be written as

$$\boldsymbol{x}^T A^T \boldsymbol{y} = \boldsymbol{x}^T (A^T \boldsymbol{y}) = \boldsymbol{x} \cdot (A^T \boldsymbol{y})$$

and we conclude that

$$(A\boldsymbol{x}) \cdot \boldsymbol{y} = \boldsymbol{x} \cdot (A^T \boldsymbol{y}), \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n. \tag{1.8}$$

There is a direct relation between the Euclidean norm and the Euclidean inner product which we will use in what follows:

$$||\boldsymbol{x}|| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}.$$

Definition 1.9. The Euclidean metric on $\mathbb{R}^n$ is a distance function defined for all $\boldsymbol{x}$, $\boldsymbol{y}$ by

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}|| = \left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{\frac{1}{2}}.$$

Using the inner product, this definition can be stated as follows:

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})}.$$

Proposition 1.10. *The Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y})$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is a metric in the sense of Definition 1.1.*

Proof. To check the first axiom, we note that since $|x_i - y_i|^2 \geq 0$ for all $i = 1, \ldots, n$, the sum is also non-negative.

Assume that $d(\boldsymbol{x}, \boldsymbol{y}) = 0$. This is equivalent to

$$\sum_{i=1}^{n} |x_i - y_i|^2 = 0.$$

Since all the terms in this sum are non-negative, this can happen if and only if $|x_i - y_i|^2 = 0$ for all i, which in turn is equivalent to $|x_i - y_i| = 0$. Hence $x_i = y_i$ for $i = 1, \ldots, n$, and $\boldsymbol{x} = \boldsymbol{y}$.

To prove symmetry, we note that $|x_i - y_i| = |y_i - x_i|$, and so

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

The triangle inequality requires a bit more work. Take $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathbb{R}^n$ and let $\boldsymbol{a} = \boldsymbol{x} - \boldsymbol{y}$, $\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{z}$. In this notation, the triangle inequality takes the form:

$$||\boldsymbol{a} + \boldsymbol{b}|| \leq ||\boldsymbol{a}|| + ||\boldsymbol{b}||.$$

In coordinates, this is equivalent to:

$$\left(\sum (a_i + b_i)^2\right)^{\frac{1}{2}} \leq \left(\sum a_i^2\right)^{\frac{1}{2}} + \left(\sum b_i^2\right)^{\frac{1}{2}},$$

with all sums for $i = 1, \ldots, n$. Since all three sums are non-negative, this is equivalent on squaring both sides to

$$\sum a_i^2 + 2\sum a_i b_i + \sum b_i^2 \leq \sum a_i^2 + 2\left(\sum a_i^2\right)^{\frac{1}{2}} \left(\sum b_i^2\right)^{\frac{1}{2}} + \sum b_i^2.$$

To show that this inequality holds, it is sufficient to prove that for all $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{R}^n$ we have

$$\left|\sum a_i b_i\right| \leq \left(\sum a_i^2\right)^{\frac{1}{2}} \left(\sum b_i^2\right)^{\frac{1}{2}}.$$

This is an important result known as the *Cauchy-Schwarz* inequality.

To prove it, consider the quadratic polynomial $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \sum (ta_i + b_i)^2 = t^2 \sum a_i^2 + 2t \sum a_i b_i + \sum b_i^2.$$

Recall that the discriminant of the quadratic polynomial $ax^2 + bx + c$ is $b^2 - 4ac$ and that for $a \neq 0$, this discriminant is zero if and only if the polynomial has a double root. It is positive if the polynomial has two distinct real roots, and negative if it has two distinct complex conjugate roots.

Clearly $f(t) \geq 0$ for all t, so f has at most one real root and hence the discriminant $(2\sum a_i b_i)^2 - 4\sum a_i^2 \sum b_i^2$ is non-positive. Equivalently,

$$\left(2\sum a_i b_i\right)^2 \leq 4\sum a_i^2 \sum b_i^2.$$

Taking roots, this gives the proof of the Cauchy-Schwarz inequality. The triangle inequality for the Euclidean norm now follows. □

Definition 1.11. The n-dimensional Euclidean space $\mathbb{E}^n$ is the space $\mathbb{R}^n$ equipped with the Euclidean metric.

1.3. **Euclidean isometries.**

In this course, we will introduce many examples of metric spaces which may not be known to you, but to get used to the new way of thinking about measuring distance it makes sense to investigate in detail the situation that you know quite well. In this section we will study the geometry of the Euclidean space, and will derive a description of Euclidean isometries. By definition, a *Euclidean isometry* is a map $\phi : \mathbb{E}^n \to \mathbb{E}^n$ which preserves the Euclidean distance. In other words, a Euclidean isometry is an isometry of the Euclidean space.

First recall the notion of *orthogonal matrix*, which you know from Linear Algebra. An $n \times n$ matrix A is orthogonal if any only if $A^T A = I_n$, where I_n is the $n \times n$ identity matrix. It follows that orthogonal matrices are invertible, and $A^{-1} = A^T$. The set of all $n \times n$ orthogonal matrices forms a group (under the usual matrix product), called the *orthogonal group* $O(n)$.

Exercise 1.12. (1) Show that $\det A = \pm 1$ for any $n \times n$ orthogonal matrix A.
(2) For any $n \in \mathbb{N}$ give an example of an $n \times n$ matrix B such that $\det(B) = 1$ but B is not orthogonal.

Example 1.13. As we will see later every Euclidean isometry can be obtained using two 'building blocks': translations and orthogonal transformations.

- Given any vector $\boldsymbol{t} \in \mathbb{E}^n$, the corresponding *translation* is the function $f_{\boldsymbol{t}} : \mathbb{E}^n \to \mathbb{E}^n$ defined by $f_{\boldsymbol{t}}(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{t}$, for all $\boldsymbol{x} \in \mathbb{E}^n$.
- If $A \in O(n)$ is an orthogonal matrix, the corresponding *orthogonal transformation* is the function $f_A : \mathbb{E}^n \to \mathbb{E}^n$ defined by $f_A(\boldsymbol{x}) = A\boldsymbol{x}$, for all $\boldsymbol{x} \in \mathbb{E}^n$.

Example 1.14. In $\mathbb{R}^2$, orthogonal transformations are rotations about the origin $\mathbf{0}$ or reflections in a line through $\mathbf{0}$.

A rotation through the angle θ in the anti-clockwise direction transforms the basis unit vector $\boldsymbol{e}_1 = (1, 0)$ to $\boldsymbol{e}_1' = (\cos\theta, \sin\theta)$. The other standard basis vector $\boldsymbol{e}_2 = (0, 1)$ is transformed to $\boldsymbol{e}_2' = (-\sin\theta, \cos\theta)$. The vectors $\boldsymbol{e}_1'$ and $\boldsymbol{e}_2'$ are still orthonormal. The corresponding 2×2

$$A_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \tag{1.15}$$

is clearly orthogonal and has determinant 1. It can be proved (and is left as an exercise) that any orthogonal 2×2 matrix A with $\det A = 1$ is of the form $A = A_\theta$ for some θ.

Now let's discuss reflections. The reflection in the first coordinate axis is given by

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ -x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The matrix

$$B_0 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

is orthogonal, with $\det B_0 = -1$.

More generally, the reflection in an axis through $\mathbf{0}$ making the angle θ with the positive x-axis is given by $\boldsymbol{x} \mapsto B_\theta \boldsymbol{x}$, where

$$B_\theta = \begin{pmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{pmatrix}. \tag{1.16}$$

It is easy to see that B_θ is orthogonal and has determinant -1. Moreover, note that $B_\theta = A_\theta B_0 A_{-\theta}$, or, equivalently, $B_\theta A_\theta = A_\theta B_0$ (geometrically, this means that the reflection in any axis L can be obtained by first rotating L to the x-axis, then reflecting in the x-axis, and then rotating the x-axis back to L).

Conversely, one can show that any orthogonal 2×2 matrix B such that $\det B = -1$ is of the form $B = B_\theta$ for some θ.

Proposition 1.17. *The linear transformation $f_A : \mathbb{E}^n \to \mathbb{E}^n$, given by an orthogonal matrix $A \in O(n)$, preserves the scalar product and the norm on $\mathbb{E}^n$. This means that for all $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{E}^n$*

$$(A\boldsymbol{x}) \cdot (A\boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y} \text{ and } ||A\boldsymbol{x}|| = ||\boldsymbol{x}||.$$

Proof. For all $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{E}^n$, using the associativity of matrix multiplication and the fact that $A^TA = I_n$, we have

$$\begin{aligned}(A\boldsymbol{x}) \cdot (A\boldsymbol{y}) &= (A\boldsymbol{x})^T(A\boldsymbol{y}) \\ &= (\boldsymbol{x}^TA^T)(A\boldsymbol{y}) \\ &= \boldsymbol{x}^T(A^TA)\boldsymbol{y} \\ &= \boldsymbol{x}^TI_n\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y} \\ &= \boldsymbol{x} \cdot \boldsymbol{y}.\end{aligned}$$

The second part is now clear, as

$$||A\boldsymbol{x}|| = \sqrt{(A\boldsymbol{x}) \cdot (A\boldsymbol{x})} = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}} = ||\boldsymbol{x}||,$$

for any $\boldsymbol{x} \in \mathbb{E}^n$. □

From this we derive our first example of a Euclidean isometry.

Theorem 1.18. *Let A be an $n \times n$ orthogonal matrix and let $f_A : \mathbb{E}^n \to \mathbb{E}^n$ be the linear transformation given by*

$$f_A(\boldsymbol{x}) = A\boldsymbol{x}$$

for all $\boldsymbol{x} \in \mathbb{E}^n$. Then f_A is a Euclidean isometry.

Proof. For every $\boldsymbol{x}, \boldsymbol{y}$ in $\mathbb{E}^n$ we have

$$d(f_A(x), f_A(y)) = ||A\boldsymbol{x} - A\boldsymbol{y}|| = ||A(\boldsymbol{x} - \boldsymbol{y})|| = ||\boldsymbol{x} - \boldsymbol{y}|| = d(\boldsymbol{x}, \boldsymbol{y}).$$

□

This is an example of a more general transformation, which we will now study.

Definition 1.19. A *Euclidean transformation* of the space $\mathbb{E}^n$ is a map $f : \mathbb{E}^n \to \mathbb{E}^n$ of the form

$$f = f_{A,\boldsymbol{t}} : \boldsymbol{x} \mapsto A\boldsymbol{x} + \boldsymbol{t}$$

where A is an $n \times n$ orthogonal matrix and $\boldsymbol{t} \in \mathbb{E}^n$ is any vector. This is a composition of an orthogonal transformation $\boldsymbol{x} \mapsto A\boldsymbol{x}$ with a translation by $\boldsymbol{t}$: $\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{t}$.

Exercise 1.20. Euclidean transformations of the space $\mathbb{E}^n$ form a group under composition, called the *Euclidean group*, which is denoted $E(n)$.

Example 1.21. We have already seen simple examples of Euclidean transformations of $\mathbb{E}^2$: rotations and reflections. The more general Euclidean transformations expand our available transformations quite considerably.

First, we have seen that rotations through θ about the origin $\mathbf{0}$ are given by the matrix A_θ, defined in (1.15).

Now, the *rotation by θ about a point $\boldsymbol{c} \in \mathbb{R}^2$* can be constructed as follows. For a given $\boldsymbol{x} \in \mathbb{E}^2$, translate it so that the centre of rotation becomes $\mathbf{0}$, rotate through θ using the matrix A_θ, translate back. The translation from $\boldsymbol{c}$ to $\mathbf{0}$ is done using the vector $-\boldsymbol{c}$, which sends a general $\boldsymbol{x} \in \mathbb{R}^2$ to $\boldsymbol{x} - \boldsymbol{c}$ and the point $\boldsymbol{c}$ is mapped to $\mathbf{0}$. Thus the whole sequence of operations can be described using the formula

$$\boldsymbol{x} \mapsto A_\theta(\boldsymbol{x} - \boldsymbol{c}) + \boldsymbol{c} = A_\theta\boldsymbol{x} + \boldsymbol{t}$$

where $\boldsymbol{t} = \boldsymbol{c} - A_\theta\boldsymbol{c}$.

Secondly, if L is any line in $\mathbb{R}^2$ we can define the reflection in L as follows. If L passes through the origin, we simply use the definition given in Example 1.14. Otherwise, choose any point $\boldsymbol{p} \in L$ and let $\boldsymbol{b} \in \mathbb{R}^2$ be its position vector. Let $\theta \in [0, \pi)$ be the angle between the positive x-axis and L, measured anti-clockwise

($\theta = 0$ is L is parallel to the x-axis), and let $B_\theta \in O(2)$ be the matrix defined in (1.16). We define the *reflection* $\pi_L : \mathbb{E}^2 \to \mathbb{E}^2$, in L, by the following formula:

$$\pi_L(\boldsymbol{x}) = B_\theta(\boldsymbol{x} - \boldsymbol{b}) + \boldsymbol{b} = B_\theta \boldsymbol{x} + \boldsymbol{t}, \text{ for all } \boldsymbol{x} \in \mathbb{E}^2,$$

where $\boldsymbol{t} = \boldsymbol{b} - B_\theta \boldsymbol{b} \in \mathbb{R}^2$. One can show that this definition is independent of the choice of the point $\boldsymbol{p} \in L$.

The last type of an isometry of $\mathbb{E}^2$, which we have not yet considered, are glide reflections. A *glide reflection* of $\mathbb{E}^2$ is defined as the composition of a reflection in a line L with a translation by some non-zero vector $\boldsymbol{t}$, parallel to L. Thus glide reflections of $\mathbb{E}^2$ can be given by the formula

$$\boldsymbol{x} \mapsto B_\theta \boldsymbol{x} + \boldsymbol{t},$$

where $\theta \in [0, \pi)$ and the angle between the positive x-axis and the vector $\boldsymbol{t} \in \mathbb{R}^2 \backslash \{\boldsymbol{0}\}$ is either θ or $\pi + \theta$.

Remark 1.22. There is a classification theorem for the isometries of the Euclidean plane $\mathbb{E}^2$ stating that any such isometry is either a translation, a rotation, a reflection or a glide reflection.

Theorem 1.23. *Every Euclidean transformation $f_{A,\boldsymbol{t}} : \mathbb{E}^n \to \mathbb{E}^n$ is a Euclidean isometry.*

Proof. We have already checked in Theorem 1.18 that $f_{A,\boldsymbol{0}} = f_A$ is an isometry. Let us prove that translations are Euclidean isometries. For any $\boldsymbol{t} \in \mathbb{E}^n$ we have that

$$\begin{aligned} d(f_{\boldsymbol{t}}(\boldsymbol{x}), f_{\boldsymbol{t}}(\boldsymbol{y})) & \\ &= d(\boldsymbol{x} + \boldsymbol{t}, \boldsymbol{y} + \boldsymbol{t}) = ||(\boldsymbol{x} + \boldsymbol{t}) - (\boldsymbol{y} + \boldsymbol{t})|| \\ &= ||\boldsymbol{x} - \boldsymbol{y}|| = d(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

As the composition of two isometries, $f_{A,\boldsymbol{t}} = f_{\boldsymbol{t}} \circ f_A$ is also an isometry (see Exercise 1.5). □

It is quite remarkable that the converse of this statement is also true, as this gives a complete description of Euclidean isometries. As a preliminary step, we prove the following.

Theorem 1.24. *Let $g : \mathbb{E}^n \to \mathbb{E}^n$ be a Euclidean isometry fixing the origin, so that $g(\boldsymbol{0}) = \boldsymbol{0}$. Then g is an orthogonal transformation.*

Note that we do not assume here that g is linear, but only that it preserves the Euclidean distance and that it fixes the origin $\boldsymbol{0}$ of $\mathbb{E}^n$.

Proof. We proceed in stages.

First we prove that g preserves norms. As g is a Euclidean isometry, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{E}^n$ we have

$$d(g(\boldsymbol{x}), g(\boldsymbol{y})) = d(\boldsymbol{x}, \boldsymbol{y})$$

or in other words

$$||g(\boldsymbol{x}) - g(\boldsymbol{y})|| = ||\boldsymbol{x} - \boldsymbol{y}||, \quad \text{for all } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{E}^n. \tag{1.25}$$

From this we have

$$||g(\boldsymbol{x})|| = ||g(\boldsymbol{x}) - \boldsymbol{0}|| = ||g(\boldsymbol{x}) - g(\boldsymbol{0})|| = ||\boldsymbol{x} - \boldsymbol{0}|| = ||\boldsymbol{x}||, \ \forall\, \boldsymbol{x} \in \mathbb{E}^n. \tag{1.26}$$

We now prove that g preserves the scalar product on $\mathbb{E}^n$. In exercises you will have proved that for all $\boldsymbol{u}$, $\boldsymbol{v}$ in $\mathbb{E}^n$

$$\boldsymbol{u} \cdot \boldsymbol{v} = \tfrac{1}{2}(||\boldsymbol{u}||^2 + ||\boldsymbol{v}||^2 - ||\boldsymbol{u} - \boldsymbol{v}||^2).$$

Therefore, in view of (1.26) and (1.25), we have

$$\begin{aligned} g(\boldsymbol{x}) \cdot g(\boldsymbol{y}) &= \tfrac{1}{2}(||g(\boldsymbol{x})||^2 + ||g(\boldsymbol{y})||^2 - ||g(\boldsymbol{x}) - g(\boldsymbol{y})||^2) \\ &= \tfrac{1}{2}(||\boldsymbol{x}||^2 + ||\boldsymbol{y}||^2 - ||\boldsymbol{x} - \boldsymbol{y})||^2) \\ &= \boldsymbol{x} \cdot \boldsymbol{y}. \end{aligned}$$

In particular, g maps orthogonal vectors to orthogonal vectors.

Let us choose the standard orthonormal basis $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ of $\mathbb{E}^n$. Then $\boldsymbol{e}'_1 = g(\boldsymbol{e}_1), \ldots, \boldsymbol{e}'_n = g(\boldsymbol{e}_n)$ is another orthonormal basis in $\mathbb{E}^n$. Let A be the $n \times n$ matrix with column vectors $\boldsymbol{e}'_i$, $i = 1, \ldots, n$. From Linear Algebra, we know that A is an orthogonal matrix, so the function $f_A : \mathbb{E}^n \to \mathbb{E}^n$, $\boldsymbol{x} \mapsto A\boldsymbol{x}$, is an orthogonal transformation.

Clearly, the transformation f_A is invertible and $f_A^{-1} = f_{A^{-1}}$ is also an orthogonal transformation, hence it is an isometry of $\mathbb{E}^n$ by Theorem 1.18. Therefore the map $h = f_A^{-1} \circ g : \mathbb{E}^n \to \mathbb{E}^n$ is also a Euclidean isometry (see Exercise 1.5). Evidently, $h(\mathbf{0}) = \mathbf{0}$ and $h(\boldsymbol{e}_i) = \boldsymbol{e}_i$ for each $i = 1, \ldots, n$.

Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{E}^n$, then $x_i = \boldsymbol{x} \cdot \boldsymbol{e}_i$. The i-th coordinate of $h(\boldsymbol{x})$ is given by

$$h(\boldsymbol{x})_i = h(\boldsymbol{x}) \cdot \boldsymbol{e}_i = h(\boldsymbol{x}) \cdot h(\boldsymbol{e}_i) = \boldsymbol{x} \cdot \boldsymbol{e}_i = x_i,$$

for all $i = 1, \ldots, n$ (here we used the fact that h, being a Euclidean isometry, preserves scalar products, as we proved above). Thus all coordinates of $h(\boldsymbol{x})$ are the same as the corresponding coordinates of $\boldsymbol{x}$, so $h(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{E}^n$. This implies that h is the identity transformation on $\mathbb{E}^n$, and hence $f_A^{-1} \circ g = \mathrm{Id}_{\mathbb{E}^n}$. It follows that $g = f_A = f_{A,0}$, as claimed. □

Exercise 1.27. Prove that a Euclidean isometry that fixes $\mathbf{0}$ preserves angles between vectors.

Theorem 1.28. *Every Euclidean isometry of $\mathbb{E}^n$ is a Euclidean transformation $f_{A,\boldsymbol{t}}$, for some orthogonal matrix $A \in O(n)$ and some vector $\boldsymbol{t} \in \mathbb{E}^n$.*

Proof. Let f be an isometry of $\mathbb{E}^n$ and let $\boldsymbol{t} = f(\mathbf{0}) \in \mathbb{E}^n$. Then $g = f_{-\boldsymbol{t}} \circ f : \boldsymbol{x} \mapsto f(\boldsymbol{x}) - \boldsymbol{t}$ is an isometry of $\mathbb{E}^n$ fixing $\mathbf{0}$, by Theorem 1.23 and Corollary 1.5. Hence $g = f_A$, for some orthogonal matrix A, by Theorem 1.24, so $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \boldsymbol{t} = A\boldsymbol{x} + \boldsymbol{t}$, for all $\boldsymbol{x} \in \mathbb{E}^n$. Thus $f = f_{A,\boldsymbol{t}}$ is a Euclidean transformation. □

This theorem is very important, as it gives a complete characterisation of Euclidean isometries of the Euclidean space $\mathbb{E}^n$. In the case of $\mathbb{E}^2$, as explained in Examples 1.14 and 1.21, the Euclidean isometries of $\mathbb{E}^2$ are of the following types: translations, rotations, reflections and glide reflections.

Remark 1.29. Note that this theorem implies that Euclidean isometries are bijections. Indeed, we now know that any Euclidean isometry f is a Euclidean transformation $f = f_{A,\boldsymbol{t}}$, and any such transformation has an inverse, as they form a group.

1.4. **Geodesics.**

It is quite clear from the definition that the Euclidean distance between two points $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{E}^n$ is the length of the straight line segment that connects them. It is natural to ask why the straight line has been singled out in this way and we provide an outline of the answer in two dimensions. The general case can be treated in a similar way, but to make our reasoning precise we would have to rely on the Calculus of Variations.

A smooth curve C in $\mathbb{E}^2$ is a set of points $(x(t), y(t))$, where $t \in [0, 1]$, and x, y are smooth functions from $[0, 1]$ to $\mathbb{E}^2$ (i.e., $x(t)$ and $y(t)$ are differentiable and have

continuous derivatives on $[0,1]$). As you have seen in Calculus, the length of the curve C, from $t=0$ to $t=1$, is given by the formula

$$\int_0^1 \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

For instance, if we parametrise the unit circle $C = \mathbb{S}^1$ by $x(t) = \cos 2\pi t$ and $y(t) = \sin 2\pi t$ for $0 \leq t \leq 1$ then a simple integration shows that it has length (or circumference)

$$\int_0^1 \sqrt{(-2\pi \sin 2\pi t)^2 + (2\pi \cos 2\pi t)^2}\, dt = \int_0^1 2\pi\, dt = 2\pi.$$

Exercise 1.30. Let $\boldsymbol{a}, \boldsymbol{b}$ be two points in $\mathbb{E}^2$. Show that the length of the Euclidean segment joining these two points is equal to $||\boldsymbol{a} - \boldsymbol{b}||$.

Let us define the distance between two points in $\mathbb{E}^2$ to be the length of the shortest curve between them. We do not know at this point if this distance is the same as the Euclidean distance. In fact, as we prove now, it is, and this is a very important property of Euclidean geometry.

In the proof of the next theorem, we will rely on the following fact, which is left as an exercise.

Exercise 1.31. Let $\boldsymbol{p}_1, \boldsymbol{p}_2$ and $\boldsymbol{q}_1, \boldsymbol{q}_2$ be two pairs of distinct points in $\mathbb{E}^2$. Then there is an isometry of $\mathbb{E}^2$ sending $\boldsymbol{p}_i$ to $\boldsymbol{q}_i$ for $i = 1, 2$ if and only if $d(\boldsymbol{p}_1, \boldsymbol{p}_2) = d(\boldsymbol{q}_1, \boldsymbol{q}_2)$.

Theorem 1.32. *The shortest curve between two points $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{E}^2$ is the line-segment joining them, and its length is the Euclidean distance $d(\boldsymbol{a}, \boldsymbol{b}) = ||\boldsymbol{a} - \boldsymbol{b}||$ between them.*

Idea of the proof. Let us sketch the proof in the case when the two points are $\boldsymbol{a} = \boldsymbol{0} = (0,0)$ and $\boldsymbol{b} = (b, 0)$ for some $b > 0$ (one can reduce the proof to this case using Exercise 1.31). Let C be any curve connecting $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{E}^2$, and let $(x(t), y(t))$, $t \in [0,1]$, be a smooth parametrisation of C. Then

$$\text{length}(C) = \int_0^1 \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt \geq \int_0^1 \left|\frac{dx}{dt}\right|\, dt,$$

with equality if and only if $dy/dt = 0$, so the shortest curves from $\boldsymbol{a}$ to $\boldsymbol{b}$ must have y constant. This is possible since both $\boldsymbol{a}$ and $\boldsymbol{b}$ have y-coordinate equal to 0. So the shortest curve lies along the x-axis. It therefore has length

$$\int_0^1 \left|\frac{dx}{dt}\right|\, dt,$$

which is equal to the total variation in $|x|$ for $0 \leq t \leq 1$. This is minimised when x increases monotonically from 0 to b, that is, the curve follows the line-segment from $\boldsymbol{a}$ to $\boldsymbol{b}$. In this case the length of the curve is

$$\int_0^1 \frac{dx}{dt}\, dt = [x(t)]_0^1 = x(1) - x(0) = b - 0 = b = ||\boldsymbol{a} - \boldsymbol{b}||.$$

Now, according to Exercise 1.30, $||\boldsymbol{a} - \boldsymbol{b}||$ is precisely the length of the Euclidean segment $[\boldsymbol{a}, \boldsymbol{b}]$ joining $\boldsymbol{a}$ with $\boldsymbol{b}$ in $\mathbb{E}^2$. Hence $lenght(C) \geq length([\boldsymbol{a}, \boldsymbol{b}])$, as claimed. □

Definition 1.33. We say that a curve C (in any metric space X) is a *geodesic* in X if, for any $\boldsymbol{a}, \boldsymbol{b} \in C$, the shortest path in X from $\boldsymbol{a}$ to $\boldsymbol{b}$ is the segment of C between $\boldsymbol{a}$ and $\boldsymbol{b}$.

Theorem 1.32 shows that the geodesics in $\mathbb{E}^2$ are the straight lines, and a similar argument proves the same result for Euclidean spaces $\mathbb{E}^n$ of any dimension. Thus if we measure the distance using the Euclidean metrics, the shortest distance between two points is measured along the straight line connecting them.

2. Spherical geometry

2.1. **Spherical distance.**

As we live on the surface of a ball, it is clearly of interest to study the geometry of the sphere. While on the small scale the Earth looks flat, and we can use this approximation to apply the two-dimensional Euclidean geometry in our daily life, on the large scale this is most definitely not true. To plot the course of a long-distance flight between cities on different continents, for example, we need to know how to measure distances on the sphere.

Definition 2.1. The *unit sphere* $\mathbb{S}^n$ in the Euclidean space $\mathbb{E}^{n+1}$ is the set of all points $\boldsymbol{x} \in \mathbb{E}^{n+1}$ whose Euclidean distance to the origin $\mathbf{0}$ is 1:

$$\mathbb{S}^n = \{\boldsymbol{x} \in \mathbb{E}^{n+1} \mid d(\boldsymbol{x}, \mathbf{0}) = ||\boldsymbol{x}|| = 1\}.$$

In this section we will study the geometry of the unit sphere $\mathbb{S}^2$ in the three-dimensional Euclidean space $\mathbb{E}^3$, and our first objective is to define a suitable metric on $\mathbb{S}^2$. We could use the Euclidean distance, but this does not conform to the geometry of our space: a plane travelling long-distance is not allowed to tunnel through the Earth! A more natural way to do this is as follows. First observe that any two points $\boldsymbol{x}$ and $\boldsymbol{y}$ on the sphere $\mathbb{S}^2$ lie on a great circle through them. As you know, a *great circle* is the intersection of $\mathbb{S}^2$ with a plane in $\mathbb{E}^3$ through the origin $\mathbf{0}$ in $\mathbb{E}^3$. If the two points are antipodal, that is $\boldsymbol{x} = -\boldsymbol{y}$, which happens for instance when $\boldsymbol{x}$ and $\boldsymbol{y}$ are the poles on the Earth, there are infinitely many such circles; otherwise the circle is unique.

When $\boldsymbol{x}$ and $\boldsymbol{y}$ are distinct non-antipodal points, the great circle C through them is divided into two segments. Since the radius of the circle C is 1 (as is the radius of the sphere), the length of each of the segments is equal to the angle at $\mathbf{0}$ (measured in radians) subtended on the segment.

Definition 2.2. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2$. The *spherical distance* $d_{\mathbb{S}^2}(\boldsymbol{x}, \boldsymbol{y})$ between $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined to be the length of the shorter of the two segments of the great circle C through the two points. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are antipodal, all such segments have equal length π. Equivalently, the spherical distance is given by the formula

$$d_{\mathbb{S}^2}(\boldsymbol{x}, \boldsymbol{y}) = \cos^{-1}(\boldsymbol{x} \cdot \boldsymbol{y}), \quad \text{for } \boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^2,$$

which describes the Euclidean angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

It follows from the definition that the spherical distance between points on $\mathbb{S}^2$ is a number between 0 and π. It is zero when the two points are identical, and equals π when they are antipodal.

We will now check that the spherical distance satisfies the axioms of Definition 1.1. It is clear from the definition that $d_{\mathbb{S}^2}(\boldsymbol{x}, \boldsymbol{y})$ is non-negative, and that it equals zero if and only if $\boldsymbol{x} = \boldsymbol{y}$. It is also clearly symmetric, as the great circle through $\boldsymbol{x}$ to $\boldsymbol{y}$ is the same circle as the one through $\boldsymbol{y}$ to $\boldsymbol{x}$. The proof of the triangle inequality requires a bit more understanding of the geometry of the sphere. While it is possible to provide a direct proof of the triangle inequality from the definition, our approach will use an identification of geodesics for this metric. For this we need to recall the spherical coordinates in $\mathbb{E}^3$. The standard coordinate vectors in $\boldsymbol{e}^3$ will be denoted $\boldsymbol{e}_1 = (1, 0, 0)$, $\boldsymbol{e}_2 = (0, 1, 0)$, and $\boldsymbol{e}_3 = (0, 0, 1)$.

The position of a point $\boldsymbol{v}$ in $\mathbb{E}^3$ is given in the spherical coordinates by three numbers (r, ϕ, θ). The first coordinate $r \in [0, \infty)$ is the Euclidean distance from $\boldsymbol{v}$ to the origin: $r = d(\boldsymbol{v}, \mathbf{0})$. The second coordinate $\phi \in [0, \pi]$ is the angle between the z-axis and $\boldsymbol{v}$; while $\theta \in [0, 2\pi)$ is the angle in the (x, y)-plane from the positive x axis to the projection of the vector $\boldsymbol{v}$ onto the (x, y)-plane, measured anti-clockwise (see Figure 1).

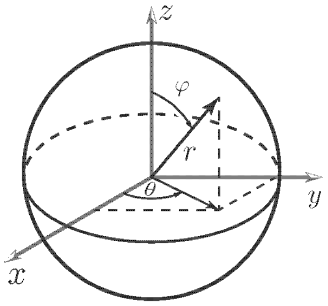


Figure 1: Spherical coordinates in $\mathbb{E}^3$. On $\mathbb{S}^2$ we have $r = 1$.

If $\boldsymbol{v} \in \mathbb{S}^2$, then $r = 1$, and so any point on $\mathbb{S}^2$ has the coordinates $(1, \phi, \theta)$. A simple calculation will show that this point has the Cartesian coordinates

$$(\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi).$$

To translate this coordinate system into one that you know from Geography, and regarding the Earth's surface as a sphere, the latitude of a point is $\phi' = \pi/2 - \phi$ in the North hemisphere and $\phi' = \phi - \pi/2$ in the South hemisphere (for this reason ϕ is sometimes called the *colatitude*). The *longitude east* is the angle θ, while the *longitude west* is $2\pi - \theta$ (so that $\theta = 0$ on the Greenwhich meridian).

Theorem 2.3. *The shortest curve between two points $\boldsymbol{a}$ and $\boldsymbol{b}$ in $\mathbb{S}^2$ is a segment of a great circle joining them. The length of that segment is equal to the spherical distance $d_{\mathbb{S}^2}(\boldsymbol{a}, \boldsymbol{b})$.*

Proof. We sketch here the main idea of the proof. Using appropriate rotations we can arrange the points $\boldsymbol{a}$ and $\boldsymbol{b}$ so that $\boldsymbol{a}$ is at the North Pole $\boldsymbol{n} = \boldsymbol{e}_3$, and so that $\boldsymbol{b}$ lies on the line $\theta = 0$ (the 'Greenwich meridian'). A smooth curve on the sphere $\mathbb{S}^2$ has coordinates $(1, \phi(t), \theta(t))$, where $t \in [0, 1]$, while the length of the curve is given by the formula:

$$L = \int_0^1 \sqrt{\left(\frac{d\phi}{dt}\right)^2 + \sin^2\phi \left(\frac{d\theta}{dt}\right)^2}\, dt$$

Hence

$$L \geq \int_0^1 \frac{d\phi}{dt} dt$$

and the equality takes place when $\frac{d\theta}{dt} = 0$, so that θ is constant. It follows that the shortest curve will be contained in the great circle from $\boldsymbol{a}$ to $\boldsymbol{b}$, and the length of the corresponding segment is given by

$$L(C) = \int_0^1 \frac{d\phi}{dt} dt = \phi(1) - \phi(0) = \phi = d_{\mathbb{S}^2}(\boldsymbol{a}, \boldsymbol{b}).$$

□

This theorem shows therefore that the geodesics in $\mathbb{S}^2$ are segments of great circles. We can now prove the triangle inequality for the spherical distance, which states that for all points $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ in $\mathbb{S}^2$ we have

$$d_{\mathbb{S}^2}(\boldsymbol{x}, \boldsymbol{z}) \leq d_{\mathbb{S}^2}(\boldsymbol{x}, \boldsymbol{y}) + d_{\mathbb{S}^2}(\boldsymbol{y}, \boldsymbol{z}).$$

Assume that this is not true, so that we can find three points $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ such that

$$d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{z}) > d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{y}) + d_{\mathbb{S}^2}(\boldsymbol{y},\boldsymbol{z}).$$

This means that the path from $\boldsymbol{x}$ to $\boldsymbol{z}$ that goes through $\boldsymbol{y}$ is shorter than the path from $\boldsymbol{x}$ to $\boldsymbol{z}$ that gives the spherical distance $d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{z})$. But we have proved that $d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{z})$ is the length of the shortest path connecting these two points, and we have a contradiction.

Example 2.4. Let us compute the spherical distance between New York and Sydney. New York has latitude 41° north and longitude 74° west, to the nearest degree, so its colatitude is $\phi = 90° - 41° = 49°$, and $\theta = -74°$. If we take the Earth's approximate radius of 3960 miles as one unit, then the Earth's surface is represented by $\mathbb{S}^2$ and New York corresponds to the point

$$\begin{aligned}\boldsymbol{x} &= (\cos(-74°)\sin(49°), \sin(-74°)\sin(49°), \cos(49°))\\ &= (0.208, -0.725, 0.656).\end{aligned}$$

Sydney has latitude 34° south, that is, −34°, and longitude 151° east, so here $\phi = 90° - (-34°) = 124°$ and $\theta = 151°$; it is therefore represented by the point

$$\begin{aligned}\boldsymbol{y} &= (\cos(151°)\sin(124°), \sin(151°)\sin(124°), \cos(124°))\\ &= (-0.725, 0.402, -0.559).\end{aligned}$$

Then

$$\begin{aligned}\boldsymbol{x}\cdot\boldsymbol{y} &\approx (0.208 \times -0.725) + (-0.725 \times 0.402) + (0.656 \times -0.559)\\ &\approx -0.809,\end{aligned}$$

so the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is approximately

$$\cos^{-1}(-0.809) \approx 2.51 \text{ radians}.$$

This is equal to the spherical distance in units of 3960 miles, so a direct flight between New York and Sydney should cover approximately $2.51 \times 3960 = 9937$ miles. There are online calculators[1] that will compute the distance between cities on Earth, and plot the corresponding geodesic. These are fun to play with.

2.2. **Spherical isometries.**

As we remarked before, there are two ways of measuring distance between points on the sphere: the spherical distance and the Euclidean distance. Let's compare the two. If $\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^2$ then the spherical distance is given by $d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{y}) = \cos^{-1}(\boldsymbol{x}\cdot\boldsymbol{y})$, hence

$$\boldsymbol{x}\cdot\boldsymbol{y} = \cos\left(d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{y})\right), \ \forall\, \boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^2.$$

On the other hand, for the Euclidean distance between $\boldsymbol{x},\boldsymbol{y} \in \mathbb{S}^2$ we have

$$d(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x}-\boldsymbol{y}|| = \sqrt{(\boldsymbol{x}-\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})}.$$

If we expand the second expression and use that $||\boldsymbol{x}|| = ||\boldsymbol{y}|| = 1$ on $\mathbb{S}^2$, we get

$$d(\boldsymbol{x},\boldsymbol{y}) = \sqrt{2(1-\boldsymbol{x}\cdot\boldsymbol{y})} = \sqrt{2\big(1-\cos\left(d_{\mathbb{S}^2}(\boldsymbol{x},\boldsymbol{y})\right)\big)}.$$

From this we can deduce that any map of $\mathbb{S}^2$ that preserves one of these distances, also preserves the other. This uses the fact that the function $\sqrt{2-\cos\alpha}$ is bijective for $0 \le \alpha \le \pi$.

It feels that the isometries of the sphere $\mathbb{S}^2$ should be closely linked to the isometries of the Euclidean space $\mathbb{E}^3$.

Theorem 2.5. *Every isometry of $\mathbb{S}^2$ extends to a unique isometry of $\mathbb{E}^3$.*

Proof. Exercise. □

[1] For example, https://www.greatcirclemap.com

This simple result gives us a way to describe the isometries of the sphere $\mathbb{S}^2$.

Theorem 2.6. *Every isometry of $\mathbb{S}^2$ is induced by an orthogonal matrix $A \in O(3)$. It is therefore either*

(1) *a rotation around a line L through $\mathbf{0}$ when $\det A = 1$; or*
(2) *a rotation about L through $\mathbf{0}$ composed with a reflection in the plane orthogonal to L, when $\det A = -1$.*

Proof. We will provide a sketch of this result. By the Theorem 2.5, every isometry of the sphere arises as the restriction of an isometry of $\mathbb{E}^3$. Any such isometry of $\mathbb{E}^3$ must fix $\mathbf{0}$, and so is given by an orthogonal transformation by Theorem 1.24, which in turn is induced by a matrix $A \in O(3)$. Every such matrix has the determinant ± 1, and the matrices with the determinant 1 represent rotations of $\mathbb{E}^3$, while those with determinant -1 are compositions of rotations with reflections. □

Definition 2.7. The matrices of determinant 1 form a subgroup of $O(3)$ which is called the *special orthogonal group* and is denoted $SO(3)$.

3. Normed vector spaces

3.1. Norms.

A very important and rich source of examples of metric spaces is provided by normed vector spaces. We have seen the example of the Euclidean norm, and this section extends this idea to more general spaces.

Let V be a vector space over the reals $\mathbb{R}$ or the complex numbers $\mathbb{C}$; we will use $\mathbb{F}$ to denote either of these fields. For convenience, we will mostly be concerned with real vector spaces, but everything we say will work for general normed vector spaces.

Definition 3.1. A *normed vector space* over a field $\mathbb{F}$ is a vector space V which is equipped with a map, called the norm, $||-|| : V \to \mathbb{R}$ that satisfies the following properties:

(1) for every $\boldsymbol{x} \in V$, $||\boldsymbol{x}|| \geq 0$, and $||\boldsymbol{x}|| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0} \in V$;
(2) for all $\alpha \in \mathbb{F}$ and $\boldsymbol{x} \in V$, $||\alpha \boldsymbol{x}|| = |\alpha| ||\boldsymbol{x}||$;
(3) for all $\boldsymbol{x}, \boldsymbol{y} \in V$, $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$.

Given a norm $||-||$ define a map $d : V \times V \to \mathbb{R}$ by

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||$$

for any two elements $\boldsymbol{x}, \boldsymbol{y} \in V$.

We record the following simple result for future reference.

Proposition 3.2. *Every normed vector space $(V, ||-||)$ is a metric space when equipped with the distance function*

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}|| \quad \textit{for all } \boldsymbol{x}, \boldsymbol{y} \in V.$$

Proof. We check the axioms of metric stated in Definition 1.1. If for some $\boldsymbol{x}$ and $\boldsymbol{y}$ in V, $d(\boldsymbol{x}, \boldsymbol{y}) = 0$, then $||\boldsymbol{x} - \boldsymbol{y}|| = 0$ which, by condition (1) from Definition 3.1 happens if and only if $\boldsymbol{x} - \boldsymbol{y} = \boldsymbol{0}$, or $\boldsymbol{x} = \boldsymbol{y}$ in V.

Now, for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in V, condition (2) from Definition 3.1 tells us that

$$||\boldsymbol{y} - \boldsymbol{x}|| = ||(-1)(\boldsymbol{x} - \boldsymbol{y})|| = |-1| ||\boldsymbol{x} - \boldsymbol{y}|| = ||\boldsymbol{x} - \boldsymbol{y}||,$$

which proves that $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$. So d is symmetric.

Finally, we prove the triangle inequality as follows. For $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z}$ in V write $\boldsymbol{a} = \boldsymbol{x} - \boldsymbol{y}$, $\boldsymbol{b} = \boldsymbol{y} - \boldsymbol{z}$ and $\boldsymbol{c} = \boldsymbol{x} - \boldsymbol{z}$. Then, by the triangle inequality for the norm (condition (3) from Definition 3.1), we have

$$||\boldsymbol{c}|| = ||\boldsymbol{a} + \boldsymbol{b}|| \leq ||\boldsymbol{a}|| + ||\boldsymbol{b}||$$

which is the same as

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

This gives the triangle inequality for d, hence it is, indeed, a metric on V. □

Let us now explore some of the most important examples of normed vector spaces, which, thanks to Proposition 3.2, can now be regarded as metric spaces.

3.2. Norms on $\mathbb{R}^n$.

Here we will introduce a very important family of norms on $\mathbb{R}^n$, indexed by a real number $p \geq 1$.

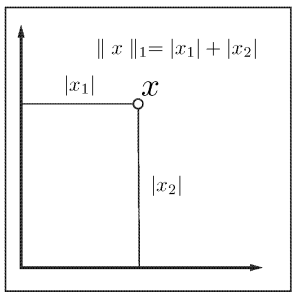
Definition 3.3. For all $p \geq 1$ and all $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, define the *$p$-norm* of $\boldsymbol{x}$ by

$$||\boldsymbol{x}||_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Theorem 3.4. *The function* $||-||_p \to \mathbb{R}$ *defined in Definition 3.3 is a norm on* $\mathbb{R}^n$.

Proof. The proof is given in the exercises. □

Proposition 3.2 allows us to regard the space $\mathbb{R}^n$, endowed with a norm $||-||_p$, as a metric space equipped with the metric $d_p(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_p$, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. The following special cases are important.



When $n = 1$, that is when we consider the space of real numbers, all these norms reduce to the absolute value on $\mathbb{R}$: $||\boldsymbol{x}||_p = |\boldsymbol{x}|$, for all $\boldsymbol{x} \in \mathbb{R}$ and all $p \geq 1$.

The associated metric is called the *standard metric* on $\mathbb{R}$.

When $p = 2$ we recover the Euclidean norm introduced in Section 1.2. We will denote this $||-||_2$ and the associated Euclidean metric will be denoted d_2.

Another important special case is the norm obtained for $p = 1$, when we have

$$||\boldsymbol{x}||_1 = \sum_{i=1}^{n} |x_i|.$$

The corresponding metric on $\mathbb{R}^n$ is given by

$$d_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} |x_i - y_i|$$

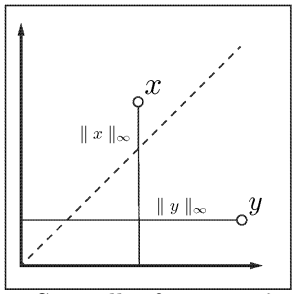
for any two $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. In the case when $n = 2$, so the two-dimensional real plane, this metric is called the *taxi-cab* metric. You should draw some pictures to figure out why.

3.2.1. *The max-norm.*

Definition 3.5. We define the *max-norm* (or the ∞-*norm*, or the *sup-norm*) on $\mathbb{R}^n$ by the formula

$$||\boldsymbol{x}||_\infty = \max\{|x_1|, \ldots, |x_n|\}, \quad \text{for all } \boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n.$$

The reason for this notation will become clear soon.



Proposition 3.6. *The max-norm* $||-||_\infty$ *is a norm on* $\mathbb{R}^n$.

Proof. Let $\boldsymbol{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. First, it is clear that $||\boldsymbol{x}||_\infty \geq 0$, since the maximum is taken over a set of non-negative numbers. If $||\boldsymbol{x}||_\infty > 0$, then there exists $k \in \{1, \ldots, n\}$ such that $|x_k| > 0$, which means that $\boldsymbol{x} \neq \boldsymbol{0}$. It is clear that if $\boldsymbol{x} = \boldsymbol{0}$ then $||\boldsymbol{x}||_\infty = 0$, so we conclude that $||\boldsymbol{x}||_\infty = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.

Secondly, for every $\lambda \in \mathbb{R}$ and all $\boldsymbol{x} \in \mathbb{R}^n$ we have

$$\begin{aligned} ||\lambda \boldsymbol{x}||_\infty &= \max\{|\lambda x_1|, \ldots, |\lambda x_n|\} \\ &= |\lambda| \max\{|x_1|, \ldots, |x_n|\} = |\lambda| ||\boldsymbol{x}||_\infty. \end{aligned}$$

Finally, for all $\boldsymbol{x}$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ in $\mathbb{R}^n$ and all $i = 1, \ldots, n$, we have

$$|\boldsymbol{x}_i + \boldsymbol{y}_i| \leq |x_i| + |y_i| \leq ||\boldsymbol{x}||_\infty + ||\boldsymbol{y}||_\infty.$$

Hence, by taking the maximum of the left-hand side over all $i = 1, \dots, n$, we obtain the triangle inequality

$$||\boldsymbol{x} + \boldsymbol{y}||_\infty \leq ||\boldsymbol{x}||_\infty + ||\boldsymbol{y}||_\infty, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n.$$

Hence, $|| - ||_\infty$ is a norm on $\mathbb{R}^n$. □

The associated *max-metric* on $\mathbb{R}^n$ will be denoted d_∞.

3.3. **Spaces of continuous functions.**

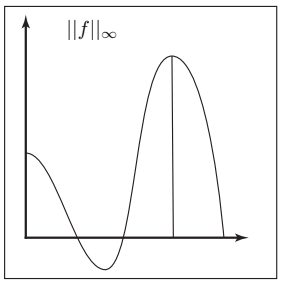
To provide further examples of how flexible the metric space structure is, let us now consider the space $\mathcal{C}[a, b]$, of continuous real-valued functions on a closed interval $[a, b]$. This is a vector space over the field of the real numbers with respect to the following operations:

- *Pointwise addition*: given any $f, g \in \mathcal{C}[a, b]$ the sum $f + g$ is defined by
$$(f + g)(x) = f(x) + g(x), \text{ for all } x \in [a, b];$$
- *Multiplication by a scalar*: given $f \in \mathcal{C}[a, b]$ and $\lambda \in \mathbb{R}$, we define the function λf by
$$(\lambda f)(x) = \lambda f(x), \quad \text{for all } x \in [a, b].$$

You should recall from Analysis that the pointwise sum of two continuous functions is continuous, as is the product of a continuous function by a scalar. It is easy to see that $\mathcal{C}[a, b]$ satisfies all of the required axioms (with the constant zero function $\mathbf{0}$ as its zero vector), thus it is a vector space over $\mathbb{R}$ with respect to these two operations.

Definition 3.7. For every $f \in \mathcal{C}[a, b]$, we define the *supremum norm* (or the ∞-*norm*) $|| - ||_\infty$ by

$$||f||_\infty = \sup\{|f(x)| \mid x \in [a, b]\}.$$



You will recall from the Analysis course that a continuous function on a closed interval is bounded, which implies that the supremum norm is finite. We will come back to this point later in the course.

Proposition 3.8. *The vector space $\mathcal{C}[a, b]$, of continuous real-valued functions on a closed interval $[a, b]$, is a normed vector space when equipped with the norm $|| - ||_\infty$.*

Consequently, the space $(\mathcal{C}[a, b], d_\infty)$ is a metric space with respect to the supremum metric

$$d_\infty(f, g) = ||f - g||_\infty, \ \forall f, g \in \mathcal{C}[a, b].$$

Proof. First, as $|f(x)| \geq 0$ for all $x \in [a, b]$, $||f||_\infty \geq 0$. If $||f||_\infty > 0$, then there must exist an $x \in [a, b]$ such that $|f(x)| > 0$; this implies that $f \neq \mathbf{0}$. Since $||f||_\infty$ is clearly zero for a function identically equal to zero, we have $||f||_\infty = 0$ if and only if $f = \mathbf{0}$.

Secondly, using properties of the supremum of a set of real numbers, for all $\lambda \in \mathbb{R}$ and $f \in \mathcal{C}[a, b]$ we can write

$$\begin{aligned} \sup\{|\lambda f(x)| \mid x \in [a, b]\} &= \sup\{|\lambda||f(x)| \mid x \in [a, b]\} \\ &= |\lambda|\{|f(x)| \mid x \in [a, b]\}. \end{aligned}$$

Thus, $||\lambda f||_\infty = |\lambda|||f||_\infty$.

Finally, for all $f, g \in \mathcal{C}[a, b]$ and all $x \in [a, b]$ we have

$$|f(x) + g(x)| \leq |f(x)| + |g(x)| \leq ||f||_\infty + ||g||_\infty.$$

Taking the supremum of the left hand side, we get the required triangle inequality

$$||f+g||_\infty \leq ||f||_\infty + ||g||_\infty, \ \forall\, f, g \in \mathcal{C}[a,b].$$

□

With this result in hand, we can now regard $\mathcal{C}[a,b]$ as a metric space with respect to the *supremum metric* d_∞.

This result has a useful and natural extension. A real-valued function $f : \mathbb{R} \to \mathbb{R}$ is called *bounded* if and only if there exists $R > 0$ such that $|f(x)| < R$ for all $x \in \mathbb{R}$. Let us denote by $\mathcal{B}(\mathbb{R})$ the set of all *bounded functions* $f : \mathbb{R} \to \mathbb{R}$. Evidently, for each bounded function $f \in \mathcal{B}(\mathbb{R})$ the supremum norm $||f||_\infty$ is finite (where the supremum is taken over all of $\mathbb{R}$). As before, we can show that the space of bounded functions is also a normed space, hence a metric space.

It is also worth noting, that the space $\mathcal{C}[a,b]$ admits a family of p-norms, which are analogues of those discussed for $\mathbb{R}^n$ in Section 3.2.

Definition 3.9. Let $p \geq 1$ be a real number. For every $f \in \mathcal{C}[a,b]$ define

$$||f||_p = \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}.$$

The norm $||-||_p$ is called the *L^p-norm.*

This is a norm on $\mathcal{C}[a,b]$, and so it gives rise to a metric on the space of continuous functions $\mathcal{C}[a,b]$. We will use most often the norms for $p = 1$, illustrated in Figure 2, and for $p = 2$.

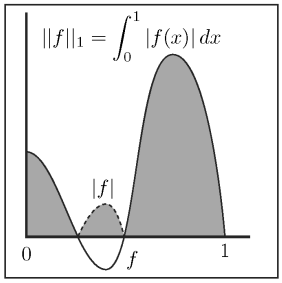


Figure 2: The L^1-norm on $\mathcal{C}[0,1]$.

4. General Metric Spaces

We have now seen many examples of spaces with metrics, but, with the exception of the spherical geometry, all our examples were those of vector spaces equipped with a norm. However, the notion of a metric, given in Definition 1.1 is much more general than that, and in this section we discuss some examples that do not come from normed vector spaces.

We start with the following simple but useful definition.

Definition 4.1. Let X be a non-empty set. Then X can be equipped with the distance function called the *discrete metric* as follows. For all $x, y \in X$, define

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

To check that this is a metric, we note that the first two axioms are obviously true. To check the third, assume that there exist points x, y, z in X such that

$$\delta(x,z) > \delta(x,y) + \delta(y,z).$$

Given that the function δ only takes 0 or 1 as its values, this can only happen if $\delta(x,z) = 1$ and $\delta(x,y) + \delta(y,z) = 0$. But this would imply that $x \neq z$ and $x = y = z$, which is a contradiction.

Exercise 4.2. Show that the discrete metric on $\mathbb{R}$ does not come from any norm.

Thus every set can be equipped with a metric, the discrete metric. We will now devote some time to specific examples of metrics which capture interesting properties of spaces on which they are defined.

4.1. Other metrics in $\mathbb{R}^2$.

We now introduce a few more metrics on $\mathbb{R}^2$ which do not arise in any of the ways discussed so far.

Definition 4.3 (The sunflower metric)**.** The *sunflower metric* on $\mathbb{R}^2$ (or the *SNCF metric* named after the French railway network in which Paris is a key hub) and is defined as follows. If $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ and there is a line in $\mathbb{R}^2$ that passes through $\boldsymbol{x}$, $\boldsymbol{y}$, and the origin $\mathbf{0}$ then we put

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2.$$

If there is no such line, we put

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x}||_2 + ||\boldsymbol{y}||_2.$$

Thus to travel between any two points, not lying on the same ray originating at $\mathbf{0}$ in $\mathbb{R}^2$, we have to go through the origin, and the distance is calculated as the length of the path that we have to take. It is clear that this metric is not the same as the Euclidean distance, as there will be points in $\mathbb{R}^2$ which are very close in the Euclidean distance but quite far away in the sunflower metric. Draw some pictures to experiment.

Definition 4.4 (The lift metric)**.** Another example is called the *lift metric*. If $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$ then

$$d(\boldsymbol{x}, \boldsymbol{y}) = \begin{cases} |x_1 - y_1| & \text{if } x_2 = y_2 \\ |x_1| + |x_2 - y_2| + |y_1| & \text{otherwise.} \end{cases}$$

The intuition behind this metric is that of how we travel in a building equipped with a vertical lift, but no stairs. Then we can travel between any two points on the same floor along the straight line connecting them, but to get from a point on

one floor to a point on another (hence different second coordinates) we first need to get to the lift (the second coordinate axis), travel in the lift between the floors, and then walk again along the required floor to get to the other point.

Again, this metric is different from the Euclidean metric or the p-metrics on $\mathbb{R}^2$, for $p \geq 1$.

Exercise 4.5. Prove that the sunflower metric and the lift metric are indeed metrics on $\mathbb{R}^2$ (i.e., they satisfy Definition 1.1). For each of these metrics determine whether it arises from a norm on $\mathbb{R}^2$ (cf. Proposition 3.2).

4.2. **Length spaces.**

We have seen in previous sections that the Euclidean and the spherical metrics can be defined in terms of geodesics, which are shortest lines connecting points. In more generality, there is an interesting class of metric spaces, where distance can be measured using the notion of path length. In broad terms, this works as follows.

Let X be a metric space and let $\gamma : [0,1] \to X$ be a continuous path (the precise meaning of this will be explained later, but for now your intuition of what a path is should suffice.) The *length* of γ is, by definition, the supremum

$$\ell(\gamma) = \sup \left\{ \sum_i d(\gamma(t_i), \gamma(t_{i+1})) \right\},$$

taken over all partitions $0 = t_0 < t_1 < \cdots < t_N = 1$ of the interval $[0,1]$. This number can be infinite. If $\ell(\gamma) < \infty$ then the path γ is said to be *rectifiable*.

Definition 4.6. A metric space X is a *length space* if the distance between any two points of X is equal to the infimum of the lengths of the continuous paths joining them in X.

Graphs are an interesting class of length spaces, and Figure 3 provides an illustration. Recall that a *graph* Γ consists of set of *vertices* $V\Gamma$ and a set of edges $E\Gamma$

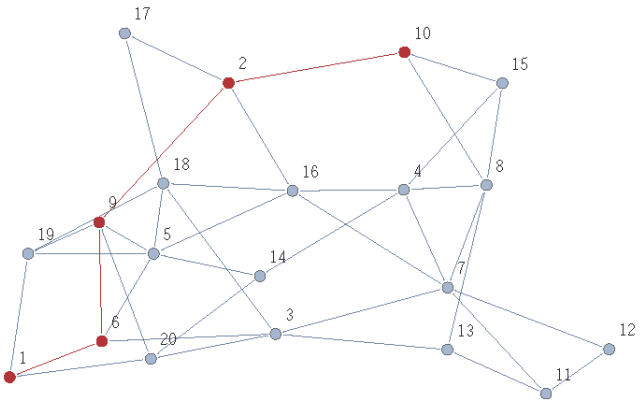


Figure 3: A shortest path between two vertices in a graph.

between the vertices. A (continuous) path in Γ is basically a finite sequence of edges $e_1, e_2, \ldots, e_k$, such that e_i and e_{i+1} are adjacent, for all $i = 1, \ldots, k-1$. The *length* of this path is the number k, of the edges involved in it. The distance between two vertices on a graph is defined as the infimum of the lengths of paths connecting these two vertices (in a general graph, there may be many possible paths between any two vertices, and there may be more than one shortest path). This distance is called the *path metric* on Γ, and it is a metric on $V\Gamma$, provided Γ is connected (i.e., any two vertices of Γ can be connected by a path).

Exercise 4.7. Prove that the path metric on any connected graph Γ satisfies the triangle inequality.

So far we have seen the following examples of length spaces:

(1) The Euclidean space $\mathbb{E}^n$, with the Euclidean metric. The shortest path between two points is a straight line segment. The Euclidean space has the property that there is a unique shortest path between any two points.
(2) The sphere $\mathbb{S}^2$, with a spherical metric. The spherical distance between two points is given as the length of a segment of a great circle connecting the two points, which is also the shortest path connecting them. We have seen that if the two points on the sphere are not antipodal then there is a unique shortest path connecting them. For antipodal points, there are infinitely many paths of the same length joining them.
(3) Connected graphs, equipped with the path metric.

4.3. **The Hamming metric.**

This metric was defined by Hamming in 1950 in his work on Information Theory. This distance was introduced to measure similarity between strings of symbols, which are drawn from a finite set. The *Hamming distance* between two finite strings of equal length is defined to be the number of places where the two strings differ. Equivalently, it is the minimum number of substitutions that need to be made to transform one string to another.

Here are some examples:

nose rose	distance: 1
1101101 1011011	distance: 4
123456789 987654321	distance: 8

We can give the following formal definition.

Definition 4.8. Let F be a non-empty set and let F^n denote the space of n-tuples of elements of F. Let δ be the discrete metric on the set F defined in Example 4.1. Then for all $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ in F^n define the Hamming distance by

$$d_H(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \delta(x_i, y_i).$$

Using the properties of the discrete metric it is now easy to check that the Hamming distance is indeed a metric. This is left as an exercise.

4.4. **Small-world networks.**

You may have heard of the idea *Six degrees of separation*, which is a hypothesis that there is a link between any two people on Earth through a chain of mutual acquaintances of length of no more than six. While it is difficult to test the veracity of this statement for all people on the planet, special cases have been tried and you can find it amusing to read up on this. There are variants of this concept that are easier to verify. One of them is the collaboration distance, which is a path metric on the space of all active scientists. Two people (vertices in the graph) are connected (there exists an edge between them) if and only if they are coauthors on

a published paper. MathSciNet, the publication database of the American Mathematical Society, provides a tool for computing the collaboration distance between any two mathematicians whose papers are in the database. It is popular to compute the collaboration distance from a given mathematician to Paul Erdös, who was a prolific Hungarian mathematician with over 1500 published papers to his name. Again, MathSciNet will provide tools for computing that distance, which is known as the *Erdös number* of the chosen mathematician. If no such path exists, the Erdös number is infinite.

There are amusing other applications of this idea. For example, the *Bacon number* is the collaboration distance to the actor Kevin Bacon, computed through a chain of actors who have appeared together in a movie. You can of course combine the distances: there are quite a lot of people with a finite Erdös number and a finite Bacon number.

Possibly the best claim to fame is to have a finite *Erdös-Bacon-Sabbath* number, which is the sum of the two numbers already explained and the Sabbath number, which is the collaboration distance to the rock group Black Sabbath. Amazingly, a sizeable group of people has a finite Erdös-Bacon-Sabbath number, and includes Stephen Hawking, Condoleeza Rice, Douglas Adams, Imogen Heap, Brian May, Mayim Bialik, Colin Firth, Marika Taylor, and others. The internet will provide further information on the subject, should you require it.

More seriously, these kind of considerations have led Watts and Strogarz to the identification of a class of graphs, called *small worlds*. A graph with N vertices is a small world if a typical path distance between a pair of randomly chosen vertices is

$$L \sim \log N.$$

Many real-life networks exhibit this property: social networks, as discussed above, the internet, Wikipedia, brain neurons, etc. Again, I invite you to investigate this further.

4.5. **Metrics and norms.** In this section, we asked the question whether a particular distance function on a vector space V is induced by a norm on V. In other words, we need to decide if, given a metric $d : V \times V \to \mathbb{R}$ on a vector space V, does there exist a norm $|| - || : V \to \mathbb{R}$ such that for all $\boldsymbol{x}, \boldsymbol{y} \in V$

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||.$$

To answer this question, let us first investigate the properties of a metric that is defined by a norm on V. We have the following.

Proposition 4.9. *Let V be a vector space over a field $\mathbb{F} = \mathbb{R}, \mathbb{C}$ equipped with a norm $|| - ||$. Then the metric d defined by*

$$d(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x}, \boldsymbol{y}||$$

is

- *Homogeneous; i.e., for every $\alpha \in \mathbb{F}$ and all $\boldsymbol{x}, \boldsymbol{y} \in V$ we have*
$$d(\alpha\boldsymbol{x}, \alpha\boldsymbol{y}) = |\alpha| d(\boldsymbol{x}, \boldsymbol{y}).$$
- *Translation invariant; this means that for all $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{t}$ in V*
$$d(\boldsymbol{x} + \boldsymbol{t}, \boldsymbol{y} + \boldsymbol{t}) = d(\boldsymbol{x}, \boldsymbol{y}).$$

Proof. We use the properties of norm. For the first point, we have that

$$d(\alpha\boldsymbol{x}, \alpha\boldsymbol{y}) = ||\alpha\boldsymbol{x} - \alpha\boldsymbol{y}|| = |\alpha|||\boldsymbol{x} - \boldsymbol{y}|| = |\alpha| d(\boldsymbol{x}, \boldsymbol{y}).$$

For the second point, for any $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{t}$ in V we have

$$d(\boldsymbol{x} + \boldsymbol{t}, \boldsymbol{y} + \boldsymbol{t}) = ||\boldsymbol{x} + \boldsymbol{t} - \boldsymbol{y} - \boldsymbol{t}|| = ||\boldsymbol{x} - \boldsymbol{y}|| = d(\boldsymbol{x}, \boldsymbol{y}).$$

□

Hence, if we want to prove that a given metric does not come from a norm, we can check if it fails one or both of the above properties. If it does, there is no norm that defines that metric. So, for example, considering the discrete metric δ on $\mathbb{R}$, if $\alpha \neq 0$ and $x, y \in \mathbb{R}$ such that $x \neq y$ then $\alpha x \neq \alpha y$, so $\delta(\alpha x, \alpha y) = 1$ but $|\alpha|\delta(x, y) = |\alpha|$. Hence if α is such that $|\alpha| \neq 1$, we have that the homogeneity condition does not hold.

Exercise 4.10. Is the discrete metric translation invariant?

You can now check if the sunflower metric defined in Definition 4.3 or the lift metric from Definition 4.4 arise from a norm on $\mathbb{R}^2$.

Exercise 4.11. Let d be a homogeneous and translation invariant metric (in the sense of Proposition 4.9) on a vector space V, and define a map $||-|| : V \to \mathbb{R}$ by

$$||\boldsymbol{x}|| = d(\boldsymbol{x}, \mathbf{0})$$

for all $\boldsymbol{x} \in V$. Prove that $||-||$ is a norm.

5. More on metric spaces

5.1. **Basic structure of metric spaces.**

Definition 5.1. Let X be a metric space equipped with a distance function d. Let $r > 0$, and let $a \in X$. Then the *open ball* $\mathcal{B}_r(a)$ of radius r around a is defined by

$$\mathcal{B}_r(a) = \{x \in X \mid d(a, x) < r.\}$$

The open ball $\mathcal{B}_r(a)$ is sometimes called an *r-neighbourhood* of the point a. While we shall use mostly open balls, it will at times be convenient to use the following notions as well.

Definition 5.2. The *closed ball* of radius r around a is defined to be

$$\overline{\mathcal{B}}_r(a) = \{x \in X \mid d(a, x) \leq r, \}$$

while the *sphere* of radius r around a is

$$\mathcal{S}_r(a) = \{x \in X \mid d(a, x) = r\}.$$

Example 5.3. Consider the set of real numbers $\mathbb{R}$ equipped with the usual absolute value metric $|-|$. Then an open ball of radius r around a real number a is simply the open interval $(a - r, a + r)$, while the closed ball of radius r around a is the closed interval $[a-r, a+r]$. The sphere of radius r around x is the two-element set $\{a - r, a + r\}$.

In fact, every open interval (a, b) is an open ball around its centre $(a+b)/2$ with radius $(b-a)/2$. Similarly, every closed interval $[a, b]$ can be thought of as a closed ball.

Example 5.4. We have defined a number of different metrics on the space $\mathbb{R}^n$. In $\mathbb{R}^2$ we can easily draw the unit ball (of radius 1 around the origin) corresponding to the metrics d_1, d_2, and d_∞ (the taxi-cab metric, the Euclidean metric, and the max metric, respectively), and the result is as Figure 4.

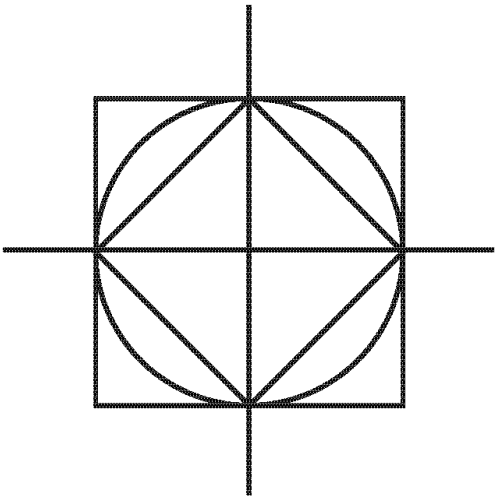


Figure 4: Unit circles in $\mathbb{R}^2$ with respect to the metrics d_1, d_2, and d_∞ (looking from inside out)

Example 5.5. Let X be a space with the discrete metric (see Definition 4.1). Then the distance between any two points is 1 and so the closed ball $\overline{\mathcal{B}}_1(a)$ of radius one around any point $a \in X$ contains all of X. The open ball $\mathcal{B}_1(a)$ of the same radius around a contains only one point, the centre a itself, as this is the only point in X whose distance from a is strictly smaller than 1. Finally, the sphere $\mathcal{S}_1(a)$ of the radius one around a consists of the complement of the set $\{a\}$ in X, that is

$$\mathcal{S}_1(a) = X \setminus \{a\}$$

Example 5.6. Let us consider the space $\mathcal{C}[-1,1]$ of continuous functions on the interval $[-1,1]$ equipped with the metric d_∞ (see Section 3.3). Let f be any function from $\mathcal{C}[-1,1]$ and let $r > 0$. A function $g \in \mathcal{C}[-1,1]$ is an element of the open ball $\mathcal{B}_r(f)$ of radius r around f if and only if $d_\infty(f,g) < r$; that is, when

$$\sup\{|f(x) - g(x)| \mid x \in [-1,1]\} < r.$$

This, in turn, implies that $f(x) - r < g(x) < f(x) + r$ for all $x \in [-1,1]$. We can visualise this set by drawing a strip of width $2r$ around the graph of f as in Figure 5. This is done by shifting the graph of the function f up and down by r. Then a function g belongs to $\mathcal{B}_r(f)$ if and only if its graph falls into this strip.

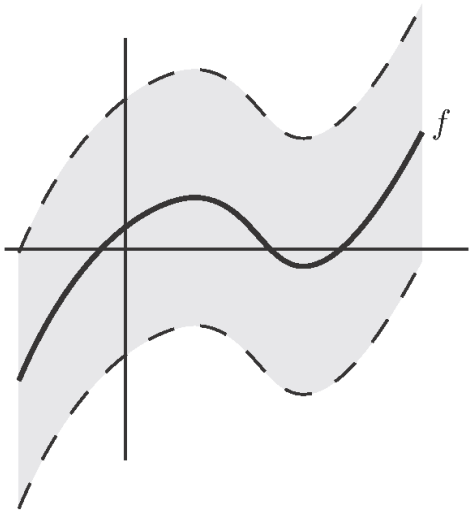


Figure 5: A ball around a function f in the metric space $(\mathcal{C}[a,b], d_\infty)$.

Here is a useful modification of the notions introduced so far.

Definition 5.7. Let (X,d) be a metric space. A *metric subspace* of X is a subset Y of X equipped with the metric d restricted to Y. The resulting metric on Y is called the *induced metric*.

Exercise 5.8. Show that a metric subspace (Y,d), of a metric space (X,d), is a metric space itself.

So, for example, the integers, rational numbers and the irrational numbers are all subspaces of the set of the real numbers equipped with the absolute value metric. It is not difficult to modify the definitions of our basic objects to metric subspaces. Indeed, if $a \in Y \subseteq X$, then the ball $\mathcal{B}_r^Y(a)$ of radius r around a in Y is given by $\mathcal{B}_r^Y(a) = \mathcal{B}_r^X(a) \cap Y$, see Figure 6.

We have to be a bit careful when translating some notions to subspaces. For example, in $\mathbb{R}$, the ball $\mathcal{B}_1(3)$ is given by the interval $(2, 4)$. However, if we take our space Y to be $Y = [2.5, 5) \subset \mathbb{R}$, then the ball $\mathcal{B}_1^Y(3) = [2.5, 4)$.

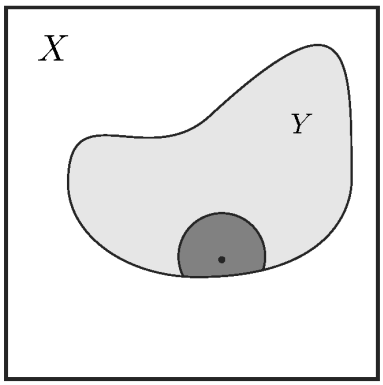


Figure 6: A ball in a subspace Y of a metric space X

Exercise 5.9. Consider two balls $\mathcal{B}_r(x)$ and $\mathcal{B}_s(y)$ in a metric space (X, d). Assume that $\mathcal{B}_r(x) \subseteq \mathcal{B}_s(y)$. Does it follow that $r \leq s$?

5.2. **Diameter.**

Definition 5.10. Let A be a non-empty subset of a metric space (X, d). The *diameter of* A, $\operatorname{diam}(A)$, is defined as

$$\operatorname{diam}(A) = \sup\{d(x, y) \mid x, y \in A\}.$$

By convention, the diameter of the empty set is zero.

We say that a subset $A \subseteq X$ is *bounded* if $\operatorname{diam}(A)$ is finite.

We record some simple properties of the diameter of a set. The first follows directly from the definition.

Proposition 5.11. *If $B \subseteq A$ then $\operatorname{diam}(B) \leq \operatorname{diam}(A)$.*

The following is also an easy consequence of the definition.

Proposition 5.12. *A subset A of a metric space (X, d) is bounded if and only if there exists a point $x \in X$ and a real number r such that $A \subseteq \mathcal{B}_r(x)$.*

Proof. First we note that any ball $\mathcal{B}_r(a)$, where $a \in X$ and $r > 0$, is a bounded set, since for every $x, y \in \mathcal{B}_r(a)$,

$$d(x, y) \leq d(x, a) + d(a, y) < 2r.$$

Thus $\operatorname{diam}(\mathcal{B}_r(a)) \leq 2r$. Any subset of a ball is then bounded, by the previous proposition.

Conversely, let A be bounded and take $x \in A$. Then for every $y \in A$, $d(x, y) \leq \operatorname{diam}(A)$, and so taking $r = \operatorname{diam}(A) + 1$ we have that $A \subseteq \mathcal{B}_r(x)$. □

Exercise 5.13. Let T be a solid triangle in $\mathbb{E}^2$ (i.e., T includes both the boundary and the interior of the triangle). Prove that the diameter of T is equal to the maximum of the side lengths of T.

6. Continuous maps of metric spaces

We have seen examples of isometries, which are distance-preserving maps between metric spaces. In this chapter we will consider a more general class of continuous maps of metric spaces. Here is a very simple example of a map that does not preserve distances.

Take any $n \geq 2$ and let X to be the space $\mathbb{R}^n$ equipped with the d_1-metric (which is defined using the norm $||-||_1$, see Section 3.2), and let Y be the space $\mathbb{R}^n$ equipped with the Euclidean metric $d_2(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Since the set on which the two metrics are defined is the same (in both cases it is $\mathbb{R}^n$), we can consider the identity map $\mathrm{Id} : X \to Y$, that is the map given by $\mathrm{Id}(\boldsymbol{v}) = \boldsymbol{v}$, for all $\boldsymbol{v} \in \mathbb{R}^n$. The map Id is not an isometry in this case. Indeed, let $\boldsymbol{x} = (1, 1, \ldots, 1) \in \mathbb{R}^n$ and let $\mathbf{0} \in \mathbb{R}^n$ be the zero vector, as before. Then

$$d_1(\boldsymbol{x}, \mathbf{0}) = ||\boldsymbol{x} - \mathbf{0}||_1 = \sum_{i=1}^{n} |1 - 0| = n,$$

while

$$\begin{aligned} d_2\big(\mathrm{Id}(\boldsymbol{x}), \mathrm{Id}(\mathbf{0})\big) = d_2(\boldsymbol{x}, \mathbf{0}) &= ||\boldsymbol{x} - \mathbf{0}||_2 \\ &= \left(\sum_{i=1}^{n} |1 - 0|^2\right)^{1/2} = \sqrt{n}. \end{aligned}$$

6.1. Lipschitz maps.

We will begin with maps that distort distances in a controlled way. Let (X, d_X) and (Y, d_Y) be two metric spaces.

Definition 6.1. A map $f : X \to Y$ satisfies the *Lipschitz condition* (we also call such f a *Lipschitz map*) if there exists a constant $L > 0$ such that

$$d_Y(f(x), f(y)) \leq L d_X(x, y), \text{ for all } x, y \in X.$$

Note that every isometry is a Lipschitz map (with $L = 1$), but not conversely, as we will see from the examples below.

Example 6.2. For any $m \in \mathbb{N}$, consider the function $f : [0, 1] \to [0, 1]$ defined by $f(x) = x^m$, $x \in [0, 1]$, where the interval $[0, 1]$ is endowed with the standard absolute value metric from $\mathbb{R}$. Let us show that this function is Lipschitz with the constant $L = m$. Indeed, note that for any $x, y \in [0, 1]$ we have

$$\begin{aligned} x^m - y^m &= (x - y)(x^{m-1} + x^{m-2}y + \cdots + xy^{m-2} + y^{m-1}) \\ &= (x - y) \sum_{i=0}^{m-1} x^{m-1-i} y^i. \end{aligned}$$

Since $|x^{m-1-i}y^i| \leq 1$ for all $i = 0, \ldots, m-1$, when $x, y \in [0, 1]$, we can deduce that

$$\begin{aligned} |x^m - y^m| &= |x - y| \left| \sum_{i=0}^{m-1} x^{m-1-i} y^i \right| \\ &\leq |x - y| \sum_{i=0}^{m-1} |x^{m-1-i} y^i| \leq m|x - y|. \end{aligned}$$

It follows that f is a Lipschitz function with the constant $L = m$.

Exercise 6.3. Show that the function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$ is not Lipschitz, when considered as a map from the entire $\mathbb{R}$ to itself.

Example 6.4. Consider the sine function $\sin : \mathbb{R} \to \mathbb{R}$ as a map of metric spaces, where $\mathbb{R}$ is equipped with the usual absolute value metric $|-|$. Let $x, y \in \mathbb{R}$. Then, by the Mean Value Theorem,

$$|\sin x - \sin y| = |x - y||\cos c|$$

for some $c \in \mathbb{R}$. This implies that $|\sin x - \sin y| \leq |x - y|$, and so the function sin satisfies the Lipschitz condition with $L = 1$. (One does not need to use the Mean Value Theorem to establish this inequality.)

Example 6.5. We have seen that the identity map $\mathrm{Id} : (\mathbb{R}^n, d_1) \to (\mathbb{R}^n, d_2)$ is not an isometry. However, it does satisfy the Lipschitz condition. Indeed, take any $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Then

$$\begin{aligned}(||\boldsymbol{x}||_1)^2 &= \left(\sum_{i=1}^n |x_i|\right)^2 = \sum_{i=1}^n \sum_{j=1}^n |x_i||x_j| \\ &= \sum_{i=1}^n |x_i|^2 + \sum_{i \neq j} |x_i||x_j| \\ &\geq \sum_{i=1}^n |x_i|^2 = (||\boldsymbol{x}||_2)^2.\end{aligned}$$

Thus $||\boldsymbol{x}||_2 \leq ||\boldsymbol{x}||_1$ for all $\boldsymbol{x} \in \mathbb{R}^n$. It follows that for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

$$d_2(\mathrm{Id}(\boldsymbol{x}), \mathrm{Id}(\boldsymbol{y})) = d_2(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2 \leq ||\boldsymbol{x} - \boldsymbol{y}||_1 = d_1(\boldsymbol{x}, \boldsymbol{y}).$$

In other words, the identity map Id is Lipschitz with $L = 1$.

Example 6.6. A more interesting example of a Lipschitz map may be obtained as follows. We consider the space $\mathcal{C}[0, 1]$ of continuous functions on the closed interval $[0, 1]$ equipped with the d_∞-metric. Define a map $\Phi : \mathcal{C}[0, 1] \to \mathcal{C}[0, 1]$ by

$$\Phi(f)(x) = \int_0^x f(t)\, dt$$

for all $f \in \mathcal{C}[0, 1]$. The Fundamental Theorem of Calculus (Theorem 9.23 in the Analysis notes) guarantees that the function $\Phi(f)$ is differentiable, and so, in particular, it is continuous on $[0, 1]$.

We will show that this map Φ is Lipschitz. For this, we need to prove that there exists a constant $L > 0$ such that for all $f, g \in \mathcal{C}[0, 1]$

$$d_\infty(\Phi(f), \Phi(g)) \leq L d_\infty(f, g).$$

For that, we need to understand how to estimate the distance on the left of this inequality. That is, we need to consider

$$d_\infty(\Phi(f), \Phi(g)) = \sup\{|\Phi(f)(x) - \Phi(g)(x)| \mid x \in [0, 1]\}.$$

In order to find an estimate for this expression, we shall first consider the value $|\Phi(f)(x) - \Phi(g)(x)|$ for all $x \in [0, 1]$.

First, for all $x \in [0, 1]$ and for every two functions $f, g \in \mathcal{C}[0, 1]$ we have

$$|\Phi(f)(x) - \Phi(g)(x)| = \left| \int_0^x (f(t) - g(t))\, dt \right|.$$

Secondly, we recall for all $h \in \mathcal{C}[a, b]$

$$\left| \int_a^b h(t)\, dt \right| \leq \int_a^b |h(t)|\, dt;$$

this was proved in Theorem 9.20 and Exercise 9.21 of the Analysis notes.

Hence, for all $x \in [0,1]$ we have

$$|\Phi(f)(x) - \Phi(g)(x)| = \left| \int_0^x (f(t) - g(t))\, dt \right| \leq \int_0^x |f(t) - g(t))|\, dt.$$

And since $|f(t) - g(t))| \leq d_\infty(f,g)$ for all $t \in [0,1]$, we can write

$$|\Phi(f)(x) - \Phi(g)(x)| \leq \int_0^x d_\infty(f,g)\, dt = x\, d_\infty(f,g) \leq d_\infty(f,g).$$

As this is valid for all $x \in [0,1]$, we can now take the supremum of the left-hand side over all $x \in [0,1]$ to get

$$d_\infty(\Phi(f), \Phi(g)) \leq d_\infty(f,g).$$

Hence the map Φ is a Lipschitz map with $L = 1$.

The same argument, with minor changes, works for the space of functions $\mathcal{C}[a,b]$ (with $0, 1$ replaced by some $a, b \in \mathbb{R}$, $a < b$). In this case we get

$$d_\infty(\Phi(f), \Phi(g)) \leq (b-a)\, d_\infty(f,g).$$

This again is a Lipschitz function, with $L = b - a$.

Example 6.7. Let us consider again the space of continuous functions $\mathcal{C}[0,1]$ with the supremum metric d_∞ and define a map $F : \mathcal{C}[0,1] \to \mathbb{R}$ by evaluation at a particular point, say $\frac{1}{2}$. Thus, for every $f \in \mathcal{C}[0,1]$, we define

$$F(f) = f(\tfrac{1}{2}).$$

Then for every two functions $f, g \in \mathcal{C}[0,1]$ we have

$$|F(f) - F(g)| = |f(\tfrac{1}{2}) - g(\tfrac{1}{2})| \leq d_\infty(f,g).$$

Thus the map F is also a Lipschitz map with Lipschitz constant $L = 1$.

6.2. **Continuity.**

Definition 6.8. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is *continuous at a point* $a \in X$ if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for any $x \in X$ with $d_X(x,a) < \delta$ we have $d_Y(f(x), f(a)) < \varepsilon$.

In other words, for every positive real number ε there exists a positive real number δ such that for every $x \in X$

$$x \in \mathcal{B}^X_\delta(a) \;\Rightarrow\; f(x) \in \mathcal{B}^Y_\varepsilon(f(a)),$$

where $\mathcal{B}^X$ and $\mathcal{B}^Y$ denote the open balls in X and Y respectively. We say that f is *continuous* if it is continuous at every point $a \in X$.

Remark 6.9. Consider $\mathbb{R}$, endowed with the usual absolute value norm and metric. It is easy to see that a function $f : \mathbb{R} \to \mathbb{R}$ is continuous at a point $x \in \mathbb{R}$ in the sense of Definition 6.8 if and only if this function is continuous at x in the standard sense (see Definition 6.5 from the Analysis notes).

The following theorem introduces a rich source of examples of continuous maps of metric spaces.

Theorem 6.10. *Let (X, d_X) and (Y, d_Y) be metric spaces and let $f : X \to Y$ be a Lipschitz map with constant $L > 0$. Then f is continuous.*

Proof. Let $a \in X$ and $\varepsilon > 0$. Since for every $x \in X$,

$$d_Y(f(x), f(a)) \leq L d_X(x,a),$$

we can take $\delta = \varepsilon / L$ to ensure that if $d_X(x,a) < \delta$ then $d_Y(f(x)), f(a)) < \varepsilon$. $\square$

It follows that maps considered in Examples 6.4, 6.5, 6.6, and 6.7 are all continuous. In fact the Lipschitz condition provides a stronger control over continuity than that required by Definition 6.8. This is because the δ selected in the proof above only depends on ε and will work for all $x \in X$. We say that functions that satisfy the Lipschitz condition are uniformly continuous, in the following sense.

Definition 6.11. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is *uniformly continuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in X$, if $d_X(x, y) < \delta$ then $d_Y(f(x), f(y)) < \varepsilon$.

Theorem 6.12. *Let (X, d_X), (Y, d_Y) and (Z, d_Z) be metric spaces, and let $f : X \to Y$ and $g : Y \to Z$ be continuous maps. Then the composite map $g \circ f : X \to Z$ is continuous.*

Proof. Let $x \in X$ and take any $\varepsilon > 0$. Since the map $g : Y \to Z$ is continuous at $f(x) \in Y$, there exists a $\rho > 0$ such that for all $t \in Y$ with $d_Y(t, f(x)) < \rho$, we have that $d_Z(g(t), g(f(x))) < \varepsilon$. As the map $f : X \to Y$ is also continuous, there exists a $\delta > 0$ such that for all $s \in X$, if $d_X(s, x) < \delta$ then $d_Y(f(s), f(x)) < \rho$.

Combined together, the above two inequalities imply that if $d_X(y, x) < \delta$ then

$$d_Z\big((g \circ f)(s), (g \circ f)(x))\big) = d_Z(g(f(s)), g(f(x))) < \varepsilon,$$

and so the composition $g \circ f : X \to Z$ is continuous at x. The latter works for all $x \in X$, hence $g \circ f$ is continuous on X. □

Definition 6.13. A *homeomorphism* of metric spaces is a bijection $f : X \to Y$ such that both f and the inverse map $f^{-1} : Y \to X$ are continuous.

Two metric spaces X and Y are said to be *homeomorphic* if there exists a homomorphism $f : X \to Y$ between them.

Example 6.14. Any two closed intervals $[a, b]$, $[c, d]$ in $\mathbb{R}$ are homeomorphic, provided $a < b$ and $c < d$. This can be seen using a simple affine map $f : [a, b] \to [c, d]$, defined by

$$f(x) = \frac{d - c}{b - a} x + \frac{cb - da}{b - a}, \text{ for all } x \in [a, b].$$

This map also restricts to provide a homeomorphism between any two non-empty open intervals (a, b) and (c, d) in $\mathbb{R}$.

Proposition 6.15. *The composition of two homeomorphisms is a homeomorphism. More precisely, if $f : X \to Y$ and $g : Y \to Z$ are homeomorphisms of metric spaces (X, d_X), (Y, d_Y) and (Z, d_Z) then $g \circ f : X \to Z$ is also a homeomorphism.*

Proof. Since a composition of bijections is again a bijection, it is clear that $g \circ f$ is a bijection from X to Z. By the assumptions, the maps f, f^{-1}, g and g^{-1} are continuous, hence, in view of Theorem 6.12, the compositions $g \circ f$ and $f^{-1} \circ g^{-1}$ are also continuous. It remains to observe that $f^{-1} \circ g^{-1} = (g \circ f)^{-1}$, so $g \circ f : X \to Z$ is a homeomorphism, as required. □

Example 6.16. The open interval $(-\pi/2, \pi/2)$ is homeomorphic to the space $\mathbb{R}$ with the homeomorphism provided by the tangent map, which is invertible on this interval:

$$\tan : (-\pi/2, \pi/2) \to \mathbb{R}$$

Combining this with the first example, and using Proposition 6.15, we conclude that any non-empty open interval is homeomorphic to the real line.

It is not always easy to determine if two spaces are homeomorphic. For instance, is the open interval $(0, 1)$ homeomorphic to the closed interval $[0, 1]$? Or the half-open interval $[0, 1)$? Is the real line $\mathbb{R}$ homeomorphic to the plane $\mathbb{R}^2$? A lot of

questions of this kind can be answered using a generalized version of the intermediate value theorem, which relies on the notion of connectedness that we will discuss later in the course.

The following class of maps provides many examples of homeomorphisms of metric spaces.

Definition 6.17. A map $f : X \to Y$ of metric spaces is *bi-Lipschitz* if there exists a constant $K \geq 1$ such that for all $x, y \in X$

$$\frac{1}{K} d_X(x, y) \leq d_Y(f(x), f(y)) \leq K d_X(x, y). \tag{6.18}$$

In particular, a bi-Lipschitz map is Lipschitz.

Theorem 6.19. *Let $f : X \to Y$ be a surjective bi-Lipschitz map of metric spaces. Then f is a homeomorphism.*

Proof. We first need to show that f is injective, which (since we assumed that it is surjective) will imply that f has an inverse. So let $f(x) = f(y)$ for some $x, y \in X$. Then $d_Y(f(x), f(y)) = 0$ and by the left hand inequality in Definition 6.17 we have

$$\frac{1}{K} d_X(x, y) \leq d_Y(f(x), f(y)) = 0$$

which implies $d_X(x, y) = 0$, so that $x = y$.

Thus f is bijective, so there exists an inverse function $f^{-1} : Y \to X$. We need to show that both f and f^{-1} are continuous. Given that f is a Lipschitz map, it is continuous by Theorem 6.10.

Now consider arbitrary $s, t \in Y$, and let $x, y \in X$ be such that $x = f^{-1}(s)$ and $y = f^{-1}(t)$. Then, by the left-hand inequality in (6.18), we have

$$\begin{aligned} \frac{1}{K} d_X(f^{-1}(s), f^{-1}(t)) = \frac{1}{K} d_X(x, y) &\leq d_Y(f(x), f(y)) \\ &= d_Y(f(f^{-1}(s)), f(f^{-1}(t))) = d_Y(s, t), \end{aligned}$$

which is equivalent to

$$d_X(f^{-1}(s), f^{-1}(t)) \leq K d_Y(s, t), \ \forall\, s, t \in Y.$$

This means that f^{-1} is a Lipschitz map, and so it is continuous by Theorem 6.10. Thus f is a homeomorphism. □

7. Sequences in metric spaces

7.1. **Limit of a sequence.**

Let (X, d) be a metric space. A sequence in X is any map $a : \mathbb{N} \to X$ from the natural numbers to X. The value of the function a at $n \in \mathbb{N}$ is normally denoted $a_n \in X$ and called the *general term* of the sequence. A sequence is normally denoted $\{a_n\}$ (or $\{a_n\}_{n\in\mathbb{N}}$, or (a_n), or $(a_n)_{n\in\mathbb{N}}$, etc.). A *subsequence* of the sequence $\{a_n\}$ is a restriction of the function a to an infinite subset of $\mathbb{N}$; we can think of a subsequence as being obtained from a sequence $\{a_n\}$ by throwing out some of its terms. A subsequence is typically denoted $\{a_{n_k}\}$, where $\{n_k\}$, $k \in \mathbb{N}$, is a strictly increasing sequence of natural numbers (see Chapter 4 of the Analysis notes).

Definition 7.1. We say that a sequence $\{x_n\}$ in a metric space X *converges to a point* $x \in X$ if for every $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that for all $n > K$,

$$d(x_n, x) < \varepsilon.$$

In this case we write

$$\lim_{n\to\infty} x_n = x,$$

or $x_n \to x$ as $n \to \infty$.

Equivalently, we can say that x_n converges to x if for every positive ε the open ball $\mathcal{B}_\varepsilon(x)$ contains all but finitely many terms of the sequence.

If a sequence $\{x_n\}$ has no limit we shall say that it *diverges*. It is important to note that the convergence or divergence of a sequence depends on the set X and on the metric d.

Consider the sequence $\{\frac{1}{n}\}$ in the metric space $\mathbb{R}$ equipped with the standard metric; it converges to 0. However, if we consider the same sequence but take the space X to be the open interval $(0, 1)$ (with the induced metric from $\mathbb{R}$) then it does not converge, as 0 is then not an element of that space.

Keeping the same sequence $\{\frac{1}{n}\}$ in $\mathbb{R}$ but changing the metric to the discrete metric δ defined in Definition 4.1 (recall that it can be defined on any set) we see that the sequence does not converge. This is because for every two distinct elements of the sequence we have that $\delta(\frac{1}{n}, \frac{1}{m}) = 1$.

Proposition 7.2. *A convergent sequence has a unique limit.*

Proof. Assume that a sequence $\{x_n\}$ in a metric space (X, d) converges and has two distinct limits x and x' in X. Then $d(x, x') > 0$, so we can choose $\varepsilon = d(x, x')/2 > 0$. Then there must exist $K \in \mathbb{N}$ such that for all $n > K$, $d(x_n, x) < \varepsilon$ and $d(x_n, x') < \varepsilon$. Thus for every $n > K$ the triangle inequality yields

$$d(x, x') \le d(x, x_n) + d(x_n, x') < 2\varepsilon = d(x, x'),$$

which is a contradiction. Therefore the sequence $\{x_n\}$ can have at most one limit in X. □

Definition 7.3. A sequence $\{x_n\}$ in a metric space X is bounded if the set $\{x_n \mid n \in \mathbb{N}\}$, of its terms, is a bounded subset of X.

Proposition 7.4. *Every convergent sequence $\{x_n\}$ in a metric space (X, d) is bounded.*

Proof. If x is the limit of the sequence $\{x_n\}$ and $\varepsilon = 1 > 0$, then the ball $\mathcal{B}_1(x)$, in X, contains all terms x_n with $n > K$, for some fixed number $K \in \mathbb{N}$. Let $R \ge 0$ be the maximum of the (finitely many) numbers $d(x_i, x)$, where $i = 1, \dots, K$, and take $r = R + 1 \ge 1$. Then the ball $\mathcal{B}_r(x)$ contains all terms of the sequence $\{x_n\}$, which is therefore bounded. □

7.2. **Convergence in the space of continuous functions.**

Convergence of sequences of real numbers is at the centre of Real Analysis. Here we will discuss in more detail convergence in the space of continuous functions $\mathcal{C}[a,b]$ equipped with the supremum metric d_∞ (see Section 3.3). Our main result here will be a classical theorem of Weierstrass that every continuous function on a closed interval can be approximated by polynomials.

Let $[a,b]$ be a non-empty interval in $\mathbb{R}$. Restating the definitions of convergence of a sequence in metric space in the case of the particular metric space $(\mathcal{C}[a,b], d_\infty)$ we can say that a sequence $\{f_n\}$ of functions from $\mathcal{C}[a,b]$ converges to $f \in \mathcal{C}[a,b]$, with respect to the metric d_∞, if

$$\forall\, \varepsilon > 0\ \exists\, N \in \mathbb{N} \text{ such that } d_\infty(f_n, f) < \varepsilon,\ \forall\, n > N.$$

Using the definition of the supremum metric, we can rewrite the last condition as follows:

$$d_\infty(f_n, f) < \varepsilon \ \Leftrightarrow\ \sup\{|f_n(x) - f(x)| \mid x \in [a,b]\} < \varepsilon.$$

The latter is equivalent to the existence of some ε_0, $0 < \varepsilon_0 < \varepsilon$, such that

$$|f_n(x) - f(x)| \leq \varepsilon_0 \text{ for all } x \in [a,b].$$

It is important to realise that this condition is required to hold for all $x \in [a,b]$ as soon as $n > N$, where N is determined by the choice of ε only. In Real Analysis this kind of convergence of a sequence of functions is called the *uniform convergence*; if this condition holds, we say that the sequence $\{f_n\}$ of functions converges *uniformly* to f and write $f_n \xrightarrow{u} f$, as $n \to \infty$.

This is contrasted with *pointwise* convergence. We say that a sequence f_n of functions in $[a,b]$ converges to a function f *pointwise* if for every $x \in [a,b]$ the sequence of real numbers $\{f_n(x)\}$ converges to the number $f(x)$. Thus convergence at a given x is determined without reference to any other points of the interval $[a,b]$. If the sequence $\{f_n\}$ converges pointwise to f we write $f_n \xrightarrow{p} f$, as $n \to \infty$.

We are now in a position to build on the notions and facts introduced so far to state a very important classical theorem of Weierstrass. For convenience, we will consider the space $(\mathcal{C}[0,1], d_\infty)$, but the result holds for the space of continuous functions on any closed interval $[a,b]$.

A subset A of a metric space (X,d) is *dense* if for every $x \in X$ and each $\varepsilon > 0$ there exists $a \in A$ such that $d(a,x) < \varepsilon$ (cf. Definition 11.17 below).

Theorem 7.5. *The space of polynomials on $[0,1]$ is dense in the space $(\mathcal{C}[0,1], d_\infty)$. That is, for every continuous function $f \in \mathcal{C}[0,1]$ and for every $\varepsilon > 0$ there exists a polynomial p such that*

$$d_\infty(f,p) = ||f - p||_\infty < \varepsilon.$$

At the heart of this theorem is an approximation statement: every continuous function on the unit interval can be approximated with an arbitrary accuracy by a polynomial. An important proof of this theorem was provided by Bernstein, which is constructive, in that it exhibits a sequence of polynomials that can be used to approximate a given continuous function. For a continuous function f, a *Bernstein polynomial of degree n* is given by the formula

$$p_n(x) = \sum_{k=0}^{n} \binom{n}{k} f(k/n) x^k (1-x)^k.$$

Because of the importance of this result, the theorem deserves a separate statement.

Theorem 7.6. *Let f be a continuous function in $\mathcal{C}[0,1]$. Then the sequence of Bernstein polynomials $\{p_n\}$, of f, converges uniformly to f.*

7.3. **Sequences and continuity.**

We now want to connect the earlier definition of continuity of a map with an equivalent property defined in terms of sequences, which is sometimes useful in applications.

Theorem 7.7. *Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is continuous at a point $x_0 \in X$ if, and only if, for every sequence $\{x_n\}$, of points of X, converging to x_0 we have $\lim_{n\to\infty} f(x_n) = f(x_0)$.*

Proof. "⇒" Let $\{x_n\}$ be a sequence of points of X converging to x_0. Since f is continuous, for every $\varepsilon > 0$ there exists $\delta > 0$ such that if $d_X(x, x_0) < \delta$ then $d_Y(f(x), f(x_0)) < \varepsilon$. As x_n converges to x_0, there exists $N \in \mathbb{N}$ such that for all $n > N$, $d_X(x_0, x_n) < \delta$, which implies that for all $n > N$, $d_Y(f(x_n), f(x_0)) < \varepsilon$. Therefore

$$\lim_{n\to\infty} f(x_n) = f(x_0).$$

"⇐" Conversely, let us assume that f is not continuous at x_0. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$ there exists $x_n \in X$ which satisfies the inequalities

$$d_X(x_n, x_0) < \frac{1}{n} \text{ and } d_Y(f(x_n), f(x_0)) \geq \varepsilon.$$

But this means that we have a sequence $\{x_n\}$ converging to x_0, such that the sequence $\{f(x_n)\}$ does not converge to $f(x_0)$. □

It is important to remember that the notion of a limit of a sequence in a metric space depends on the chosen metric. Similarly, the choice of a metric is crucial in the definition of continuity. Let us investigate this a bit more.

Example 7.8. Consider the space of continuous functions $\mathcal{C}[0, 1]$ on the unit interval, and equip it with two different metrics: the d_∞-metric, coming from the supremum norm from Definition 3.7 and the d_1-metric coming from the L^1-norm from Definition 3.9 (see Section 3.3).

In Example 6.7 we looked at the *evaluation function* $F : \mathcal{C}[0, 1] \to \mathbb{R}$ which sends each function $f \in \mathcal{C}[0, 1]$ to its value at $1/2$. (The choice of $1/2$ here is not important, there is an evaluation function for every point $x \in [0, 1]$.) We proved that this map

$$F : (\mathcal{C}[0, 1], d_\infty) \to \mathbb{R}$$

is Lipschitz, and hence continuous, when the space $\mathcal{C}[0, 1]$ is equipped with the d_∞-metric. However, this is not true when we consider the same space but measure distances using the metric d_1. To prove this, we need to find a function $f \in \mathcal{C}[0, 1]$ and a sequence $f_n \to f$ which converges to f in this metric, but is such that the sequence of values $F(f_n)$ of the function F does not converge to $F(f)$.

Let us take $f = 1$, the constant function that takes the value 1 for all $x \in [0, 1]$. For every $n \in \mathbb{N}$ we define the function $f_n : [0, 1] \to \mathbb{R}$ as follows:

$$f_n(x) = \begin{cases} 1, & 0 \leq x \leq \frac{1}{2} - \frac{1}{n} \\ n(\frac{1}{2} - x), & \frac{1}{2} - \frac{1}{n} < x \leq \frac{1}{2} \\ n(x - \frac{1}{2}), & \frac{1}{2} \leq x < \frac{1}{2} + \frac{1}{n} \\ 1, & \frac{1}{2} + \frac{1}{n} \leq x \leq 1. \end{cases}$$

While the definition might be complicated, the basic idea is not, and is illustrated in Figure 7: the function f_n is a constant in the interval $[0, 1]$ except in a small neighbourhood $(\frac{1}{2} - \frac{1}{n}, \frac{1}{2} + \frac{1}{n})$ around $\frac{1}{2}$ where it decreases linearly to 0 at $\frac{1}{2}$ and then increases to 1, again in a linear fashion.

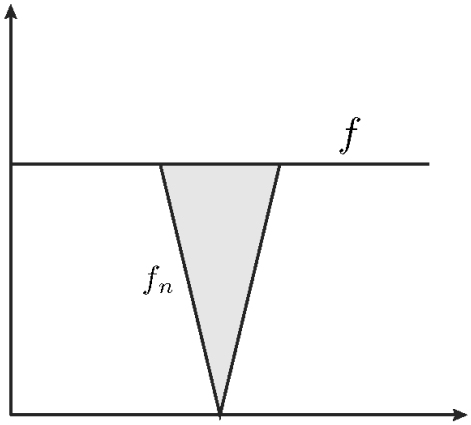


Figure 7: The sequence f_n converges to f in the metric d_1.

We will prove that $f_n \to f$, as $n \to \infty$ in the metric d_1. For this we compute the distance between the two functions:

$$d_1(f_n, f) = ||f_n - f||_1 = \int_0^1 |f_n(x) - f(x)|\, dx$$
$$= \int_{\frac{1}{2}-\frac{1}{n}}^{\frac{1}{2}+\frac{1}{n}} |f_n(x) - f(x)|\, dx = \frac{1}{n}$$

The distance between the two functions is simply the area of the triangle shaded grey in Figure 7, and it is easy to see directly that it equals $\frac{1}{n}$, as stated. As this tends to 0, as $n \to \infty$, we have that $f_n \to f$ in the metric d_1.

However, $F(f) = 1$, but $F(f_n) = 0$ for all n, so the sequence $F(f_n)$ does not converge to $F(f)$. In view of Theorem 7.7, we conclude that the function F is not continuous when the space $\mathcal{C}[0,1]$ is equipped with the metric d_1.

7.4. **Cauchy sequences and completeness.**

Definition 7.9. A sequence $\{x_n\}$ of points in a metric space (X, d) is called a *Cauchy sequence* if for every $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that for all $m, n \in \mathbb{N}$ with $m, n > K$, we have $d(x_n, x_m) < \varepsilon$.

In the case of sequences of real numbers the next proposition was proved in the Analysis module (see Theorem 4.25 in the Analysis notes).

Proposition 7.10. *Every convergent sequence in a metric space (X, d) is a Cauchy sequence.*

Proof. Suppose that $x_n \to x$, as $n \to \infty$. For all $m, n \in \mathbb{N}$ the triangle inequality gives

$$d(x_m, x_n) \leq d(x_n, x) + d(x, x_m).$$

Let $\varepsilon > 0$. Given that the sequence x_n converges to x, there exists $K \in \mathbb{N}$ such that for all $n > K$ and $m > K$, $d(x_n, x) < \varepsilon/2$ and $d(x, x_m) < \varepsilon/2$. This means that, for all $m, n > K$,

$$d(x_m, x_n) < \varepsilon,$$

and so the sequence $\{x_n\}$ is Cauchy. □

The converse of this statement is not true for general metric spaces. The simplest example is probably the following.

Example 7.11. Let $X = (0,1)$ considered as a subspace of $\mathbb{R}$ with the standard metric. Consider again the sequence $\{1/n\}$ in X. This sequence converges to 0 in $\mathbb{R}$, and so it is a Cauchy sequence in $\mathbb{R}$ by Proposition 7.10. But 0 is not an element of $(0,1)$, so this sequence does not converge in $(0,1)$ (e.g., by Proposition 7.2).

Definition 7.12. A metric space (X,d) is called *complete* if every Cauchy sequence in X converges to a point in X.

From your Analysis module, you know that $\mathbb{R}$, equipped with the standard metric, is complete (see Theorem 5.8 from the Analysis notes). However, this is no longer true for the subspace $\mathbb{Q}$, of rational numbers, in $\mathbb{R}$.

Example 7.13. The space $\mathbb{Q}$, equipped with the induced metric from $\mathbb{R}$, is not complete. To see this, consider the sequence of positive rational numbers constructed as follows. For every n, there exists a positive rational number q_n such that $|2 - q_n^2| < 1/n$. Let us check that the sequence $\{q_n\}$ is Cauchy. Of course we can also assume that $1 \leq q_n$ for all $n \in \mathbb{N}$. Then, for all $n, m \in \mathbb{N}$, we have

$$|q_n^2 - q_m^2| = |q_n - q_m||q_n + q_m| \geq 2|q_n - q_m|.$$

On the other hand,

$$|q_n^2 - q_m^2| \leq |q_n^2 - 2| + |q_m^2 - 2| \leq \tfrac{1}{n} + \tfrac{1}{m} \leq 2\tfrac{1}{\min\{m,n\}}.$$

Thus

$$|q_n - q_m| \leq \tfrac{1}{\min\{m,n\}}.$$

Now let $\varepsilon > 0$ and choose $K \in \mathbb{N}$ so that $1/K < \varepsilon$. Then for all $n, m \geq K$,

$$|q_n - q_m| \leq \tfrac{1}{\min\{m,n\}} \leq \tfrac{1}{K} < \varepsilon.$$

Thus $\{q_n\}$ is a Cauchy sequence of rational numbers. But the sequence $\{q_n\}$ does not converge in $\mathbb{Q}$.

Indeed, assume that

$$\lim_{n\to\infty} q_n = \frac{a}{b} \in \mathbb{Q}, \quad \text{where } a \in \mathbb{Z} \text{ and } b \in \mathbb{N}.$$

If $|\frac{a^2}{b^2} - 2| = r > 0$, then there exists $K \in \mathbb{N}$ such that for all $n > K$, $|q_n^2 - 2| < r/2$ and $|\frac{a^2}{b^2} - q_n^2| < r/2$. Then

$$|\tfrac{a^2}{b^2} - 2| \leq |q_n^2 - 2| + |\tfrac{a^2}{b^2} - q_n^2| < r = |\tfrac{a^2}{b^2} - 2|,$$

leading to a contradiction. Hence $\frac{a^2}{b^2} = 2$, which is impossible in $\mathbb{Q}$, as $\sqrt{2} \notin \mathbb{Q}$.

We will now introduce another important example of a complete space.

7.5. Completeness of $\mathcal{C}[a,b]$.

Let $[a,b]$ be a non-empty interval in $\mathbb{R}$.

Theorem 7.14. *The space $\mathcal{C}[a,b]$ of continuous functions on $[a,b]$ is complete when regarded as a metric space with the metric d_∞.*

Proof. Let $\{f_n\}$ be a Cauchy sequence of functions in $\mathcal{C}[a,b]$. This means that for every $\varepsilon > 0$ we can find a natural number N such that for all $n, m > N$, we have

$$d_\infty(f_n, f_m) < \varepsilon.$$

As we pointed out above, this can be restated by saying that for any $\varepsilon > 0$ and all $n, m > N$ the following inequality holds

$$|f_n(x) - f_m(x)| < \varepsilon, \quad \text{for all } x \in [a,b]. \tag{7.15}$$

If we now choose a particular $x \in [a,b]$, then (7.15) shows that for each $x \in [a,b]$ the sequence $\{f_n(x)\}$, of real numbers, is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, this sequence has a limit, which we denote $f(x)$: $f(x) = \lim\limits_{n\to\infty} f_n(x)$.

This can be done for every $x \in [a,b]$ and so defines a function $f : [a,b] \to \mathbb{R}$. We need to show that $f_n \xrightarrow{u} f$ as $n \to \infty$, and that f is continuous.

Take any $\varepsilon > 0$, then, since the sequence of functions $\{f_n\}$ is Cauchy in the metric space $(\mathcal{C}[a,b], d_\infty)$, there exists $K \in \mathbb{N}$ such that

$$|f_n(x) - f_m(x)| < \varepsilon/2, \quad \text{for all } x \in [a,b], \text{ provided } m,n > K. \tag{7.16}$$

Let us now show that if $n > K$ then $|f_n(x) - f(x)| < \varepsilon$ for all $x \in [a,b]$. Indeed, given any $x \in [a,b]$, $\lim\limits_{m\to\infty} f_m(x) = f(x)$, hence there is $M \in \mathbb{N}$ (which will, in principle, depend on x) such that $|f_m(x) - f(x)| < \varepsilon/2$ for all $m > M$. Now we set $l = \max\{K, M\} + 1 \in \mathbb{N}$. Then $l > M$, so $|f_m(x) - f(x)| < \varepsilon/2$; moreover, since $l > K$, in view of (7.16) we have $|f_n(x) - f_l(x)| < \varepsilon/2$ whenever $n > K$. The triangle inequality implies

$$|f_n(x) - f(x)| \le |f_n(x) - f_l(x)| + |f_l(x) - f(x)| \le \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon,$$

provided $n > K$. Since the above is true for an arbitrary $x \in [a,b]$, as long as $n > K$, we can conclude that the sequence $\{f_n\}$ uniformly converges to f. As we know, the latter is equivalent to $f_n \to f$ as $n \to \infty$ in the metric space $(\mathcal{C}[a,b], d_\infty)$.

We now need to show that f is continuous. This was actually proved in your Analysis module: see Theorem 11.3 from the Analysis notes. For convenience, we reproduce the argument below.

Recall that a function $f : [a,b] \to \mathbb{R}$ is continuous at a point $x \in [a,b]$ if and only if for every sequence $\{x_n\}$, converging to x in $\mathbb{R}$, the sequence $f(x_n)$ converges to $f(x)$ (see Theorem 7.7).

For arbitrary $x \in [a,b]$ and $n,m \in \mathbb{N}$ the triangle inequality gives

$$\begin{aligned} |f(x) - f(x_n)| \le |f(x) - f_m(x)| + |f_m(x) - f_m(x_n)| \\ + |f_m(x_n) - f(x_n)|. \end{aligned}$$

Given that x and x_n, for all $n \in \mathbb{N}$, are points of the interval $[a,b]$ and $f_n \xrightarrow{u} f$ on that interval, the first and third terms in this inequality can be made smaller than any $\varepsilon/3$, for all large enough $m \in \mathbb{N}$ and all $x \in [a,b]$. Assume that we have chosen such an m. Now each f_m is continuous, so for all n greater than some natural number N, the middle term will be smaller than $\varepsilon/3$ (since $\lim\limits_{n\to\infty} f_m(x_n) = f_m(x)$ by Theorem 7.7). Putting all of this together, we see that for all $n > N$

$$|f(x) - f(x_n)| < \varepsilon,$$

thus $\lim\limits_{n\to\infty} f(x_n) = f(x)$. Since this holds for every $x \in [a,b]$ and each sequence $\{x_n\}$, converging to x, we can conclude that f is continuous. $\square$

8. Contraction Mapping Theorem

In this Chapter we introduce a beautiful application that brings together various ideas developed so far to provide a simple and powerful result.

8.1. **Contractions.**

We assume that (X, d) is a metric space.

Definition 8.1. A function $f : X \to X$ is a *contraction* if there exists a real number k, $0 < k < 1$, such that for all $x, y \in X$

$$d(f(x), f(y)) \leq kd(x, y).$$

Thus a contraction is, in particular, a Lipschitz map, and so it is continuous by Theorem 6.10

The main result of this section is the following.

Theorem 8.2 (Contraction Mapping Theorem)**.** *Let (X, d) be a complete metric space and let $f : X \to X$ be a contraction. Then f has a unique fixed point in X. In other words, there exists a unique point $x \in X$ such that $f(x) = x$.*

Proof. An important part of the proof is the following iteration procedure. Let x_0 be any point of X, and let $x_1 = f(x_0)$, $x_2 = f(x_1)$, and so on. Thus we obtain a sequence $\{x_n\}$ of points of X where

$$x_n = f(x_{n-1}) = \underbrace{(f \circ \cdots \circ f}_{n})(x_0).$$

Using the definition of a contraction it is not difficult to estimate the distance between consecutive terms in the sequence. Indeed, we have

$$d(x_2, x_1) = d(f(x_1), f(x_0)) \leq kd(x_1, x_0),$$

while

$$d(x_3, x_2) = d(f(x_2), f(x_1)) \leq kd(x_2, x_1) \leq k^2 d(x_1, x_0).$$

This suggests that in general

$$d(x_n, x_{n+1}) \leq k^n d(x_1, x_0), \quad \text{for all } n \in \mathbb{N} \cup \{0\}, \tag{8.3}$$

which we can easily check by induction:

$$d(x_{n+1}, x_n) = d(f(x_n), f(x_{n-1})) \leq kd(x_n, x_{n-1}) \leq k^n d(x_1, x_0).$$

Given that $k < 1$ this implies that the distance between consecutive terms decreases to zero as n tends to infinity. Indeed, we can prove the following

Lemma 8.4. *The sequence $\{x_n\}$ constructed by the above iterative procedure is a Cauchy sequence.*

Proof. We need to estimate the distance between arbitrary terms x_m and x_n of the sequence. Let us assume that $m < n$ and write $n = m+p$, for some natural number $p > 0$. Then, in view of (8.3), we have

$$\begin{aligned} d(x_m, x_{m+p}) &\leq d(x_m, x_{m+1}) + \cdots + d(x_{m+p-1}, x_{m+p}) \\ &\leq (k^m + k^{m+1} + \cdots + k^{m+p-1}) d(x_1, x_0) \\ &\leq d(x_1, x_0) k^m \sum_{i=0}^{\infty} k^i. \end{aligned}$$

We have assumed that $0 < k < 1$, therefore the geometric series $\sum_{i=0}^{\infty} k^i$ converges to $1/(1-k)$ and $k^m \to 0$, as $m \to \infty$. This implies that

$$d(x_m, x_{m+p}) \leq \frac{d(x_1, x_0)}{1-k} k^m \to 0, \text{ as } m \to \infty. \tag{8.5}$$

Thus for every $\varepsilon > 0$ there exists $K \in \mathbb{N}$ such that for all $m > K$ and $p \in \mathbb{N}$,

$$d(x_m, x_{m+p}) < \varepsilon.$$

This proves that $\{x_n\}$ is a Cauchy sequence. □

We have assumed that X is a complete metric space, therefore the Cauchy sequence $\{x_n\}$ will have a limit x in X:

$$x = \lim_{n\to\infty} x_n = \lim_{n\to\infty} f(x_{n-1}).$$

We claim that x is a fixed point of f. Indeed, since $\{x_n\}$ tends to x, as $n \to \infty$, and f is continuous (as it is a contraction), we know, by Theorem 7.7, that

$$f(x) = \lim_{n\to\infty} f(x_n) = \lim_{n\to\infty} x_{n+1} = x.$$

It remains to show that this fixed point is unique. Assume that $x' \in X$ is another point fixed by the map f. Then

$$d(x, x') = d(f(x), f(x')) \leq k d(x, x') < d(x, x'),$$

which is impossible. This finishes the proof of the theorem. □

Given a function $f : X \to X$, we will use $f^n : X \to X$ to denote the *n-th iterate of* f, i.e., the composition of f with itself n times: $f^n = \underbrace{f \circ \cdots \circ f}_{n}$.

The following observation is a simple consequence of the estimate (8.5) from the proof of the Contraction Mapping Theorem.

Exercise 8.6. Let $f : X \to X$ be a contraction of a complete metric space (X, d), with contraction constant $k \in (0, 1)$. Prove that for every $x_0 \in X$ the sequence $\{x_n\}$, defined by $x_n = f^n(x_0)$, $n \in \mathbb{N}$, converges to the fixed point $x \in X$ of f exponentially fast. More precisely, show that there is a constant $C > 0$ such that $d(x_n, x) \leq C\, k^n$, for all $n \in \mathbb{N}$.

In fact, we can generalize Theorem 8.2 to a more general situation, when the map f is not a contraction but some iterate of f is. This will be used in the proof of Picard's theorem below.

Lemma 8.7. *Let X be a complete metric space and $f : X \to X$ be such that the n-iterate f^n is a contraction, for some $n \in \mathbb{N}$. Then f has a unique fixed point on X.*

Proof. By Theorem 8.2, f^n has a unique fixed point $x \in X$. Then $f(x) = f(f^n(x)) = f^n(f(x))$, and so $f(x)$ is another fixed point of f^n. By the uniqueness, we must have that $f(x) = x$. Since any fixed point of f is also a fixed point of f^n, this fixed point must be unique. □

8.2. **Applications of the Contraction Mapping Theorem.**

The Contraction Mapping Theorem is a tremendously powerful result that can be applied in a large variety of situations. We indicate some of the more important examples.

Example 8.8. Let us begin with the simple problem of calculating $\sqrt{2}$. In other words, we want to find rational approximations of the positive number x (obviously we can assume that $x > 1$), which satisfies the equation $x^2 = 2$. A little thought shows that this equation is equivalent to

$$\frac{1}{2}\left(x + \frac{2}{x}\right) = x \tag{8.9}$$

Define a function $\Phi : [1, \infty) \to [1, \infty)$ by

$$\Phi(x) = \frac{1}{2}\left(x + \frac{2}{x}\right).$$

Then the equation (8.9) becomes simply an equation for a fixed point of Φ:

$$\Phi(x) = x.$$

Given that the half-line $X = [1, \infty)$ is a complete space, once we check that Φ is a contraction, the Contraction Mapping Theorem will imply that Φ has a unique fixed point. In other words, we will show that $\sqrt{2}$ exists and is unique. This seems disappointing, as this was known for a very long time. The novelty of this approach lies in the iteration procedure which will be used to provide a sequence of approximations of this number.

Let us then take arbitrary $x, y \in X = [1, \infty)$. We have

$$|\Phi(x) - \Phi(y)| = \frac{1}{2}\left|x + \frac{2}{x} - y - \frac{2}{y}\right| = \frac{|x-y|}{2}\left|1 - \frac{2}{xy}\right| \leq \frac{1}{2}|x-y|.$$

Thus indeed, Φ is a contraction and as we have seen in the proof of the Contraction Mapping Theorem, the sequence obtained by taking iterations of Φ applied to *any* element of X will converge to the unique fixed point of Φ, i.e., to $\sqrt{2}$. Let's start at 1 so that $x_1 = \Phi(1) = 3/2$; $x_2 = \Phi(3/2) = 17/12$, $x_3 = \Phi(17/12) = 577/408$.

This way we obtain a numerical procedure that would be very easy to implement on a computer and which will generate rational approximations of $\sqrt{2}$.

To give you some idea of how fast this iteration converges, here is $\sqrt{2}$ computed to one hundred decimal places by Mathematica:

$$1.41421356237309504880168872420969807856967187537694807$$
$$31766797379907324784621070388503875343276415 73.$$

Start with $x_0 = 1$ as before, and iterate the map Φ.

$$x_1 = 1.5;$$
$$x_2 = 1.41(6);$$
$$x_3 = 1.41421568627450980392156862745098039215686274$$
$$509803921568627450980392156862745098039215686274$$
$$509803921568627450980392156862745098039215686274$$
$$156862745098039;$$

$$x_4 = 1.41421356237468991062629557889013491011655962211574404458490501920005437183538926835899004315764434 0;$$

$$x_5 = 1.41421356237309504880168962350253024361498192577619742849828949862319582422892362178494183673583 0357;$$

$$x_6 = 1.41421356237309504880168872420969807856967187537723400156101313311326525563033997853178716125071 0475.$$

After just *seven* iterations we get

$$\begin{aligned} x_7 = 1.&41421356237309504880168872420\\ &96980785696718753769480731766\\ &79737990732478462107038850387\\ &534327641602, \end{aligned}$$

which is already remarkably close, while eight iterations suffice to match $\sqrt{2}$ to one hundred decimal places:

$$\begin{aligned} x_8 = 1.&41421356237309504880168872420 9\\ &69807856967187537694807317667 9\\ &73799073247846210703885038753 4\\ &327641573. \end{aligned}$$

One of the main applications of the Contraction Mapping Theorem is to solving differential equations, as we now illustrate.

Example 8.10. A typical initial value problem is stated as follows. Prove that there exists a unique differentiable function $f \in \mathcal{C}[0,1]$ which satisfies the equation

$$f'(t) = \sin t\, f(t), \tag{8.11}$$

subject to the condition $f(0) = 0$.

Recalling the Fundamental Theorem of Calculus, we can write

$$f(s) = f(0) + \int_0^s f'(t)\, dt, \text{ for any } s \in [0,1].$$

We now use our differential equation (8.11), which provides an expression for $f'(t)$, and the initial condition, which sets $f(0) = 0$, to obtain the following integral equation

$$f(s) = \int_0^s \sin t\, f(t)\, dt.$$

This integral equation is equivalent to the original initial value problem. The point of doing this is that we now can represent this integral equation as a fixed-point problem for a suitably defined map. And so we define the map $\Phi : \mathcal{C}[0,1] \to \mathcal{C}[0,1]$ by the formula

$$\Phi(g)(s) = \int_0^s \sin t\, g(t)\, dt, \text{ for any } g \in \mathcal{C}[0,1],$$

where $s \in [0,1]$. Let us show that Φ is a contraction, with respect to the supremum metric d_∞ on $\mathcal{C}[0,1]$, defined in Section 3.3.

Indeed, for any $f, g \in \mathcal{C}[0,1]$ and an arbitrary $s \in [0,1]$, we have

$$\begin{aligned} |\Phi(f)(s) - \Phi(g)(s)| &= \left| \int_0^s \sin t\, f(t)\, dt - \int_0^s \sin t\, g(t)\, dt \right| \\ &= \left| \int_0^s \sin t\, (f(t) - g(t))\, dt \right| \\ &\le \int_0^s |\sin t|\, |f(t) - g(t)|\, dt. \end{aligned}$$

If we now use that $|\sin t| \leq t$ for all $t \geq 0$, to get

$$|\Phi(f)(s) - \Phi(g)(s)| \leq \int_0^s t|f(t) - g(t)|\, dt \leq d_\infty(f,g) \int_0^s t\, dt$$
$$= \frac{s^2}{2} d_\infty(f,g) \leq \frac{1}{2} d_\infty(f,g), \text{ as } s^2 \leq 1 \text{ when } s \in [0,1].$$

This proves that Φ is a contraction with $k = \frac{1}{2}$, and so we conclude from the Contraction Mapping Theorem (in view of Theorem 7.14) that there exists a unique solution of the initial value problem (8.11).

The iteration procedure introduced in the proof of the Contraction Mapping Theorem allows us to find approximate solutions of this equation. Beginning, for instance, with $f_0(t) = 1$, for all $t \in [0,1]$, we find

$$f_1(s) = \Phi(f_0)(s) = \int_0^s \sin t\, dt = -\cos s + 1, \ \ s \in [0,1].$$

Similarly, $f_2 = \Phi(f_1)$ can be calculated as follows

$$f_2(s) = \int_0^s \sin t\,(1 - \cos t)\, dt = 1 - \cos s - \int_0^s \sin t \cos t\, dt$$
$$= \frac{3}{4} - \cos s + \frac{1}{4}\cos 2s, \ \ s \in [0,1].$$

And so on. We will develop this in detail in the next chapter.

9. Picard's theorem

In this section we present an important application of the Contraction Mapping Theorem to finding solutions of differential equations.

Let

$$\frac{dy}{dx} = F(x, y) \tag{9.1}$$

be a differential equation subject to the initial condition $y(0) = c \in \mathbb{R}$, where we assume that $F : [0, 1] \times \mathbb{R} \to \mathbb{R}$ is a continuous function.

Theorem 9.2. *Consider the initial value problem stated in (9.1). If, for some $L > 0$, the function F satisfies the Lipschitz condition*

$$|F(x, y_1) - F(x, y_2)| \leq L|y_1 - y_2| \text{ for all } x \in [0, 1] \text{ and all } y_1, y_2 \in \mathbb{R},$$

then the initial condition problem (9.1) has a unique solution $y \in \mathcal{C}[0, 1]$.

Proof. By the Fundamental Theorem of Calculus, the initial condition problem (9.1) is equivalent to the integral equation

$$y(x) = c + \int_0^x F(t, y(t))\, dt.$$

The same theorem also implies that the function y, defined this way, is continuous on the interval $[0, 1]$ and is differentiable on the open interval $(0, 1)$. Using this observation, we can define a map $\Phi : \mathcal{C}[0, 1] \to \mathcal{C}[0, 1]$ that sends a continuous function $y \in \mathcal{C}[0, 1]$ to the function

$$\Phi(y)(x) = c + \int_0^x F(t, y(t))\, dt.$$

A key conclusion from this simple manipulation is that the initial condition problem (9.1) is equivalent to finding a fixed point of the map Φ.

To check if Φ is a contraction (with respect to the supremum metric d_∞ on $\mathcal{C}[0, 1]$), let $y, z \in \mathcal{C}[0, 1]$. Then for all $x \in [0, 1]$ we have

$$\begin{aligned} |\Phi(y)(x) - \Phi(z)(x)| &= \left| \int_0^x (F(t, y(t)) - F(t, z(t)))\, dt \right| \\ &\leq \int_0^x |F(t, y(t)) - F(t, z(t))|\, dt \\ &\leq L \int_0^x |y(t) - z(t)|\, dt, \end{aligned} \tag{9.3}$$

where we have used the Lipschitz condition in the statement of the theorem. Given that

$$|y(t) - z(y)| \leq d_\infty(y, z), \text{ for all } t \in [0, 1],$$

we have that, for every $x \in [0, 1]$,

$$\begin{aligned} |\Phi(y)(x) - \Phi(z)(x)| &\leq L \int_0^x |y(t) - z(t)|\, dt \\ &\leq L d_\infty(y, z) \int_0^x dt = xL\, d_\infty(y, z). \end{aligned} \tag{9.4}$$

It follows that

$$|\Phi(y)(x) - \Phi(z)(x)| \leq L d_\infty(y, z), \text{ for all } x \in [0, 1].$$

And so, after taking the supremum of the left hand side over all $x \in [0, 1]$, we deduce that

$$d_\infty(\Phi(y), \Phi(z)) \leq L\, d_\infty(y, z).$$

Unfortunately, this does not imply that Φ is a contraction, as L does not have to be smaller than 1. Let us check if iterating Φ helps to improve things. For this we apply again the map Φ to $\Phi(y)$ and $\Phi(z)$. As $\Phi(y)$ and $\Phi(z)$ are continuous functions, then, by the estimate from (9.3), we have

$$|\Phi^2(y)(x) - \Phi^2(z)(x)| \leq L \int_0^x |\Phi(y)(t) - \Phi(z)(t)|\, dt, \tag{9.5}$$

for all $x \in [0,1]$. We now consider the integrand in the above formula. Using the estimate derived in equation (9.4) we know that

$$|\Phi(y)(t) - \Phi(z)(t)| \leq Lt\, d_\infty(y,z),$$

for all $t \in [0,1]$, so we can substitute this into the integral on the right hand side of inequality (9.5) to get

$$\begin{aligned} |\Phi^2(y)(x) - \Phi^2(z)(x)| &\leq L \int_0^x |\Phi(y)(t) - \Phi(z)(t)|\, dt \\ &\leq L^2\, d_\infty(y,z) \int_0^x t\, dt \\ &\leq L^2\, \frac{t^2}{2}\, d_\infty(y,z) \leq \frac{L^2}{2}\, d_\infty(y,z). \end{aligned}$$

Using induction, we can prove that for all $n \in \mathbb{N}$,

$$d_\infty(\Phi^n(y), \Phi^n(z)) \leq \frac{L^n}{n!}\, d_\infty(y,z). \tag{9.6}$$

Recalling that $\frac{L^n}{n!} \to 0$, as $n \to \infty$, for all sufficiently large enough $n \in \mathbb{N}$ we will have $\frac{L^n}{n!} < 1$, and hence, in view of (9.6), the n-th iterate Φ^n of Φ is a contraction (with respect to d_∞). The result now follows from Lemma 8.7 and Theorem 7.14. □

It is important to note that the choice of the interval $[0,1]$ in the Picard's Theorem above was for convenience of notation, and the same argument carries over to the case to any interval $[a,b]$ in $\mathbb{R}$.

Example 9.7. Consider the initial value problem

$$\frac{dy}{dx} = x(y+x),$$

where $x \in [0,1]$ and $y(0) = 0$.

In the notation of Picard's theorem, we have $F(x,y) = x(y+x)$. Then

$$|F(x,y_1) - F(x,y_2)| = |x||y_1 - y_2| \leq |y_1 - y_2|,$$

since $x \in [0,1]$. Thus F satisfies the Lipschitz condition of Theorem 9.2, and so our initial value problem has a unique solution.

For differentiable functions there is a straightforward way to check if the Lipschitz condition holds.

Proposition 9.8. *Let us consider again the initial value problem stated in equation (9.1). If the partial derivative $\partial F/\partial y$ exists and is bounded on $[0,1] \times \mathbb{R}$ then the problem (9.1) has a unique solution $y \in \mathcal{C}[0,1]$.*

Proof. For each $x \in [0,1]$ and $y_1, y_2 \in \mathbb{R}$, the Mean Value Theorem from Analysis tells us that there exists $c \in \mathbb{R}$ such that

$$|F(x,y_1) - F(x,y_2)| = \left|\frac{\partial F}{\partial y}(x,c)\right| |y_1 - y_2|.$$

Now, since the partial derivative $\partial F/\partial y$ is bounded, there exists $L > 0$ such that $|\frac{\partial F}{\partial y}(x, c)| \leq L$ for all $x \in [0, 1]$ and all $c \in \mathbb{R}$. Hence we can deduce that

$$|F(x, y_1) - F(x, y_2)| \leq L|y_1 - y_2|, \ \forall\, x \in [0, 1].$$

Thus Picard's theorem applies and yields the desired result. □

Exercise 9.9. Prove that the initial value problem given by the equation

$$\frac{dy}{dx} = \frac{1}{x + e^y}$$

for $x \in [1, R]$, for some real number $R > 1$, together with the initial condition $y(1) = 0$ has a unique solution.

Example 9.10. Consider the initial value problem

$$\frac{dy}{dx} = 3y^{2/3}, \quad y(0) = 0, \quad x \in [0, 1].$$

This has two solutions, $y(0) = 0$ for all x and $y(x) = x^3$. In this case, the Lipschitz condition is not satisfied (you should check it). Here $F(x, y) = 3y^{2/3}$ and $\partial F/\partial y = 2y^{-1/3}$, which is unbounded (again, an exercise).

Example 9.11. Let us return again to Example 9.7. We will set up an iteration procedure to approximate the solution. We can begin with the simplest function that satisfies the initial condition, which is $y_0 = 0$. Then

$$y_1(x) = \Phi(y_0) = \int_0^x t^2 dt = \frac{1}{3}x^3.$$

Hence

$$y_2(x) = \Phi(y_1) = \int_0^x \left(\frac{1}{3}t^4 + t^2\right) dt = \frac{1}{3 \times 5}x^5 + \frac{1}{3}x^3.$$

Continuing this way we can derive the general solution

$$y(x) = \frac{x^3}{1 \times 3} + \frac{x^5}{1 \times 3 \times 5} + \frac{x^7}{1 \times 3 \times 5 \times 7} + \dots + \frac{x^{2n+1}}{1 \times 3 \times 5 \cdots \times (2n+1)} + \dots$$

10. Topological spaces

Often one is interested in metric spaces up to homeomorphism, i.e., two metric spaces X and Y are considered "the same" if there exists a homeomorphism $f\colon X \to Y$. In this case, the precise metric on X and Y is not very relevant and it is helpful to express continuity in a way that no longer references the metric directly. For this we make the following definition.

Definition 10.1. A subset U of a metric space (X, d) is *open* if for every $x \in U$ there exists $\varepsilon > 0$ such that $\mathcal{B}_\varepsilon(x) \subseteq U$.

Proposition 10.2. *A map $f\colon X \to Y$ between metric spaces is continuous if and only if the preimages of open sets are open.*

Proof. Assume that the preimage of every open set in Y is open in X. Then for every $a \in X$ and each $\varepsilon > 0$ the preimage $V = f^{-1}(\mathcal{B}^Y_\varepsilon(f(a)))$, of the open ball $\mathcal{B}^Y_\varepsilon(f(a))$ in Y, is open in X. Therefore, by Definition 10.1, there exists a $\delta > 0$ such that $\mathcal{B}^X_\delta(a) \subseteq V$. But then if $x \in \mathcal{B}^X_\delta(a)$, $f(x) \in f(V) \subseteq \mathcal{B}^Y_\varepsilon(f(a))$. So f is continuous at a, by Definition 6.8, for all $a \in X$.

"$\Rightarrow$" Now assume that f is continuous, and let U be any open subset of Y. Consider any $a \in f^{-1}(U)$, so that $f(a) \in U$. Since U is open, there exists $\varepsilon > 0$ such that $\mathcal{B}^Y_\varepsilon(f(a)) \subseteq U$. As f is continuous at a there exists $\delta > 0$ such that if $x \in \mathcal{B}^X_\delta(a)$ then $f(x) \in \mathcal{B}^Y_\varepsilon(f(x))$. But this implies that $\mathcal{B}^X_\delta(a) \subseteq f^{-1}(U)$.

Thus we have shown that for every $a \in f^{-1}(U)$, $f^{-1}(U)$ contains an open ball around a. Thus $f^{-1}(U)$ is open by definition. □

The idea of topology is to study spaces only be prescribing the open sets. Metric spaces will be an important example of topological spaces, but we will also see many examples that do not come from metric spaces.

While topological spaces are less intuitive than metric spaces, often proofs will become easier. For example using the above reformulation of continuity, it is very easy to show that the composition of continuous maps is continuous. If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then for $U \subseteq Z$ open, $g^{-1}(U)$ and thus also $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$ are open. Thus $g \circ f$ is continuous.

10.1. **Topology.**

Definition 10.3. Let X be a set. A family $\mathcal{T}$ of subsets of X is called a *topology on X* if it satisfies all of the following conditions:

(1) $X, \emptyset \in \mathcal{T}$;
(2) the union of any collection of subsets from $\mathcal{T}$ is in $\mathcal{T}$;
(3) the intersection of any finite number of subsets from $\mathcal{T}$ is in $\mathcal{T}$.

If $\mathcal{T}$ is a topology on X then $(X, \mathcal{T})$ is called a *topological space*. Elements of $\mathcal{T}$ are called the *open sets* of $(X, \mathcal{T})$.

In other words, a topology on a set X is a family of subsets of X, which we call open sets, such that X itself and the empty set $\emptyset$ are open, an arbitrary union of open sets is open, and the intersection of finitely many open sets is open. From now on, when we say that a subset U of X is open, we mean that U is an element of $\mathcal{T}$.

Given a topology on a set X it is not always easy to check if a particular subset of X is open. The following criterion is useful in this regard.

Proposition 10.4. *A subset U of X is open if and only if for every $x \in U$ there is an open set U_x such that $x \in U_x \subseteq U$.*

Proof. If U is open, then we take $U_x = U$ for every $x \in U$.

Conversely, if the condition holds, then we shall prove that

$$U = \bigcup_{x \in U} U_x. \tag{10.5}$$

Indeed, if $y \in U$, then there is an open set U_y in X such that $y \in U_y$, hence $y \in \bigcup_{x \in U} U_x$, and thus $U \subseteq \bigcup_{x \in U} U_x$. Conversely, as $U_x \subseteq U$ for all $x \in U$, then the union of these sets is contained in U. Thus $\bigcup_{x \in U} U_x \subseteq U$, so we have proved the equality (10.5). It follows that U is open since it is a union of open sets. □

If x is a point of a topological space $(X, \mathcal{T})$, any open set containing x is said to be an *open neighborhood of* x. Thus, Proposition 10.4 can be restated by saying that a subset $U \subseteq X$ is open if and only if with each point $x \in U$, some open neighborhood of x is contained in U.

In practice, instead of specifying the family of sets that form the topology, it is easier to identify a smaller family (of open neighborhoods) from which each open set can be created by taking unions.

Definition 10.6. Let $(X, \mathcal{T})$ be a topological space. A family $\mathfrak{B} \subseteq \mathcal{T}$ is said to be a *basis* for the topology $\mathcal{T}$ if every open set in $\mathcal{T}$ is a union of sets from $\mathfrak{B}$.

We will see an example of an application of this idea in the next section.

10.2. **Metric topology.**

Metric spaces provide an important example of topological spaces, while on the other hand, topological methods are very useful in organising research into metric spaces. In this section we will begin the study of the topology of metric spaces and will return to this as we develop more notions in topological spaces.

Let (X, d) be a metric space. Let $\mathfrak{B}_m$ be the family of all open metric balls in X, that is all the sets $\mathcal{B}_r(x)$ where $x \in X$ and $r > 0$.

Definition 10.7. The *metric topology* on the metric space X, denoted $\mathcal{T}_m$, is the topology created from the basis $\mathfrak{B}_m$. That is, in the metric topology $\mathcal{T}_m$, a subset $U \subseteq X$ is open if and only if U is a union of a family of open balls.

In this case we will say that the metric d *induces the topology* $\mathcal{T}_m$ on X.

In other words, in the metric topology, the basic open neighborhoods of points of X are open balls around these points. In particular, an open ball in a metric space is an open subset. The choice of open balls is justified by the following property, which leads on to a more practical description of an open set.

Lemma 10.8. *Let (X, d) be a metric space and let $\mathcal{B}_r(y)$ be an open ball in X, for some $y \in X$ and $r > 0$. For every $x \in \mathcal{B}_r(y)$ there exists $\varepsilon > 0$ such that $\mathcal{B}_\varepsilon(x) \subseteq \mathcal{B}_r(y)$.*

Proof. For x and y as in the statement, put $\varepsilon = r - d(x, y) > 0$. Take any $z \in \mathcal{B}_\varepsilon(x)$, so that $d(z, x) < \varepsilon$. Then

$$d(z, y) \leq d(z, x) + d(x, y) < \varepsilon + d(x, y) = r - d(x, y) + d(x, y) = r.$$

This means that $z \in \mathcal{B}_r(y)$, and the lemma is proved. □

Proposition 10.9. *A subset U of a metric space (X, d) is open in the metric topology if and only if for every $x \in U$ there exists $r > 0$ such that $\mathcal{B}_r(x) \subseteq U$, i.e. the notion of open subsets of metric spaces agrees with our previous definition.*

Proof. If U is open, then, according to Definition 10.7, it is the union of open balls:

$$U = \bigcup_{\alpha \in I} \mathcal{B}_{r_\alpha}(y_\alpha).$$

So for every $x \in U$ there exists $\beta \in I$ such that $x \in \mathcal{B}_{r_\beta}(y_\beta)$. By Lemma 10.8 there exists $\varepsilon > 0$ such that

$$x \in \mathcal{B}_\varepsilon(x) \subseteq \mathcal{B}_{r_\beta}(y_\beta) \subseteq U.$$

Conversely, if for every $x \in U$ there is an open ball $\mathcal{B}_{r_x}(x)$, contained in U, then U is the union of the balls $\cup_{x\in U}\mathcal{B}_{r_x}(x)$, and so it is open in the metric topology. □

Theorem 10.10. *Let (X, d) be a metric space. The metric topology $\mathcal{T}_m$ on X, given by Definition 10.7, is indeed a topology in the sense of Definition 10.3.*

Proof. Obviously X is the union of open balls around all of its points and the empty set is the union of an empty family of balls, hence $X, \emptyset \in \mathcal{T}_m$. It is also clear that $\mathcal{T}_m$ is closed under taking arbitrary unions. Thus it remains to check condition (3) from Definition 10.3.

Let $U_1, \dots, U_n \in \mathcal{T}_m$ be open sets, we need to show that $U = \bigcap_{i=1}^n U_i$ is also an open set in $\mathcal{T}_m$. Take any $x \in U$, then $x \in U_i$, for each $i = 1, \dots, n$, so, by Proposition 10.9, there exists $\varepsilon_i > 0$ such that $\mathcal{B}_{\varepsilon_i}(x) \subseteq U_i$, $i = 1, \dots, n$.

Observe that for $\varepsilon = \min\{\varepsilon_1, \dots, \varepsilon_n\} > 0$, we have $\mathcal{B}_\varepsilon(x) \subseteq \mathcal{B}_{\varepsilon_i}(x) \subseteq U_i$, for each $i = 1, \dots, n$. It follows that $\mathcal{B}_\varepsilon(x) \subseteq \bigcap_{i=1}^n U_i = U$. Therefore U is open in $\mathcal{T}_m$ by Proposition 10.9, so (3) holds and $\mathcal{T}_m$ is indeed a topology on X. □

Example 10.11. Every non-empty open interval (a, b) is an open ball in $\mathbb{R}$, equipped with the standard absolute value metric: this is the ball at $(a + b)/2$ with radius $r = (a - b)/2$. The converse is also clear, so such intervals form a basis of the metric topology on $\mathbb{R}$. This metric topology on $\mathbb{R}$ is called the *standard topology on* $\mathbb{R}$.

Note that the closed interval $[a, b]$ is not open in this topology. Indeed, any ball $\mathcal{B}_r(a) = (a - r, a + r)$ has a non-empty intersection with the complement of the interval, so the interval is not open, according to Proposition 10.9. The same argument shows that the half-open interval $[a, b)$ is not open.

Remark 10.12. In the third defining axiom of a topology from Definition 10.3, we require that the intersection of any *finite* family of open sets is open. In fact, we can check that if X is a metric space with the metric topology, then the intersection of infinitely many open sets need not be open. To see this, take U_n to be the open interval $(-1 - \frac{1}{n}, 1 + \frac{1}{n})$, $n \in \mathbb{N}$. Then

$$\bigcap_{n\in\mathbb{N}} U_n = [-1, 1].$$

Indeed, the interval $[-1, 1]$ is a subset of every U_n, and so it belongs to the intersection. To see that the intersection actually is the same as this interval, let $y > 1$. Then $y = 1 + x$ for some $x > 0$, and for a large enough $m \in \mathbb{N}$ we have that

$$1 + \frac{1}{m} < 1 + x,$$

so $1 + x \notin U_m$. We prove in the same way that no number less than -1 is in the intersection.

Thus, in this case, the intersection of infinitely many open intervals is a closed interval, which is not an open set as we have seen.

10.3. **More examples of topologies.**

The definition of topology introduced at the start of this chapter is very flexible and allows many examples, which are very different to the metric topology. Here is a short list.

Example 10.13. The *discrete topology* on a set X is given by the family $\mathcal{T}$ of *all* subsets of X. Thus, in particular, every singleton set $\{x\}$, $x \in X$, is open in this topology.

Example 10.14. The *indiscrete topology* on a non-empty set X is the family $\{\emptyset, X\}$.

Exercise 10.15. Let X be a set.

(a) Check that the discrete and indiscrete topologies on a set X satisfy Definition 10.3.
(b) Show that the discrete topology on any set X is the topology induced by the discrete metric δ (see Definition 4.1).
(c) Suppose that X has at least two elements. Prove that the indiscrete topology is not a metric topology on X (i.e., there is no metric on X that would give rise to the indiscrete topology).

Example 10.16. Let X be an infinite set and let us define a family $\mathcal{T}$, of subsets of X, that contains $\emptyset$ and all *cofinite* subsets U in X, that is subsets $U \subseteq X$ such that the complement $U^c = X \setminus U$ is finite. Let us show that $\mathcal{T}$ is a topology on X. This is called the *cofinite topology*.

We shall check the defining axioms of a topology. Clearly X is cofinite in itself, so X and $\emptyset$ are both open. Also, for any collection $\{U_\alpha\}_{\alpha\in I}$, of open sets, either all of them are empty, in which case so is their union, or U_β is cofinite in X for some $\beta \in I$. Then $\bigcup_{\alpha\in I} U_\alpha$ is also cofinite, as

$$\left(\bigcup_{\alpha\in I} U_\alpha\right)^c = \bigcap_{\alpha\in I} {U_\alpha}^c \subseteq {U_\beta}^c.$$

Let U and V be open sets. If one of them is empty, then so is $U \cap V$. Otherwise both U and V are cofinite subsets in X, so U^c and V^c are finite. Since

$$(U \cap V)^c = U^c \cup V^c,$$

$(U \cap V)^c$ is finite, i.e., $U \cap V$ is cofinite in X, and, hence, it is open.

We will later show (see Example 14.4) that the cofinite topology on an infinite set X is never induced by a metric.

As one can define many topologies on a given set, it is natural to ask if they can be compared in a useful way.

Definition 10.17. We say that topology $\mathcal{T}_1$ on a set X is *coarser* than topology $\mathcal{T}_2$ on the same set if $\mathcal{T}_1 \subseteq \mathcal{T}_2$, i.e., every set, which is open with respect to $\mathcal{T}_1$, is also open with respect to $\mathcal{T}_2$. In this case we say that the topology $\mathcal{T}_2$ is *finer* than $\mathcal{T}_1$.

Example 10.18. The one-element set $X = \{0\}$ has only one topology: the discrete topology which is the same as the indiscrete topology. The open sets are $\{\emptyset, X\}$.

The two element set $X = \{0, 1\}$ admits four different topologies:

- the indiscrete topology $\mathcal{T}_i = \{\emptyset, X\}$;
- two intermediate topologies:
$$\mathcal{T}_t = \{\emptyset, \{0\}, \{0, 1\}\} \quad \text{and} \quad \mathcal{T}_S = \{\emptyset, \{1\}, \{0, 1\}\};$$
- the discrete topology
$$\mathcal{T}_d = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}.$$

The *Sierpiński topology* on $\{0, 1\}$ is the topology $\mathcal{T}_S = \{\emptyset, \{1\}, \{0, 1\}\}$.

11. Closed sets, interior, and closure

11.1. **Closed sets.**

Definition 11.1. We say that a subset V of a topological space $(X, \mathcal{T})$ is *closed* if its complement $V^c = X \setminus V$ is open.

The next proposition is a straightforward consequence of de Morgan's laws, which we have already used in the previous chapter:

$$\left(\bigcup_{\alpha\in I} U_\alpha\right)^c = \bigcap_{\alpha\in I} {U_\alpha}^c \quad \text{and} \quad \left(\bigcap_{\alpha\in I} U_\alpha\right)^c = \bigcup_{\alpha\in I} {U_\alpha}^c.$$

Proposition 11.2. *Let $(X, \mathcal{T})$ be a topological space.*

(1) *X and $\emptyset$ are closed sets.*
(2) *The union of any finite family of closed sets is closed.*
(3) *The intersection of any family of closed subsets of X is closed.*

Example 11.3. Consider the space of real numbers $\mathbb{R}$ with the standard topology (see Example 10.11). We have seen that the open subsets of $\mathbb{R}$ are the unions of open intervals. It follows that for any $a, b \in \mathbb{R}$, $a < b$, the half-infinite intervals $(-\infty, a)$ and (b, ∞) are open (as $(-\infty, a) = \bigcup_{n\in\mathbb{N}}(a - n, a)$ and $(b, \infty) = \bigcup_{n\in\mathbb{N}}(b, b + n)$).

Now, the complement of the closed interval $[a, b]$ is

$$[a, b]^c = (-\infty, a) \cup (b, \infty),$$

which is open, as a union of two open sets. Thus the interval $[a, b]$ is a closed set in the sense of our definition.

The set of integers $\mathbb{Z}$ is a closed subset of $\mathbb{R}$. This is because the complement of $\mathbb{Z}$ in $\mathbb{R}$, given by

$$\mathbb{Z}^c = \bigcup_{m\in\mathbb{Z}} (m, m + 1),$$

is open as a union of open intervals.

11.2. **Interior and closure.**

Definition 11.4. Let $(X, \mathcal{T})$ be a topological space, and let $A \subseteq X$ be a subset. A point $x \in A$ called an *interior point* of A if there exists an open subset U_x, of X, such that $x \in U_x$ and $U_x \subseteq A$. In other words, $x \in A$ is an interior point, if some open neighborhood of x lies in A.

The set of all interior points of A is called the *interior of A* and is denoted A^o.

When A is a subset of a metric space (X, d) then Proposition 10.9 implies that a point $x \in A$ is an interior point of A if and only if there exists an $r > 0$ such that the open ball $\mathcal{B}_r(x)$ is contained in A . This property is often taken as the definition of the interior point of a subset of a metric space.

Proposition 11.5. *Suppose that A is a subset of a topological space $(X, \mathcal{T})$. Then the interior A^o is the union of all open subsets of A, and, hence, it is open in X.*

Proof. Let $\mathcal{U}$ denote the union of all open subsets of A

$$\mathcal{U} = \bigcup_{U\subseteq A,\, U\in\mathcal{T}} U.$$

Then $x \in \mathcal{U}$ if and only if there exists a subset $U \subseteq A$, which is open in X and contains x. This is equivalent to saying that x is an interior point of A. Thus $A^o = \mathcal{U}$, and so the interior A^o is open in X as a union of open sets. □

It follows from this statement that A^o is the largest open subset of A in the following sense. If $U \subseteq A$ is open, then $U \subseteq A^o \subseteq A$.

The following result is useful in practice.

Proposition 11.6. *A subset A of a topological space X is open if and only if $A = A^o$.*

Proof. If $A = A^o$ then A is open by Proposition 11.5. If A is open, then every point of A is an interior point, and so $A = A^o$. □

Exercise 11.7. Let A and B be subsets of a topological space X. Prove that

$$(A \cap B)^o = A^o \cap B^o.$$

By analogy to the definition of an interior point we say that $x \in X$ is an *exterior point* of A if it is an interior point of the complement A^c of A; this is to say that there exists an open set $U \subseteq X$ such that $x \in U$ and $U \cap A = \emptyset$.

We will now identify points which in some sense are 'close' to a given set.

Definition 11.8. A point x in a topological space $(X, \mathcal{T})$ is an *adherent point* of a subset $A \subset X$ if every open subset $U \subseteq X$, containing x, has a non-empty intersection with A.

The set of all adherent points of A is called the *closure of A* and is denoted $\overline{A}$.

It is clear from the definition that $A \subset \overline{A}$. Indeed, if $x \in A$, then of course $x \in U \cap A$ for any open set U containing x, and so x is an adherent point of A. Furthermore, an exterior point of A is not an adherent point of A.

Proposition 11.9. *The closure of any subset $A \subseteq X$ is a closed subset of X.*

Proof. We need to show that $\overline{A}^c = X \setminus \overline{A}$ is open. So, suppose that $x \in \overline{A}^c$; this means that $x \in X$ is not an adherent point of A. Thus there must exist an open subset U of X such that $x \in U$ and $U \cap A = \emptyset$. But then every element y of U is not an adherent point of A either, and so $U \cap \overline{A} = \emptyset$, i.e., U is an open neighborhood of x contained in $\overline{A}^c$. Since we found such a neighborhood for each point $x \in \overline{A}^c$, we can use Proposition 10.4 to conclude that $\overline{A}^c$ is open, hence $\overline{A}$ is closed in X. □

The closure of the set A and the interior of the set A enjoy properties that are in some sense dual to each other in the sense of the following proposition.

Proposition 11.10. *Let A be a subset of a topological space $(X, \mathcal{T})$. Then*

$$(A^o)^c = \overline{A^c}, \qquad (\overline{A})^c = (A^c)^o.$$

Proof. We will prove the first part, leaving the second as exercise. A point x is not an interior point of A (and so belongs to $(A^o)^c$) if and only if every open set U containing x has a non-empty intersection with A^c, which is equivalent to saying that x is an adherent point of A^c, and so $x \in \overline{A^c}$. Thus $(A^o)^c = \overline{A^c}$. □

The following statement is dual to Proposition 11.5.

Proposition 11.11. *Let $(X, \mathcal{T})$ be a topological space and let $A \subseteq X$. The closure $\overline{A}$ is the intersection of all closed subsets V containing A in X.*

Proof. Exercise. □

It follows that the closure of a set A is the *smallest* closed subset of X containing A in the sense that if $V \subset X$ is closed and $A \subseteq V$ then $\overline{A} \subseteq V$.

Example 11.12. Consider $X = \mathbb{R}$ with its metric topology and $A = (0, 1)$. Then $0 \in \mathbb{R}$ is an adherent point of A. Indeed, if $U \subseteq \mathbb{R}$ is an open subset of $\mathbb{R}$ containing 0, then, by Proposition 10.9, there exists $r > 0$ such that $(-r, r) \subseteq U$. Given that this interval has a non-empty intersection with A for all $r > 0$, it follows that any open set U, containing 0, has a non-empty intersection with A. The same argument shows that 1 is also an adherent point of A.

Now any real number $x \notin [0, 1]$ is an exterior point of the interval $(0, 1)$, as every such x is contained in $\mathbb{R} \setminus [0, 1] = (-\infty, 0) \cup (1, \infty)$, which is open, as a union of two open sets.

Thus $\overline{(0, 1)} = [0, 1]$ in $\mathbb{R}$.

Observe that if a point x is an element of the closure $\overline{A}$, of A, but is not in A, then the intersection of every open set, containing x, with A will contain a point of A distinct from x. If $x \in A$, then it can happen that there exists an open set U that contains the point x but it does not contain any other point of A. In such a case, we call x an *isolated point* of A.

Exercise 11.13. Show that for any subsets A, B of a topological space $(X, \mathcal{T})$,

$$\overline{A \cup B} = \overline{A} \cup \overline{B}.$$

Proposition 11.14. *Let A and B be subsets of a topological space X.*

(1) *If $A \subseteq B$ then $\overline{A} \subseteq \overline{B}$;*
(2) *A is closed if and only if $A = \overline{A}$;*
(3) $\overline{\overline{A}} = \overline{A}$.

Proof. Exercise. □

The following natural notion complements our definitions of an interior and exterior point of a set. As usual, we assume that $(X, \mathcal{T})$ is a topological space.

Definition 11.15. A point $x \in X$ is a *boundary point* of a subset A of X if $x \in \overline{A}$ and $x \in \overline{A^c}$. The *boundary of A* is the set of all boundary points of A, and is denoted ∂A. In other words, the boundary ∂A consists of all points $x \in X$ such that each open neighborhood of x intersects both A and A^c.

By definition, we have that

$$\partial A = \overline{A} \cap \overline{A^c},$$

hence ∂A is a closed subset of X, as an intersection of two closed subsets (see Proposition 11.9). We also have that

$$\partial A = \partial(A^c) \quad \text{and} \quad \partial A = \overline{A} \setminus A^o.$$

Example 11.16. Considering again the open interval $A = (0, 1)$ in $\mathbb{R}$ we have seen that $\overline{A} = [0, 1]$ and, since A is open, $A^o = A$. Thus the boundary of A is

$$\partial A = [0, 1] \setminus (0, 1) = \{0, 1\}.$$

Alternatively, $A^c = (-\infty, 0] \cup [1, \infty)$, and this is a closed subset of $\mathbb{R}$, since A is open. Thus $\overline{A^c} = A^c$, and

$$\partial A = A^c \cap \overline{A} = \big((-\infty, 0] \cup [1, \infty)\big) \cap [0, 1] = \{0, 1\}.$$

To conclude this section, let us consider the space $\mathbb{R}$ together with the subspace of rational numbers $\mathbb{Q}$, where both are considered with the standard metric topology on $\mathbb{R}$. If r is a real number, then every open interval $(r - s, r + s)$ will contain rational number; this follows from the Archimedean property of the reals. Thus every real number r is an adherent point of $\mathbb{Q}$, and so $\overline{Q} = \mathbb{R}$. It is useful to have the following terminology to describe this situation.

Definition 11.17. A subset A of a topological space $(X, \mathcal{T})$ is *dense* in X if $\overline{A} = X$.

This is equivalent to saying that every open subset U of X has a non-empty intersection with A.

11.3. **Closed subsets of metric spaces.**

We will make a brief return to metric spaces as there are a few interesting points we can make about closure and closed sets in this case.

Let (X, d) be a metric space, and suppose that $A \subseteq X$. Let $x \in \overline{A}$, so x is an adherent point of A. This means that every open set containing x has a non-empty intersection with A. In particular, for every open ball $\mathcal{B}_r(x)$, $\mathcal{B}_r(x) \cap A \neq \emptyset$.

Now assume that a point $x \in \overline{A}$ is not an isolated point of A. Then every open ball $\mathcal{B}_r(x)$ contains an element y of A which is distinct from x. We use this as follows. For each $n \in \mathbb{N}$, let $x_n \in A$ be a point from $\mathcal{B}_{1/n}(x)$ distinct from x. Thus we create a sequence of points $\{x_n\}$ in $A \setminus \{x\}$ such that $d(x_n, x) < 1/n$. It follows from the definition of the limit of a sequence, that $x_n \to x$ as $n \to \infty$. Thus we have the following.

Proposition 11.18. *Let A be a subset of a metric space (X, d) and $x \in \overline{A}$. Assume that x is not an isolated point of A. Then there exists a sequence $\{x_n\}$ of points in $A \setminus \{x\}$ which converges to x.*

The above proposition tells us that any point $x \in \overline{A}$ which is not an isolated point of A is a limit point of A in the following sense.

Definition 11.19. Let A be a subset of a metric space (X, d). An element $x \in X$ is a *limit point of A* if every open neighborhood of x contains of point of A other than x itself, in other words, if U is any open subset of X containing x then $U \cap A \setminus \{x\} \neq \emptyset$.

Proposition 11.20. *Let A be a subset of a metric space (X, d), and let $\{x_n\}$ be a sequence of elements of A which converges to a point $x \in X$. Then x is an adherent point of A, i.e., $x \in \overline{A}$. If A is closed then $x \in A$.*

Proof. Exercise. □

There is one last point we need to make here, which concerns closed subsets of complete spaces.

Proposition 11.21. *Let (X, d) be a complete metric space and let $A \subseteq X$ be a subset of X, equipped with the induced metric from X. If A is closed in the metric topology on X, then it is also complete as a metric space.*

Proof. Let $\{x_n\}$ be a Cauchy sequence of elements of A. Since the metric on A is induced from X, this is also a Cauchy sequence in X. As the space X is complete, this sequence converges to a point $x \in X$. The point x is an element of the closure $\overline{A}$ of A by Proposition 11.20, hence $x \in A$ because A is closed in X. Thus the Cauchy sequence $\{x_n\}$ has a limit in A, so A is complete. □

12. Continuity

Recall that in our study of metric spaces we introduced the following notion of continuity of maps between metric spaces (see Definition 6.8):

Let (X, d_X) and (Y, d_Y) be two metric spaces. A function $f : X \to Y$ is continuous at a point $a \in X$ if for every ε there exists $\delta > 0$ such that $d_X(x, a) < \delta$ implies that $d_Y(f(x), f(a)) < \varepsilon$.

In terms of open metric balls, this condition can be written as

$$f(\mathcal{B}^X_\delta(a)) \subseteq \mathcal{B}^Y_\varepsilon(f(a)). \tag{12.1}$$

We want to extend the notion of continuity to topological spaces, but of course we can't use the metric definition, as general topological spaces are not equipped with a metric.

However, the condition (12.1) is a useful starting point if we remember that an open ball in a metric space is an example of an open set; this brings us closer to a definition that would work in topological spaces. Let's try to simply re-write the definition introduced for metric spaces using open sets in place of balls. And so a function $f : X \to Y$ of topological spaces would be continuous at a point $a \in X$ if and only if for any open neighborhood U of $f(a)$ in Y, there exists an open neighborhood V of a in X such that

$$f(V) \subseteq U.$$

How would we find the set V? It is clear that V is a subset of the preimage $f^{-1}(U)$ of U, and if we knew that $f^{-1}(U)$ is open, the problem would be solved. Let us then agree on the following definition.

Definition 12.2. Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces. A function $f : X \to Y$ is *continuous* if for every $U \in \mathcal{T}_Y$, $f^{-1}(U) \in \mathcal{T}_X$. In other words, for every open subset U of Y its preimage $f^{-1}(U)$ is open in X.

Note that we have already seen that this agrees with the $\varepsilon - \delta$-definition of continuity for metric spaces in Proposition 10.2, which we recall here.

Proposition 12.3. *Let (X, d_X) and (Y, d_Y) be metric spaces. Then $f : X \to Y$ is continuous in the sense of Definition 6.8 if and only if for every open subset U of Y the preimage $f^{-1}(U)$ is open in X.*

Thus, for metric spaces, the two definitions give the same class of continuous functions. From now on for continuity of functions between topological spaces we will use Definition 12.2.

Exercise 12.4. Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces.

(a) Show that a function $f : X \to Y$ is continuous if and only if the preimage of every closed subset of Y is closed in X.
(b) Suppose that either $\mathcal{T}_X$ is a discrete topology or $\mathcal{T}_Y$ is an indiscrete topology. Show that any function $f : X \to Y$ is continuous.
(c) Assume that $\mathcal{T}_X$ is an indiscrete topology on X and $\mathcal{T}_Y$ is a discrete topology on Y. Prove that the only continuous maps $f : X \to Y$ are constant maps (i.e., there must exist $y_0 \in Y$ such that $f(x) = y_0$ for all $x \in X$).

The following statement follows easily from the definition.

Proposition 12.5. *Let $f : (X, \mathcal{T}_X) \to (Y, \mathcal{T}_Y)$ and $g : (Y, \mathcal{T}_Y) \to (Z, \mathcal{T}_Z)$ be continuous maps of topological spaces. Then the composition $g \circ f : X \to Z$ is continuous.*

Proof. Let U be an open subset of Z, then $g^{-1}(U)$ is open in Y, as g is continuous, and $f^{-1}(g^{-1}(U))$ is open in X, by the continuity of f. Since

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

this shows that the composition $g \circ f : X \to Z$ is a continuous map. □

The following notion generalizes the concept of a homeomorphism between metric spaces (cf. Definition 6.13).

Definition 12.6. A *homeomorphism* of topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ is a bijection $f : X \to Y$ such that both f and f^{-1} are continuous maps.

If there exists a homeomorphisms between spaces X and Y, we call the spaces *homeomorphic*.

Remark 12.7. Note that if $f : X \to Y$ is continuous and a bijection, it does not follow that $f^{-1} : Y \to X$ is also continuous. For instance, let $\mathcal{T}_m$ be usual metric topology on $\mathbb{R}$, induced by the standard metric, and let $\mathcal{T}_d$ be the discrete topology on $\mathbb{R}$. Then the identity map $\mathrm{Id} : (\mathbb{R}, \mathcal{T}_d) \to (\mathbb{R}, \mathcal{T}_m)$ is continuous, but its inverse $\mathrm{Id} : (\mathbb{R}, \mathcal{T}_m) \to (\mathbb{R}, \mathcal{T}_d)$ is not.

Exercise 12.8. Let us consider again the two-point set from Example 10.18, so that $X = \{0, 1\}$. Define a map $f : X \to X$ by $f(0) = 1$ and $f(1) = 0$. Prove that f is a homeomorphism of topological spaces $f : (X, \mathcal{T}_t) \to (X, \mathcal{T}_s)$, where $\mathcal{T}_t$ and $\mathcal{T}_s$ are the topologies defined in Example 10.18.

13. Subspaces and products

Definition 13.1. Let $(X, \mathcal{T})$ be a topological space and A be a subset of X. Then $\mathcal{T}_A = \{A \cap U \mid U \in \mathcal{T}\}$ is a topology on A, called the *subspace topology*, or the *induced topology*. In other words, a set $V \subseteq A$ is open in the subspace topology if and only if there exists an open set $U \subseteq X$ such that $V = U \cap A$.

Example 13.2. Consider the space of reals $\mathbb{R} \subset \mathbb{R}^2$, and assume that this embedding is given by the map $x \mapsto (x, 0)$, which identifies $\mathbb{R}$ with the x-axis in $\mathbb{R}^2$. Then a subset A of $\mathbb{R}$ is open in the subspace topology induced from $\mathbb{R}^2$ if and only if there exists an open subset G of $\mathbb{R}^2$ such that $A = \mathbb{R} \cap G$.

So, for instance, the interval $(-1, 1)$ is open in this topology as this is an intersection of the open ball $\mathcal{B}_1(\mathbf{0})$ with the x-axis.

Example 13.3. The standard metric topology on the set $\mathbb{Q}$ of rational numbers is the same as the subspace topology induced by the inclusion $\mathbb{Q} \subset \mathbb{R}$.

Exercise 13.4. Let $A = \{0\} \cup \{1/n \mid n \in \mathbb{N}\}$. Considering A as a subset of $\mathbb{R}$ with the subspace topology, identify the open and closed subsets of A in the subspace topology. This is dealt with in detail in the problem sheets. To get you started, consider the following.

For every $n \in \mathbb{N}$, $1/n \in (\frac{1}{n+1}, \frac{1}{n-1})$, which is an open subset of $\mathbb{R}$, and $1/n \in (\frac{1}{n+1}, \frac{1}{n-1}) \cap A = \{1/n\}$. Thus every singleton set $\{1/n\}$ is open in the subspace topology on A, and hence so is any union of these sets, which means that every subset of A, not containing 0 is open.

Now for every $r > 0$, there exist $K \in \mathbb{N}$ such that for all $n > K$, $1/n \in (-r, r)$. After intersecting this interval with the set A we conclude that every cofinite subset containing zero is open in A.

A subset V of A is closed if and only if its complement is open. In particular, every subset containing 0 is closed. Moreover, our discussion of the open subsets of A indicates that every cofinite subset of A is open. Hence, every finite subset of A is closed. In particular, $\{0\}$ is a closed subset of A.

Proposition 13.5. *Let $(X, \mathcal{T})$ be a topological space and let $A \subseteq X$ be a subset of X equipped with the subspace topology. Then the inclusion map $i : A \to X$ defined by $i(a) = a$, for all $a \in A$, is continuous.*

Proof. By the definition of the inclusion map, for every subset U in X, $i^{-1}(U) = U \cap A$. In particular, if $U \in \mathcal{T}$ then $i^{-1}(A) \in \mathcal{T}_A$, by the definition of $\mathcal{T}_A$. Hence i is continuous. □

An important and frequently used application of this fact is the following.

Corollary 13.6. *Let $f : X \to Y$ be a continuous map of topological spaces and $A \subseteq X$. Then the restriction $f|_A : A \to Y$ is continuous when A is equipped with the induced topology from X.*

Proof. By definition, the restriction $f|_A = f \circ i$, and so it is continuous as a composition of continuous maps (see Proposition 12.5). □

A complementary (and still useful) result to this corollary is the following.

Proposition 13.7. *Suppose that X and Z are topological spaces. Let $A \subseteq X$ be equipped with the subspace topology, and let g be a map $g : Z \to A$. Then g is continuous if and only if the map $i \circ g : Z \to X$ is continuous.*

Proof. When A is equipped with the subspace topology, the inclusion map $i : A \to X$ is continuous, by Proposition 13.5. If $g : Z \to A$ is a continuous function, then so is the composition $i \circ g : Z \to X$, by Proposition 12.5.

Conversely, assume that $i \circ g : Z \to X$ is continuous and let $V \subseteq A$ be open. By definition of the subspace topology there exists an open set $U \subseteq X$ such that $V = U \cap A = i^{-1}(U)$. But by the continuity of the function $i \circ g$, the set

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

is open in Z. Since $g^{-1}(i^{-1}(U)) = g^{-1}(V)$, we conclude that g is continuous. □

We now turn to the products of topological spaces. Recall, that the *product* $X \times Y$, of sets X and Y, is defined as the set of all pairs $\{(x, y) \mid x \in X, y \in Y\}$. In various applications it is useful to be able to say when a map in or from a product of topological spaces is continuous. For example, a path γ in $\mathbb{R}^2$ may be described parametrically by specifying two functions x and y such that $\gamma(t) = (x(t), y(t))$, and we say that γ is continuous if and only if x and y are continuous. We can generalize this in the following way.

Definition 13.8. Let $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ be topological spaces. The *product topology* $\mathcal{T}_{X\times Y}$ on $X \times Y$ is the topology in which every open set is a union of subsets of the form $U \times V \subseteq X \times Y$, where $U \in \mathcal{T}_X$ and $V \in \mathcal{T}_Y$. In other words, a basis for the product topology consists of products of open sets from X with open sets from Y.

Proposition 13.9. *The family $\mathcal{T}_{X\times Y}$ is a topology on $X \times Y$.*

Proof. It is clear that $\emptyset = \emptyset \times \emptyset$ and $X \times Y$ are elements of $\mathcal{T}_{X\times Y}$. Typical elements W_1 and W_2 of $\mathcal{T}_{X\times Y}$ have the form:

$$W_1 = \bigcup_{i\in I} U_{1i} \times V_{1i}, \quad W_2 = \bigcup_{j\in J} U_{2j} \times V_{2j},$$

where all the sets U_{1i}, U_{1j} belong to $\mathcal{T}_X$, V_{1i}, V_{1j} are in $\mathcal{T}_Y$, and I, J are some indexing sets. Note that for every $i \in I$,$j \in J$, the intersection

$$(U_{1i} \times V_{1j}) \cap (U_{2j} \times V_{2j}) = (U_{1i} \cap U_{2j}) \times (V_{1i} \cap V_{2j})$$

is a product of open sets from X and Y ($U_{1i} \cap U_{2j}$ is open in X since the intersection of two open sets in X is open, as $(X, \mathcal{T}_X)$ is a topological space). It follows that

$$W_1 \cap W_2 = \left(\bigcup_{i\in I} U_{1i} \times V_{1i}\right) \cap \left(\bigcup_{j\in J} U_{2j} \times V_{2j}\right) = \bigcup_{i\in I}\bigcup_{j\in J} (U_{1i} \cap U_{2j}) \times (V_{1i} \cap V_{2j}) \in \mathcal{T}_{X\times Y}.$$

Thus the family $\mathcal{T}_{X\times Y}$ is closed under finite intersection.

It remains to observe that $\mathcal{T}_{X\times Y}$ is closed under arbitrary unions, by definition, hence it is a topology on $X \times Y$. □

Define the *projections* $\pi_X : X \times Y \to X$ and $\pi_Y : X \times Y \to Y$ by $\pi_X((x, y)) = x$ and $\pi_Y((x, y)) = y$, for all $(x, y) \in X \times Y$.

Proposition 13.10. *If the product space $X \times Y$ is equipped with the product topology, then the two coordinate projections π_X and π_Y are continuous.*

Moreover, any map $f : Z \to X \times Y$, where Z is a topological space, is continuous if and only if both the maps $\pi_X \circ f : Z \to X$ and $\pi_Y \circ f : Z \to Y$ are continuous.

Proof. For every open subset U of X, $\pi_X^{-1}(X) = U \times Y$, which is open in $X \times Y$, and so π_X is continuous. Similarly for π_Y.

For the second part, if $(Z, \mathcal{T}_Z)$ is a topological space and f is continuous, then both compositions $\pi_X \circ f$ and $\pi_Y \circ f$ are continuous by Proposition 12.5.

Conversely, assume that the two maps $\pi_X \circ f$ and $\pi_Y \circ f$ are continuous, and we need to show that f is continuous. To do this it is sufficient to prove that for every set S that belongs to the basis $\mathfrak{B}$ of the product topology, the preimage $f^{-1}(S)$ is open in Z (this is because the preimage of a union is the union of preimages). Recall that the basis $\mathfrak{B}$ consists of sets of the form $S = U \times V$, where U is open in X and V is open in Y. Then

$$U \times V = (U \times Y) \cap (X \times V) = \pi_X^{-1}(U) \cap \pi_Y^{-1}(V).$$

Now, the preimage of the set on the right under f is given by

$$f^{-1}(U \times V) = f^{-1}(\pi_X^{-1}(U) \cap \pi_Y^{-1}(V)) = f^{-1}(\pi_X^{-1}(U)) \cap f^{-1}(\pi_Y^{-1}(V)).$$

The set on the right is open in Z, by the assumption on the continuity of the composite maps $\pi_X \circ f$ and $\pi_Y \circ f$. Thus f is continuous. □

14. Hausdorff spaces

Definition 14.1. We say that a topological space $(X, \mathcal{T})$ is *Hausdorff* if any two distinct points in X have disjoint open neighborhoods. In other words, for any $x, y \in X$, $x \neq y$, there are $U, V \in \mathcal{T}$ such that $x \in U$, $y \in V$ and $U \cap V = \emptyset$.

The Hausdorff condition helps in distinguishing different points of a space. There are many spaces where this condition is not satisfied, for example, take a set X with at least 2 points and with the indiscrete topology $\{\emptyset, X\}$. Then for every $x \neq y$, there are no disjoint open sets separating these two points, as they would need to be non-empty, hence equal to X.

One reason why the Hausdorff condition is useful is the following observation.

Proposition 14.2. *Let $(X, \mathcal{T})$ be a Hausdorff topological space. Then for every $x \in X$ the singleton subset $\{x\}$ is closed in X.*

Proof. Since X is Hausdorff, for every $y \in X \setminus \{x\}$ there is an open subset $V \subset X$ such that $y \in V$ and $x \notin V$. The latter yields that $V \subseteq X \setminus \{x\}$, so we can conclude that $X \setminus \{x\}$ is open in X by Proposition 10.4. Therefore $\{x\}$ is closed in X, as claimed. □

Proposition 14.3. *Any metric space (X, d), endowed with the metric topology, is Hausdorff.*

Proof. For every two points $x \neq y$ in X, the balls $\mathcal{B}_r(x)$ and $\mathcal{B}_r(y)$, where $r = d(x, y)/2 > 0$ are open and disjoint.

Indeed, if z belongs to the intersection of the two balls, then, by the triangle inequality,

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

which is a contradiction. □

Example 14.4. Let X be an infinite set and let $\mathcal{T}$ be the cofinite topology on X, defined in Example 10.16, which consists of $\emptyset$ and all *cofinite* subsets $U \subseteq X$, i.e., the subsets U with finite complement $U^c = X \setminus U$. This topology is not Hausdorff.

To see this, let x and y be distinct points of X, and let U, V be open sets in X containing x and y respectively. Clearly, U and V are both non-empty, hence their complements U^c and V^c must be finite. But then $(U \cap V)^c = U^c \cup V^c$ is also finite. Since X is infinite, $U \cap V$ must be infinite, hence non-empty.

Combining this example with Proposition 14.3, we can conclude that the cofinite topology on any infinite set cannot be induced by any metric on this set. Moreover, this example also shows that the converse of Proposition 14.2 is false, because it is easy to see that all subsets $\{x\}$ are closed in X with respect to the cofinite topology.

We record the following easy, but useful facts.

Proposition 14.5.
(1) *Every subspace of a Hausdorff space is Hausdorff.*
(2) *The topological product $X \times Y$ (i.e., the cartesian product equipped with the product topology) is Hausdorff if and only if both X and Y are Hausdorff.*
(3) *If $f : X \to Y$ is an injective continuous map of topological spaces and Y is Hausdorff then so is X.*
(4) *If topological spaces X and Y are homeomorphic then X is Hausdorff if and only if Y is Hausdorff. In other words, the Hausdorff condition is a topological property.*

Proof. Exercise. □

15. Connected spaces

Definition 15.1. A *partition* $\{A, B\}$ of a topological space $(X, \mathcal{T})$ is a pair of non-empty open disjoint subsets $A, B \subset X$ such that $X = A \cup B$.

A topological space X is said to be *disconnected* if it admits a partition; otherwise X is said to be *connected*.

Proposition 15.2. *A topological space $(X, \mathcal{T})$ is disconnected if and only if there exists a continuous surjective map from X to the two-point discrete space $\{0, 1\}$.*

Proof. Assume that $\{A, B\}$ is a partition of X and define a function $f : X \to \{0, 1\}$ by

$$f(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \in B \end{cases}.$$

Thus f is the characteristic function of the subset A. The open subsets of $\{0, 1\}$, equipped with the discrete topology, are $\emptyset$, $\{0\}$, $\{1\}$ and $\{0, 1\}$, and the preimages of these sets under f are $\emptyset$, B, A and X respectively. Since all of these preimages are open, f is continuous.

Conversely, suppose that there is a continuous surjective map $f : X \to \{0, 1\}$, where $\{0, 1\}$ is equipped with the discrete topology. Then the sets $A = f^{-1}(0)$ and $B = f^{-1}(1)$ are open and form a partition of X (they are non-empty because f is surjective). □

The following is a useful criterion to establish connectedness.

Proposition 15.3. *A topological space X is connected if and only if the only subsets of X which are both open and closed are X and $\emptyset$.*

Proof. Let us prove the contrapositive of the statement.

If X is disconnected, then X has a partition $\{A, B\}$, where both subsets A, B are non-empty, open, disjoint, and $A \cup B = X$. Then $B^c = A$, so A is also closed. Since both A and B are non-empty, A does not equal X or $\emptyset$.

Conversely, if $A \subseteq X$ is both open and closed, and is neither X nor $\emptyset$, then its complement in X, A^c, is closed and open. Moreover, $A^c \neq \emptyset, X$. Therefore $\{A, A^c\}$ is a partition of X, so X is disconnected. □

We shall now discuss some connected subsets of the real line $\mathbb{R}$, which is endowed with the usual metric topology (its basic open sets are intervals (a, b), $a, b \in \mathbb{R}$, $a < b$). Recall the following definition.

Definition 15.4. A subset $I \subseteq \mathbb{R}$ of the real line is an *interval* if for every three real numbers x, y, z such that $x < y < z$, if $x, z \in I$, then $y \in I$.

An interval $I \subseteq \mathbb{R}$ is said to be *open above* if for every $x \in I$ there exists $y \in I$ such that $x < y$. Similarly, I is *open below* if $\forall\, x \in I\ \exists\, y \in I$ such that $y < x$.

A quick run through the possibilities reveals that an interval in $\mathbb{R}$ has one of the following forms:

$$(a, b),\ [a, b),\ (a, b],\ [a, b],\ (-\infty, a),\ (-\infty, a],\ (b, \infty),\ [b, \infty),\ \mathbb{R}.$$

What is interesting, that this is sufficient to characterise all connected subsets of $\mathbb{R}$.

Theorem 15.5. *A subset of $\mathbb{R}$ is connected if and only if it is an interval.*

Proof. We first prove that any interval I is a connected subset of $\mathbb{R}$. Assume, on the contrary, that there exists a partition $\{A, B\}$ of I. Then there exist $a \in A$ and $b \in B$, and we assume (without loss of generality) that $a < b$. Since $a, b \in I$, and I is an interval, we have that the closed interval $[a, b]$ is contained in I.

Let $A' = A \cap [a,b]$ and $B' = B \cap [a,b]$. As we saw in the proof of Proposition 15.3, the subsets A and B are both open and closed in I (in the subspace topology induced from $\mathbb{R}$) Therefore the subsets A' and B' are closed in the interval $[a,b]$. As the interval $[a,b]$ is closed in $\mathbb{R}$, we deduce that the two sets A' and B' are closed in $\mathbb{R}$ as well. Let $c = \sup A'$. As A' is closed, $c \in A'$ and $c < b$, since $b \in B'$ and $A' \cap B'$ is empty. Now A' is also open in $[a,b]$, and so there exists $\delta > 0$ such that the intersection $(c-\delta, c+\delta) \cap [a,b]$ is contained in A'. In particular, as $c < b$, this intersection contains elements of A' which are greater than c, which contradicts the fact that c is the supremum of A'.

Let us now assume that $S \subseteq \mathbb{R}$ is connected but not an interval. Then by Definition 15.4 there exist $x, y \in S$ and $z \in S^c$ such that $x < z < y$. Then the subsets $(-\infty, z) \cap S$ and $(z, \infty) \cap S$ are both open in S in the subspace topology, they are both non-empty, as the first contains x and the other y, they are disjoint, and their union is S. So the set S is not connected, contradiction. □

Theorem 15.6. *Let $f : X \to Y$ be a continuous surjective map of topological spaces. If X is connected, then so is Y.*

Proof. Suppose that Y is not connected, so there exists a partition $\{A, B\}$ of Y. As f is continuous, $f^{-1}(A)$ and $f^{-1}(B)$ are open in X. Since f is surjective, $f^{-1}(A) \neq \emptyset$ and $f^{-1}(B) \neq \emptyset$. Moreover, $X = f^{-1}(Y) = f^{-1}(A \cup B) = f^{-1}(A) \cup f^{-1}(B)$, and $f^{-1}(A) \cap f^{-1}(B) = f^{-1}(A \cap B) = \emptyset$. Thus $\{f^{-1}(A), f^{-1}(B)\}$ is a partition of X, and so X is disconnected. □

Corollary 15.7. *Let $f : X \to Y$ be a homeomorphism. Then X is connected if and only if Y is connected.*

Thus connectedness is an important topological invariant of a space. Sometimes it can even help to distinguish connected spaces, such as the intervals $[0,1)$ and $(0,1)$ in $\mathbb{R}$.

Proposition 15.8. *The intervals $[0,1)$ and $(0,1)$ in $\mathbb{R}$ are not homeomorphic.*

Proof. Assume that there is a homeomorphism $f : [0,1) \to (0,1)$. Then $0 < f(0) < 1$ and the restriction $g = f|_{(0,1)}$, of f to the interval $(0,1)$, is a homeomorphism between $(0,1)$ and its image, $g((0,1)) = (0,1) \setminus \{f(1)\}$ (this can be deduced from Corollary 13.6). But $(0,1)$ is an interval, hence connected, while the disjoint open intervals $(0, f(0))$, $(f(0), 1)$ form a partition of the set $g((0,1))$, which is therefore not connected. Therefore the two sets cannot be homeomorphic by Corollary 15.7. This contradiction shows that no homeomorphism $f : [0,1) \to (0,1)$ can exist. □

Example 15.9. We shall prove that the unit circle $\mathbb{S}^1 = \{(x,y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ is connected using the fact that the real line $\mathbb{R}$ is connected. For this, let us define the map $e : \mathbb{R} \to \mathbb{R}^2$ by $e(x) = (\cos(2\pi x), \sin(2\pi x))$. Given that the topology on $\mathbb{R}^2$ can be considered to be the product topology on the set $\mathbb{R}^1 \times \mathbb{R}^1$, the map e is continuous by Proposition 13.10. It is easy to see that the image of this map is the circle $\mathbb{S}^1$, and so $\mathbb{S}^1$ is connected as is the image of a connected space $\mathbb{R}$ under the continuous surjective map e (see Theorem 15.6).

Exercise 15.10. Let $(X, \mathcal{T})$ be a topological space with two connected subsets $A, B \subseteq X$. Show that if $A \cap B \neq \emptyset$, then $A \cup B$ is a connected subset of X.

Example 15.11. For each $n \in \mathbb{N}$, the n-sphere $\mathbb{S}^n$ (see Definition 2.1) in $\mathbb{E}^{n+1}$ can be represented as the union of two closed hemispheres $A = \{(x_1, \ldots, x_{n+1}) \in \mathbb{S}^n \mid x_{n+1} \geq 0\}$ and $B = \{(x_1, \ldots, x_{n+1}) \in \mathbb{S}^n \mid x_{n+1} \leq 0\}$. It is not difficult to show that each of the hemispheres is homeomorphic to the closed unit ball $\overline{\mathcal{B}}_1(\mathbf{0})$ in $\mathbb{E}^n$ (the projection $(x_1, \ldots, x_n, x_{n+1}) \mapsto (x_1, \ldots, x_n)$ provides a homeomorphism),

hence both A and B are connected (see Example 16.6 for a proof of the fact that any closed ball in $\mathbb{E}^n$ is connected). The intersection $A \cap B$ is naturally isometric to the sphere $\mathbb{S}^{n-1}$ in $\mathbb{E}^n$, and so it is non-empty. Hence, $\mathbb{S}^n$ is connected, by Exercise 15.10.

Example 15.12. The general linear group $GL(n, \mathbb{R})$ (considered as a subset of $\mathbb{R}^{n^2}$, with the induced topology) is disconnected, as the determinant map provides a continuous surjective map $\det : GL(n, \mathbb{R}) \to \mathbb{R} \setminus \{0\}$, and the target space is disconnected. Similarly, the group $O(n)$ is disconnected, however the special linear group $SL(n, \mathbb{R})$ and the special orthogonal group $SO(n, \mathbb{R})$ are connected.

Theorem 15.13 (Intermediate Value Theorem)**.** *If $f : [a, b] \to \mathbb{R}$ is a continuous function such that $f(a) < 0$ and $f(b) > 0$ then there exists $x \in [a, b]$ such that $f(x) = 0$.*

Proof. If $f(x) \neq 0$ for all $x \in [a, b]$ then define $f^*(x) = f(x)/|f(x)|$. This is continuous, and the values of $f^*(x)$ are ± 1, so f^* is a continuous surjective map onto the two element set $\{-1, 1\}$, which is a discrete space. (Note that $f^*(a) = -1$ and $f^*(b) = 1$.) But this is impossible as $[a, b]$ is connected, so there must exist $x \in [a, b]$ such that $f(x) = 0$. □

We record here the following useful result.

Proposition 15.14. *Let A be a connected subset of a topological space $(X, \mathcal{T})$. Then for any subset $B \subseteq X$ satisfying $A \subseteq B \subseteq \overline{A}$, B is connected. In particular, the closure of a connected set is connected.*

Proof. Let $f : B \to \{0, 1\}$ be a continuous surjective function from B onto the two-point discrete space $\{0, 1\}$. The restriction of f to A is continuous, and since A is connected, it must be constant; so let us assume that $f(x) = 0$ for all $x \in A$. Now suppose that there exists $y \in B$ such that $f(y) = 1$. Since the space $\{0, 1\}$ is discrete, the set $\{1\}$ is open, hence the preimage $f^{-1}(\{1\})$ is open in B, and so there exists an open subset U of X such that $f^{-1}(\{1\}) = B \cap U$. The point y is an element of $\overline{A}$ by assumption, and so it is an adherent point of A. Since $y \in U$ and U is open, the intersection $U \cap A$ is non-empty, by the definition of an adherent point. Thus there exists $z \in U \cap A$, and so $f(z) = 1$. But this contradicts the assumption that $f(x) = 0$ for all elements x of A. Hence $f(y) = 0$ for all $y \in B$, so B is connected. □

16. Path-connected spaces

16.1. **Continuous paths.**

Definition 16.1. A *continuous path* in a topological space $(X, \mathcal{T})$ is a continuous map $\gamma : [a, b] \to X$, where $a < b$ and the interval $[a, b]$ in $\mathbb{R}$ is equipped with the standard (subspace) topology.

In the previous chapter we discussed the notion of connectedness of topological spaces. However, to show that a topological space is connected it is often easier to prove that it satisfies the following stronger property.

Definition 16.2. A space X is said to be *path-connected* if any two points of X can be joined by a continuous path. That is, for any points $x, y \in X$ there exists a continuous path $\gamma : [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$.

A subset A of the topological space X is *path-connected* if any two points x and y of A can be joined by a continuous path whose image is contained in A. This means that there exists a continuous path $\gamma : [a, b] \to X$, whose value at a is x, at b it is y, and $\gamma(t) \in A$ for all $t \in [a, b]$.

Note that the above definition only needs to be checked for distinct points $x, y \in X$, because if $x = y$ then x can be joined with itself by a constant path, which is obviously continuous in any topological space.

Let us start with comparing this new notion to that of connectedness.

Theorem 16.3. *Any path-connected subset of a topological space is connected.*

Proof. Arguing by contradiction, assume that X is path-connected but not connected. Then, according to Proposition 15.2, there exists a surjective continuous function $g : X \to \{0, 1\}$ (the set $\{0, 1\}$ is assumed to have the discrete topology). So there exist $x, y \in X$ such that $g(x) = 0$ and $g(y) = 1$. As X is path-connected, there is a continuous path $\gamma : [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$, which gives a continuous surjective map $g \circ \gamma : [a, b] \to \{0, 1\}$, contradicting the connectedness of the interval $[a, b]$ (cf. Theorem 15.5). □

We will now give an example showing that the converse of Theorem 16.3 does not always hold.

Example 16.4. Consider the subset of $\mathbb{R}^2$ defined to be the union of the interval $X_1 = \{(0, y) \mid y \in [-1, 1]\}$ and the graph of the function $\cos(\pi/x)$ for $x \in (0, 1]$. Then this set is connected but not path-connected. Here is how to prove it. Let $X_2 = \{(x, \cos(\pi/x)) \mid x \in (0, 1]\}$ and let $X = X_1 \cup X_2$.

We first prove that X is connected. Note that X_2 is path-connected. Indeed, for any distinct two points $(x_1, y_1), (x_2, y_2) \in X_2$, we have $y_i = \cos(\pi/x_i)$, $i = 1, 2$. Evidently we can suppose that $x_1 < x_2$. Then the continuous path $\gamma : [x_1, x_2] \to X_2$, given by $\gamma(t) = (t, \cos(\pi/t))$, joins (x_1, y_1) with (x_2, y_2). Thus X_2 is path-connected, and so it is connected by Theorem 16.3.

We shall now show that X is contained in the closure of X_2, which implies that X is connected. Consider arbitrary $(0, y) \in X_1$ and $\varepsilon > 0$. Let N be an even natural number such that such that $1/N < \varepsilon$. Since N is even, we have $\cos(N\pi) = 1$ and $\cos((N + 1)\pi) = -1$. Then, by the Intermediate Value Theorem, there exists $t \in [1/(N + 1), 1/N]$ such that $\cos(\pi/t) = y$. It follows that the distance from the point $(t, \cos(\pi/t)) \in X_2$ to $(0, y)$ in $\mathbb{R}^2$ is at most $1/N < \varepsilon$, and since this is true for an arbitrary ε, the point $(0, y)$ is a limit point of the set X_2. Therefore, $X_1 \subseteq \overline{X_2}$, and since $X_2 \subseteq \overline{X_2}$, by definition of the closure, we deduce that $X = X_1 \cup X_2 \subseteq \overline{X_2}$. Thus X is connected by Proposition 15.14.

We will now prove that X is not path-connected. Assume then that there exists a continuous path $\gamma : [a, b] \to X$ such that $\gamma(a) = (0, 1) \in X_1$ and $\gamma(b) = (1, -1) \in$

X_2. The function γ has coordinate functions γ_1, γ_2 so that $\gamma = (\gamma_1, \gamma_2)$, where γ_1 and γ_2 are continuous real-valued functions on $[a, b]$. Moreover, $\gamma_1(a) = 0$, and $\gamma_1(b) = 1$.

The set $\gamma_1^{-1}(\{0\})$ is a closed (and bounded) subset of $[a, b]$, since γ_1 is continuous and $\{1\}$ is closed in $\mathbb{R}$. Thus it contains its supremum t^*, which satisfies $t^* < b$, since $\gamma_1(b) = 1 \neq 0$. We will prove that the function γ_2 is not continuous at t^*.

If t is such that $t^* < t \leq 1$, then $\gamma_1(t) > 0$, and $\gamma(t) \in X_2$, with $\gamma_2(t) = \cos(\pi/\gamma_1(t))$. Take any $\delta > 0$ which is such that $t^* + \delta \leq 1$. Since $\gamma_1(t^* + \delta) > 0$, there exists a large even N such that

$$\gamma_1(t^*) = 0 < \frac{1}{N+1} < \frac{1}{N} < \gamma_1(t^* + \delta).$$

It now follows from the IVT that there exist u, v in $[t^*, t^* + \delta]$ such that $\gamma_1(u) = 1/(N+1)$ and $\gamma_1(v) = 1/N$. But then $\gamma_2(u) = \cos(N+1)\pi = -1$ and $\gamma_2(v) = \cos N\pi = 1$. Thus if $\gamma_2(t^*) \geq 0$, $|\gamma_2(t^*) - \gamma_2(u)| \geq 1$, while if $\gamma_2(t^*) \leq 0$, we have that $|\gamma_2(t^*) - \gamma_2(v)| \geq 1$. This implies that the function γ_2 is not continuous at t^*. This contradiction shows that no such path γ exists and X is not path-connected.

Having established this result, we can characterise all path-connected subsets of $\mathbb{R}$.

Proposition 16.5. *A subset S of $\mathbb{R}$ is path-connected if and only if it is an interval.*

Proof. Note first that an interval $I \subseteq \mathbb{R}$ is path-connected. Indeed, let $x < y$, and $x, y \in I$ and define $\gamma : [x, y] \to I$ by $\gamma(t) = t$. Then γ is obviously continuous and joins x with y. Now for all t such that $x < t < y$, $x < \gamma(t) < y$, so $\gamma(t) \in I$, by the definition of an interval.

Conversely, if the subset S is path-connected, then it is connected, and so is an interval by Theorem 15.5 □

Example 16.6. In some cases it is much easier to prove that a topological space is path-connected than just connected. For example, this is true for $\mathbb{R}^n$, endowed with the standard topology, or, more generally, for any convex subset A of $\mathbb{R}^n$ (endowed with the subspace topology).

Indeed, if $A \subseteq \mathbb{R}^n$ is convex, then for any two points $\boldsymbol{x}, \boldsymbol{y} \in A$ we can define the continuous path $\gamma : [0, 1] \to \mathbb{R}^n$ to be any continuous parametrization of the Euclidean line segment joining $\boldsymbol{x}$ with $\boldsymbol{y}$. This Euclidean line segment lies in A since A is convex, so A is path-connected, and hence, it is connected by Theorem 16.3.

We have the following easy analogue of Theorem 15.6.

Proposition 16.7. *Let $f : X \to Y$ be a continuous surjective map of topological spaces. If X is path-connected, then so is Y.*

Proof. Consider any $y_1, y_2 \in Y$, then, by the surjectivity of f, there exist points $x_1, x_2 \in X$ such that $f(x_1) = y_1$ and $f(x_2) = y_2$. Since X is path-connected, there exists a continuous path $\gamma : [a, b] \to X$ such that $\gamma(a) = x_1$ and $\gamma(b) = x_2$. But then the composition $f \circ \gamma : [a, b] \to Y$ is a continuous path joining y_1 with y_2 in Y. Hence Y is path-connected. □

As before, this has the following corollary.

Corollary 16.8. *Assume that X and Y are homeomorphic topological spaces. Then X is path-connected if and only if Y is.*

Proof. To prove the statement we simply need to apply Proposition 16.7 to the map f and its inverse f^{-1}. □

Example 16.9. The interval $[0,1]$ in $\mathbb{R}$ is not homeomorphic to the square $[0,1] \times [0,1]$ in $\mathbb{R}^2$.

Indeed, assume that $f : [0,1] \to [0,1] \times [0,1]$ is a homeomorphism. Then the restriction of this map to $[0,1/2) \cup (1/2,1]$ is again a homeomorphism with its image $[0,1] \times [0,1] \setminus \{f(1/2)\}$. However, the set $[0,1/2) \cup (1/2,1]$ is not path-connected, whereas the square with the point $f(1/2)$ removed still is. Hence no such homeomorphism exists.

A similar argument shows that the real line $\mathbb{R}$ is not homeomorphic to the plane $\mathbb{R}^2$.

16.2. **Some topological consequences of the Intermediate Value Theorem.**

Theorem 16.10 (One-dimensional Brouwer Theorem)**.** *Let $f : [a,b] \to [a,b]$ be a continuous function. Then there exists $c \in [a,b]$ such that $f(c) = c$, i.e., c is a fixed point of f.*

Proof. If $f(a) = a$ or $f(b) = b$, there is nothing to prove. So assume that $a < f(a)$ and $f(b) < b$ and define $g : [a,b] \to \mathbb{R}$ by $g(x) = f(x) - x$. Then $g(a) > 0$ while $g(b) < 0$, so there exists $c \in [a,b]$ such that $g(c) = 0$. But then $f(c) = c$. □

An easy variant of this result is the following.

Theorem 16.11 (One-dimensional Borsuk-Ulam Theorem)**.** *Let $\mathbb{S}^1$ be the unit circle in the plane and let $f : \mathbb{S}^1 \to \mathbb{R}$ be a continuous function. Then there exists a point $z \in \mathbb{S}^1$ such that $f(z) = f(-z)$. In other words, the function f maps at least one pair of diametrically opposite points of the circle to the same point of $\mathbb{R}$.*

Proof. Define $g : [0, 2\pi] \to \mathbb{R}$ by $g(t) = f((\cos t, \sin t)) - f((-\cos t, -\sin t))$. Note that g is continuous, as a composition of continuous functions, and $g(\pi) = f((-1,0)) - f((1,0)) = -g(0)$, so that 0 is in the interval between $g(\pi)$ and $g(0)$. Hence the IVT implies that there is a point $c \in [0, 2\pi]$ such that $g(c) = 0$, or $f(z) = f(-z)$, where $z = (\cos c, \sin c) \in \mathbb{S}^1$. □

17. Compact spaces

In Proposition 15.8 we used connectedness to show that the open interval $(0,1)$ in $\mathbb{R}$ is not homeomorphic to the half-closed interval $[0,1)$, and a similar argument would show that $(0,1)$ is not homeomorphic to the closed interval $[0,1]$. In this chapter we will discuss another important property of topological spaces, which can help to distinguish a closed interval from a non-closed one.

17.1. Sequential compactness.

Let us begin by recalling the Bolzano-Weierstrass theorem, which you have seen in the Analysis course.

Theorem 17.1. *Every bounded sequence $\{x_n\}$ of real numbers contains a convergent subsequence.*

Using this theorem, we discover the following property of closed intervals in $\mathbb{R}$.

Theorem 17.2. *Let $[a,b]$ be a closed interval in $\mathbb{R}$. Then every sequence x_n of elements of $[a,b]$ contains a subsequence that converges to a point $x \in [a,b]$.*

Proof. Since $[a,b]$ is a bounded subset of $\mathbb{R}$, the sequence x_n is bounded, and so by the Bolzano-Weierstrass it contains a subsequence that converges to $x \in \mathbb{R}$. As $[a,b]$ is a closed subset of $\mathbb{R}$, $x \in [a,b]$ by Proposition 11.20. □

It is easy to see that open or half-open intervals do not have this property. Indeed, consider the sequence $\{x_n\}$, $x_n = 1/n$, $n \geq 2$, in the interval $(0,1)$. This sequence converges in $\mathbb{R}$ to 0, which is not an element of $(0,1)$. Because the sequence x_n converges, all its subsequence converge to the same limit, and so the sequence x_n does not contain a subsequence that converges to an element of $(0,1)$.

This is an important distinction, and it is worth recording this property as a definition. You have seen a variant of this in the Analysis course (cf. Definition 5.6 in the Analysis lecture notes).

Definition 17.3. A metric space (X,d) is called *sequentially compact* if every sequence $\{x_n\}$ of elements of X contains a subsequence $\{x_{n_k}\}$ which converges to a point in X.

A subset A of a metric space (X,d) is *sequentially compact* if (A,d) is a sequentially compact metric space.

Thus closed intervals of $\mathbb{R}$ are sequentially compact by Theorem 17.2, while open and half-open intervals are not.

Exercise 17.4. Let A be a sequentially compact subset of a metric space (X,d) and let B be a closed subset of A. Prove that B is also sequentially compact.

We have the following initial characterisation of sequentially compact sets in metric spaces.

Theorem 17.5. *A sequentially compact subset K of a metric space (X,d) is closed and bounded.*

Proof. Let y be an element of the closure $\overline{K}$ of K. This means that y is an adherent point of K, so for every $n \in \mathbb{N}$ there is a point $x_n \in K$ such that $d(x_n, y) < 1/n$. Obviously the latter implies that the sequence $\{x_n\}$ converges to y in X. As K is sequentially compact, this sequence has a subsequence which converges to a point $x \in K$. But all subsequences of a convergent sequence converge to the same point (cf. Proposition 7.2), so $x = y$ and, hence, $y \in K$. Thus $K = \overline{K}$ is closed by Proposition 11.9.

Assume that K is not bounded. Pick a point x_0 in K and choose $x_1 \in K$ such that $d(x_1, x_0) > 1$ (such a point x_2 exists, as X is unbounded). Choose $x_2 \in K$

such that $d(x_2, x_0) > 2$, and in general, for each $n \in \mathbb{N}$ we can select a point $x_n \in K$ such that $d(x_n, x_0) > n$.

Let us show that any ball of finite radius in X contains only finitely many elements of the sequence $\{x_n\}$. Indeed, for any $z \in X$ and any $R > 0$ choose $N \in \mathbb{N}$ so that $N \geq d(z, x_0) + R$. Then, by the triangle inequality, for all $n \in \mathbb{N}$ with $n > N$ we have

$$d(x_n, z) \geq d(x_n, x_0) - d(z, x_0) > n - d(z, x_0) > N - d(z, x_0) \geq R.$$

Thus $x_n \notin \mathcal{B}_R(z)$ for all $n > N$.

Therefore no subsequence of the sequence $\{x_n\}$ can be bounded (cf. Proposition 5.12), and hence, by Proposition 7.4, no such subsequence converges in X. This implies that K is not sequentially compact, contradicting our assumption. □

The converse is not true in general metric spaces, but is true in $\mathbb{R}^n$ as we will see. Theorem 17.9 below describes an important property of compact sets which is very helpful in deciding compactness of sets in metric spaces. Before we state it, we need to define a useful metric on the Cartesian product of metric spaces.

Definition 17.6. Let (X, d_X) and (Y, d_Y) be metric spaces. We define the *product metric* $d_{X\times Y}$ on $X \times Y$ by the formula

$$d_{X\times Y}((x_1, y_1), (x_2, y_2)) = d_X(x_1, x_2) + d_Y(y_1, y_2),$$

for all (x_1, y_1) and (x_2, y_2) in $X \times Y$.

Lemma 17.7. *Using the notation of Definition 17.6, $d_{X\times Y}$ is metric on $X \times Y$ in the sense of Definition 1.1.*

Proof. Exercise. □

We will need to understand the behaviour of sequences in $X \times Y$, as summarised in the following proposition.

Proposition 17.8. *Let (X, d_X) and (Y, d_Y) be metric spaces. A sequence (x_n, y_n) converges to a point (x, y) in the product metric space $(X \times Y, d_{X\times Y})$ if and only if $x_n \to x$ in the space X and $y_n \to y$ in the space Y, as $n \to \infty$.*

Proof. By an exercise from Problem Sheet 4, we know that $(x_n, y_n) \to (x, y)$, as $n \to \infty$, in $X \times Y$ if and only if

$$d_{X\times Y}((x_n, y_n), (x, y)) \to 0, \text{ as } n \to \infty.$$

Using the definition of the product metric we have that

$$d_{X\times Y}((x_n, y_n), (x, y)) = d_X(x_n, x) + d_Y(y_n, y),$$

and so, for every $n \in \mathbb{N}$,

$$d_X(x_n, x) \leq d_{X\times Y}((x_n, y_n), (x, y)) \text{ and } d_Y(y_n, y) \leq d_{X\times Y}((x_n, y_n), (x, y)).$$

So, if the sequence $\{(x_n, y_n)\}$ converges to the point $(x, y) \in X \times Y$, then

$$0 \leq d_X(x_n, x) \leq d_{X\times Y}((x_n, y_n), (x, y)) \to 0, \text{ as } n \to \infty,$$

from which it follows that $d(x_n, x)$ converges to 0, hence $\{x_n\}$ converges to x. The same argument shows that y_n converges to y.

Conversely, if $x_n \to x$ and $y_n \to y$, as $n \to \infty$, then

$$d_{X\times Y}((x_n, y_n), (x, y)) = d_X(x_n, x) + d_Y(y_n, y) \to 0$$

and so $(x_n, y_n) \to (x, y)$, as claimed. □

Theorem 17.9. *Let (X, d_X) and (Y, d_Y) be metric spaces and let $X \times Y$ be their Cartesian product regarded as a metric space with the product metric. Then $X \times Y$ is sequentially compact if and only if X and Y are sequentially compact.*

Proof. Assume that $X \times Y$ is sequentially compact. Let $\{x_n\}$ be a sequence in X. If Y is empty, then $X \times Y$ is also empty, and, hence, sequentially compact. Otherwise, choose a point $z \in Y$ and create a sequence $\{\boldsymbol{a}_n\}$ in $X \times Y$, with $\boldsymbol{a}_n = (x_n, z) \in X \times Y$, $n \in \mathbb{N}$. Since $X \times Y$ is sequentially compact, this sequence contains a subsequence $\{(x_{n_k}, z)\}$ which converges to a point (x, y) in $X \times Y$. By Proposition 17.8, this implies that the sequence $\{x_{n_k}\}$, which is a subsequence of $\{x_n\}$, converges to x in X. Thus X is sequentially compact. Obviously, the same argument works for Y.

Now assume that X and Y are sequentially compact. Let $\{\boldsymbol{a}_n\}$ be a sequence in $X \times Y$, so $\boldsymbol{a}_n = (x_n, y_n)$, $n \in \mathbb{N}$. By the sequential compactness of X, the sequence $\{x_n\}$, of elements of X, contains a subsequence $\{x_{n_k}\}$ which converges to a point $x \in X$. Now consider the subsequence $\{y_{n_k}\}$ of the sequence $\{y_n\}$. This is again a sequence in Y and so it contains a subsequence $\{y_{n_{k_l}}\}$ which converges to a point y in Y. Since any subsequence of a convergent sequence converges to the same point, $\{x_{n_{k_l}}\} \to x$ as $l \to \infty$. Thus the subsequence $\left\{(x_{n_{k_l}}, y_{n_{k_l}})\right\}$, of the original sequence $\{\boldsymbol{a}_n\}$, converges to the point $(x, y) \in X \times Y$, by Proposition 17.8. Therefore $X \times Y$ is sequentially compact. □

This theorem extends by induction to any finite Cartesian product of sequentially compact metric spaces. We also have the following important corollary.

Corollary 17.10. *Any subset of $\mathbb{R}^n$ of the form*

$$I = [a_1, b_1] \times \cdots \times [a_n, b_n]$$

is sequentially compact.

Proof. Let $\mathbb{R}^n$ be equipped with the metric d_1, defined in Subsection 3.2. It is easy to see that this is precisely the product metric (cf. Definition 17.6) on $\mathbb{R}^n$, when it is considered as the Cartesian product of n copies of $\mathbb{R}$. The set I is the Cartesian product of closed intervals in $\mathbb{R}$, each of which is sequentially compact, by Theorem 17.2, and so I is sequentially compact by Theorem 17.9, with respect to the metric d_1.

By an exercise from Problem Sheet 3, we know that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $d_2(\boldsymbol{x}, \boldsymbol{y}) \leq d_1(\boldsymbol{x}, \boldsymbol{y})$. It follows that any sequence of points $\{\boldsymbol{x}_n\}$ in I, converging to some $\boldsymbol{x} \in I$ with respect to the metric d_1, also converges to the same $\boldsymbol{x}$ with respect to d_2. Therefore, the space I is also sequentially compact with respect to the standard Euclidean metric d_2 on $\mathbb{R}^n$. □

Definition 17.11. Any set in $\mathbb{R}^n$ which is the product of closed intervals from $\mathbb{R}$ will be called a *cube*.

We can now state the following characterisation of compact subsets of $\mathbb{R}^n$.

Theorem 17.12 (Heine-Borel)**.** *A subset of $\mathbb{R}^n$ is sequentially compact if and only if it is closed and bounded.*

Proof. We have seen in Theorem 17.5 that a sequentially compact subset of a metric space is closed and bounded, so we only need to prove the converse.

Let K be a closed and bounded subset of $\mathbb{R}^n$. As K is bounded, it is contained in some cube I, which is sequentially compact by Corollary 17.10. As K is closed, it is sequentially compact by Exercise 17.4. □

One application of sequential compactness is the following fact.

Theorem 17.13. *If (X, d) is a sequentially compact metric space then it is complete in the sense of Definition 7.12.*

Proof. Suppose that $\{x_n\}$ is a Cauchy sequence in X. Since X is sequentially compact, some subsequence $\{x_{n_k}\}$ of this sequence must converge to a point $x \in X$. Let us show that $x_n \to x$, as $x \to \infty$.

Consider any $\varepsilon > 0$. Since $\lim\limits_{k\to\infty} x_{n_k} = x$, there is $K \in \mathbb{N}$ such that

$$d(x_{n_k}, x) < \frac{\varepsilon}{2}, \text{ whenever } k > K. \tag{17.14}$$

On the other hand, since the sequence $\{x_n\}$ is Cauchy, there exists $N \in \mathbb{N}$ such that

$$d(x_n, x_m) < \frac{\varepsilon}{2}, \text{ whenever } m, n > N. \tag{17.15}$$

Choose any $k \in \mathbb{N}$ so that $k > K$ and $n_k > N$. Then for any $n > N$, combining the estimates (17.15), (17.14) with the triangle inequality, we get

$$d(x_n, x) \le d(x_n, x_{n_k}) + d(x_{n_k}, x) < \varepsilon.$$

This shows that the sequence $\{x_n\}$ converges to the point $x \in X$. Hence X is complete. □

17.2. **Compact spaces.**

We want to extend the notion of sequential compactness to topological spaces, but immediately we face problems. First, sequences are not so useful in topological spaces, as they can be too wild. Secondly, we have seen that compact sets in metric spaces are closed and bounded, and while, of course, it makes sense to talk about closed subsets, we can't really define bounded subsets of general topological spaces. All we can use are open sets.

One example of a non-compact set is the space of real numbers $\mathbb{R}$, as it is not bounded. The sequence $\{n\}_{n\in\mathbb{N}}$ can be used to demonstrate the non-compactness of $\mathbb{R}$ as it is an example of a sequence with no convergent subsequences, as they are all unbounded.

We now want to find a family of open sets that covers $\mathbb{R}$ in a way that mimics the behaviour of the above sequence. This means, that it can't be reduced to a smaller family that still covers $\mathbb{R}$. For this, let us consider the open intervals $U_n = (n - 3/4, n + 3/4)$, $n \in \mathbb{Z}$. Then clearly

$$\bigcup_{n\in\mathbb{Z}} U_n = \mathbb{R}.$$

However, each of the sets U_n contains precisely one integer n, so no smaller subfamily will be sufficient to cover $\mathbb{R}$. This is one motivation for the following definition.

Definition 17.16. Let V be a subset of a topological space $(X, \mathcal{T})$. An *open cover of* V is a family $\{U_i\}_{i\in I}$, of open subsets from X, whose union contains V:

$$V \subseteq \bigcup_{i\in I} U_i.$$

A *subcover* of an open cover $\{U_i\}_{i\in I}$ of V is a subfamily $\{U_j\}_{j\in J}$, for some $J \subseteq I$, which still covers V. Such a subcover is said to be *finite* if it consists of finitely many sets.

A given subset of a topological space can admit both finite and infinite covers. In the beginning of the section we saw an example of an infinite open cover of the real line $\mathbb{R}$. A finite open cover of $\mathbb{R}$ might be $\mathbb{R}$ itself, the half line $(0, \infty)$ together with the half line $(-\infty, 1)$, etc.

Bounded intervals also admit both finite and infinite open covers. Take the open interval $(0, 1)$ in $\mathbb{R}$, which we know to be not sequentially compact. Consider an infinite open cover $\{I_n\}_{n\ge 3}$, given by the family of increasing nested intervals

$I_n = (1/n, 1 - 1/n)$, for $n \geq 3$. No smaller subfamily of this cover will suffice to cover this set, as any finite selection of these intervals will be contained in the largest one, and none of the intervals I_n is the same as $(0, 1)$.

On the other hand, the family I_n is not an open cover of the *closed* interval $[0, 1]$, because the union of the intervals I_n is the open interval $(0, 1)$.

Definition 17.17. Let $(X, \mathcal{T})$ be a topological space. X is said to be *compact* if every open cover of X has a finite subcover.

A subset V of X is *compact*, if V is compact as a topological space with the subspace topology from X (this means that any cover of V by open subsets from X has a finite subcover).

Example 17.18. (a) If X is a finite set, then $(X, \mathcal{T})$ is compact for any topology $\mathcal{T}$ on X. Indeed, suppose that $X = \{x_1, \dots, x_k\}$ and $\mathcal{U} = \{U_i\}_{i \in I}$ is an open cover of X. Then for each $j = 1, \dots, k$ there exists $i_j \in I$ such that $x_j \in U_{i_j}$. It follows that $X \subseteq \bigcup_{j=1}^{k} U_{i_j}$, i.e., $\{U_{i_j}\}_{j=1}^{k}$ is a finite subcover of the cover $\mathcal{U}$.

(b) Let X be any set and let $\mathcal{T}$ be the cofinite topology on X, defined in Example 10.16. Let us show that the topological space $(X, \mathcal{T})$ is compact. Indeed, consider any open cover $\mathcal{U} = \{U_i\}_{i \in I}$ of X. If X is empty then $X \subseteq U_i$ for all $i \in I$, so any set $U_i \in \mathcal{U}$ forms a finite subcover of X. Thus we can assume that X is non-empty, hence $U_{i_0} \neq \emptyset$ for some $i_0 \in I$. From the definition of the cofinite topology it follows that the complement of U_{i_0} in X is finite, so $X \setminus U_{i_0} = \{x_1, \dots, x_k\}$, for some non-negative integer k. Since X is covered by $\mathcal{U}$, we can argue as before to find $i_1, \dots i_k \in I$ such that $x_j \in U_{i_j}$ for each $j = 1, \dots, k$. It follows that $\{U_{i_j}\}_{j=0}^{k}$ is a finite subcover of X in $\mathcal{U}$. Thus $(X, \mathcal{T})$ is compact.

Exercise 17.19. Let X be an infinite set, endowed with the discrete topology $\mathcal{T}$ (see Example 10.13). Show that $(X, \mathcal{T})$ is not compact.

In light of Definition 17.17 and the discussion above we can see that the open interval $(0, 1)$ (equipped with the subspace topology from $\mathbb{R}$) is not compact, nor is the real line $\mathbb{R}$. This agrees with our previous results on sequential compactness, where we saw that these sets are not sequentially compact. It clearly is interesting to work out if these two notions agree where they both make sense, that is for metric spaces.

Theorem 17.20. *A metric space (X, d) is compact if and only if it is sequentially compact.*

Proof. "⇒" Let us assume that a metric space (X, d) is compact, and let $\{x_n\}$ be a sequence of elements of X. For each $k \in \mathbb{N}$ define the set $F_k = \{x_n \mid n \geq k\}$. Let $\overline{F_k}$ denote the closure of F_k in X, and let $U_k = \overline{F_k}^c$ be the complement of $\overline{F_k}$ in X, $k \in \mathbb{N}$. Then U_k is an open subset of X and

$$U_k \subseteq U_l, \text{ whenever } k \leq l, \ k, l \in \mathbb{N}, \tag{17.21}$$

because $F_k \supseteq F_l$, by definition, and hence $\overline{F_k} \supseteq \overline{F_l}$ by Proposition 11.14.(1).

Let us show that $\bigcap_{k \in \mathbb{N}} \overline{F_k} \neq \emptyset$. Indeed, otherwise we would have

$$X = \left(\bigcap_{k \in \mathbb{N}} \overline{F_k} \right)^c = \bigcup_{k \in \mathbb{N}} U_k,$$

which would means that $\mathcal{U} = \{U_k \mid k \in \mathbb{N}\}$ is an open cover of X. The compactness of X would then yield that there exist natural numbers $k_1 < \dots < k_m$ such that $\{U_{k_i} \mid i = 1, \dots, m\}$ is a finite subfamily of $\mathcal{U}$ covering X. Thus, in view of (17.21), we would have

$$X \subseteq \bigcup_{i=1}^{m} U_{k_i} = U_{k_m}.$$

The latter would yield that $\overline{F_{k_m}} = (U_{k_m})^c = \emptyset$, and so $F_{k_m} = \emptyset$, contradicting the construction of these sets.

Hence there is a point $x \in \bigcap_{k\in\mathbb{N}} \overline{F_k}$. This means that any open ball centred at x must intersect each of the sets F_k, $k \in \mathbb{N}$. Therefore we can construct a strictly increasing sequence $\{n_s\}_{s=1}^{\infty}$, of natural numbers, such that $x_{n_s} \in \mathcal{B}_{1/s}(x)$ in X, i.e., $d(x_{n_s}, x) < 1/s$. It follows that $\{x_{n_s}\}$ is a subsequence of the original sequence $\{x_n\}$, which converges to x in X. Hence X is sequentially compact.

"$\Leftarrow$" The proof of the converse statement is a bit more involved and relies on a precise analysis of coverings of a space. This material is included here for completeness and for your interest, but will not be required for the exam.

Definition 17.22. Let (X, d) be a metric space and let $\mathcal{U}$ be a cover (not necessarily open) of a subset A of X. A real number $\varepsilon > 0$ is said to be a *Lebesgue number* for $\mathcal{U}$ if for every $a \in A$ the ball $\mathcal{B}_\varepsilon(a)$ is contained in one of the sets from $\mathcal{U}$.

It is clear that if ε is a Lebesgue number then so is any positive number $\delta \leq \varepsilon$.

Example 17.23. Consider the cover of $\mathbb{R}$ using open intervals $U_m = (m-1/4, m+5/4)$ for all $m \in \mathbb{Z}$. Then this is an open cover of $\mathbb{R}$ with the Lebesgue number $\varepsilon = \frac{1}{4}$.

We will need the following statement.

Proposition 17.24. *Every open cover $\mathcal{U}$ of a sequentially compact metric space (X, d) has a (positive) Lebesgue number.*

Proof. If a cover $\mathcal{U}$ of X has no Lebesgue number, then for every $n \in \mathbb{N}$, there exists a point $x_n \in X$ such that the ball $\mathcal{B}_{1/n}(x_n)$ is not contained in a single set of the cover $\mathcal{U}$. As the space X is sequentially compact, the sequence $\{x_n\}$ contains a convergent subsequence $\{x_{n_k}\}$ which converges to a point $x \in X$. Since $\mathcal{U}$ is a cover of X, there exists $U \in \mathcal{U}$ such that $x \in U$. As U is open, there exists $\varepsilon > 0$ such that $\mathcal{B}_{2\varepsilon}(x) \subseteq U$. On the other hand, as $\{x_{n_k}\}$ converges to x, there exists N such that for all $k \geq N$, $x_{n_k} \in \mathcal{B}_\varepsilon(x)$.

Choose $k \in \mathbb{N}$ large enough so that $1/n_k < \varepsilon$. Let $y \in \mathcal{B}_{1/n_k}(x_{n_k})$. Then

$$d(y, x) \leq d(y, x_{n_k}) + d(x_{n_k}, x) < 1/n_k + \varepsilon < 2\varepsilon,$$

which shows that $\mathcal{B}_{1/n_k}(x_{n_k}) \subseteq \mathcal{B}_{2\varepsilon}(x) \subseteq U$, which is a contradiction. □

We now need to establish when a family of balls of a given size can cover the space X. We shall use the following notion.

Definition 17.25. A subset S of a metric space (X, d) is called an *ε-net* (where ε is a positive real number), if the family of open balls $\{\mathcal{B}_\varepsilon(x)\}_{x\in S}$, is a cover of X.

An example of an ε-net in $\mathbb{R}$ may be given by the subset $S = \mathbb{Z}$ of the integers, for any $\varepsilon > 1/2$.

Proposition 17.26. *If X is a sequentially compact metric space, then for every $\varepsilon > 0$ there exists a finite ε-net in X.*

Proof. Arguing by contradiction, assume that there exists a positive ε for which X has no finite ε-net. Choose a point x_1 in X. Since X does not admit a finite ε-net, in particular it cannot be covered with a single ball and so there exists a point $x_2 \in X$ such that $d(x_1, x_2) \geq \varepsilon$. Assume that we have selected this way a set $\{x_1, \ldots, x_n\}$ of points such that $d(x_i, x_j) \geq \varepsilon$, for all i, j, $1 \leq i < j \leq n$. Since X has no finite ε-net, there exists x_{n+1} such that its distance to every point x_i is greater than or equal to ε. Thus, by induction, we can construct an infinite sequence $\{x_n\}$ with the property that the distance between any two of its terms is at least ε. Clearly such a sequence has no Cauchy subsequence, and so it has no

convergent subsequence. This contradicts the assumption that X is sequentially compact. □

Finally, we can finish the proof of Theorem 17.20. We need to show that every sequentially compact metric space X is compact.

Let $\mathcal{U}$ be an open cover of X, then there exists a Lebesgue number $\varepsilon > 0$ for $\mathcal{U}$, by Proposition 17.24, and a finite ε-net S by Proposition 17.26. Suppose that $S = \{x_1, \ldots, x_k\}$. The definition of the Lebesgue number implies that for every $i = 1, \ldots, k$ there exists a single set $U_i \in \mathcal{U}$ such that $\mathcal{B}_\varepsilon(x_i) \subseteq U_i$.

But then the finite family of balls $\{\mathcal{B}_\varepsilon(x_i)\}_{i=1}^k$ is a cover of X, and as each $\mathcal{B}_\varepsilon(x_i)$ is contained in a set U_i, the family $\{U_1, \ldots, U_k\}$ is a finite subcover of X. Thus X is compact, which ends the proof of Theorem 17.20. □

Corollary 17.27. *Any closed interval $[a, b]$ in $\mathbb{R}$ (or, more generally, any cube in $\mathbb{R}^n$) is compact.*

Proof. Any closed interval in $\mathbb{R}$ is sequentially compact by Theorem 17.2 (and any cube in $\mathbb{R}^n$ is sequentially compact by Corollary 17.10). Thus the claim follows from Theorem 17.20. □

Proposition 17.28. *Let X be a compact topological space. Then every closed subset A of X is compact.*

Proof. Since A is closed, its complement $A^c = X \setminus A$ is open. Let $\mathcal{U}$ be any cover of A by open subsets of X. If we add A^c to this family we obtain an open cover of the space X. But X is compact, so this cover of X has a finite subcover containing finitely many sets $U_1, \ldots, U_n \in \mathcal{U}$ and, perhaps, A^c. Thus $A \subseteq \bigcup_{j=1}^n U_j \cup A^c$, and since $A \cap A^c = \emptyset$, we can remove A^c to obtain a finite subcover $\{U_j \mid j = 1, \ldots, n\}$, of $\mathcal{U}$, which still covers A. Hence A is compact. □

In view of Theorem 17.5 we could expect that any compact subset K of a topological space $(X, \mathcal{T})$ closed in X. However, this requires a further assumption.

Proposition 17.29. *Let K be a compact subset of a Hausdorff topological space $(X, \mathcal{T})$. Then K is closed in X.*

Proof. Let $y \in K^c = X \setminus K$. Using the fact that the space X is Hausdorff, for every $x \in K$ there exist disjoint open subsets U_x and V_x of X such that $x \in U_x$, $y \in V_x$. Then the family $\{U_x\}_{x \in K}$ forms and open cover of K, and since K is compact, it has a finite subcover $U_{x_1}, \ldots, U_{x_n}$. But then, the intersection $V = V_{x_1} \cap \cdots \cap V_{x_n}$ is open, by the definition of a topology, and contains the point y. Moreover, for every $i = 1, \ldots, n$, the set V is disjoint from U_{x_i} and so V and K are disjoint. Thus $V \subseteq K^c$, and so K^c contains an open neighborhood of y. Since this holds for all $y \in K^c$, we can use Lemma 10.4 to conclude that K^c is open in X. Hence K is closed, as the complement of an open set. □

17.3. **Tychonoff's theorem.**

We have proved in Theorem 17.9 that if X and Y are two metric spaces, then their Cartesian product $X \times Y$ is sequentially compact if and only if both X and Y are sequentially compact.

This theorem is a special case of Tychonoff's theorem, which we now state.

Theorem 17.30 (Tychonoff's theorem)**.** *A topological product $X \times Y$ of non-empty spaces X and Y is compact if and only if both X and Y are compact.*

It is clear that this theorem generalises by induction to finite products of compact spaces.

Proof. (The presented proof of Tychonoff's theorem follows Ken Brown's Cornell lecture notes.)

"$\Rightarrow$" Suppose that $X \times Y$ is compact. Let us show that X is also compact (the proof that Y is compact can be done similarly). Suppose that $\mathcal{U} = \{U_i \mid i \in I\}$ is an open cover of X. Then, by the definition of the topology on $X \times Y$ (see Definition 13.8), each subset $U_i \times Y$ is open in $X \times Y$. Since $\mathcal{U}$ covers X, the family $\{U_i \times Y \mid i \in I\}$ covers $X \times Y$. So, by compactness, there must exist $i_1, \dots, i_k \in I$ such that the subfamily $\{U_{i_j} \times Y \mid j = 1, \dots, k\}$ covers $X \times Y$. Since Y is non-empty, the latter implies that $\{U_{i_j} \mid j = 1, \dots, k\}$ is a finite subcover of X in $\mathcal{U}$. Hence X is compact.

"$\Leftarrow$" Let us start with recalling our definition of compactness. A space X is compact if it has the following property: *if a family $\mathcal{U}$, of open sets, covers the space X then a finite subcollection of $\mathcal{U}$ covers X.* It is convenient to use the contrapositive of this condition: *if a family $\mathcal{U}$, of open sets, is such that no finite subcollection of $\mathcal{U}$ covers X then $\mathcal{U}$ does not cover X.* The proof follows this idea.

So, let us assume that X and Y are compact and $\mathcal{W}$ is a family of open sets in $X \times Y$ with the property that no finite subcollection of $\mathcal{W}$ covers $X \times Y$. Our goal is to show that $\mathcal{W}$ does not cover $X \times Y$. We will do this using the following two claims.

Claim 17.31. *There exists $x_0 \in X$ such that for any open neighborhood U of x_0 in X, the open set $U \times Y$ is not finitely covered by $\mathcal{W}$.*

Proof. If the claim is false, then every $x \in X$ has an open neighbourhood U_x such that $U_x \times Y$ is finitely covered by $\mathcal{W}$. The compactness of X implies then that there is a finite subset $S \subseteq X$ such that $X \subseteq \bigcup_{x \in S} U_x$. Hence the finite subfamily $\{U_x \times Y \mid x \in S\}$, of covers $X \times Y$. Since each of the sets $U_x \times Y$ is finitely covered by $\mathcal{W}$, we can deduce that $X \times Y$ is covered by a finite subcollection of $\mathcal{W}$, which contradicts our hypothesis. □

Claim 17.32. *There exists $y_0 \in Y$ such that no open set $U \times V$ containing (x_0, y_0) is finitely covered by the family $\mathcal{W}$.*

Proof. If this statement is false, then for every $y \in Y$ there is a finitely covered open set $U_y \times V_y$ containing (x_0, y). The compactness of Y implies that there exists a finite subset $F \subseteq Y$ such that $Y = \bigcup_{y \in F} V_y$. If we now define U to be the intersection of all sets U_y, $U = \bigcap_{y \in F} U_y$, then U is open in X, $x_0 \in U$ and the set

$$U \times Y = \bigcup_{y \in F} U \times V_y \subseteq \bigcup_{y \in F} U_y \times V_y$$

is finitely covered by $\mathcal{W}$, which contradicts Claim 17.31. □

The theorem now follows from Claim 17.32. Indeed, note that the point $z = (x_0, y_0) \in X \times Y$ is such that no basic open set (i.e., a set of the form $U \times V$, where U is open in X and V is open in Y), which contains it, is finitely covered by the

family $\mathcal{W}$. In particular, this implies that no basic open set which contains z can be contained in any set $W \in \mathcal{W}$, which further implies that $z \notin \bigcup_{W\in\mathcal{W}} W$. We conclude that $\mathcal{W}$ does not cover the space $X \times Y$, as required. □

Tychonoff's theorem is known in far greater generality and it is this general statement that is very useful in all manner of applications.

Theorem 17.33 (Tychonoff)**.** *Given an arbitrary family $\{X_i\}_{i\in I}$ of compact topological spaces, their product*

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

is compact with respect to the product topology.

We need to explain how to define the product of spaces indexed by an arbitrary indexing set I and what is the product topology on it. So let us take a family of sets X_i, for $i \in I$. Let X be the union of all these sets:

$$X = \bigcup_{i\in I} X_i.$$

Then, the *Cartesian product* $\prod_{i\in I} X_i$ is defined by

$$\prod_{i\in I} X_i = \{f : I \to X |\ \forall\, i \in I,\ f(i) \in X_i\}.$$

To get a feeling for what this means, let us consider the Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$. In this case the indexing set is $I = \{1, 2\}$, $X_1 = \mathbb{R}$, and $X_2 = \mathbb{R}$. The union of these sets is $X = X_1 \cup X_2 = \mathbb{R} \cup \mathbb{R} = \mathbb{R}$. Then the Cartesian product $\mathbb{R} \times \mathbb{R}$ consists of all functions $f : I \to X$ such that $f(1) \in X_1 = \mathbb{R}$, and $f(2) \in X_2 = \mathbb{R}$. Thus a function of this type is simply a pair of real numbers $(f(1), f(2))$, which can be arbitrary. If we denote this pair in the usual way $(f(1), f(2)) = (x, y)$, then the above definition applied to this special case simply says that $\mathbb{R} \times \mathbb{R}$ is the set of all pairs of real numbers (x, y), which agrees with the usual definition of the Cartesian product.

The *product topology* on $P = \prod_{i\in I} X_i$ is defined as follows. A subset $U \subseteq P$ is open if it is a union of products $\prod_{i\in I} U_i$, where for each $i \in I$, U_i is an open subset of X_i, and $U_i \neq X_i$ for only finitely many indices $i \in I$. It is easy to check that this is indeed a topology on P, and, when I is finite, it coincides with the topology given by Definition 13.8.

17.4. **Continuity and compactness.**

The following theorem shows that compactness is preserved by continuous maps. In particular, compactness is a topological property.

Theorem 17.34. *Let $f : X \to Y$ be a continuous map of topological spaces. If K is a compact subset of X then its image $f(K)$ is compact in Y.*

Proof. Let $\{U_\alpha\}_{\alpha\in I}$ be an open cover of the set $f(K)$ in Y. Since each of the sets U_α is open in Y and f is continuous, the inverse image $f^{-1}(U_\alpha)$ is open in X, for every $\alpha \in I$. The family $\{f^{-1}(U_\alpha)\}$ forms an open cover of K. Indeed, for any $x \in K$, $f(x)$ lies in $f(K)$. Since the family $\{U_\alpha\}$ is a cover of K, $f(x)$ must belong to one of those sets, say U_β, for some $\beta \in I$. But this means that $x \in f^{-1}(U_\beta)$; hence these sets form an open cover of K.

Since K is compact there exists a finite subcover $\{f^{-1}(U_j) \mid j = 1, \ldots, n\}$ of K, which implies that the family $\{U_j \mid j = 1, \ldots, k\}$ forms a finite subcover of $f(K)$. Thus $f(K)$ is compact. □

Corollary 17.35. *Let $f : X \to Y$ be a continuous function from a compact topological space X to a metric space Y. Then the image $f(X)$ is a closed and bounded subset of Y.*

Proof. The set $f(X)$, of values of f in Y, is compact by Theorem 17.34, hence it is closed and bounded in Y by Theorem 17.5. □

Another simple corollary of Theorem 17.34 is the following.

Corollary 17.36. *Let X and Y be homeomorphic topological spaces. Then they are either both compact or both non-compact.*

This corollary gives another way to prove that a closed, bounded interval in $\mathbb{R}$ is not homeomorphic to an open/half-open interval, as one is compact and the other is not.

Proposition 17.37. *Suppose that K is a non-empty compact topological space and $f : K \to \mathbb{R}$ is a continuous function. Then f attains its supremum, i.e., there exists $x \in K$ such that*

$$f(x) = \sup_{p \in K} \{f(p)\}.$$

Proof. The image $f(K)$ is a non-empty closed and bounded subset of $\mathbb{R}$ by Corollary 17.35, and therefore it has a supremum $s \in \mathbb{R}$. Then there exists a sequence $\{y_n\}$, of points in $f(K)$, that converges to s. But $f(K)$ closed, so $s \in f(K)$, by Proposition 11.20. The latter is equivalent to saying that there exists a point $x \in K$ such that $f(x) = s$, as claimed. □

One of the most useful applications of this result is the following min-max theorem (cf. Theorem 6.17 from your Analysis notes).

Theorem 17.38. *Let $f : [a, b] \to \mathbb{R}$ be a continuous map, where $a, b \in \mathbb{R}$, $a \leq b$. Then f is bounded above and below and attains its bounds. This implies that $f([a, b]) = [s, t]$, where $s \in \mathbb{R}$ is the infimum and $t \in \mathbb{R}$ is the supremum of the set $f([a, b])$.*

Proof. Since the interval $[a, b]$ is connected, and f is continuous, $f([a, b])$ has to be connected by Theorem 15.6, and so it must be an interval in $\mathbb{R}$, by Theorem 15.5. As $[a, b]$ is compact (see Corollary 17.27), $f([a, b])$ is a closed and bounded subset of $\mathbb{R}$ by Corollary 17.35. Hence $f([a, b]) = [s, t]$ where s is the infimum, and t is the supremum of the set $f([a, b])$. □

18. Quotient topology

This section is devoted to the development of an important topological tool, where more complicated spaces are constructed from simpler ones 'by identification'. For example, if we take a square and glue together opposite sides (preserving direction, more on this later) we will obtain the torus in $\mathbb{R}^3$. It is well known that the Möbius strip is obtained from gluing two sides of a rectangle where one side was rotated through 180° before gluing.

A convenient way to describe such constructions is by using equivalence relations. We recall the relevant notions.

Definition 18.1. An *equivalence relation* $\sim$ on a set X is a subset $R_\sim \subseteq X \times X$ such that

(1) ($\sim$ is *symmetric*) for all $x \in X$, $(x, x) \in R_\sim$ (i.e., $x \sim x$);
(2) ($\sim$ is *reflexive*) for all $x, y \in X$, if $(x, y) \in R_\sim$ then $(y, x) \in R_\sim$ (i.e., if $x \sim y$ then $y \sim x$);
(3) ($\sim$ is *transitive*) for all $x, y, z \in X$, if $(x, y) \in R_\sim$ and $(y, z) \in R_\sim$ then $(x, z) \in R_\sim$ (i.e., if $x \sim y$ and $y \sim z$ then $x \sim z$).

An equivalence relation determines a partition of X into disjoint subsets, called the *equivalence classes*. The equivalence class $[x]$, of an element $x \in X$, is the set of all $y \in X$ such that $x \sim y$ (i.e., $(x, y) \in R_\sim$).

The set of all equivalence classes of an equivalence relation $\sim$ is denoted $X/\sim$. There is a canonical surjective map $p : X \to X/\sim$, defined by $x \mapsto [x]$, for all $x \in X$.

When X is a topological space, it is natural to ask if we can define a topology on the quotient space $X/\sim$, which would make that map continuous.

Definition 18.2. Suppose that $(X, \mathcal{T})$ is a topological space and $\sim$ is an equivalence relation on X. The *quotient topology* $\widetilde{\mathcal{T}}$ on $Y = X/\sim$ consists of all subsets $U \subseteq Y$ such that $p^{-1}(U)$ is open in X.

Proposition 18.3. *Let $(X, \mathcal{T})$ be a topological space and let $\sim$ be an equivalence relation on X. Then the quotient topology $\widetilde{\mathcal{T}}$ is indeed a topology on the quotient space $X/\sim$. Moreover, the quotient map $p : X \to X/\sim$ is continuous (when X and $X/\sim$ are equipped with the topologies $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ respectively).*

Proof. We need to check the axioms of topology. First, by definition, $p^{-1}(X/\sim) = X$, so $X \in \widetilde{\mathcal{T}}$. Also, $p^{-1}(\emptyset) = \emptyset$, so again $\emptyset \in \widetilde{\mathcal{T}}$.

Let U and V be elements of $\widetilde{\mathcal{T}}$; since $p^{-1}(U \cap V) = p^{-1}(U) \cap p^{-1}(V)$, and $p^{-1}(U)$, $p^{-1}(V)$ belong to $\mathcal{T}$, we have that $U \cap V \in \widetilde{\mathcal{T}}$.

Similarly, for any family of subsets U_α in $\widetilde{\mathcal{T}}$, indexed by some set I,

$$p^{-1}\left(\bigcup_{\alpha \in I} U_\alpha\right) = \bigcup_{\alpha \in I} p^{-1}(U_\alpha),$$

which belongs to $\widetilde{\mathcal{T}}$ since every set $p^{-1}(U_\alpha)$ does.

Continuity of p follows from the definition of $\widetilde{\mathcal{T}}$: if $U \in \widetilde{\mathcal{T}}$ then by construction $p^{-1}(U) \in \mathcal{T}$. □

Exercise 18.4. Suppose that $(X, \mathcal{T})$ is a topological space and $\sim$ is an equivalence relation on X. Show that the quotient topology $\widetilde{\mathcal{T}}$ is the finest topology which makes the quotient map $p : X \to Y = X/\sim$ continuous. In other words, if $\mathcal{S}$ is any other topology on Y such that p is continuous then $\mathcal{S} \subseteq \widetilde{\mathcal{T}}$ (i.e., $\mathcal{S}$ is coarser than $\mathcal{T}$ and $\mathcal{T}$ is finer than $\mathcal{S}$ in the sense of Definition 10.17).

Before we look at examples, let us consider the following general results.

Definition 18.5. We say that a map $f : X \to Y$ of topological spaces is a *quotient map* if f is surjective and satisfies the following condition: a subset $U \subseteq Y$ is open in Y if and only if $f^{-1}(U)$ is open in X.

A quotient map is continuous, by definition. When we consider this notion in the context of the quotient topology on the set of equivalence classes of an equivalence relation, then it is clear that the map $p : X \to X/\sim$ is a quotient map in the sense of this definition.

This definition is particularly useful when we combine it with the results that describe the behaviour of compactness, connectedness and path-connectedness with respect to surjective continuous maps.

Proposition 18.6. *Let $f : X \to Y$ be a quotient map. If X is compact, then so is Y. Similarly, if X is connected, then so is Y. If X is path-connected, then so is Y.*

Proof. This statement is simply a summary of results obtained earlier for surjective continuous maps between topological spaces (see Theorems 17.34, 15.6 and 16.7). □

We will also need the following simple statements.

Proposition 18.7. *If $p : X \to Y$ is a quotient map and $g : Y \to Z$ is a map of topological spaces, then g is continuous if and only if $g \circ p$ is continuous.*

Proof. Exercise. □

The following "inverse function theorem" will be useful in our discussion of geometric examples.

Proposition 18.8. *Suppose that X is a compact topological space and Y is a Hausdorff topological space. Then any continuous bijection $f : X \to Y$ is a homeomorphism.*

Proof. We need to show that $f^{-1} : Y \to X$ is continuous. Suppose $V \subseteq X$ is closed, then, since X is compact, V is compact by Proposition 17.28. And so $f(V)$ is compact in Y, by Theorem 17.34. Since Y is Hausdorff, a compact subset of Y is closed (by Proposition 17.29), hence $f(V)$ is closed in Y.

Thus $(f^{-1})^{-1}(V) = f(V)$ is closed in Y for every closed subset V in X, hence f^{-1} is continuous, by Exercise 12.4.(a). □

Let us now look at some examples.

Example 18.9. The circle can be obtained by gluing the endpoints of a closed interval, which can be regarded as an equivalence relation, where the only non-trivial equivalence is that of the two end points, and no other two points are equivalent.

To be precise, let us consider the interval $[0, 2\pi]$ and identify $0 \sim 2\pi$. Then, by the construction of quotient topology, the circle acquires a topology.

There are a number of possible explicit maps that would enable us to identify the quotient space with the unit circle in the plane. For example, define $f : [0, 2\pi] \to \mathbb{S}^1$ by $f(t) = (\cos t, \sin t)$. Then f is continuous (since the compositions of f with the two coordinate projection maps are continuous, f is surjective, it is injective on the interior $(0, 2\pi)$ of the domain interval, and $f(0) = f(2\pi)$. Thus f induces a bijective map $g : [0, 2\pi]/\sim \to \mathbb{S}^1$, such that $f = g \circ p$, where $p : [0, 2\pi] \to [0, 2\pi]/\sim$ is the canonical quotient map. Given that the quotient space $[0, 2\pi]/\sim$ is compact and the circle $\mathbb{S}^1$ is Hausdorff (as a subspace of the metric space $\mathbb{R}^2$), the map g is a homeomorphism.

Example 18.10. Let X be the unit square $X = [0,1] \times [0,1]$ in $\mathbb{R}^2$, and let us define an equivalence relation on X by $(0,t) \sim (1,t)$ for all $t \in [0,1]$, then the quotient space $X/\sim$ is can be identified with the cylinder, as in Figure 8.

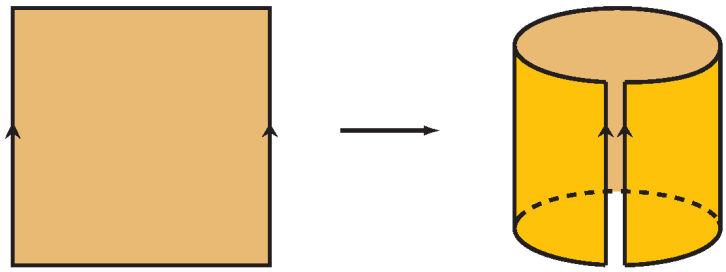


Figure 8: The cylinder.

Example 18.11. Starting with the same unit square X as before, but introducing the relation $(0,t) \sim (1,1-t)$ produces the Möbius strip, as in Figure 9. Note that because of the twist of the ends required by the equivalence relation, the resulting shape has only one side.

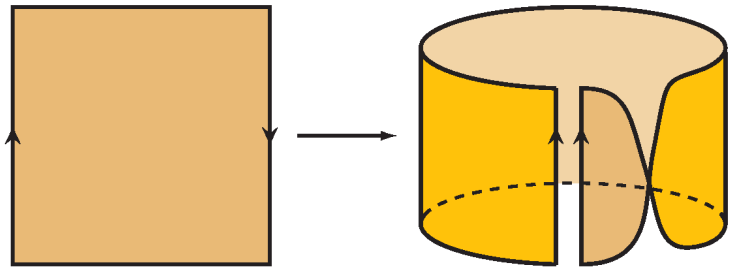


Figure 9: The Möbius strip.

Example 18.12. The standard sphere $\mathbb{S}^2$ in $\mathbb{R}^3$ can also be obtained as a quotient space; a possible equivalence relation on the square X is suggested in Figure 10. Specifically, this is given by $(s,0) \sim (0,s)$ and $(t,1) \sim (1,t)$.

Example 18.13. Another interesting example is the torus, which is identified with the space of equivalence classes of the relation $(0,s) \sim (1,s)$ and $(t,0) \sim (t,1)$ on the unite square in $\mathbb{R}^2$. The four corners of the square are all identified, and the illustration in Figure 11 shows a neighbourhood of the resulting point in the quotient space, as determined by the quotient topology.

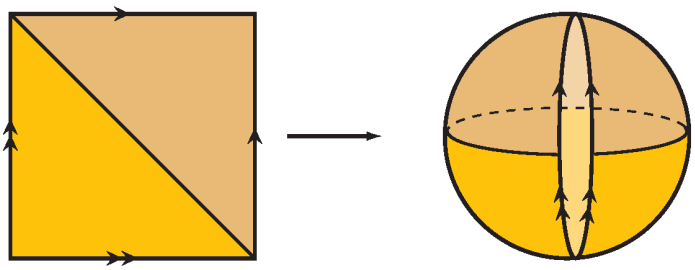


Figure 10: The sphere.

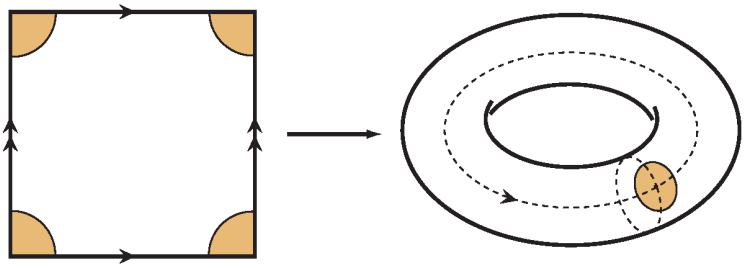


Figure 11: The torus.

An explicit parametrisation of the torus, which builds on our example of the circle, can be given by a map $f : [0, 2\pi] \times [0, 2\pi] \to \mathbb{R}^3$, defined by

$$f(s, t) = ((a + r\cos t)\cos s, (a + r\cos t)\sin s, r\sin t).$$

Here we imagine rotating a circle in the xOz-plane, with centre at $(a, 0, 0)$ and radius r (assumed to be smaller than a) around the z-axis.

Example 18.14. Finally, reversing the orientation of the right vertical edge of the square X before gluing them produces the Klein bottle. Specifically, this is defined using the equivalence relation $(s, 0) \sim (s, 1)$ and $(0, t) \sim (1, 1 - t)$ for all $s, t \in [0, 1]$. This is shown in Figure 12.

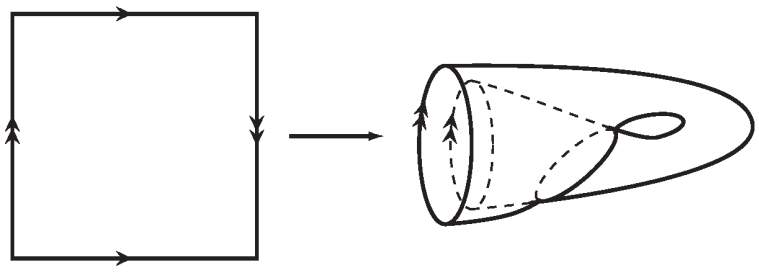


Figure 12: The Klein bottle.