A2: Metric Spaces

Alan Lauder[1]

[1] Aside from Chapter 1, these are mildly edited versions of notes written by Ben Green and Kevin McGerty which were in turn based upon notes due to Richard Earl. Many thanks to them all, and also Dmitry Belyaev, Panos Papazoglou and Melanie Rupflin.

Contents

CHAPTER 1

Differentiability in $\mathbf{R}^2$

We shall use the usual Euclidean notion of size and distance on $\mathbf{R}^n$. So, for example, with $\underline{x} = (x, y) \in \mathbf{R}^2$ we have $|\underline{x}| = \sqrt{x^2 + y^2}$. We are primarily interested in functions on $\mathbf{R}^2$, but in many places it is easier just to work in general dimension.

For $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^m$ we of course write $f(\underline{x}) \to \underline{b}$ as $\underline{x} \to \underline{a}$ to mean

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ such that } \forall \underline{x} \in \Omega \text{ with } 0 < |\underline{x} - \underline{a}| < \delta \text{ we have } |f(\underline{x}) - \underline{b}| < \varepsilon.$$

And when this holds we write $\lim_{\underline{x} \to \underline{a}} f(\underline{x}) = \underline{b}$. (So everything as usual, but just with "underlines" to indicate vectors.)

1.1. The total derivative

We begin with a definition.

DEFINITION 1.1.1. We say that a function $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^m$ is *differentiable at* $\underline{a} \in \Omega$ if there exists a linear map $L : \mathbf{R}^n \to \mathbf{R}^m$ such that

$$\lim_{\underline{h} \to \underline{0}} \frac{f(\underline{a} + \underline{h}) - f(\underline{a}) - L\underline{h}}{|\underline{h}|} = \underline{0}. \tag{1.1}$$

We call L the *(total) derivative* of f at $\underline{a}$ and denote it $df(\underline{a})$. We say f is *differentiable* in Ω if it is differentiable at every point $\underline{a} \in \Omega$.

We divided by $|\underline{h}|$ rather than $\underline{h}$ here because of course it does not make sense to divide by a vector. (But when $f : \mathbf{R}^2 \to \mathbf{R}^2$ and you identify $\mathbf{R}^2$ with $\mathbf{C}$, then this is fine, but leads to a much stronger condition and the whole subject of *complex analysis.*) The matrix for the linear map L with respect to the usual bases of $\mathbf{R}^n$ and $\mathbf{R}^m$ is called the *Jacobian matrix.* (The matrix just contains the partial derivatives, as in Prelims Multivariable Calculus: see for example, equation(1.5).)

We shall shortly see that L is unique when it exists. But where did this definition come from? Let us assume that f is differentiable at $\underline{a}$ and do some rearranging. Define

$$R(\underline{h}) := \frac{f(\underline{a} + \underline{h}) - f(\underline{a}) - L\underline{h}}{|\underline{h}|} \in \mathbf{R}^m.$$

So we then have

$$f(\underline{a} + \underline{h}) - f(\underline{a}) = L\underline{h} + R(\underline{h})|\underline{h}| \tag{1.2}$$

and when $\underline{h} \neq \underline{0}$

$$\frac{f(\underline{a}+\underline{h}) - f(\underline{a})}{|\underline{h}|} = L\left(\frac{\underline{h}}{|\underline{h}|}\right) + R(\underline{h})$$

where now we have $R(\underline{h}) \to \underline{0}$ as $\underline{h} \to \underline{0}$. So then letting $\underline{h} \to \underline{0}$ and noting L is independent of $\underline{h}$ one has for $\underline{h} \neq \underline{0}$ that

$$\lim_{\underline{h}\to 0} \frac{f(\underline{a}+\underline{h}) - f(\underline{a})}{|\underline{h}|} = L \lim_{\underline{h}\to\underline{0}} \left(\frac{\underline{h}}{|\underline{h}|}\right) \tag{1.3}$$

provided both of these two limits exists. And indeed one exists if and only if the other does.

Suppose now that $n = 1$. Then we have $\underline{h} = h$ and $\underline{h}/|\underline{h}| = 1$ or -1 according to whether $h > 0$ or $h < 0$. Either way our limit on each side exists, and after taking into account a sign one gets

$$\lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = L.$$

And hey presto! we see that L is the usual derivative, but now interpreted as a linear map on $\mathbf{R}$.

With a sense of false confidence we now look at $n = 2$. And now one sees that

$$\lim_{\underline{h}\to\underline{0}} \left(\frac{\underline{h}}{|\underline{h}|}\right)$$

does not exist. Here $\underline{h}/|\underline{h}|$ is a vector of unit length in the direction of $\underline{h} \neq 0$. And since we allowed complete freedom in letting $\underline{h} \to \underline{0}$ we can take a sequence of such $\underline{h}$ going to zero for which the corresponding normalised vectors $\underline{h}/|\underline{h}|$ are anything you like at all.

So the moral here is that our usual definition of the derivative is fine when $n = 1$, but wholly inadequate for $n > 1$.

1.2. Directional derivatives

But we can salvage something from the ashes. Let $\underline{n} \in \mathbf{R}^n$ be a unit vector.

DEFINITION 1.2.1. We say that a function $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^m$ is *differentiable at* $\underline{a} \in \Omega$ *in direction* $\underline{n}$ if

$$\lim_{\lambda\to 0} \frac{f(\underline{a}+\lambda\underline{n}) - f(\underline{a})}{\lambda}$$

exists.

Assume that f is differentiable at $\underline{a}$. And in (1.3) restrict to those $\underline{h} = \lambda\underline{n}$ for $\lambda \in \mathbf{R}$. Well now

$$\lim_{\underline{h}\to\underline{0}} \left(\frac{\underline{h}}{|\underline{h}|}\right) = \underline{n}$$

does exist. And so we find

$$\lim_{\lambda\to 0} \frac{f(\underline{a}+\lambda\underline{n}) - f(\underline{a})}{\lambda} = L\underline{n}. \tag{1.4}$$

That is, the derivative in direction $\underline{n}$ exists and is just given by applying our matrix L to the unit vector $\underline{n}$. (That $\underline{n}$ was a unit vector was unimportant; we just need it is non-zero, as one can easily check that only the direction would matter.)

Note in all this I made no mention of the dimension of the codomain $\mathbf{R}^m$. The reason is that if $m > 1$ one can write f as an n-tuple of functions $(f_1, \cdots, f_m)$ each with codomain $\mathbf{R}$. And it is easy to check that our definitions apply "component-wise", i.e., f is differentiable at a point, or has some directional derivative, if and only if all the functions $f_1, \cdots, f_m$ do. This is just because a vector $\underline{y} \in \mathbf{R}^m$ tends to zero exactly when all its entries do. So we may as well assume that $m = 1$ in what follows.

Here is the million dollar question:

If all of the directional derivatives exist at $\underline{a}$ does the total derivative exist?

When $n > 1$, no. We shall see some counterexamples to this shortly. But let us keep our momentum going and prove a partial converse which is true.

1.3. Continuous partial derivatives give differentiability

We shall restrict for simplicity and notational convenience to $n = 2$, and also $m = 1$, and take $\Omega = \mathbf{R}^2$. So we are considering functions $f : \mathbf{R}^2 \to \mathbf{R}$. I will write $\underline{a} = (a, b)$ and $\underline{h} = (h, k)$, and $f(a, b)$ rather than $f((a, b))$ etc. For our partial converse I will only need to consider two directions, given of course by the unit vectors $(1, 0)$ and $(0, 1)$. I will write these directional derivatives when they exist as $f_x(a, b)$ and $f_y(a, b)$. (They are also denoted $\frac{\partial f}{\partial x}(a, b)$ etc but this requires more effort to write.) We will of course follow convention and call these special directional derivatives the *partial derivatives.*

Assume briefly $f : \mathbf{R}^2 \to \mathbf{R}$ is differentiable at a point $\underline{a} = (a, b)$. We have the linear map $L = df(\underline{a}) : \mathbf{R}^2 \to \mathbf{R}$, the derivative. And we see from (1.4) that

$$df(\underline{a}) : (1, 0) \mapsto f_x(a, b),\ df(\underline{a}) : (0, 1) \mapsto f_y(a, b).$$

So writing $df(\underline{a})$ as a 1×2 matrix with respect to the usual basis (the Jacobian matrix), we can see it is just

$$df(\underline{a}) = (f_x(a, b), f_y(a, b)). \tag{1.5}$$

Note this proves that $df(\underline{a})$ is unique, since the partial derivatives are uniquely defined (and the same argument works for general dimension). And this little calculation also inspires the next proof.

THEOREM 1.3.1. *Suppose $f : \mathbf{R}^2 \to \mathbf{R}$ has continuous partial derivatives. Then f is differentiable in $\mathbf{R}^2$ with derivative at $\underline{a} = (a, b)$ given by $df(\underline{a}) = (f_x(a, b), f_y(a, b))$.*

Proof. Let $(a,b)\in\Omega$ and $(h,k)\in\mathbf{R}^2$. Define

$$L := (f_x(a,b), f_y(a,b)).$$

We need to show that

$$\lim_{(h,k)\to\underline{0}} \frac{f(a+h,b+k)-f(a,b)-L(h,k)}{|(h,k)|} = 0. \tag{1.6}$$

Our proof uses the telescoping sum

$$f(a+h,b+k)-f(a,b) = (f(a+h,b+k)-f(a+h,b)) + (f(a+h,b)-f(a,b)).$$

Note in each bracket we move only one of the arguments. And by the Mean Value Theorem in one variable from Prelims Analysis: there exists $\theta_h,\theta_k\in(0,1)$ such that

$$\begin{array}{rcl} f(a+h,b+k)-f(a+h,b) & = & f_y(a+h,b+\theta_k k)k \\ f(a+h,b)-f(a,b) & = & f_x(a+\theta_h h,b)h. \end{array}$$

(Note the function f is differentiable in each argument when one fixes the other, i.e., the partial derivatives exist. That is why the MVT can be applied.) So

$$f(a+h,b+k)-f(a,b) = f_y(a+h,b+\theta_k k)k + f_x(a+\theta_h h,b)h.$$

Also we have (thinking of (h,k) as a column vector so our matrix for L can act)

$$L(h,k) = f_x(a,b)h + f_y(a,b)k.$$

So the numerator in (1.6) is just

$$(f_x(a+\theta_h h,b)-f_x(a,b))h + (f_y(a+h,b+\theta_k k)-f_y(a,b))k$$

$$= (f_x(a+\theta_h h,b)-f_x(a,b), f_y(a+h,b+\theta_k k)-f_y(a,b))\cdot(h,k).$$

This is the dot product of two vectors and we shall apply the Cauchy-Schwarz inequality to deduce its absolute value is at most

$$\sqrt{(f_x(a+\theta_h h,b)-f_x(a,b))^2+(f_y(a+h,b+\theta_k k)-f_y(a,b))^2}\times\sqrt{h^2+k^2}.$$

So to show (1.6) we must show that

$$\sqrt{(f_x(a+\theta_h h,b)-f_x(a,b))^2+(f_y(a+h,b+\theta_k k)-f_y(a,b))^2}\to 0$$

as $(h,k)\to\underline{0}$. But as $(h,k)\to\underline{0}$ we have that the argument $(a+\theta_h h,b)$ of f_x tends to (a,b) and since f_x is continuous on $\mathbf{R}^2$ we get that $f_x(a+\theta_h h,b)-f_x(a,b)\to 0$. And likewise $(a+h,b+\theta_k k)\to(a,b)$ and by continuity of f_y we see the second term in the sum tends to zero. This completes the proof. $\square$

Note that the partial derivative f_x itself is a function from (some subset of) $\mathbf{R}^2$ to $\mathbf{R}$. So we may consider its partial derivatives with respect to x and y, which I will denote f_{xx} and f_{xy}. In Prelims calculus you often used that $f_{xy} = f_{yx}$. On problem sheet 1 you will prove a sufficient condition for this equality to be true, and see that it is not always so.

The proof of Theorem 1.3.1 in fact works for any $f : \Omega \subseteq \mathbf{R}^2 \to \mathbf{R}$ with continuous partial derivatives, provided one assumes Ω is open (Definition 1.6.1); that is, every point $\underline{a} \in \Omega$ lies in a disk contained within Ω. One needs this assumption in the proof simply to ensure f is defined on the line segments implicit in our use of the telescoping sum and Mean Value Theorem.

1.4. Example, counterexample and some geometric intuition

Let us focus now on $n = 1$ or 2 and $m = 1$. So we have a map $f : \Omega \subseteq \mathbf{R} \to \mathbf{R}$ or $f : \Omega \subseteq \mathbf{R}^2 \to \mathbf{R}$ which is differentiable at a point $\underline{a}$. The reason for doing this is that we can now visualise the graph of the function

$$G_f = \{(x, f(x)) : x \in \Omega \subseteq \mathbf{R}\} \quad \text{or} \quad G_f = \{(x, y, f(x, y)) : (x, y) \in \Omega \subseteq \mathbf{R}^2\}$$

in $\mathbf{R}^2$ or $\mathbf{R}^3$ near the point $(\underline{a}, f(\underline{a}))$.

Take $n = 1$ first. Then we have from (1.2) that

$$f(a + h) = f(a) + Lh + R(h)|h|$$

where $L = f_x(a)$ is the derivative and $R(h) \to 0$ as $h \to 0$. So we recover that the *tangent line* $f(a) + Lh$ is a good approximation to the function $f(x)$ near $x = a$.

Consider now $n = 2$. We take as above $\underline{a} = (a, b)$ and $\underline{h} = (h, k)$, and recall from (1.5) our derivative $df(\underline{a}) = L$ is just the matrix $(f_x(a, b), f_y(a, b))$. We see now that

$$f(a + h, b + k) = f(a, b) + f_x(a, b)h + f_y(a, b)k + R(h, k)|(h, k)|. \tag{1.7}$$

And now the *tangent plane* $f(a, b) + f_x(a, b)h + f_y(a, b)k$ is a good approximation to the function $f(x, y)$ near (a, b).

Put slightly differently, (1.7) is the “first-order Taylor expansion” around (a, b), as in the one-variable case. And as $R(h, k) \to 0$ as $|(h, k)| \to 0$, then the remainder $R(h, k)|(h, k)|$ goes to zero faster than linearly. So the linear function $f(a, b) + f_x(a, b)h + f_y(a, b)k$ provides a good numerical approximation near (a, b).

Recall when you first learned calculus, computing the derivative of a function directly from the definition was rather hard work. The same is true in this higher dimensional case. Working directly from Definition 1.1.1 would be an effort. But our description of the matrix L in terms of partial derivatives, as in (1.5), makes life much easier.

EXAMPLE 1.4.1. Let

$$f(x,y) = \sqrt{1-(x^2+y^2)}$$

and $\Omega = \{(x,y) : x^2+y^2 \leqslant 1\} \subseteq \mathbf{R}^2$. Then obviously our graph G_f is just the top half of a sphere of radius one with centre the origin. And if we compute the matrix of partial derivatives (f_x, f_y) it is

$$\left(\frac{-x}{\sqrt{1-(x^2+y^2)}}, \frac{-y}{\sqrt{1-(x^2+y^2)}}\right).$$

These are defined and continuous except on the boundary $x^2+y^2=1$ of Ω. So Theorem 1.3.1 tells us the function is differentiable there, and the derivative at such a point (a,b) is just the evaluation $(f_x(a,b), f_y(a,b))$. And this gives an explicit description of the tangent plane at any such point (which I won't write down). (Note that on the boundary the function is not especially badly behaved, it is just that its derivative becomes "infinite", in the same way a function like $x^{1/3}$ on $\mathbf{R}$ does at the origin, and our present set-up is not robust enough to handle this.)

EXAMPLE 1.4.2. Let $f : \mathbf{R}^2 \to \mathbf{R}$ be defined by

$$f(x,y) := \begin{cases} \frac{xy(x+y)}{x^2+y^2} & \text{when } (x,y) \neq (0,0) \\ 0 & \text{for } (x,y) = (0,0). \end{cases}$$

For a unit vector $\underline{n}$ note that $f(\lambda\underline{n}) = \lambda f(\underline{n})$. (That is, the rational function $f(x,y)$ is *homogeneous of degree* 1.) Hence

$$\lim_{\lambda\to 0} \frac{f((0,0)+\lambda\underline{n}) - f(0,0)}{\lambda} = f(\underline{n}).$$

Thus all the direction derivatives exist at $(0,0)$, and are just equal to the value of the function in that direction. But if f were differentiable at $(0,0)$ then by (1.4) there would exists a *linear* map $L : \mathbf{R}^2 \to \mathbf{R}$ such that $L\underline{n} = f(\underline{n})$. And this is not the case, since f is not linear. So f is not differentiable at $(0,0)$.

You can use the 3D calculator at *www.desmos.com/3d* to see the graph of this function, or any other.

1.5. Continuity

I am sure you all remember the $\varepsilon - \delta$ definition of continuity from Prelims Analysis II. Let me just recall it for you, but in a rather unusual manner.

We shall say that a function $f : \mathbf{R} \to \mathbf{R}$ is "not continuous" at a point $a \in \mathbf{R}$ if

$$\exists\varepsilon > 0 \text{ such that } \forall\delta > 0, \exists x \in \mathbf{R} \text{ with } |x-a| < \delta \text{ but } |f(x)-f(a)| \geqslant \varepsilon.$$

This seems a pretty good way of capturing that the function f is not continuous in the intuitive sense, since our statement suggests it "jumps by ε" at a. And you might also remember how to negate logical statements, by sweeping through them and switching $\forall$ and $\exists$. So let's do that. We find our function is *not* "not continuous" at a when

$$\forall\varepsilon > 0, \exists\delta > 0 \text{ such that } \forall x \in \mathbf{R} \text{ with } |x-a| < \delta \text{ we have } |f(x) - f(a)| < \varepsilon.$$

Let us call such a function *continuous* at a, since writing not not continuous continually would be tiresome.

I find this a much more intuitive approach to the definition, since really it is the property of being "not continuous", i.e. a function jumping, that we have a feeling for and should capture in a logical manner. But anyway let us stick with convention:

Definition 1.5.1. A function $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^m$ is *continuous at* $\underline{a} \in \Omega$ if

$$\forall\varepsilon > 0, \exists\delta > 0 \text{ such that } \forall \underline{x} \in \Omega \text{ with } |\underline{x}-\underline{a}| < \delta \text{ we have } |f(\underline{x}) - f(\underline{a})| < \varepsilon.$$

Note this equivalent to $f(\underline{x}) - f(\underline{a}) \to \underline{0}$ as $\underline{x} \to \underline{a}$.

Theorem 1.5.2. *Assume $f : \Omega \subseteq \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $\underline{a} \in \Omega$. Then f is continuous at $\underline{a}$.*

Proof. Let the derivative be $L : \mathbf{R}^n \to \mathbf{R}^m$. Then from (1.2) we have

$$f(\underline{a} + \underline{h}) - f(\underline{a}) = L\underline{h} + R(\underline{h})|\underline{h}|$$

where $R(\underline{h}) \to \underline{0}$ as $\underline{h} \to \underline{0}$. Taking $\underline{h} \to \underline{0}$ on the righthand side gives zero, and hence

$$f(\underline{a} + \underline{h}) - f(\underline{a}) \to 0 \text{ as } \underline{h} \to \underline{0}.$$

□

You shall see on problem sheet 1 though, that even if *all* the directional derivatives exist at a point, the function is not necessarily continuous at that point.

1.6. Some topology in $\mathbf{R}^2$

In this section we shall introduce some properties of sets in $\mathbf{R}^2$ which you will need for the complex analysis course. We revisit all of these properties in greater depth and generality later in the course. (Note I will drop the underlines from vectors in most of this section to reduce clutter on notation.)

1.6.1. Path-connected and connected sets. Let $r > 0$ and $x \in \mathbf{R}^2$. We call

$$B(x, r) := \{y \in \mathbf{R}^2 | |x - y| < r\} \text{ and } \overline{B}(x, r) := \{y \in \mathbf{R}^2 | |x - y| \leqslant r\}$$

the (respectively) *open* and *closed balls* of radius r around x.

Definition 1.6.1. We say that a set $U \subseteq \mathbf{R}^2$ is *open* if for all $x \in U$ there exists $r > 0$ such that $B(x, r) \subseteq U$. We say that $U \subseteq \mathbf{R}^2$ is *closed* if the complement $\mathbf{R}^2 \backslash U$ is open. Given $a \in \mathbf{R}^2$, we call any open set $U \subseteq \mathbf{R}^2$ containing a an *open neighbourhood* of a.

Notice that $\emptyset$ and $\mathbf{R}^2$ itself are open. And they are also both closed: in particular, it is not the case that a set is closed when it is not open. It can be both, or neither.

The collection of all open sets in $\mathbf{R}^2$ gives what is called a *topology* on $\mathbf{R}^2$. That is, a collection of subsets containing $\emptyset$ and $\mathbf{R}^2$ which is closed under finite intersections and arbitrary unions. The properties of sets we shall introduce below depend only upon this topology, in the following sense. Suppose we used a *different* notion of distance on $\mathbf{R}^2$ which defined the *same* collection of open sets. Then we could use that different notion of distance in our definitions below, and end up defining the same property of sets.

We first look at the notion of paths and path-connectedness.

Definition 1.6.2. Let $U \subseteq \mathbf{R}^2$ and $a_0, a_1 \in U$. A *path* in U between a_0 and a_1 is a continuous map $\gamma : [0, 1] \to U$ with $\gamma(0) = a_0$ and $\gamma(1) = a_1$. We say that U is *path-connected* if for every $a_0, a_1 \in U$ there exists a path between a_0 and a_1.

One thinks of the value $\gamma(t)$ for $t \in [0, 1]$ as continuously tracing the path between a_0 and a_1. The significance of paths in complex analysis is that, after identifying $\mathbf{R}^2$ and $\mathbf{C}$, you will integrate complex valued functions along them. (Though you will usually make the stronger assumption that the map γ is in fact differentiable.)

Definition 1.6.3. Let $U \subseteq \mathbf{R}^2$ and $a_0, a_1 \in U$. Let γ_0 and γ_1 be paths in U connecting a_0 and a_1. We say that the paths γ_0 and γ_1 are *homotopic* if there exists a continuous map $\Gamma : [0, 1] \times [0, 1] \to U$ with $\Gamma(s, 0) = a_0$ and $\Gamma(s, 1) = a_1$ for all $s \in [0, 1]$, and $\Gamma(0, t) = \gamma_0(t)$ and $\Gamma(1, t) = \gamma_1(t)$ for all $t \in [0, 1]$.

Here for each $s \in [0, 1]$ we get a path $\Gamma(s, t) : [0, 1] \to U$, and one thinks of the first path $\gamma_0(t) = \Gamma(0, t)$ being continuously deformed into the second $\gamma_1(t) = \Gamma(1, t)$ as s moves between 0 and 1.

Definition 1.6.4. We shall say that $U \subseteq \mathbf{R}^2$ is *simply-connected* if it is path-connected and given any points $a_0, a_1 \in U$ then any two paths in U between a_0 and a_1 are homotopic.

So "up to homotopy" there is a unique path between any a_0 and a_1 in a simply-connected set. We can put this another way. For any point $a \in U$ there is the trivial path from a to a given by $\gamma(t) = a$ for all $t \in [0, 1]$. Then a path-connected set U will be simply-connected exactly when for every $a \in U$ any path between a and a can be "shrunk to a point", i.e. is homotopic to this trivial path. (That this is equivalent to the definition is a little work to prove.)

The significance of all this for complex analysis is that it will turn out the path integrals only depend on paths up to homotopy (which makes life much easier).

Example 1.6.5. The open unit ball $B(0,1)$ is simply-connected. Given any two points $a_0, a_1 \in B(0,1)$ and any two paths connecting them, one can visualise deforming one into the other. By contrast, the open "doughnut" $B(0,3)\backslash\overline{B}(0,1)$ is not simply-connected. Following the circle of radius 2 anticlockwise from $(2,0)$ gives a path from $(2,0)$ to itself. But it is intuitively clear that there is no way to shrink this path to a point. (I do not claim either of these visually properties is especially easy to prove, though you will show the first on problem sheet 1.)

Here is a useful property of path-connectedness.

Theorem 1.6.6. *Let $U \subseteq \mathbf{R}^2$ be path-connected and $f : U \to \mathbf{R}^2$ continuous. Then $f(U)$ is path-connected.*

Proof. Let $v_0, v_1 \in f(U)$. Then $v_0 = f(u_0)$ and $v_1 = f(u_1)$ for some $u_0, u_1 \in U$. Since U is path-connected there exists a path γ joining u_0 and v_0. Then $f \circ \gamma$ gives the required path between v_0 and v_1. (I will leave it to you to prove directly from Definition 1.5.1 that the composition of two continuous maps is continuous.) □

There is a slightly less intuitive related notion in topology, namely that a set is *connected*.

Definition 1.6.7. Let $U \subseteq \mathbf{R}^2$. We shall say that U is *connected* if the following is true. For any open sets A and B in $\mathbf{R}^2$ with $U \subseteq A \cup B$ and $(A \cap B) \cap U = \emptyset$, then either $U \subseteq A$ or $U \subseteq B$.

We shall look closely at this notion later. Path-connected sets are always connected (Theorem 8.3.1). It turns out that *open* connected sets in $\mathbf{R}^2$ are always path-connected (by Theorem 8.3.2). So in complex analysis when you read an "open connected set in $\mathbf{C}$" this is just the same as it being an open path-connected set in $\mathbf{R}^2$.

Example 1.6.8. The space $\mathbf{R}^2$ itself is connected, since it is obviously path-connected. This implies that the only sets in $\mathbf{R}^2$ which are both open and closed are $\emptyset$ and $\mathbf{R}^2$. For suppose A was such a set, with $A \neq \emptyset$ or $\mathbf{R}^2$. Then $B := \mathbf{R}^2\backslash A$ is also open with $B \neq \mathbf{R}^2$. And $\mathbf{R}^2 = A \cup B$ with $A \cap B = \emptyset$ but $\mathbf{R}^2 \not\subseteq A$ or B, contradicting the fact $\mathbf{R}^2$ is connected.

And again the property of being connected is preserved under continuous maps (Lemma 8.1.6).

Theorem 1.6.9. *Let $U \subset \mathbf{R}^2$ be connected and $f : U \to \mathbf{R}^2$ be continuous. Then $f(U)$ is connected.*

1.6.2. Compactness.

We now turn to a quite different but equally important notion.

Definition 1.6.10. Let $U \subseteq \mathbf{R}^2$. We shall say that U is *compact* if given any collection $\{V_i : i \in I\}$ of open sets with $U \subseteq \cup_{i \in I} V_i$, there exists a finite subset $J \subseteq I$ such that $U \subseteq \cup_{j \in J} V_j$.

The mantra to remember here is "U is compact exactly when every open covering of U has a finite subcovering". This is not at all an intuitive notion, but the essential point really is that it is a good proxy for finiteness when dealing with infinite sets. That is to say, compactness allows you to apply to infinite sets certain arguments you would normally only apply to finite sets. Fortunately in $\mathbf{R}^2$ we can relate compactness to more intuitive notions.

Definition 1.6.11. We shall say that $U \subseteq \mathbf{R}^2$ is *bounded* if there exists $x \in \mathbf{R}^2$ and $r > 0$ such that $U \subseteq B(x, r)$.

It turns out that in $\mathbf{R}^2$ being compact is the same as being closed and bounded (Theorem 10.3.3). It is important to stress though that these two notions are not the same in more general settings.

You will recall from Prelims that a continuous function $f : U \to \mathbf{R}$ on a non-empty closed and bounded set is bounded and achieves that bound. This is a special case of a much more general theorem in topology, that the image of a compact set under a continuous map is compact. In particular, we also have the following.

Theorem 1.6.12. *Let $f : U \subseteq \mathbf{R}^2 \to \mathbf{R}$ be a continuous function where U is a non-empty compact set. Then $f(U)$ is bounded and achieves that bound. That is, there exists $b \in \mathbf{R}$ and $a \in U$ such that $|f(x)| \leqslant b$ for all $x \in U$ and furthermore $|f(a)| = b$.*

Proof. We will derive this from some theorems proved later in the course. (So you should defer reading this for now.) Since U is compact it is sequentially compact (Proposition 10.2.1). Note $|f|$ is also continuous, since it is a composition of two continuous functions. Hence $|f(U)|$ is sequentially compact (Lemma 9.3.1), and so closed and bounded (Lemma 9.2.1). As $|f(U)|$ is non-empty and bounded it has a supremum, and since it is closed this supremum lies in $|f(U)|$ (Corollary 6.2.5). (Note the supremum is a limit point of $|f(U)|$.) That is to say, there exists $b \in \mathbf{R}$ and $a \in U$ such that $|f(x)| \leqslant b$ for all $x \in U$ and furthermore $|f(a)| = b$. □

1.6.3. Notions of distance in $\mathbf{R}^2$.

For $\underline{x} = (x_1, x_2)$ and $\underline{y} = (y_1, y_2) \in \mathbf{R}^2$ define $d_2(x, y) := |\underline{x} - \underline{y}| = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$. This is just the usual Euclidean notion of distance which we have been using throughout. But it is not the only "sensible" notion of distance on $\mathbf{R}^2$. There are many, some less sensible than others. By "sensible" we mean primarily that the notion of distance defines a *metric*, a term to be made precise in the next chapter.

For example we could take

$$d_\infty(\underline{x},\underline{y}) := \max(|x_1-y_1|,|x_2-y_2|)$$

and then define the open ball of radius $r>0$ around a point $\underline{a}\in\mathbf{R}^2$ to be

$$B_\infty(\underline{x},r) := \{\underline{y}\in\mathbf{R}^2 : d_\infty(\underline{x},\underline{y})<r\}.$$

This ball would not be very good for football (rather square: draw one). But it would be perfectly good for doing topology as it defines the same open sets as our usual Euclidean metric. The reason being simply that given any point $\underline{x}$ and $r>0$, then there exists $r'<r$ such that $B_\infty(\underline{x},r')\subseteq B(\underline{x},r)$, and vice-versa. In other words, square balls can be fitted within round ones, and vice-versa.

Here is a notion of distance on $\mathbf{R}^2$ which defines a metric, but which is good for neither football nor topology (in the sense that like all metrics it defines a topology on $\mathbf{R}^2$, but this one is not of any use in geometry).

Example 1.6.13. Define

$$\delta(\underline{x},\underline{y}) := \begin{cases} |x_2-y_2| & \text{if } x_1=y_1 \\ |x_1-y_1|+|x_2|+|y_2| & \text{if } x_1\neq y_1. \end{cases}$$

This metric is effectively the distance between two points in a library of shelves parallel to the y-axis and with a connecting corridor along the x-axis only.

1.7. The Inverse function theorem

For $U,V\subseteq\mathbf{R}^2$ we write $C^1(U,V)$ for the set of functions $f:U\to V$ with continuous partial derivatives on U. Note in particular such functions are differentiable on U.

Theorem 1.7.1 (Inverse Function Theorem in $\mathbf{R}^2$). *Let $\Omega\subseteq\mathbf{R}^2$ be open, $f\in C^1(\Omega,\mathbf{R}^2)$ and $\underline{a}\in\Omega$. Suppose that $df(\underline{a}):\mathbf{R}^2\to\mathbf{R}^2$ is invertible. Then there exists an open neighbourhood U of $\underline{a}$ such that $f(U)$ is open and $f:U\to f(U)$ is a bijection with inverse $f^{-1}\in C^1(f(U),U)$. Moreover, we have $df^{-1}(f(\underline{a}))=df(\underline{a})^{-1}$.*

The inverse function theorem allows you to "locally invert" functions on $\mathbf{R}^2$, in the same way that you can "globally invert" (invertible) linear maps on $\mathbf{R}^2$.

Example 1.7.2. We show that the equations $u=x^3y+y^2, v=\ln(x+y)$ have a unique solution $x=x(u,v)$, $y=y(u,v)$ on some neighbourhood of $(6,\ln(3))$ with $x(6,\ln(3))=1$ and $y(6,\ln(3))=2$.

Define $f(x,y) = (u,v) = (x^3y + y^2, \ln(x+y))$ (for say $x+y>0$). Then to solve for x, y in terms of (u,v) it is enough to show f^{-1}exists, for then we have $(x,y) = f^{-1}(u,v)$. Now the Jacobian matrix

$$df(x,y) = \begin{pmatrix} 3x^2y & x^3+2y \\ 1/(x+y) & 1/(x+y) \end{pmatrix}$$

is continuous in a neighbourhood of $(1,2)$ and invertible there, and so the Inverse Function Theorem applies.

The intuition behind the Inverse Function Theorem is very simple. At a point $\underline{a}+\underline{h}$ near $\underline{a}$ the value $f(\underline{a}+\underline{h})$ is well approximated by the value $f(\underline{a}) + df(\underline{a})\underline{h}$. If it were in fact equal to this, then one could invert the function on a point $f(\underline{a})+\underline{k}$ near $f(\underline{a})$ simply by mapping this to $a + df(\underline{a})^{-1}\underline{k}$, provided the derivative is invertible. And the inverse function theorem tells us this is almost true. That is, there is an inverse near $f(\underline{a})$ which is well approximated by this linear map. Unfortunately the proof itself is rather complicated, and omitted from the course (and non-examinable).

This complication is mainly notational though, and it is rather straightforward to explain the main steps in the one variable case. And these carry over almost exactly to higher dimensions. So let's look at the existence and continuity of the inverse. (This is *not* the simplest proof in dimension one, but rather one that can be translated easily to the general case and suggests a useful numerical algorithm.)

Let $f : \mathbf{R} \to \mathbf{R}$ be continuously differentiable at $a \in \mathbf{R}$ with $f'(a) \neq 0$. Without any loss of generality we may assume $a = 0$, $f(0) = 0$ and $f'(0) = 1$ (think about why). Consider the function

$$\psi(x) = x - f(x). \tag{1.8}$$

Note that $\psi(0) = \psi'(0) = 0$ and since $\psi'(x)$ is continuous at 0, there exists $r > 0$ such that

$$|x| \leqslant r \Rightarrow |\psi'(x)| \leqslant \frac{1}{2}. \tag{1.9}$$

Hence by the Mean Value Theorem in one variable we see

$$|x| \leqslant r \Rightarrow |\psi(x)| \leqslant \frac{1}{2}|x| \leqslant \frac{r}{2}. \tag{1.10}$$

So now let us fix $y_0 \in \mathbf{R}$ with $|y_0| \leqslant r/2$. Define $\phi(y_0, x) : \mathbf{R} \to \mathbf{R}$ by

$$\phi(y_0, x) := x - f(x) + y_0 = \psi(x) + y_0.$$

Then by (1.10)

$$|x| \leqslant r \Rightarrow |\phi(y_0,x)| = |x - f(x) + y_0| \leqslant |\psi(x)| + |y_0| \leqslant r.$$

Thus $\phi(y_0, x)$ maps the closed ball $\overline{B}(0,r)$ to itself. Also for any $x, x' \in \overline{B}(0,r)$ using (1.9) and the Mean Value Theorem again,

$$|\phi(y_0, x) - \phi(y_0, x')| = |\psi(x) - \psi(x')| \leqslant \frac{1}{2}|x - x'|.$$

So we have a contraction mapping and by Prelims analysis / constructive mathematics, there exists a unique "fixed point" $x_0 \in \overline{B}(0,r)$ such that $\phi(y_0, x_0) = x_0$, that is $f(x_0) = y_0$.

Note that if $|y_0| < r/2$ then we have $|y_0| \leqslant r'/2$ for some $r' < r$. And replacing r by r' in the above we have $x_0 \in \overline{B}(0, r') \subset B(0,r)$. So in fact we have a map on open sets

$$g : B(0, r/2) \to U := f^{-1}(B(0, r/2)) \cap B(0,r), \quad y_0 \mapsto x_0.$$

And this gives our inverse $g : f(U) \to U$ with $f \circ g = I_{f(U)}$ and $g \circ f = I_U$.

To prove continuity of g, let $\varepsilon > 0$. And take $y_1, y_2 \in f(U) \subseteq B(0, r/2)$. Define $x_1 := g(y_1) \in U$ and $x_2 := g(y_2) \in U$, so $f(x_1) = y_1$ and $f(x_2) = y_2$. By the Mean Value Theorem and (1.9) we have

$$|\psi(x_1) - \psi(x_2)| \leqslant \frac{1}{2}|x_1 - x_2|.$$

Thus from (1.8),

$$|(x_1 - x_2) - (f(x_1) - f(x_2))| \leqslant \frac{1}{2}|x_1 - x_2|,$$

and a little thought then gives

$$|x_1 - x_2| \leqslant 2|f(x_1) - f(x_2)|.$$

So defining $\delta := \varepsilon/2$ we see that

$$|y_1 - y_2| < \delta \Rightarrow |g(y_1) - g(y_2)| < 2\delta = \varepsilon.$$

So we have proved (uniform) continuity of g on $f(U)$.

Let's turn now to an example in the two-dimensional case, and try and understand our partial proof in this context.

Example 1.7.3. Let $f : \mathbf{R}^2 \to \mathbf{R}^2$ be defined by

$$f(x,y) := (\cos(y) + x - 1, \sin(x) + y - \cos(y) - x + 1).$$

Then $f(0,0) = (0,0)$. The Jacobian matrix L here at the point (x,y) is

$$df(x,y) = \begin{pmatrix} 1 & -\sin(y) \\ \cos(x) - 1 & 1 + \sin(y) \end{pmatrix}.$$

Note that $df(0,0) = I_2$. So we are in a situation analogous to that in our partial proof for the one-dimensional case. (I cooked up this function, starting with $(x,y) \to (\cos(y) + x, \sin(x) + y)$. You might like to think how this was done.)

Inspired by our proof above, let us try to invert the function *numerically* near the point $(0,0) = f(0,0)$. So pick $(x_0, y_0) \in \overline{B}((0,0), r/2)$, and define $\phi(x_0, y_0, x, y) : \mathbf{R}^2 \to \mathbf{R}^2$ by

$$\phi(x_0, y_0, x, y) := (x, y) - f(x, y) + (x_0, y_0)$$

$$= (x - (\cos(y) + x - 1) + x_0, y - (\sin(x) + y - \cos(y) - x + 1) + y_0).$$

On $\overline{B}((0,0), r)$ this should define a contraction mapping. And so iterating it from any starting point $(s_0, t_0) \in \overline{B}((0,0), r)$ will converge to a fixed point $\phi(x_0, y_0, s_\infty, t_\infty) = (s_\infty, t_\infty)$. And we will have $f(s_\infty, t_\infty) = (x_0, y_0)$. (My apologies the notation for $\mathbf{R}^2$ in this paragraph has become inconsistent with that in our partial proof for $\mathbf{R}$.)

But the problem is we have no idea on the value for r! In the one-dimensional case we chose r small enough so that the derivative of the function $\psi(x) = x - f(x)$ was bounded by $1/2$ for $|x| \leqslant r$. And this is where we used the continuity of the derivative. So it appears in this case we need to bound the *size* of the matrix $df(x, y)$ by $1/2$ on some disk around $(0,0)$, using the continuity of its entries. To do this we need a notion of size on matrices, or linear maps. And this is really the only ingredient missing if we want to generalise our proof above to higher dimension. We will not pursue this further, but just end with some calculations.

Experimentally take $(x_0, y_0) = (0.5, 0.5)$ and start from $(s_0, t_0) = (0, 0)$. After 5 iterations we get

$$(s_5, t_5) = (0.5898\ldots, 0.4462\ldots) \text{ and } f(s_5, t_5) = (0.4919\ldots, 0.5105\ldots).$$

And more iterations give higher precision. And this works well for $(x_0, y_0) = (0.6, 0.6)$, and so on. But somewhere between taking $(x_0, y_0) = (0.79, 0.79)$ and $(0.8, 0.8)$ things go wrong. I have not looked closely enough at the example to see why: perhaps you would like to figure this out.

CHAPTER 2

Metric spaces

2.1. The real numbers and the axiom of choice

The real numbers. I will assume familiarity with the real numbers $\mathbf{R}$ as discussed at some length in the Prelims course Analysis I. I will not repeat the long list of axioms for the real numbers here. The most important properties we shall need are

- Any non-empty, bounded subset $S \subseteq \mathbf{R}$ has a least upper bound $\sup(S)$, which is a real number c such that $x \leqslant c$ for all $x \in S$, and such that if c' is any other number with this property then $c' \geqslant c$;
- Similarly, any non-empty, bounded subset $S \subseteq \mathbf{R}$ has a greatest lower bound $\inf(S)$;
- (Bolzano-Weierstrass) Any bounded sequence of real numbers has a convergent subsequence;
- Any Cauchy sequence of real numbers converges.

It might be a good idea to remind yourself of the precise meaning of these statements now, though we shall be going over the last two points in a more general context later in the course.

The Prelims course Analysis I assumed that the real numbers exist. This is not, by any means, obvious! We will also assume they exist.

The axiom of choice. The following statement, used for example in the proof of Corollary 6.1.5, seems very uncontroversial: given nonempty subsets $S_1, S_2, \dots$ of some set X, we may find a sequence $(x_n)_{n=1}^{\infty}$ with $x_n \in S_n$ for all n. One might have thought that this is the most trivial induction imaginable: pick $x_1 \in S_1$, then pick $x_2 \in S_2$, and so on. This does indeed show that there are $x_1, \dots, x_N$ with $x_n \in S_n$ for $n = 1, \dots, N$, but it does *not* show the infinitary statement about the existence of a sequence. In fact, the existence of a sequence $(x_n)_{n=1}^{\infty}$ with $x_n \in S_n$ for all n has the status of a separate axiom of mathematics, called the *axiom of countable choice.*

You can learn much more about this and, more particularly, the axiom of choice itself in the course *B1.2: Set Theory.* However, the introduction of the Wikipedia page on the Axiom of Choice is a good read.

2.2. The definition of a metric space

One of the key definitions of Analysis I was that of the *continuity* of a function. Recall that if $f\colon \mathbf{R} \to \mathbf{R}$ is a function, we say that f is continuous at $a \in \mathbf{R}$ if, for any $\varepsilon > 0$, we can find a $\delta > 0$ such that if $|x - a| < \delta$ then $|f(x) - f(a)| < \varepsilon$.

Now consider what it is about real numbers that we need in order for this definition to make sense: Really we just need, for any pair of real numbers x_1 and x_2, a measure of the *distance* between them. Thus we should be able to talk about continuous functions $f\colon X \to X$ on any set X provided it is equipped with a notion of distance. Even more generally, provided we have prescribed a notion of distance on two sets X and Y, we should be able to say what it means for a function $f\colon X \to Y$ to be continuous. In order to make this precise, we will therefore need to give a mathematically rigorous definition of what a "notion of distance" on a set X should be. This is the concept of a metric space.

Definition 2.2.1. Let X be a set. Then a *distance function* on X is a function $d\colon X \times X \to \mathbf{R}$ with the following properties:

(i) (positivity) $d(x, y) \geqslant 0$ and $d(x, y) = 0$ if and only if $x = y$;
(ii) (symmetry) $d(x, y) = d(y, x)$;
(iii) (triangle inequality) if $x, y, z \in X$ then we have $d(x, z) \leqslant d(x, y) + d(y, z)$.

The pair (X, d) consisting of a set X together with a distance function d on it is called a *metric space*.

Remark. Often we will not be quite so formal, and will refer to X (rather than the pair (X, d)) as a metric space. However, it is important to note that the same space X can have many different distances on it, and in fact that different distances on the same space X can have wildly differing properties.

Occasionally, we will be *more* formal, for instance when we have two metric spaces (X, d_X) and (Y, d_Y) and wish to make it clear which distance we are talking about.

The axioms that a distance function d is required to satisfy are very basic, and one feels that any "reasonable" notion of distance ought to satisfy these properties. This, coupled with the fact that using just these axioms one can develop a satisfactory theory of continuity of functions – as well as many other things – is the point of the definition.

Before moving on, let us record one very simple but useful equivalent form of the triangle inequality, sometimes (but not by me) known as the reverse triangle inequality.

Lemma 2.2.2. *Let x, y, z be points in a metric space. Then we have $|d(x, y) - d(x, z)| \leqslant d(y, z)$.*

Proof. This is two inequalities in one, namely the inequality $d(x,y) - d(x,z) \leqslant d(y,z)$, and the inequality $d(x,z) - d(x,y) \leqslant d(y,z)$. Both are instances of (in fact, equivalent to) the triangle inequality. □

2.3. Some examples of metric spaces

In this section we look at some examples of metric spaces. A very basic example is that of the real numbers.

Example 2.3.1. Take $X = \mathbf{R}$ and $d(x,y) = |x - y|$.

Let us generalise this to higher dimensions. In fact, there is no "obvious" generalisation. Here are several natural ones.

Example 2.3.2. Take $X = \mathbf{R}^n$. Then each of the following functions define metrics on X.

$$\begin{aligned} d_1(v,w) &= \sum_{i=1}^n |v_i - w_i|; \\ d_2(v,w) &= \big(\sum_{i=1}^n (v_i - w_i)^2\big)^{1/2} \\ d_\infty(v,w) &= \max_{i \in \{1,2,\dots,n\}} |v_i - w_i|. \end{aligned}$$

These are called the ℓ^1- ("ell one"), ℓ^2- (or Euclidean) and ℓ^∞-distances respectively. Of course, the Euclidean distance is the most familiar one.

The proof that each of d_1, d_2, d_∞ defines a distance is mostly very routine, with the exception of proving that d_2, the Euclidean distance, satisfies the triangle inequality. To establish this, recall that the Euclidean norm $\|v\|_2$ of a vector $v = (v_1, \dots, v_n) \in \mathbf{R}^n$ is

$$\|v\|_2 := \big(\sum_{i=1}^n |v_i|^2\big)^{1/2} = \langle v, v\rangle^{1/2},$$

where the inner product is given by

$$\langle v, w\rangle = \sum_{i=1}^n v_i w_i.$$

Then $d_2(v,w) = \|v - w\|_2$, and so the triangle inequality is the statement that

$$\|u - w\|_2 \leqslant \|u - v\|_2 + \|v - w\|_2.$$

This follows immediately by taking $x = u - v$ and $y = v - w$ in the following lemma.

Lemma 2.3.3. *If $x, y \in \mathbf{R}^n$ then $\|x + y\|_2 \leqslant \|x\|_2 + \|y\|_2$.*

Proof. Since $\|v\|_2 \geqslant 0$ for all $v \in \mathbf{R}^n$ the desired inequality is equivalent to

$$\|x+y\|_2^2 \leqslant \|x\|_2^2 + 2\|x\|_2\|y\|_2 + \|y\|_2^2.$$

But since $\|x+y\|_2^2 = \langle x+y, x+y\rangle = \|x\|_2^2 + 2\langle x,y\rangle + \|y\|_2^2$, this inequality is immediate from the Cauchy-Schwarz inequality, that is to say the inequality $|\langle x,y\rangle| \leqslant \|x\|_2\|y\|_2$. □

The next example is rather a routine and trivial one. However, it behaves very differently to the Euclidean examples and can often provide counterexamples to over-optimistic conjectures based on geometric intuition.

Example 2.3.4 (Discrete metric). Let X be an arbitrary set. The *discrete* metric on a set X is defined as follows:

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

The axioms for a distance function are easy to check.

Now we turn to some metrics which come up very naturally in diverse areas of mathematics. Our first example is critical in number theory, and also serves to show that metrics need not conform to one's most naïve understand of "distance".

Example 2.3.5 (2-adic metric). Let $X = \mathbf{Z}$, and define $d(x,y)$ to be 2^{-m}, where 2^m is the largest power of two dividing $x - y$. The triangle inequality holds in the following stronger form, known as the *ultrametric property*:

$$d(x,z) \leqslant \max(d(x,y), d(y,z)).$$

Indeed, this is just a rephrasing of the statement that if 2^m divides both $x-y$ and $y-z$, then 2^m divides $x-z$.

This metric is very unlike the usual distance. For example, $d(999, 1000) = 1$, whilst $d(0, 1000) = \frac{1}{8}$!

The role of 2 can be replaced by any other prime p, and the metric may also be extended in a natural way to the rationals $\mathbf{Q}$.

Metrics are also ubiquitous in graph theory:

Example 2.3.6 (path metric). Let G be a graph, that is to say a finite set of vertices V joined by edges. Suppose that G is connected, that is to say that there is a path joining any pair of distinct vertices. Define a distance d as follows: $d(v,v) = 0$, and $d(v,w)$ is the length of the shortest path from v to w. Then d is a metric on V, as can be easily checked.

They also come up in group theory:

Example 2.3.7 (Word metric). Let G be a group, and suppose that it is generated by elements a, b and their inverses. Define a distance on G as follows: $d(v, w)$ is the minimal k such that $v = wg_1 \cdots g_k$, where $g_i \in \{a, b, a^{-1}, b^{-1}\}$ for all i.

When G is finite, the word metric is a special case of the path metric – you may wish to think about why.

There are many metrics with a prominent position in computer science, for instance:

Example 2.3.8 (Hamming distance). Let $X = \{0, 1\}^n$ (the boolean cube), the set of all strings of n zeroes and ones. Define $d(x, y)$ to be the number of coordinates in which x and y differ.

Remark. In fact, one can if desired see $\{0, 1\}^n$ as a subset of $\mathbf{R}^n$, and in this case d is the restriction of one of the metrics already considered in Example 2.3.2 (you may care to contemplate which one).

It hardly need be said that metrics are ubiquitous in geometry.

Example 2.3.9 (Projective space). Consider the set $\mathbf{P}(\mathbf{R}^n)$ of one-dimensional subspaces of $\mathbf{R}^n$, that is to say lines through the origin). One way to define a distance on this set is to take, for lines L_1, L_2, the distance between L_1 and L_2 to be

$$d(L_1, L_2) = \sqrt{1 - \frac{|\langle v, w\rangle|^2}{\|v\|^2\|w\|^2}},$$

where v and w are any non-zero vectors in L_1 and L_2 respectively. It is easy to see this is independent of the choice of vectors v and w. The Cauchy-Schwarz inequality ensures that d is well-defined, and moreover the criterion for equality in that inequality ensures positivity. The symmetry property is evident, while the triangle inequality is left as an exercise.

It is useful to think of the case when $n = 2$ here, that is, the case of lines through the origin in the plane $\mathbf{R}^2$. The distance between two such lines given by the above formula is then $\sin(\theta)$ where θ is the angle between the two lines (another exercise).

2.4. Norms

In Example 2.3.2, we looked at three examples of metrics on $\mathbf{R}^n$. They are all, as it turns out, induced from *norms.* This is an important notion which we now develop in its general context.

Definition 2.4.1 (Norms). Let V be any vector space (over the reals). A function $\|\cdot\| : V \to [0, \infty)$ is called a *norm* if the following are all true:

- $\|x\| = 0$ if and only if $x = 0$;
- $\|\lambda x\| = |\lambda|\|x\|$ for all $\lambda \in \mathbf{R}$, $x \in V$;

- $\|x+y\| \leqslant \|x\| + \|y\|$ whenever $x, y \in V$.

Given a norm, it is very easy to check that $d(x,y) := \|x-y\|$ defines a metric on V. Indeed, we have already seen that when $V = \mathbf{R}^n$, $\|\cdot\|_2$ is a norm (and so the name "Euclidean norm" is appropriate) and we defined $d_2(x,y) = \|x-y\|_2$.

As we mentioned, the other metrics in Example 2.3.2 also come from norms. Indeed, d_1 comes from the ℓ^1-norm

$$\|x\|_1 := \sum_{i=1}^n |x_i|,$$

whilst d_∞ comes from the ℓ^∞-norm

$$\|x\|_\infty := \max_{i=1,\dots,n} |x_i|.$$

As the notation suggests, these are special cases of a more general family of norms, the ℓ^p-norms

$$\|x\|_p := \big(\sum_{i=1}^n |x_i|^p\big)^{1/p}.$$

It is true (but we shall not prove it in this course) that these do indeed define norms for $1 \leqslant p < \infty$. Moreover,

$$\lim_{p\to\infty} \|x\|_p = \|x\|_\infty,$$

which is how the ℓ^∞-norm comes to have its name.

The principle of turning norms into metrics is important enough that we state it as a lemma in its own right.

Lemma 2.4.2. *Let V be a vector space over the reals, and let $\|\cdot\|$ be a norm on it. Define $d : V \times V \to [0,\infty)$ by $d(x,y) := \|x-y\|$. Then (V,d) is a metric space.*

It is important to note that the converse is very far from true. For instance, the discrete metric does not arise from a norm. All metrics arising from a norm have the *translation invariance* property $d(x+z, y+z) = d(x,y)$, as well as the scalar invariance $d(\lambda x, \lambda y) = |\lambda| d(x,y)$, neither of which are properties of arbitrary metrics. Conversely one can show that a metric with these two additional properties *does* come from a norm, an exercise we leave to the reader (*Hint*: the norm must arise as $\|v\| = d(v,0)$).

We call a vector space endowed with a norm $\|\cdot\|$ a *normed space.* Whenever we talk about normed spaces it is understood that we are also thinking of them as metric spaces, with the metric being defined by $d(v,w) = \|v-w\|$.

Note that we do not assume that the underlying vector space V is finite-dimensional. Here are some examples which are not finite-dimensional (whilst we do not *prove* that they are not finite-dimensional here, it is not hard to do so and we suggest this as an exercise).

Example 2.4.3 (ℓ^p spaces). Let

$$\ell_1 = \{(x_n)_{n=1}^\infty : \sum_{n\geqslant 1} |x_n| < \infty\}$$

$$\ell_2 = \{(x_n)_{n=1}^\infty : \sum_{n\geqslant 1} x_n^2 < \infty\}$$

$$\ell_\infty = \{(x_n)_{n=1}^\infty : \sup_{n\in\mathbb{N}} |x_n| < \infty\}.$$

The sets $\ell_1, \ell_2, \ell_\infty$ are all real vector spaces, and moreover $\|(x_n)\|_1 = \sum_{n\geqslant 1}|x_n|$, $\|(x_n)\|_2 = \big(\sum_{n\geqslant 1} x_n^2\big)^{1/2}$, $\|(x_n)\|_\infty = \sup_{n\in\mathbf{N}} |x_n|$ define norms on ℓ_1, ℓ_2 and ℓ_∞ respectively. Note that ℓ_2 is in fact an inner product space where

$$\langle (x_n), (y_n)\rangle = \sum_{n\geqslant 1} x_n y_n,$$

(the fact that the right-hand side converges if (x_n) and (y_n) are in ℓ_2 follows from the Cauchy-Schwarz inequality).

The space ℓ^2 is known as *Hilbert space* and it is of great importance in mathematics.

2.5. New metric spaces from old ones

Subspaces. Suppose that (X, d) is a metric space and let Y be a subset of X. Then the restriction of d to $Y \times Y$ gives Y a metric so that $(Y, d_{|Y\times Y})$ is a metric space. We call Y equipped with this metric a *subspace*.

The word "subspace" is rather overused in mathematics. If $X = \mathbf{R}^n$, so that X is a vector space, then Y need not be a *vector* subspace – it is just a subset of X.

Let us give an example of a subspace of a metric space. If $X = \mathbf{R}$, we could take $Y = [0, 1]$, for instance, or $Y = \mathbf{Q}$ (the rationals) or $Y = \mathbf{Z}$ (the integers). (It would be perverse to *define* the usual metric on $\mathbf{Z}$ or on $\mathbf{Q}$ by restricting from $X = \mathbf{R}$. Indeed, the metric space (X, d) with $X = \mathbf{Z}$ and $d(x, y) := |x - y|$ is a much more basic object than $\mathbf{R}$.)

Product spaces. If (X, d_X) and (Y, d_Y) are metric spaces, then it is natural to try to make $X \times Y$ into a metric space. One method is as follows: if $x_1, x_2 \in X$ and $y_1, y_2 \in Y$ then we set

$$d_{X\times Y}((x_1, y_1), (x_2, y_2)) = \sqrt{d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2}$$

The use of the square mean on the right, rather than the max or the sum, is appealing since then the product $\mathbf{R} \times \mathbf{R}$ becomes the space $\mathbf{R}^2$ with the Euclidean metric. However, either of those alternative definitions results in a metric which is equivalent, in the sense made precise in Section 4.3.

Lemma 2.5.1. *With notation as above, $d_{X\times Y}$ gives a metric on $X \times Y$.*

Proof. Positivity and symmetry are obvious. Less clear is the triangle inequality. We need to prove that

$$\sqrt{d_X(x_1,x_3)^2+d_Y(y_1,y_3)^2}+\sqrt{d_X(x_3,x_2)^2+d_Y(y_3,y_2)^2} \geqslant \sqrt{d_X(x_1,x_2)^2+d_Y(y_1,y_2)^2}. \tag{2.1}$$

To make this appear less frightening, write $a_1 = d_X(x_2,x_3)$, $a_2 = d_X(x_1,x_3)$, $a_3 = d_X(x_1,x_2)$ and similarly $b_1 = d_Y(y_2,y_3)$, $b_2 = d_Y(y_1,y_3)$ and $b_3 = d_Y(y_1,y_2)$. Thus we want to show

$$\sqrt{a_2^2+b_2^2}+\sqrt{a_1^2+b_1^2} \geqslant \sqrt{a_3^2+b_3^2}. \tag{2.2}$$

To prove this, note that from the triangle inequality we have $a_1+a_2 \geqslant a_3$, $b_1+b_2 \geqslant b_3$. Squaring and adding gives

$$a_1^2+b_1^2+a_2^2+b_2^2+2(a_1a_2+b_1b_2) \geqslant a_3^2+b_3^2.$$

By Cauchy-Schwarz,

$$a_1a_2+b_1b_2 \leqslant \sqrt{a_1^2+b_1^2}\sqrt{a_2^2+b_2^2}.$$

Substituting this into the previous line gives precisely the square of (2.2), and (2.1) follows. □

2.6. Balls and boundedness

Definition 2.6.1 (Balls). Let X be a metric space. If $a \in X$ and $\varepsilon > 0$ then we define the *open ball of radius* ε to be the set

$$B(a,\varepsilon) = \{x \in X : d(x,a) < \varepsilon\}.$$

Similarly we defined the *closed ball* of radius ε about a to be the set

$$\overline{B}(a,\varepsilon) = \{x \in X : d(x,a) \leqslant \varepsilon\}.$$

Thus when $X = \mathbf{R}^3$ with the Euclidean metric we see that $B(0,1)$ really is what we understand geometrically as a ball (minus its boundary, the unit sphere), whilst $\overline{B}(0,1)$ contains the unit sphere and everything inside it.

We caution that this intuitive picture of the closed ball being the open ball "together with its boundary" is totally misleading in general. For instance, in the discrete metric on a set X, the open ball $B(a,1)$ contains only the point a, whereas the closed ball $\overline{B}(a,1)$ is the whole of X.

Definition 2.6.2. Let X be a metric space, and let $Y \subseteq X$. Then we say that Y is *bounded* if Y is contained in some open ball.

Lemma 2.6.3. *Let X be a metric space and let $Y \subseteq X$. Then the following are equivalent.*

(i) *Y is bounded;*

(ii) *Y is contained in some closed ball;*

(iii) *The set $\{d(y_1, y_2) : y_1, y_2 \in Y\}$ is a bounded subset of* $\mathbf{R}$.

Proof. That (i) implies (ii) is totally obvious. That (ii) implies (iii) follows immediately from the triangle inequality. Finally, suppose Y satisfies (iii). Then there is some K such that $d(y_1, y_2) \leqslant K$ whenever $y_1, y_2 \in Y$. If Y is empty, it is certainly bounded. Otherwise, let $a \in Y$ be an arbitrary point. Then Y is contained in $B(a, r)$ where $r = K + 1$. □

CHAPTER 3

Limits and continuity

The main purpose of introducing the idea of a metric space is that many notions familiar over $\mathbf{R}$, such as those of limit and continuous function, can be extended to metric spaces, and theorems about them proven in that context.

3.1. Basic definitions and properties

Definition 3.1.1 (Limit). Suppose that $(x_n)_{n=1}^{\infty}$ is a sequence of elements of a metric space (X, d). Let $x \in X$. Then we say that $x_n \to x$, or that $\lim_{n\to\infty} x_n = x$, if the following is true. For every $\varepsilon > 0$, there is an N such that $d(x_n, x) < \varepsilon$ for all $n \geqslant N$.

Let us bolster this definition with a couple of easy remarks. First, it is quite possible and indeed usual for a sequence $(x_n)_{n=1}^{\infty}$ to have no limit. Take, for instance, the sequence $(0, 1, 0, 1, 0, 1, \dots)$ in $\mathbf{R}$. Second, if the limit does exist then it is unique. To see this, suppose that $x_n \to a$ and $x_n \to b$, but that $a \neq b$. Let $\delta := d(a, b)$. Then, taking $\varepsilon = \delta/2$ in the definition of limit, we see that for n sufficiently large we have $d(x_n, a), d(x_n, b) < \delta/2$. But then the triangle inequality yields

$$\delta = d(a, b) \leqslant d(x_n, a) + d(x_n, b) < \delta,$$

a contradiction.

Definition 3.1.2 (Continuity). Let (X, d_X) and (Y, d_Y) be metric spaces. We say a function $f \colon X \to Y$ is continuous at $a \in X$ if for any $\varepsilon > 0$ there is a $\delta > 0$ such that for any $x \in X$ with $d_X(a, x) < \delta$ we have $d_Y(f(x), f(a)) < \varepsilon$.

We say f is *continuous* if it is continuous at every $a \in X$.

Although we will not come across it all that much in this course, it is important to note that the definition of *uniform* continuity may be extended to metric spaces as well. As for real functions, the idea is that "δ should depend only on ε".

Definition 3.1.3 (Uniform continuity). Let (X, d_X) and (Y, d_Y) be metric spaces. We say a function $f \colon X \to Y$ is uniformly continuous if for any $\varepsilon > 0$ there is a $\delta > 0$ such that for any $x, y \in X$ with $d_X(x, y) < \delta$ we have $d_Y(f(x), f(y)) < \varepsilon$.

As for functions on the reals, one may also phrase the definition of continuity in terms of limits.

Lemma 3.1.4. *Let $f : X \to Y$ be a function between metric spaces. Then f is continuous at a if and only if the following is true: for any sequence $(x_n)_{n=1}^{\infty}$ with $\lim_{n\to\infty} x_n = a$, we have $\lim_{n\to\infty} f(x_n) = f(a)$.*

Proof. Suppose first that f is continuous at a. Let $\varepsilon > 0$. Then there is a $\delta > 0$ such that for all $x \in X$ with $d(x, a) < \delta$ we have $d(f(x), f(a)) < \varepsilon$. Now let $(x_n)_{n=1}^{\infty}$ be a sequence with limit a. So, by the definition of limit, there is an $N > 0$ such that $d(a, x_n) < \delta$ for all $n \geqslant N$. But then for all $n \geqslant N$ we see that $d(f(a), f(x_n)) < \varepsilon$, so indeed $\lim_{n\to\infty} f(x_n) = f(a)$ as required.

For the other direction we show the contrapositive. Suppose f is not continuous at a. Then there is an $\varepsilon > 0$ such that for all $\delta > 0$ there is some $x \in X$ with $d(x, a) < \delta$ and $d(f(x), f(a)) \geqslant \varepsilon$. Taking $\delta = 1/n$, we see that for each n there is some $x_n \in X$ with $d(x_n, a) < 1/n$ and $d(f(x_n), f(a)) \geqslant \varepsilon$. Therefore $\lim x_n = a$, but $\lim f(x_n) \neq f(a)$. □

3.2. Function spaces

A great deal of power comes from considering the set of all functions on a space satisfying some property, such as continuity, as a metric space in its own right. In this section we consider some important examples of such spaces.

We begin with the space of bounded real-valued functions on a set X. At this stage we assume nothing about X.

Definition 3.2.1. If X is any set we define $B(X)$ to be the space of functions $f : X \to \mathbf{R}$ for which $f(X) = \{f(x) : x \in X\}$ is bounded. If $f \in B(X)$, define $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

Lemma 3.2.2. *For any set X, $B(X)$ is a vector space, and $\|\cdot\|_\infty$ is a norm.*

We leave the proof as an easy exercise.

Now we turn to the space of continuous real-valued functions, $C(X)$. To make sense of what this means we now need X to be a metric space.

Definition 3.2.3. Let X be a metric space. Then we write $C(X)$ for the space of all continuous functions $f : X \to \mathbf{R}$.

Lemma 3.2.4. *The space $C(X)$ is a vector space over $\mathbf{R}$, with pointwise addition and multiplication by scalars.*

Proof. One must check that $C(X)$ is closed under addition and scalar multiplication. We do the case of addition; scalar multiplication is left as an (easy) exercise.

Suppose that $f, g \in C(X)$, and let $\varepsilon > 0$. Let $a \in X$.

Since f is continuous at a, there is some δ_1 such that $d(x, a) < \delta_1$ implies $|f(x) - f(a)| < \varepsilon/2$.

Since g is continuous at a, there is some δ_2 such that $d(x,a) < \delta_2$ implies $|g(x) - g(a)| < \varepsilon/2$.

Take $\delta = \min(\delta_1, \delta_2)$. Then, if $d(x,a) < \delta$ we have

$$\begin{aligned}|(f+g)(x) - (f+g)(a)| &= |f(x) + g(x) - f(a) - g(a)| \\ &\leqslant |f(x) - f(a)| + |g(x) - g(a)| \\ &< \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

Therefore $f + g$ is continuous at a. □

In general, we certainly do not have $B(X) \subseteq C(X)$, and unless X is special we do not have $C(X) \subseteq B(X)$. We will discuss situations in which this *is* true later on; you will already be familiar with a nontrivial example, namely that $C([0,1]) \subseteq B([0,1])$, that is to say all continuous functions on $[0,1]$ are bounded.

Definition 3.2.5. Let X be a metric space. Write $C_b(X) := C(X) \cap B(X)$ for the space of continuous, bounded functions on X. Since $C_b(X)$ is a subspace of $B(X)$, it inherits the norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$, and we may define a metric d_∞ on $C_b(X)$ in the usual way via $d_\infty(f,g) := \|f - g\|_\infty$

A useful exercise in checking your understanding of these definitions is Example 3.2.6 below. Here, let $X = [0,1]$. Then, as we just remarked, $C(X) = C_b(X)$. Instead of writing $C([0,1])$, it is conventional to write $C[0,1]$ for the vector space of continuous (and automatically bounded) functions on $[0,1]$.

Example 3.2.6. Consider the space $C[0,1]$ together with the metric d_∞ induced from the norm $\|\cdot\|_\infty$. Let $(f_n)_{n=1}^\infty$ be a sequence of elements (functions) of this space, and let f be a further element. Then $f_n \to f$ in the metric d_∞ if, and only if, f_n converges to f uniformly.

Proof. This is essentially a tautology, but it takes a little thought to unravel all the definitions. □

CHAPTER 4

Isometries, homeomorphisms and equivalence

One learns as mathematician that, when one studies a type of structure, one should also study maps which preserve that structure. In this chapter we will look at various such notions applicable to metric spaces.

4.1. Isometries

Maps which genuinely preserve the distance function are called isometries.

Definition 4.1.1. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f \colon X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is said to be an *isometry* if

$$d_Y(f(x), f(y)) = d_X(x, y) \text{ for all } x, y \in X. \tag{4.1}$$

Remarks. An isometry is automatically injective, but not automatically surjective. For instance, the right-shift map on ℓ^2 defined by $f((x_1, x_2, x_3, \dots)) = (0, x_1, x_2, \dots)$ is an isometry, but it is not surjective.

If an isometry is surjective as well, we call it a *bijective isometry*. Some authors use the word "isometry" to mean "bijective isometry", but we have refrained from doing this so that we are consistent with Prelims course M4: Geometry. There, isometries in the case $X = Y = \mathbf{R}^n$ were discussed in considerable detail. It was shown that they all have the form $f(x) = Ax + b$ for some orthogonal matrix A; in particular, they are automatically surjective.

For any metric space X the set of all bijective isometries from X to itself is a group under composition, denoted $\mathrm{Isom}(X)$.

4.2. Homeomorphisms

The notion of isometry is rather rigid. A weaker notion is that of a homeomorphism.

Definition 4.2.1. Let $f \colon X \to Y$ be a continuous function between metric spaces X and Y. We say that f is a *homeomorphism* if it is continuous, a bijection, and if its inverse $f^{-1} : Y \to X$ is also continuous.

Remark. Note that it is possible for a map $f : X \to Y$ to be both continuous and a bijection, but for its inverse to fail to be continuous (so in this case f is not a homeomorphism). For instance, consider the spaces $X = [0, 1) \cup [2, 3]$ and $Y = [0, 2]$. Then the function

$f\colon X \to Y$ given by

$$f(x) = \begin{cases} x, & \text{if } x \in [0,1) \\ x-1, & \text{if } x \in [2,3] \end{cases}$$

is a bijection and is clearly continuous. However, its inverse $g\colon Y \to X$ is not continuous at 1 – the one-sided limits of g as x tends to 1 from below and above are 1 and 2 respectively.

The following examples illustrate the extent to which homeomorphisms are less rigid than isometries.

Example 4.2.2. The closed disk $\bar{B}(0,1)$ of radius 1 in $\mathbf{R}^2$ is homeomorphic to the square $[-1,1]\times[-1,1]$. The easiest way to see this is to inscribe the disk in the square and stretch the disk radially out to the square. One can write explicit formulas for this in the four quarters of the disk given by the lines $x \pm y = 0$ to check this does indeed give a homeomorphism.

Example 4.2.3. The open interval $(-1,1)$ is homeomorphic to $\mathbf{R}$: an explicit homeomorphism is given by $f(x) = x/(1-|x|)$, which has inverse $g(x) = x/(1+|x|)$. It follows (using translation and scaling maps) that any open interval is homeomorphic to $\mathbf{R}$. Similarly, the function $h(x) = 1/x$ shows that $(0,1)$ and $(1,\infty)$ are homeomorphic, and from this one can see that the spaces $\mathbf{R}$, (a,b), $(-\infty,a)$ and (a,∞) are all homeomorphic for any $a,b \in \mathbf{R}$ with $a < b$.

Example 4.2.4. A coffee cup (reusable, with a handle) is homeomorphic to a doughnut.

4.3. Equivalent metrics (non-examinable)

One space X can certainly support wildly different metrics. For instance, the 2-adic metric on $\mathbf{Q}$ is very different to the standard Euclidean metric. However, there is a useful notion of two metrics d_1, d_2 on the same space being equivalent.

Definition 4.3.1 (Equivalent metrics). Let X be a set, and let d, d' be two metrics on X. Then we say that the metrics d, d' are equivalent if the identity map $\iota : (X,d) \to (X,d')$ is a homeomorphism.

An easy exercise in the definitions show that this is equivalent to the following property: every open ball $B(x,\varepsilon)$ with respect to the d-metric contains an open ball $B'(x,\varepsilon')$ in the d'-metric, and vice versa.

If two metrics d, d' are equivalent then, for example, the notions of limit coincide in the two metric spaces (X,d) and (X,d'). We leave the detailed proof as an exercise.

Proposition 4.3.2. *The metrics d_1, d_2, d_∞ on $\mathbf{R}^n$ are equivalent.*

Proof. In fact, we will show that these metrics are *strongly* equivalent. Two metrics d, d' on a space X are strongly equivalent if there is a constant C such that

$$d(x,y) \leqslant Cd'(x,y) \text{ and } d'(x,y) \leqslant Cd(x,y)$$

for all $x \neq y$. We leave it as an easy exercise to show that strongly equivalent metrics are indeed equivalent (the converse is not true).

The three metrics under consideration all come from norms, and it is enough to find some constant C such that

$$\|x\| \leqslant C\|x\|' \tag{4.2}$$

for each pair $\|\cdot\|, \|\cdot\|'$ of these norms. Four such inequalities are obvious, namely

$$\|x\|_\infty \leqslant \|x\|_1 \leqslant n\|x\|_\infty$$

and

$$\|x\|_\infty \leqslant \|x\|_2 \leqslant n^{1/2}\|x\|_\infty.$$

The remaining two inequalities follow from these two, or one could use the Cauchy-Schwarz inequality to get better constants. □

CHAPTER 5

Open and closed sets

5.1. Basic definitions

The definition of an open set, which we give now, is one of the most important in the course.

Definition 5.1.1 (Open sets). If X is a metric space then we say a subset $U \subseteq X$ is *open* (or *open in* X) if for each $y \in U$ there is some $\delta > 0$ such that the open ball $B(y, \delta)$ is contained in U.

To check you have understood the definition, convince yourself of the following facts:

- The interval $(0,1)$ is open in $\mathbf{R}$, but $[0,1]$ is not;
- The rationals $\mathbf{Q}$ are not open in $\mathbf{R}$;
- If X is a set with the discrete metric, every set is open.

Note carefully that the notion of open set is a relative notion, depending on U being contained in X. Thus, while $[0,1]$ is not open in $\mathbf{R}$, it is an open set considered as a subspace of itself.

The first basic result about open sets is that open balls $B(a, \varepsilon)$ are open. Note that this is not a tautology – at this point "open ball" is just the name we gave to the sets $B(a, \varepsilon)$, and the fact that they are indeed open in the sense of Definition 5.1.1 needs to be proven.

Lemma 5.1.2. *Every open ball in an metric space is an open set.*

Proof. Let the ball be $B(a, \varepsilon)$. Let $x \in B(a, \varepsilon)$. Then $d(x, a) < \varepsilon$, so there is $\varepsilon' > 0$ so that $d(x, a) < \varepsilon - \varepsilon'$. We claim that the open ball $B(x, \varepsilon')$ is contained in $B(a, \varepsilon)$. To see this, suppose that $z \in B(x, \varepsilon')$. Then $d(z, x) < \varepsilon'$ and so by the triangle inequality $d(z, a) \leqslant d(z, x) + d(x, a) < \varepsilon' + (\varepsilon - \varepsilon') = \varepsilon$. □

The complement of an open set is a closed set.

Definition 5.1.3 (Closed sets). If X is a metric space, then a subset $F \subseteq X$ is said to be a *closed* subset of X if and only if its complement $F^c = X \backslash F$ is an open subset.

It is important to note that the property of being closed is *not* the property of not being open! In a metric space, it is possible for a subset to be open, closed, both or neither: In $\mathbf{R}$

the set $\mathbf{R}$ is open and closed, the set $(0,1)$ is open and not closed, the set $[0,1]$ is closed and not open while the set $(0,1]$ is neither.

Just as open balls are open sets, so closed balls are closed sets, and this is also a fact requiring proof.

LEMMA 5.1.4. *Every closed ball in a metric space is a closed set. In particular, singleton sets are closed.*

Proof. Let the ball be $\bar{B}(a,\varepsilon)$. We will show that the complement $\bar{B}(a,\varepsilon)^c$ is open. Let $x \in \bar{B}(a,\varepsilon)^c$. Then $d(x,a) > \varepsilon$, so there is $\varepsilon' > 0$ so that $d(x,a) > \varepsilon + \varepsilon'$. We claim that the open ball $B(x,\varepsilon')$ is contained in $\bar{B}(a,\varepsilon)^c$. To see this, suppose that $z \in B(x,\varepsilon')$. Then $d(z,x) < \varepsilon'$ and so by the triangle inequality $d(z,a) \geqslant d(x,a) - d(z,x) > (\varepsilon + \varepsilon') - \varepsilon' = \varepsilon$.

The second statement – that singleton sets are closed – follows from the observation that $\{a\} = \bar{B}(a,0)$. □

5.2. Basic properties of open sets

LEMMA 5.2.1. *Let X be metric space. Then we have*

(i) *The subsets X and $\emptyset$ are open.*
(ii) *For any indexing set I and $\{U_i : i \in I\}$ a collection of open sets, the set $\bigcup_{i\in I} U_i$ is an open set.*
(iii) *If I is finite and $\{U_i : i \in I\}$ are open sets then $\bigcap_{i\in I} U_i$ is open.*

Proof. The first claim is trivial. For the second claim, if $x \in \bigcup_{i\in I} U_i$ then there is some $i \in I$ with $x \in U_i$. Since U_i is open, some open ball $B(x,\varepsilon)$ is contained in U_i and hence in $\bigcup_{i\in I} U_i$.

Finally, for claim (iii), suppose that I is finite and that $x \in \bigcap_{i\in I} U_i$. For each $i \in I$, we have $x \in U_i$, and so some ball $B(x,\varepsilon_i)$ is contained in U_i. Set $\varepsilon := \min_{i\in I} \varepsilon_i$; then $\varepsilon > 0$ (here it is, of course, crucial that I be finite), and $B(x,\varepsilon) \subseteq B(x,\varepsilon_i) \subseteq U_i$ for all i. Therefore $B(x,\varepsilon) \subseteq \bigcap_{i\in I} U_i$. □

Remarks. (i) is in fact a special case of (ii) and (iii), taking I to be the empty set.

It is extremely important to note that, whilst the indexing set I in (ii) can be arbitrary, the indexing set in (iii) must be finite. In general, an arbitrary intersection of open sets is not open; for instance, the intervals $U_i = (-1/i, 1/i)$ are all open in $\mathbf{R}$, but their intersection $\bigcap_{i=1}^{\infty} U_i$ is just the singleton $\{0\}$, which is not an open set.

A result equivalent to Lemma 5.2.1 may be formulated in terms of closed sets, simply by taking complements and applying de Morgan's laws. We simply state the outcome.

Lemma 5.2.2. *Let X be a metric space and let $\{F_i : i \in I\}$ be a collection of closed subsets.*

(i) *The subsets X and $\emptyset$ are closed.*
(ii) *The intersection $\bigcap_{i\in I} F_i$ is a closed subset.*
(iii) *If I is finite then $\bigcup_{i\in I} F_i$ is closed.*

If X is a metric space, the collection of all open sets in X is called the *topology* of X.

5.3. Continuity in terms of open sets

An interesting and important fact is that continuity of a function may be expressed in an ε-δ-free manner using open sets.

Proposition 5.3.1. *Let X, Y be metric spaces and let $f : X \to Y$ be a map. Then f is continuous on all of X if and only if for each open subset U of Y, its preimage $f^{-1}(U)$ is open in X.*

Proof. Suppose first that f is continuous at every point, and let $U \subseteq Y$ be open; we want to show that $f^{-1}(U)$ is open. Let $a \in f^{-1}(U)$ be arbitrary. Then $f(a) \in U$, and so, since U is open, some ball $B(f(a), \varepsilon)$ also lies in U. By the definition of continuity, there is some $\delta > 0$ such that if $x \in B(a, \delta)$ then $f(x) \in B(f(a), \varepsilon)$, and therefore $f^{-1}(B(f(a), \varepsilon)) \supseteq B(a, \delta)$. Therefore $f^{-1}(U)$ contains $B(a, \delta)$, which means that $f^{-1}(U)$ is open.

Now suppose that f satisfies the open sets preimages property, and let $a \in X$. The ball $B(f(a), \varepsilon)$ is open, and so by assumption the preimage $f^{-1}(B(f(a), \varepsilon))$ is open. Since a lies in this set, it follows from the definition of open that there is some $\delta > 0$ such that $B(a, \delta) \subseteq f^{-1}(B(f(a), \varepsilon))$, whence $f(B(a, \delta)) \subseteq B(f(a), \varepsilon)$. This is what it means for f to be continuous at a. □

By taking complements, one can show the following version of Proposition 5.3.1 for closed sets: $f : X \to Y$ is continuous if and only if for each *closed* subset V of Y, its preimage $f^{-1}(V)$ is a *closed* subset of X.

Finally, it is important to take note of what Proposition 5.3.1 does *not* say, namely that a continuous function maps open sets to open sets. This is obvious since, for example, constant functions are continuous. Less obvious is the fact that it still fails even under the assumption that f is injective. For instance, the injective map $f : [0, 1) \to S^1 \subset \mathbf{C}$ given by $t \to e^{2\pi i t}$ is continuous. The set $[0, 1/2)$ is open in $[0, 1)$, but its image is not open in S^1.

5.4. Topological spaces (non-examinable)

In this section we offer a very brief taster of the course A5: Topology by discussing the notion of a topological space. One may of course observe that Proposition 5.3.1 allows one to define the notion of a continuous function without explicitly mentioning the metric X or concepts equivalent to it such as the notion of an open ball of radius δ. Of course, those notions are embedded within the definition of the notion of an open set, so this comment is a little misleading.

In the concept of a topological space, the open sets are to the fore. Thus a topological space is a set X together with a collection of sets U (which we call the open sets) satisfying certain properties. The properties we require are precisely those which we *proved* in Lemma 5.2.1, namely

(i) The subsets X and $\emptyset$ are open.

(ii) For any indexing set I and $\{U_i; i \in I\}$ a collection of open sets, the set $\bigcup_{i \in I} U_i$ is an open set.

(iii) If I is finite and $\{U_i : i \in I\}$ are open sets then $\bigcap_{i \in I} U_i$ is open in X.

Note that we have not said anything about the "geometry" of the open sets, or anything about them containing balls - indeed there are no such notions, because X is equipped with no structure.

Lemma 5.2.1 may then be phrased as follows.

LEMMA 5.4.1 (Lemma 5.2.1). *Let X be a metric space together with the open sets as defined in Definition 5.1.1. Then X is a topological space, with the same collection of open sets.*

The concept of a topological space is considerably more general than that of a metric space, and there are certainly topological spaces which do not have the structure of a metric space (are not *metrizable*). However, as a consequence of Proposition 5.3.1 we may still formulate the notion of a continuous function between two topological spaces, in such a way that when restricted to metric spaces it coincides with the usual definition.

DEFINITION 5.4.2. Suppose that X and Y are two topological spaces. Then we say that $f : X \to Y$ is continuous if and only if, for every open set $U \subseteq Y$, the inverse $f^{-1}(U)$ is open in X.

Let us emphasise that in the generality of topological spaces, there is no equivalent form of this definition in terms of εs and δs.

5.5. Subspaces

If (X, d) is a metric space, then as we noted in Section 2.5, any subset $Y \subseteq X$ is automatically also a metric space since the distance function $d\colon X \times X \to \mathbf{R}_{\geqslant 0}$ restricts to a distance function on Y. We will use the letter d for both metrics, but it is important to distinguish the balls in Y from the balls in X, because these are quite different objects.

We will write

$$B_Y(y,r) = \{z \in Y : d(z,y) < r\}$$

for the open ball about y of radius r in Y and

$$B_X(y,r) = \{x \in X : d(x,y) < r\}$$

for the open ball of radius r about y in X.

Note that $B_Y(y,r) = Y \cap B_X(y,r)$.

Similarly, the notions of a set being open in X and of being open in Y are quite different.

By way of an example, consider $X = \mathbf{R}^2$ and $Y = \mathbf{R} \times \{0\}$, that is to say Y is the x-axis. The the ball $B_X(0,1)$ is simply the open unit disc of radius 1, whilst the ball $B_Y(0,1)$ is the open unit line segment $(-1,1)$. Note carefully that $B_Y(0,1)$, whilst it is open as a subset of Y, does not look remotely like an open subset of X.

The following lemma clarifies the relationship between open sets in X and open sets in the subspace Y.

Lemma 5.5.1. *Let X be a metric space and suppose that $Y \subseteq X$. Then a subset $U \subseteq Y$ is an open subset of Y if and only if there is an open subset V of X such that $U = Y \cap V$. Similarly a subset $Z \subseteq Y$ is a closed subset of Y if and only if there is a closed subset F of X such that $Z = F \cap Y$.*

Proof. Suppose first that $U = Y \cap V$, where V is open in X. We will show that U is open in Y. Let $y \in U$. Then, since V is open, there is some $\varepsilon > 0$ such that $B_X(y,\varepsilon) \subseteq V$. Therefore

$$B_Y(y,\varepsilon) = Y \cap B_X(y,\varepsilon) \subseteq V \cap Y = U.$$

We have shown that some open ball (in Y) about y is contained in U, and therefore U is open.

In the other direction, suppose that U is an open subset of Y. Then for each $y \in U$ we may pick an open ball $B_Y(y,\varepsilon_y)$ contained in U. We have $\bigcup_{y \in U} B_Y(y,\varepsilon_y) = U$. Now define $V = \bigcup_{y\in U} B_X(y,\varepsilon_y)$. Then V, being a union of open balls in X, is open. Moreover

$$Y \cap V = Y \cap \bigcup_{y \in Y} B_X(y,\varepsilon_y) = \bigcup_{y\in Y}(Y \cap B_X(y,\varepsilon_y)) = \bigcup_{y \in Y} B_Y(y,\varepsilon_y) = U.$$

The corresponding result for closed sets follows by taking complements – we leave the detailed verification as an exercise. □

The concept of being open in a subspace is a bit confusing when you first meet it, so let us give an example. Let $X = \mathbf{R}$, and $Y = (0,1] \cup [2,3]$. Set $U = (0,1]$. Then U is not open as a subset of X – for instance, no open ball $B_X(1,\varepsilon)$ is contained in U. However, U

is open as a subset of Y. For example, the ball $B_Y(1, \frac{1}{2})$, which consists of all points of Y at distance less than $\frac{1}{2}$ from 1, is the set $(\frac{1}{2}, 1]$, and this is contained in U.

CHAPTER 6

Interiors, closures, limit points

In this chapter we explore some further concepts in the basic theory of metric spaces.

6.1. Interiors and closures

Definition 6.1.1. Let X be a metric space, and let $S \subseteq X$. The *interior* $\text{int}(S)$ of S is defined to be the union of all open subsets of X contained in S. The closure $\bar{S}$ is defined to be the intersection of all closed subsets of X containing S. The set $\bar{S} \backslash \text{int}(S)$ is known as the *boundary* of S and denoted ∂S. A set $S \subseteq X$ is said to be *dense* if $\overline{S} = X$.

It is very important to note that, in Definition 6.1.1, the notion of open and closed is here being taken in the metric space X, *not* in the subspace metric on S, which would result in trivial definitions.

Since an arbitrary union of open sets is open (Lemma 5.2.1), $\text{int}(S)$ is itself an open set, and it is clearly the unique largest open subset of X contained in S. If S is itself open then evidently $S = \text{int}(S)$.

Since an arbitrary intersection of closed sets is closed, $\bar{S}$ is the unique smallest closed subset of X containing S. If S is itself closed then evidently $S = \overline{S}$.

If $x \in \text{int}(S)$ we say that x is an *interior point* of S.

Example 6.1.2. If $S = [a, b]$ is a closed interval in $\mathbf{R}$ then its interior is just the open interval (a, b). If we take $S = \mathbb{Q} \subset \mathbf{R}$ then $\text{int}(\mathbb{Q}) = \emptyset$.

Example 6.1.3. The rationals $\mathbb{Q}$ are a dense subset of $\mathbf{R}$, as is the set $\{\frac{a}{2^n} : a \in \mathbb{Z}, n \in \mathbb{N}\}$.

Let us give a couple of simple characterisations of the closure of a set.

Lemma 6.1.4. *Let X be a metric space, and let $S \subseteq X$ be a subset. Then $a \in \bar{S}$ if and only if the following is true: every open ball $B(a, \varepsilon)$ contains a point of S.*

Proof. Suppose $a \in \bar{S}$ and let $\varepsilon > 0$. If $B(a, \varepsilon)$ does not meet S, then $B(a, \varepsilon)^c$ is a closed set containing S. Therefore $B(a, \varepsilon)^c$ contains $\bar{S}$, and hence it contains a, which is obviously nonsense.

Conversely, let $a \in X$ and suppose that every ball $B(a, \varepsilon)$ meets S. If $a \notin \bar{S}$ then, since $\bar{S}^c$ is open, there is a ball $B(a, \varepsilon)$ contained in $\bar{S}^c$, and hence in S^c, contrary to assumption. □

Remark. A particular consequence of this is that $S \subseteq X$ is dense if and only if it meets every open set in X.

COROLLARY 6.1.5. *Let X be a metric space, and let $S \subseteq X$ be a subset. Let $a \in X$. Then a lies in the closure $\bar{S}$ if and only if there is a sequence $(x_n)_{n=1}^{\infty}$ of elements of S with* $\lim_{n\to\infty} x_n = a$. *In particular, S is closed if and only if the limit of every convergent sequence $(x_n)_{n=1}^{\infty}$ of elements of S lies in S.*

Proof. We use Lemma 6.1.4. Suppose that $a \in \bar{S}$. Then by Lemma 6.1.4 every ball $B(a, 1/n)$ contains a point of S, so we may pick a sequence $(x_n)_{n=1}^{\infty}$ with $x_n \in B(a, 1/n) \cap S$. Clearly $\lim_{n\to\infty} x_n = a$.

Conversely, suppose $\lim_{n\to\infty} x_n = a$, where $x_n \in S$. If $a \notin \bar{S}$ then by Lemma 6.1.4 there must be some ball $B(a, \varepsilon)$ not meeting S. But if n is large enough then $d(x_n, a) < \varepsilon$, and so $x_n \in S \cap B(a, \varepsilon)$, contradiction. □

We conclude with a cautionary example which has often confused people when they first meet it.

EXAMPLE 6.1.6. In general, it need *not* be the case that $\bar{B}(a, \varepsilon)$ is the closure of $B(a, \varepsilon)$. Since we have seen that $\bar{B}(a, \varepsilon)$ is closed, it is always true that $\overline{B(a, \varepsilon)} \subseteq \bar{B}(a, \varepsilon)$, but the containment can be proper. Indeed, take any set X with at least two elements equipped with the discrete metric. Then if $x \in X$ we have $B(x, 1) = \overline{B(x, 1)} = \{x\}$, but $\bar{B}(x, 1)$ is the whole space X.

6.2. Limit points

This section introduces the notion of *limit points* (also known in some places as cluster points or accumulation points). The notion is a well-studied one, introduced here for cultural reference, but we will not come across it in subsequent chapters of the course.

DEFINITION 6.2.1. If X is a metric space and $S \subseteq X$ is any subset, then we say a point $a \in X$ is a *limit point* of S if any open ball about a contains a point of S other than a itself.

We will write $L(S)$ for the set of limit points of S; I am not sure that there is any completely standard notation for this. Note that we do not necessarily have $S \subseteq L(S)$, that is to say it is quite possible for a point $a \in S$ not to be a limit point of S. This occurs if there is some ball $B(a, \varepsilon)$ such that $B(a, \varepsilon) \cap S = \{a\}$, and in this case we say that a is an *isolated point* of S.

EXAMPLE 6.2.2. Take $X = \mathbf{R}$ and $S = (0, 1] \cup \{2\}$. Then $L(S) = [0, 1]$. Note in particular that 0 does not lie in S, but is a limit point; by contrast, 2 does lie in S, but it is not a limit point, so it is an isolated point.

LEMMA 6.2.3. *Let S be a subset of a metric space X. Then $L(S)$ is a closed subset of X.*

Proof. We need to show that the complement $L(S)^c$ is open. Suppose $a \in L(S)^c$. Then there is a ball $B(a, \varepsilon)$ whose intersection with S is either empty or $\{a\}$.

We claim that $B(a, \varepsilon/2) \subseteq L(S)^c$. Let $b \in B(a, \varepsilon/2)$. If $b = a$, then clearly $b \in L(S)^c$. If $b \neq a$, there is some ball about b which is contained in $B(a, \varepsilon)$, but does not contain a: the ball $B(b, \delta)$ where $\delta = \min(\varepsilon/2, d(a, b))$ has this property. This ball meets S in the empty set, and so $b \in L(S)^c$ in this case too. □

PROPOSITION 6.2.4. *Let S be a subset of a metric space X. Let $L(S)$ be its set of limit points, and $\bar{S}$ its closure. Then $\bar{S} = S \cup L(S)$.*

Proof. We first show the containment $S \cup L(S) \subseteq \bar{S}$. Obviously $S \subseteq \bar{S}$, so we need only show that $L(S) \subseteq \bar{S}$. Suppose $a \in \bar{S}^c$. Since $\bar{S}^c$ is open, there is some ball $B(a, \varepsilon)$ which lies in $\bar{S}^c$, and hence also in S^c, and therefore a cannot be a limit point of S. This concludes the proof of this direction.

Now we look at the opposite containment $\bar{S} \subseteq S \cup L(S)$. Let $a \in \bar{S}$. We saw in Lemma 6.1.5 that there is a sequence $(x_n)_{n=1}^{\infty}$ of elements of S with $x_n \to a$. If $x_n = a$ for some n then we are done, since this implies that $a \in S$. Suppose, then, that $x_n \neq a$ for all n. Let $\varepsilon > 0$. Then there exists N such that for all $n \geqslant N$ we have $x_n \in B(a, \varepsilon) \setminus \{a\}$, and they all also lie in S. It follows that a is a limit point of S, as $B(a, \varepsilon)$ contains for example $x_N \neq a$. So we are done in this case also. □

COROLLARY 6.2.5. *Let S be a subset of a metric space X. Then S is closed if and only if it contains all its limit points.*

Proof. We already remarked, in Section 6.1, that S is closed if and only if $S = \bar{S}$. The corollary is immediate from this and Proposition 6.2.4. □

CHAPTER 7

Completeness

Students may wish to remind themselves of the Prelims course M2: Analysis I, which covered some of the topics of this section in the specific case of the real numbers. Much of the theory in a general metric space is a natural generalisation of what was done there.

7.1. Basic definitions and examples

DEFINITION 7.1.1. Let $(x_n)_{n=1}^{\infty}$ be a sequence in some metric space X. Then we say that this sequence is

- *Bounded* if the set $\{x_n : n \geqslant 1\}$ is bounded in the sense of Definition 2.6.2, that is to say if all the x_n lie in some ball $B(a, R)$;
- *Cauchy* if the x_n become arbitrarily close together as $n \to \infty$, in the following sense: for every $\varepsilon > 0$, there is some N such that $d(x_n, x_m) < \varepsilon$ whenever $n, m \geqslant N$.
- *Convergent* if there is some $a \in X$ such that $\lim_{n\to\infty} x_n = a$.

If a sequence σ has any one of these properties, then any subsequence of σ also has the property. We leave the proof of this as an exercise.

The relation between the above concepts is as follows.

PROPOSITION 7.1.2. *A convergent sequence is Cauchy. A Cauchy sequence is bounded. Neither of the reverse implications holds in general.*

Proof. We begin by showing that the reverse implications do not hold, since the examples we will give serve to illustrate the concepts. Take $X = (0, 1]$. Then the sequence $x_n = 1/n$ is Cauchy, but not convergent. The sequence in which $x_n = 1$ for n odd and $x_n = 1/2$ for n even is bounded, but it is not Cauchy since there is no N such that $d(x_n, x_{n+1}) < 1/2$ for all $n \geqslant N$.

Now we show the two main implications. Suppose that $(x_n)_{n=1}^{\infty}$ is convergent, and that $\lim_{n\to\infty} x_n = a$. Let $\varepsilon > 0$. By the definition of limit, there is some N such that, if $n \geqslant N$, $d(x_n, a) < \varepsilon/2$. Now suppose that $m, n \geqslant N$. Then

$$d(x_m, x_n) \leqslant d(x_n, a) + d(x_m, a) < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

and so $(x_n)_{n=1}^{\infty}$ is Cauchy.

Now suppose that $(x_n)_{n=1}^{\infty}$ is Cauchy. Taking $\varepsilon = 1$ in the definition, we see that there is some N such that $d(x_m, x_n) < 1$ whenever $m, n \geqslant N$. In particular, all points of the sequence except (possibly) $x_1, \ldots, x_{N-1}$ lie in $B(x_N, 1)$. It follows that *all* points of the sequence lie in $B(x_N, R)$, where R is the largest value of the set $\{d(x_N, x_1) + 1, \ldots, d(x_N, x_{N-1}) + 1\}$, and so $(x_n)_{n=1}^{\infty}$ is bounded. □

Now we turn to the key definition of the chapter.

Definition 7.1.3 (Completeness). A metric space is said to be *complete* if every Cauchy sequence converges.

One of the main results of the Prelims course was that $\mathbf{R}$ is complete, and it is easy to deduce from this that $\mathbf{R}^n$ is complete also (since a sequence in $\mathbf{R}^n$ converges if and only if each of its coordinates converge).

On the other hand, we observed above that $(0, 1]$ is not complete. For much the same reason, $(0, 1)$ is not complete. Note, however, that $(0, 1)$ is homeomorphic to $\mathbf{R}$, as we showed earlier. Therefore the notion of completeness is not (necessarily) preserved under homeomorphisms.

Let V be a normed vector space with norm $\| \cdot \|$. As previously discussed, we can define a metric on V by $d(v, w) = \|v - w\|$. We say that V is complete if, when endowed with the structure of a metric space in this way, it is complete. That is, when we talk about completeness of normed spaces we implicitly assume that the obvious metric has been put on V, without necessarily mentioning it explicitly.

7.2. First properties of complete metric spaces

In this section we collect a couple of basic properties of complete metric spaces.

Lemma 7.2.1. *A subspace of a complete metric space is complete if and only if it is closed.*

Proof. Let X be a complete metric space and let $Y \subseteq X$. Suppose first that Y is closed; we will show that it is complete. Let $(y_n)_{n=1}^{\infty}$ be a Cauchy sequence in Y. Then it is also a Cauchy sequence in X. Since X is complete, it converges, say $\lim_{n\to\infty} y_n = a$. By Corollary 6.1.5, $a \in Y$.

In the other direction, suppose that Y is complete. Let $(y_n)_{n=1}^{\infty}$ be a sequence of elements of Y with $\lim_{n\to\infty} y_n = a$ for some $a \in X$. Then $(y_n)_{n=1}^{\infty}$ is certainly a Cauchy sequence in Y, and so by completeness of Y it converges to an element of Y. So by uniqueness of limits we have $a \in Y$. Thus Y contains the limits of such sequences, and by Corollary 6.1.5, Y is closed.

□

Definition 7.2.2. Let X be a metric space and $Y \subseteq X$ a non-empty subset. The *diameter* of Y, written $\operatorname{diam}(Y)$, is defined to be the supremum of the set $\{d(x,y) : x, y \in Y\}$ when this set is bounded, and infinity otherwise.

The next lemma is sometimes known as Cantor's intersection theorem.

Lemma 7.2.3. *Let X be a complete metric space and suppose that $S_1 \supseteq S_2 \supseteq \dots$ form a nested sequence of non-empty closed sets in X with the property that $\operatorname{diam}(S_n) \to 0$ as $n \to \infty$. Then $\bigcap_{n=1}^{\infty} S_n$ contains a unique point a.*

Proof. For each n, pick $x_n \in S_n$. We claim that $(x_n)_{n=1}^{\infty}$ is Cauchy. To see this, let $\varepsilon > 0$, and suppose that N is large enough that $\operatorname{diam}(S_N) < \varepsilon$. If $n, m \geqslant N$ then, since the S_i are nested, $x_n, x_m \in S_N$. By the definition of diameter, $d(x_n, x_m) \leqslant \operatorname{diam}(S_N) < \varepsilon$.

Since X is complete, we have $\lim_{n\to\infty} x_n = a$ for some a. For each i, the nesting property of the sets S_i implies that we have $x_n \in S_i$ for all $n \geqslant i$. Therefore, since S_i is closed, Corollary 6.1.5 tells us that $a \in S_i$. Since this is true for all i, we have $a \in \bigcap_{i=1}^{\infty} S_i$.

To show that a is unique, suppose that $b \in \bigcap_{i=1}^{\infty} S_i$. Then $d(a,b) \leqslant \operatorname{diam}(S_i)$ for all i. Since $\operatorname{diam}(S_i) \to 0$, we have $d(a,b) = 0$ and so $a = b$. □

What if we drop the condition $\operatorname{diam}(S_i) \to 0$? We certainly could not expect a to be unique since, for instance, we could take all the S_i to be the whole space X. Somewhat surprisingly at first sight, the intersection $\bigcap_{i=1}^{\infty} S_i$ may even be empty. For instance, take $S_i = [i, \infty) \subset \mathbf{R}$.

7.3. Completeness of function spaces

In this section we show that two natural spaces of functions give rise to complete metric spaces.

For the first result, recall that if X is a set then $B(X)$ denotes the normed vector space of bounded functions $f : X \to \mathbf{R}$, with norm $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

Theorem 7.3.1. *Let X be any set. Then $B(X)$ is complete.*

Proof. Let $(f_n)_{n=1}^{\infty}$ be a Cauchy sequence in $B(X)$. Then for each x the sequence $(f_n(x))_{n=1}^{\infty}$ is a Cauchy sequence of real numbers (convincing yourself of this is a good exercise to check you have understood the definitions). Since $\mathbf{R}$ is complete, each such sequence has a limit, and we write $f(x)$ for this limit. That is, $\lim_{n\to\infty} f_n(x) = f(x)$.

We claim that f is a bounded function. To see this, take $\varepsilon = 1$ in the definition of Cauchy sequence. This gives an N such that, if $n, m \geqslant N$, $\sup_x |f_n(x) - f_m(x)| \leqslant 1$. In particular, $|f_N(x) - f_n(x)| \leqslant 1$ for all $n \geqslant N$ and for all $x \in X$. Taking the limit as $n \to \infty$, it follows that $|f_N(x) - f(x)| \leqslant 1$ for all x. Since f_N is a bounded function, so is f.

Finally, we need to show that $f_n \to f$ in the norm $\| \cdot \|_\infty$ (at the moment we have only shown pointwise convergence). The argument is a simple modification of the preceding one. Let $\varepsilon > 0$, and let N be such that, if $n, m \geqslant N$, $|f_n(x) - f_m(x)| \leqslant \varepsilon$ for all $x \in X$. For each fixed $n \geqslant N$ and $x \in X$, we may let $m \to \infty$, obtaining that $|f_n(x) - f(x)| \leqslant \varepsilon$. That is, for all $n \geqslant N$ we have $\|f_n - f\|_\infty \leqslant \varepsilon$. It follows that $f_n \to f$ in the $\| \cdot \|_\infty$-norm. □

For the second result, recall that if X is a metric space then $C_b(X)$ denotes the normed vector space of bounded continuous functions $f : X \to \mathbf{R}$, again with norm $\|f\|_\infty = \sup_{x \in X} |f(x)|$.

Theorem 7.3.2. *Let X be a metric space. Then $C_b(X)$ is complete.*

Proof. We have shown in Theorem 7.3.1 that $B(X)$ is complete, so by Lemma 7.2.1 it is enough to show that $C_b(X)$ is a closed subset of $B(X)$.

By Corollary 6.1.5, it suffices to show that if $(f_n)_{n=1}^\infty$ is a sequence of elements of $C_b(X)$ converging in the $\| \cdot \|_\infty$-norm to some $f \in B(X)$, then $f \in C_b(X)$, or in other words f is continuous.

Let $a \in X$, and let $\varepsilon > 0$. Since $f_n \to f$ in the $\| \cdot \|_\infty$-norm, there is some n such that $\|f_n - f\|_\infty \leqslant \varepsilon/3$. Since f_n is continuous, there is a $\delta > 0$ such that $|f_n(x) - f_n(a)| < \varepsilon/3$ for all $x \in B(a, \delta)$. But then for $x \in B(a, \delta)$ we have

$$\begin{aligned} |f(x) - f(a)| &\leqslant |f(x) - f_n(x)| + |f_n(x) - f_n(a)| + |f_n(a) - f(a)| \\ &< \varepsilon/3 + \varepsilon/3 + \varepsilon/3 = \varepsilon. \end{aligned}$$

It follows that f is continuous at a, and since a was arbitrary, f is a continuous function on X. □

Remark. You may have the impression that you have seen something like this argument before, and indeed that is the case. In Prelims: Analysis II you saw that a uniform limit of continuous functions on $\mathbf{R}$ is continuous, and our task here was essentially the same, but in the setting of a general metric space.

7.4. The contraction mapping theorem

The final topic of this section is a classic theorem about fixed points of certain maps from a metric space to itself. We will discuss the result for its own intrinsic interest, but it has important applications to the solutions of differential equations, as you will see in the course A1 : Differential Equations.

Let us begin with a couple of definitions.

Definition 7.4.1. Let (X, d_X) and (Y, d_Y) be metric spaces and suppose that $f \colon X \to Y$. We say that f is a *Lipschitz* map (or is *Lipschitz continuous*) if there is a constant $K \geqslant 0$

such that

$$d_Y(f(x), f(y)) \leqslant K d_X(x, y).$$

If $Y = X$ and $K \in [0, 1)$ then we say that f is a *contraction mapping* (or simply a *contraction*).

An easy exercise is to check that every Lipschitz map is continuous, and to give an example of a continuous map between metric spaces which is *not* Lipschitz.

It is very important to note, in the definition of a contraction, that it says something stronger than that $d(f(x), f(y)) < d(x, y)$, namely that there is a constant $K < 1$ such that $d(f(x), f(y)) \leqslant K d(x, y)$ for all x, y.

Theorem 7.4.2 (Contraction mapping theorem). *Let X be a nonempty complete metric space and suppose that $f \colon X \to X$ is a contraction. Then f has a unique fixed point, that is, there is a unique $x \in X$ such that $f(x) = x$.*

Proof. We begin by showing that there cannot be two fixed points. Suppose that $f(x_1) = x_1$ and that $f(x_2) = x_2$. Then we have

$$d(x_1, x_2) = d(f(x_1), f(x_2)) \leqslant K d(x_1, x_2).$$

Since $d(x_1, x_2) \geqslant 0$ and $K < 1$, we are forced to conclude that $d(x_1, x_2) = 0$ and hence that $x_1 = x_2$.

Now we show that there is a fixed point. The proof is constructive (and may be used in practical situations to find fixed points numerically). The idea is as follows. Pick an arbitrary $x_0 \in X$, and form the sequence of iterates $x_1 := f(x_0)$, $x_2 := f(x_1)$, and so on. We claim that (no matter which x_0 we started with) the sequence $(x_n)_{n=1}^{\infty}$ converges to some limit x, and that $f(x) = x$.

To show that $(x_n)_{n=1}^{\infty}$ converges, it suffices to show that it is Cauchy, since X is complete. To do this, first observe that by repeated use of the contraction property and the definition of the sequence $(x_n)_{n=1}^{\infty}$ we have

$$d(x_n, x_{n-1}) \leqslant K d(x_{n-1}, x_{n-2}) \leqslant K^2 d(x_{n-2}, x_{n-3}) \leqslant \dots \leqslant K^{n-1} d(x_0, x_1)$$

(you could prove this formally by induction if you wanted). Therefore if $n > m$ we have

$$\begin{aligned} d(x_n, x_m) &\leqslant d(x_n, x_{n-1}) + \cdots + d(x_{m+1}, x_m) \\ &\leqslant (K^{n-1} + K^{n-2} + \cdots + K^m) d(x_0, x_1) \\ &\leqslant K^m (1 + K + K^2 + \dots) d(x_0, x_1) = C K^m, \end{aligned}$$

where $C = d(x_0, x_1)/(1 - K)$ (by summing the geometric series).

It follows that if $n, m \geqslant N$ then $d(x_m, x_n) \leqslant C K^N$.

Since $K < 1$, for any $\varepsilon > 0$ there is some N such that $CK^N < \varepsilon$, and therefore $(x_n)_{n=1}^{\infty}$ is indeed a Cauchy sequence.

Since X is complete, $x_n \to x$ for some $x \in X$. To complete the proof we must show that $f(x) = x$. This is quite straightforward. Indeed, since f is continuous we have

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

This finishes the proof. □

Remarks. Over the years, many people have lost a mark in exam questions for forgetting that X must be non-empty.

Let us conclude by giving a couple of examples to show that the hypotheses of the theorem are necessary. First, we observe that the weaker condition that $d(f(x), f(y)) < d(x, y)$ for all $x \neq y$ is *not* sufficient. For instance, it may be checked that the function $f\colon [1,\infty) \to [1,\infty)$ defined by $f(x) = x + 1/x$ has this property, but it obviously has no fixed points.

More obviously, the requirement that X is complete is important. For instance, if we define $f\colon (0,1) \to (0,1)$ by $f(x) = x/2$ then clearly f is a contraction, but f has no fixed points in $(0,1)$.

CHAPTER 8

Connectedness and path-connectedness

In this section we try to understand what makes a space "connected". We will consider two natural approaches to this question, and show that for reasonably nice spaces the two notions in fact coincide. In particular, the two notions of connectedness coincide for open subsets of the complex plane.

8.1. Connectedness

The concept of *connectedness* formulates the intuitive idea of a space which cannot be split into two "separated" pieces.

DEFINITION 8.1.1. We say that a metric space is *disconnected* if we can write it as the disjoint union of two nonempty open sets. We say that a space is *connected* if it is not disconnected.

If X is written as a disjoint union of two nonempty open sets U and V then we say that these sets *disconnect* X.

If $X = [0,1] \cup [2,3] \subset \mathbf{R}$ then we have seen that both $[0,1]$ and $[2,3]$ are open in X. Since X is their disjoint union, X is disconnected.

It is a little harder to give a nontrivial example of a connected space. Later on, we will show that all intervals in $\mathbf{R}$ are connected.

The following lemma gives some equivalent ways to formulate the concept of a connected space.

LEMMA 8.1.2. *Let X be a metric space. Then the following are equivalent.*

(i) *X is connected.*

(ii) *If $f\colon X \to \{0,1\}$ is a continuous function then f is constant.*

(iii) *The only subsets of X which are both open and closed are X and $\emptyset$.*

(*Here the set $\{0,1\}$ is viewed as a metric space via its embedding in $\mathbf{R}$, or equivalently with the discrete metric.*)

Proof. (i) $\Rightarrow$ (ii): Let X be connected, and let $f\colon X \to \{0,1\}$ be a continuous function. The singleton sets $\{0\}$ and $\{1\}$ are both open in $\{0,1\}$ and so both $f^{-1}(0)$ and $f^{-1}(1)$ are open subsets of X. They are clearly disjoint, and their union is X. Therefore one of them must be empty, which means that f is constant.

(ii) $\Rightarrow$ (iii): Suppose that $A \subseteq X$ is both open and closed. Then A^c is open (and closed), and so the function $f : X \to \{0, 1\}$ defined by $f(x) = 1$ for $x \in A$ and $f(x) = 0$ for $x \in A^c$ (that is, the characteristic function of A) is continuous. Assuming (ii), it must be constant. If it takes the constant value 1, then $A = X$. If it takes the constant value 0, then $A = \emptyset$.

(iii) $\Rightarrow$ (i): Suppose that $X = U \cup V$ with U, V open and disjoint. Then $U^c = V$ is open, so U is also closed. Thus U is both open and closed, and hence (assuming (iii)) is either X or $\emptyset$. Similarly for V. Hence there is no way to disconnect $f(X)$. □

Frequently one has a metric space X and a subset Y of it whose connectedness or otherwise one wishes to ascertain. To this end, it is useful to record the following lemma.

Lemma 8.1.3. *Let X be a metric space, and let $Y \subseteq X$ be a subset, considered as a metric space with the metric induced from X. Then Y is connected if and only if the following is true. If U, V are open subsets of X, and $U \cap V \cap Y = \emptyset$, then whenever $Y \subseteq U \cup V$, either $Y \subseteq U$ or $Y \subseteq V$.*

Proof. The key point here is to recall that the open sets in Y are precisely the sets of the form $U \cap Y$, where U is open in X. This was proven in Lemma 5.5.1. Take a pair $U \cap Y$, $V \cap Y$ of such open sets. They disconnect Y if and only if

(i) They are disjoint, thus $U \cap V \cap Y = \emptyset$;
(ii) They cover Y, which is equivalent to $Y \subseteq U \cup V$;
(iii) Neither is empty.

Thus Y is connected if and only if (i) and (ii) imply that one of $U \cap Y$, $V \cap Y$ is empty or equivalently that $Y \subseteq V$ or $Y \subseteq U$. □

We now turn to some basic properties of the notion of connectedness. These broadly conform with one's intuition about how connected sets should behave, but of course proof is required in each case.

Lemma 8.1.4 (Sunflower lemma). *Let X be a metric space. Let $\{A_i : i \in I\}$ be a collection of connected subsets of X such that $\bigcap_{i \in I} A_i \neq \emptyset$. Then $\bigcup_{i \in I} A_i$ is connected.*

Proof. We use the alternative characterisation of connectedness given in Lemma 8.1.2 (ii). Suppose that $f \colon \bigcup_{i \in I} A_i \to \{0, 1\}$ is continuous. We must show that f is constant. Pick $x_0 \in \bigcap_{i \in I} A_i$. Then if $x \in \bigcup_{i \in I} A_i$ there is some i for which $x \in A_i$. But then the restriction of f to A_i is constant since A_i is connected, so that $f(x) = f(x_0)$ as $x, x_0 \in A_i$. But since x was arbitrary, it follows that f is constant as required. □

Lemma 8.1.5 (Connectedness and closures). *Let X be a metric space. If $A \subseteq X$ is connected then if B is such that $A \subseteq B \subseteq \bar{A}$, the set B is also connected.*

Proof. We use the criterion for a subspace to be connected from Lemma 8.1.3. Suppose that $B \subseteq U \cup V$ where U and V are open in X and $U \cap V \cap B = \emptyset$. Then certainly $A \subseteq U \cup V$ and $A \cap U \cap V = \emptyset$. Hence, since A is connected, either $A \subseteq U$ or $A \subseteq V$. Without loss of generality, $A \subseteq U$, and since $A \cap U \cap V = \emptyset$ this means that $A \subseteq V^c$. However, V^c is closed and so taking closures we obtain $\bar{A} \subseteq \bar{V^c} = V^c$. In particular $B \subseteq V^c$ and so, since $B \subseteq U \cup V$, we must have $B \subseteq U$. We have verified the criterion (Lemma 8.1.3) for a subspace to be connected. □

Lemma 8.1.6 (Connected image of a connected set). *Let X be a connected metric space, and let $f : X \to Y$ be continuous. Then $f(X)$ is connected.*

Proof. We may as well suppose that f is surjective (otherwise replace Y by $f(X)$). Suppose that U and V are disjoint open subsets of Y with $U \cup V = Y$. Then $f^{-1}(U)$ and $f^{-1}(V)$ are disjoint open subsets of X with $f^{-1}(U) \cup f^{-1}(V) = X$. Since X is connected one of them, say $f^{-1}(U)$, is empty. Therefore U is empty.

It follows that there is no way to disconnect Y. □

A simple corollary is that (unlike completeness) the property of connectedness is preserved under homeomorphisms.

Connected components. A consequence of the Sunflower Lemma is that, for each $x \in X$, there is a unique maximal connected subset of X containing x, which contains all other such sets (take the union of all connected subsets of X containing x). This is called the *connected component* of X containing x.

Proposition 8.1.7 (Connected components). *The connected components of a metric space partition the space. A space is connected if and only if it has a unique connected component.*

Proof. Let X be the space, and for $x \in X$ write $\Gamma(x)$ for the connected component containing x. Suppose that $\Gamma(x)$ and $\Gamma(y)$ are not disjoint, say $a \in \Gamma(x) \cap \Gamma(y)$. We wish to show that they coincide, which is what it means for them to partition the space. By the Sunflower Lemma, $\Gamma(x) \cup \Gamma(y)$ is connected. By the definition of connected component, $\Gamma(x)$ must contain this set, which of course means that $\Gamma(y) \subseteq \Gamma(x)$. Similarly $\Gamma(x) \subseteq \Gamma(y)$, and so $\Gamma(x) = \Gamma(y)$.

The second statement is obvious. □

As promised we now show that intervals in $\mathbf{R}$ are connected.

Theorem 8.1.8. *Any interval $[x, y]$ in $\mathbf{R}$ is connected.*

Proof. We show that $[x, y]$ is connected using Lemma 8.1.3. Suppose $[x, y] \subseteq U \cup V$ where U and V are open subsets of $\mathbf{R}$ with $[x, y] \cap U \cap V = \emptyset$, but that we do not have $[x, y] \subseteq U$ or $[x, y] \subseteq V$. Thus $[x, y] \cap U$ and $[x, y] \cap V$ are both non-empty. Without loss of generality we may assume $x \in [x, y] \cap U$ and $y \in [x, y] \cap V$. (The latter is not so obvious: as $[x, y] \cap V \neq \emptyset$, there exists $y' \in [x, y]$ with $y' \in V$. Note $y' \neq x \in U$. We *replace* $[x, y]$ by the possibly smaller interval $[x, y']$: Then still $[x, y'] \subseteq U \cup V$, with $[x, y'] \cap U \cap V = \emptyset$, and moreover $[x, y'] \nsubseteq U$ and $[x, y'] \nsubseteq V$, as $x \in U$ and $y' \in V$. Note we shall reach a contradiction, which shows our assumption on the *original* interval $[x, y]$ was false.) Note that as $[x, y] \cap U \cap V = \emptyset$ this tells us $y \notin U$ and $x \notin V$.

Now define $S = \{z \in [x, y] : z \in U\}$. Then S is non-empty and bounded and so $c = \sup(S)$ exists. Note $c \in [x, y]$. Since $[x, y] \subseteq U \cup V$, we have either $c \in U$ or $c \in V$.

Assume $c \in U$. Since $y \notin U$ we have $c \neq y$. So as U is open, there is some interval $[c, c + \varepsilon)$ contained in U and also in $[x, y]$. This means that $[c, c + \varepsilon) \subseteq S$, which contradicts the fact that $c = \sup(S)$ (for instance, $c + \varepsilon/2$ lies in S and is bigger than c).

If $c \in V$ then $c \neq x$ and so, since V is open, there is some interval $(c - \varepsilon, c]$ contained in V and also in $[x, y]$. In particular, $[c - \varepsilon/2, c]$ is disjoint from S, which contradicts the fact that $c = \sup(S)$ (for instance, $c - \varepsilon/2$ as an upper bound for S, and is smaller than c).

These two contradictions show that we were wrong to assume that neither $[x, y] \subseteq U$ or $[x, y] \subseteq V$. Therefore $[x, y]$ is connected. □

To finish this section, let us remark that the intermediate value theorem is an almost immediate consequence of Theorem 8.1.8 and Lemma 8.1.6. Indeed, suppose $f : [a, b] \to \mathbf{R}$ is continuous. Then, since $[a, b]$ is connected, $f([a, b])$ is connected. Therefore this latter set is an interval and in particular it contains every c with between $f(a)$ and $f(b)$.

8.2. Path-connectedness

We now turn to a different, but equally intuitive, notion of what it means for a set to be connected: that one should be able to “continuously move” from any point to another. Here is the precise definition.

Definition 8.2.1 (Path connectedness). Let X be a metric space. Then we say that X is path-connected if the following is true: for any $a, b \in X$ there is a continuous map $\gamma : [0, 1] \to X$ with $\gamma(0) = a$ and $\gamma(1) = b$.

A continuous map $\gamma : [0, 1] \to X$ is called a *path*. To develop the basic theory of path-connectedness, we introduce a couple of simple operations on paths.

Given two paths γ_1, γ_2 in X such that $\gamma_1(1) = \gamma_2(0)$ we can form the *concatenation* $\gamma_1 \star \gamma_2$ of the two paths to be the path

$$\gamma_1 \star \gamma_2(t) = \begin{cases} \gamma_1(2t), & 0 \leqslant t \leqslant 1/2 \\ \gamma_2(2t-1), & 1/2 \leqslant t \leqslant 1. \end{cases}$$

We leave it as an easy exercise to show carefully that $\gamma_1 \star \gamma_2$ is continuous, and hence really is a path.

If $\gamma\colon [0,1] \to X$ is a path, then the *opposite* path γ^- is defined by $\gamma^-(t) = \gamma(1-t)$.

Lemma 8.2.2. *Let X be a metric space. Define a relation $\sim$ on X as follows: $a \sim b$ if and only if there is a path $\gamma : [0,1] \to X$ with $\gamma(0) = a$ and $\gamma(1) = b$. Then $\sim$ is an equivalence relation.*

Proof. To show that $a \sim a$, use the path γ which takes the constant value a. To show that $a \sim b$ implies $b \sim a$, take a path γ from a to b and consider its opposite path γ^-. Finally, to show transitivity, use the join of two paths. □

The equivalence classes into which this relation partitions X are called the *path-components* of X.

8.3. Connectedness and path-connectedness

In the final part of this chapter, we explore the link between connectedness and path-connectedness. The key points to be covered are as follows:

- Path-connectedness implies connectedness;
- Connectedness does not imply path-connectedness in general, but it does in normed vector spaces.

Theorem 8.3.1. *A path-connected metric space is connected.*

Proof. Suppose that X is path-connected, and let $f : X \to \{0,1\}$. We claim that f is constant, which is enough to establish connectedness of X by Lemma 8.1.2 (ii). Let $a, b \in X$. Since X is path-connected, there is a path $\gamma\colon [0,1] \to X$ such that $\gamma(0) = a$ and $\gamma(1) = b$. Consider the composition $f \circ \gamma$. This is a continuous function from $[0,1]$ to $\{0,1\}$ and hence, since $[0,1]$ is connected by Theorem 8.1.8, it is constant. Therefore $f(a) = (f \circ \gamma)(0) = (f \circ \gamma)(1) = f(b)$. Since a and b were arbitrary, this implies that f is indeed constant. □

Theorem 8.3.2. *A connected open subset of a normed space is path-connected.*

Proof. Write X for the connected open set. The key observation is that any path-component of X is open. To see this, suppose that P is a path-component of X, and let $a \in P$. Since

X is open, there is a ball $B(a, \varepsilon)$ contained in X. Let b be a point in this ball. We can now write down an explicit path γ between a and b, namely $\gamma(t) = (1-t)a + tb$. This is easily seen to be continuous, and its image is contained in $B(a, \varepsilon)$ since

$$\|\gamma(t) - a\| = t\|a - b\| \leqslant \|a - b\| = d(a, b) < \varepsilon$$

for all t. Therefore b lies in the same path-component P.

With this observation in place, the theorem follows easily. Indeed, the path-components partition X, and so if there was more than one of them we could write X as a disjoint union of non-empty open sets, contrary to the assumption that X is connected. □

We now return to the topic of path-connectedness, and give a famous example.

Theorem 8.3.3. *There is a connected subset of* $\mathbf{R}^2$ *which is not path-connected.*

Proof. (Non-examinable) There is a classic example, known as the *Topologist's sine-curve.* This is the set $A \subseteq \mathbf{R}^2$ given by

$$\{(0, y) : -1 \leqslant y \leqslant 1\} \cup \{(x, \sin(1/x) : x \in (0, 1]\}.$$

Why is A connected? It is quite easy to convince oneself that $A = \bar{E}$, where $E = \{(x, \sin(1/x) : x \in (0, 1]\}$. However, E is connected, being the image of the connected set $(0, 1]$ under a continuous map, and so the connectedness of A is immediate from Lemma 8.1.5.

Why is A not path-connected? It is "intuitively clear" that there is no path $\gamma : [0, 1] \to A$ with $\gamma(0) = (0, 0)$ and $\gamma(1) = (1, \sin(1))$, but we must prove this. Suppose we have such a path γ. Write ℓ for the vertical line $\{0\} \times [-1, 1]$, thus $A = E \cup \ell$. Since ℓ is closed in A, $\gamma^{-1}(\ell)$ is closed, and in particular contains its supremum t. Thus $\gamma(t) \in \ell$, whilst $\gamma(u) \in E$ for all $u > t$.

Let $p_Y : \mathbf{R}^2 \to \mathbf{R}$ be projection onto the y-coordinate, i.e. $p_Y(x, y) = y$. Since p_Y is continuous, so is the composition $p_Y \circ \gamma : [0, 1] \to \mathbf{R}$. Thus there is some $\delta > 0$ such that

$$|p_Y(\gamma(u_1)) - p_Y(\gamma(u_2))| \leqslant 1 \text{ for all } u_1, u_2 \in [t, t + \delta]. \tag{8.1}$$

Now let p_X be projection onto the x-coordinate, i.e. $p_X(x, y) = x$. The composition $p_X \circ \gamma$ is continuous, and so by the intermediate value theorem and the fact that $p_X(\gamma(t + \delta)) > 0$, $(p_X \circ \gamma)[t, t + \delta]$ contains some interval $[0, c]$, $c > 0$.

However, as x ranges over $(0, c]$, $\sin(1/x)$ takes all values in $[-1, 1]$ (infinitely often), so there are $u_1, u_2 \in [t, t + \delta]$ such that $p_Y(\gamma(u_1)) = 1$, $p_Y(\gamma(u_2)) = -1$. This contradicts (8.1). □

CHAPTER 9

Sequential compactness

In this chapter (and in Chapter 9) we will be talking a lot about sequences and subsequences, so let us be clear about what these concepts are. If X is some space, let $\sigma = (x_n)_{n=1}^{\infty} = (x_1, x_2, \dots)$ be a sequence of elements of X. Any sequence of the form $\sigma' = (x_{n_k})_{k=1}^{\infty}$, where $n_1 < n_2 < n_3 < \dots$, is called a subsequence of σ. For instance, $(x_1, x_4, x_9, x_{16}, \dots)$ is a subsequence of $(x_1, x_2, x_3, x_4, \dots)$.

9.1. Definitions

In this chapter we study metric spaces which satisfy the metric-space analogue of the Bolzano-Weierstrass property. Recall what the Bolzano-Weierstrass property of **R** is: any bounded sequence has a convergent subsequence. More precisely, if $(x_n)_{n=1}^{\infty}$ is a sequence of elements in some closed bounded interval $[a, b]$, there is a subsequence of the x_n which converges to some $c \in [a, b]$.

There is an obvious way to generalise this notion to subsets of metric spaces, and the resulting notion is called sequential compactness.

Definition 9.1.1 (Sequential compactness). Let X be a metric space. Then X is said to be *sequentially compact* if any sequence of elements in X has a convergent subsequence.

Example 9.1.2. The closed interval $[0, 1]$ is sequentially compact, by the Bolzano-Weierstrass theorem.

The open interval $(0, 1)$ is not sequentially compact. For instance, the sequence $x_n = 1/n$ has no convergent subsequence in this space.

The set of rational numbers in $[0, 1]$ is not sequentially compact – for instance, the sequence $0.1, 0.14, 0.141, 0.1415, \dots$ consisting of decimal approximations to $\pi - 3$ has no convergent subsequence.

Finally, the real line **R** is not sequentially compact. For instance, the sequence $x_n = n$ has no convergent subsequence in this space.

9.2. Closure and boundedness properties

In this section we prove a couple of basic lemmas about sequentially compact spaces.

Lemma 9.2.1. *A sequentially compact subspace of a metric space is closed and bounded.*

Proof. Let X be the space and Y the sequentially compact subspace.

Suppose first that Y is not closed. Then $\bar{Y} \setminus Y$ is nonempty. Let a be a point in this set. By Lemma 6.1.5, there is a sequence $(y_n)_{n=1}^{\infty}$ of elements of Y with $\lim_{n\to\infty} y_n = a$. Then any subsequence of $(y_n)_{n=1}^{\infty}$ converges to a and hence, by the uniqueness of limits, does not converge to an element of Y. Therefore Y cannot be sequentially compact.

Suppose next that Y is not bounded. Pick an arbitrary point $y_0 \in Y$, and pick a sequence $(y_n)_{n=1}^{\infty}$ such that $d(y_0, y_n) \geqslant n$ for all n. Suppose there is a subsequence $(y_{n_k})_{k=1}^{\infty}$ converging to b. Then for k sufficiently large we have $d(y_{n_k}, b) < 1$, which implies that

$$d(y_0, b) \geqslant d(y_0, y_{n_k}) - d(y_{n_k}, b) \geqslant n_k - 1.$$

Since $n_k \to \infty$ as $k \to \infty$, whilst $d(y_0, b)$ is a fixed finite quantity, this is a contradiction. □

The converse is not true – for instance, take $X = Y = (0, 1)$ (noting that Y is closed *as a subset of* X).

Lemma 9.2.2. *A closed subset of a sequentially compact metric space is sequentially compact.*

Proof. Let X be the space and Y the closed subspace. Consider a sequence $(y_n)_{n=1}^{\infty}$ of elements of Y. It is also a sequence of elements of X and so, by sequential compactness of X, has a subsequence converging to some $a \in X$. However, Y is closed, so the limit of any convergent sequence of elements of Y lies in Y. In particular, $a \in Y$. □

9.3. Continuous functions on sequentially compact spaces

Sequential compactness has some nice properties with respect to continuous maps.

Lemma 9.3.1. *The image of a sequentially compact metric space under a continuous map is sequentially compact.*

Proof. Let X be sequentially compact, and suppose that $f : X \to Y$ is continuous. Let $\sigma = (f(x_n))_{n=1}^{\infty}$ be a sequence of elements of $f(X)$. The sequence (x_n) contains a convergent subsequence (x_{n_k}) say, with $x_{n_k} \to a$ as $k \to \infty$ for some $a \in X$. But then, since f is continuous, we have $f(x_{n_k}) \to f(a)$, and so $\sigma' = (f(x_{n_k}))_{k=1}^{\infty}$ is a convergent subsequence of σ. □

As a consequence of Lemma 9.2.1, we see that continuous function f from a sequentially compact metric space X to $\mathbf{R}$ has closed and bounded image, so in particular f is bounded and attains its bounds.

Another consequence of Lemma 9.3.1 is if X and Y are homeomorphic metric spaces and if X is sequentially compact, then so is Y.

Proposition 9.3.2. *A continuous function from a sequentially compact metric space to* $\mathbf{R}$ *is uniformly continuous.*

Proof. Let X be a sequentially compact metric space, and suppose that $f : X \to \mathbf{R}$ is continuous but not uniformly continuous. Then there exists some $\varepsilon > 0$ such that for each $n \in \mathbf{N}$ we may find $a_n, b_n \in X$ such that $d(a_n, b_n) < 1/n$ but $|f(a_n) - f(b_n)| \geqslant \varepsilon$. Since X is sequentially compact, $(a_n)_{n=1}^{\infty}$ has a subsequence, $(a_{n_k})_{k=1}^{\infty}$ converging to some point ℓ. Consider the corresponding sequence $(b_{n_k})_{k=1}^{\infty}$. Since $d(a_{n_k}, b_{n_k}) \leqslant 1/n_k \to 0$, it follows that b_{n_k} also converges to ℓ as $k \to \infty$.

Relabelling (to avoid double subscripts) we may now assume we have sequences $(a_n)_{n=1}^{\infty}$, $(b_n)_{n=1}^{\infty}$ with $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = \ell$ and $|f(a_n) - f(b_n)| \geqslant \varepsilon$ for all n.

Now f is continuous at ℓ, so there is a $\delta > 0$ such that for all $x \in X$ with $d(\ell, x) < \delta$, we have $|f(\ell) - f(x)| < \varepsilon/2$. If n is sufficiently large, we have $d(\ell, a_n), d(\ell, b_n) < \delta$ and hence

$$\varepsilon \leqslant |f(a_n) - f(b_n)| \leqslant |f(a_n) - f(\ell)| + |f(\ell) - f(b_n)| < \varepsilon/2 + \varepsilon/2 < \varepsilon,$$

which is a contradiction.

We were therefore wrong to assume that f is not uniformly continuous. □

9.4. Product spaces

Recall that if (X, d_X) and (Y, d_Y) are metric spaces then their Cartesian product $X \times Y$ can be equipped with a metric $d_{X\times Y}$ by setting

$$d_{X\times Y}((x_1, y_1), (x_2, y_2)) = \sqrt{d_X(x_1, x_2)^2 + d_Y(y_1, y_2)^2}.$$

The main result of this section, Proposition 9.4.2 below, is that the product of two sequentially compact spaces is compact. Before proving this, we note an important lemma.

Lemma 9.4.1. *Let X and Y be metric spaces. A sequence $((x_n, y_n))_{n=1}^{\infty}$ in $X \times Y$ converges if and only if $(x_n)_{n=1}^{\infty}$ converges in X and $(y_n)_{n=1}^{\infty}$ converges in Y.*

Proof. The projection maps $p_X \colon X \times Y \to X$ and $p_Y \colon X \times Y \to Y$ are continuous. In fact it is easy to see that they are Lipschitz continuous with Lipschitz constant 1. It follows that if $\lim_{n\to\infty}(x_n, y_n) = (a, b)$ then

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} p_X(x_n, y_n) = p_X(a, b) = a,$$

and similarly $\lim_{n\to\infty} y_n = b$.

Conversely, if $x_n \to a$ and $y_n \to b$ then

$$d_{X\times Y}((x_n, y_n), (a, b)) = \sqrt{d_X(x_n, a)^2 + d_Y(y_n, b)^2} \to 0$$

as $n \to \infty$ and so $(x_n, y_n) \to (a, b)$ as $n \to \infty$, as required. □

Now we prove that the product of two sequentially compact spaces is compact, with apologies for using a rather unpleasant triple subscript notation in the argument.

PROPOSITION 9.4.2. *The product of two sequentially compact metric spaces is sequentially compact.*

Proof. Let $((x_n, y_n))_{n=1}^{\infty}$ be a sequence in $X \times Y$. As X is sequentially compact, the sequence $\sigma = (x_n)_{n=1}^{\infty}$ in X has a convergent subsequence $\sigma' = (x_{n_k})_{k=1}^{\infty}$, with $x_{n_k} \to a$ as $k \to \infty$. Now consider the sequence $(y_{n_k})_{k=1}^{\infty}$ in Y. Since Y is sequentially compact this in turn has a convergent subsequence $(y_{n_{k_r}})_{r=1}^{\infty}$, say $y_{n_{k_r}} \to b$ as $r \to \infty$. Let σ'' be the corresponding subsequence of xs, that is to say $\sigma'' = (x_{n_{k_r}})_{r=1}^{\infty}$. Then σ'' is a subsequence of σ', and so it converges to a.

By the previous Lemma it follows that $(x_{n_{k_r}}, y_{n_{k_r}}) \to (a, b)$ as $r \to \infty$, and so we have exhibited a convergent subsequence of $((x_n, y_n))_{n=1}^{\infty}$. Therefore $X \times Y$ is sequentially compact. □

A corollary of this is the following result, which is often called the Bolzano-Weierstrass theorem (being a generalisation of the version on $\mathbf{R}$).

COROLLARY 9.4.3 (Bolzano-Weierstrass). *Any closed and bounded subset of $\mathbf{R}^n$ is sequentially compact.*

Proof. Let $X \subseteq \mathbf{R}^n$ be the set. Since X is bounded, it is contained in some cube $[-M, M]^n$. The Bolzano-Weierstrass theorem on $\mathbf{R}$ implies that $[-M, M]$ is sequentially compact, and therefore by Proposition 9.4.2, $[-M, M]^n$ is sequentially compact. Since X is closed, it is sequentially compact by Lemma 9.2.2. □

9.5. Sequentially compact equals complete and totally bounded

As a warm-up to the main business of this section, we prove the following.

PROPOSITION 9.5.1. *A sequentially compact metric space X is complete and bounded. The converse is not true in general.*

Proof. Suppose that X is sequentially compact. We have already shown that X is bounded in Lemma 9.2.1. Let us now show that X is complete. Suppose that $(x_n)_{n=1}^{\infty}$ is a Cauchy sequence in X. Since X is sequentially compact, $(x_n)_{n=1}^{\infty}$ has a convergent subsequence $(x_{n_k})_{k=1}^{\infty}$. Suppose that $\lim_{k\to\infty} x_{n_k} = a$. We claim that in fact $\lim_{n\to\infty} x_n = a$.

Let $\varepsilon > 0$. Then, since $(x_n)_{n=1}^{\infty}$ is Cauchy, there is some N such that for all $n, m \geqslant N$ we have $d(x_n, x_m) < \varepsilon/2$. Since $\lim_{k\to\infty} x_{n_k} = a$, we may find a k such that $n_k \geqslant N$ and $d(x_{n_k}, a) < \varepsilon/2$. But then if $n \geqslant N$ we have

$$d(x_n, a) \leqslant d(x_n, x_{n_k}) + d(x_{n_k}, a) < \varepsilon/2 + \varepsilon/2 = \varepsilon,$$

as required.

To show that the converse is not true in general, consider the following example. Take $C_b(\mathbf{R})$ to be the normed space of continuous bounded functions on the real line equipped as usual with the $\|\cdot\|_\infty$-norm and the associated metric. Let $X = \bar{B}(0,1)$ (that is, functions having sup norm bounded by 1). Then X is bounded. Moreover, by Lemma 7.2.1 as X is closed and $C_b(\mathbf{R})$ is complete (Theorem 7.3.2) we have that X is complete. Define a function $\phi : \mathbf{R} \to \mathbf{R}$ by

$$\phi(t) = \begin{cases} 2t+1, & -1/2 \leqslant t \leqslant 0; \\ 1-2t, & 0 \leqslant t \leqslant 1/2 \end{cases}$$

and $\phi(t) = 0$ for $t \notin [-1/2, 1/2]$. For each $n \in \mathbf{N}$ set $f_n(t) = \phi(t-n)$ (we might call this sequence of functions a "moving bump"). All of the functions f_n lie in $X = \bar{B}(0,1)$. However, if $n \neq m$ then $f_n(n) = 1$, whilst $f_m(n) = 0$, so $\|f_n - f_m\|_\infty = 1$. Thus the sequence $(f_n)_{n=1}^{\infty}$ has no Cauchy subsequence, and hence certainly no convergent subsequence. □

Remark. The Bolzano-Weierstass theorem (Corollary 9.4.3) implies that the converse *is* true for subsets of $\mathbf{R}^n$.

It turns out that there is a stronger notion of boundedness called *total boundedness* which – together with completeness - implies sequential compactness and in fact is equivalent to it.

Definition 9.5.2. A metric space is said to be *totally bounded* if, for any $\varepsilon > 0$, it may be covered by finitely many open balls of radius ε.

Here is one of the more substantial theorems of the course.

Theorem 9.5.3. *A metric space is sequentially compact if and only if it is complete and totally bounded.*

Proof. Suppose first that we have a space X which is sequentially compact. We have already shown in Proposition 9.5.1 that X is complete. Let us now show that it is totally bounded. Suppose X is *not* totally bounded, and let ε be such that there is no way to cover X by finitely many open balls of radius ε.

Using a greedy algorithm, we select an infinite sequence $(x_n)_{n=1}^{\infty}$ of elements of X which are separated by at least ε, that is to say $d(x_i, x_j) \geqslant \varepsilon$ whenever $i \neq j$.

To do this, suppose that $x_1, \dots, x_n$ have already been selected. By assumption, the balls $B(x_i, \varepsilon)$ do not cover X, and so we may select a point $x_{n+1} \in X$ which does not lie in any of these balls, and therefore $d(x_i, x_{n+1}) \geqslant \varepsilon$ for $i = 1, \dots, n$.

It is clear that such a sequence has no convergent subsequence, and so we were wrong to assume that X is not totally bounded.

We turn now to the more substantial direction of the theorem, which is to show that a complete and totally bounded metric space X is sequentially compact. Let σ be a sequence of elements of X. We will use the total boundedness assumption for balls of radii $1, \frac{1}{2}, \frac{1}{4}, \dots$. Thus, for each nonnegative integer m there is a finite collection of open balls $B_1^{(m)}, \dots, B_{k_m}^{(m)}$ of radius 2^{-m} which cover X.

Start with the balls $B_1^{(0)}, \cdots, B_{k_0}^{(0)}$ of radius 1. One of these balls contains infinitely many elements of the sequence σ. Write B_0 for the ball with this property, and let $\sigma^{(0)}$ be the infinite subsequence of σ of elements contained in this ball.

Now look at the balls $B_1^{(1)}, \dots, B_{k_1}^{(1)}$ of radius $\frac{1}{2}$. One of *these* balls contains infinitely many elements of the new subsequence $\sigma^{(0)}$. Write B_1 for such a ball, and let $\sigma^{(1)}$ be the infinite subsequence of $\sigma^{(0)}$ of elements contained in it.

Continue in the obvious fashion, producing new subsequences $\sigma^{(2)}, \sigma^{(3)}, \dots$ with $\sigma^{(r)}$ contained in B_r and a subsequence of $\sigma^{(r-1)}$.

Now consider the sequence σ^* obtained by a diagonal argument: the ith element of σ^* is taken to be the ith element of $\sigma^{(i)}$. Clearly σ^* is a subsequence of σ and, if we write $\sigma^* = (x_n)_{n=1}^{\infty}$, we have $x_n \in B_r$ for all $n \geqslant r$.

It is now clear that σ^* is a Cauchy sequence. Indeed, given $\varepsilon > 0$, let N be such that $2^{-N} < \varepsilon/2$. If $n, m \geqslant N$ then x_n, x_m both lie in B_N, which is a ball of radius 2^{-N}, and hence $d(x_n, x_m) < \varepsilon/2 + \varepsilon/2 = \varepsilon$.

Finally, since X is complete the sequence σ^* converges. We have shown that σ, which was an arbitrary sequence in X, has a convergent subsequence, and therefore X is sequentially compact. □

Remark. Observe that the argument in fact shows that any sequence in a totally bounded metric space has a subsequence which is Cauchy. We only used completeness right at the end.

CHAPTER 10

Compactness

10.1. Open covers and the definition of compactness

In this final chapter of the metric spaces part of the course, we come to one of the most powerful and important notions in all of mathematics: compactness.

Let us start by giving the definition.

Definition 10.1.1. Let X be a metric space and $\mathcal{U} = \{U_i : i \in I\}$ a collection of open subsets of X. We say that $\mathcal{U}$ is an *open cover* of X if $X = \bigcup_{i \in I} U_i$. If $J \subseteq I$ is a subset such that $X = \bigcup_{i \in J} U_i$ then we say that $\{U_i : i \in J\}$ is a *subcover* of $\mathcal{U}$ and if $|J| < \infty$ then we say that it is a *finite subcover*.

Definition 10.1.2 (Compactness). A metric space is said to be compact if every open cover has a finite subcover.

Example 10.1.3. The real line $\mathbf{R}$ is not compact. For instance, the open cover $\bigcup_{n \in \mathbf{N}} (-n, n)$ has no finite subcover.

Motivation. It is quite hard to motivate the definition of compactness when one first sees it. Indeed, von Neumann's famous quote "... *in mathematics you don't understand things. You just get used to them*" is quite apposite. Nonetheless, a couple of comments are in order. First of all, it turns out that compactness and sequential compactness are the same concept in metric spaces. We prove this in Sections 10.2 and 10.3 below (with the second of these being non-examinable). Second, the notion of compactness looks rather natural in the context of topological spaces, since it talks about open sets in a very basic way. Whilst the notion of sequential compactness can also be formulated in topological spaces, it is somehow less basic and, in this more general situation, *not* equivalent to compactness.

Subspaces. Sometimes, we will have a metric space X and a subspace $Y \subseteq X$, and we wish to talk about whether Y is compact. In this context, by convention an open cover $\mathcal{U}$ of Y is a collection $\{U_i : i \in I\}$ of open subsets *of* X, such that $Y \subseteq \bigcup_{i \in I} U_i$. A subcollection $\{U_i : i \in J\}$ is called a subcover if $Y \subseteq \bigcup_{i \in J} U_i$.

Then Y is compact if and only if every open cover has a finite subcover. The reason this notion is the same as the previous one (which was "internal to Y", making no reference to open sets in X) is Lemma 5.5.1, which says that open sets in Y are the same thing as open sets in X intersected with Y.

It should be said that this abuse of nomenclature of using the phrase "open cover" in two slightly different ways can be a touch confusing when you first see it.

10.2. Compactness implies sequential compactness

Proposition 10.2.1. *A compact metric space is sequentially compact.*

We isolate a lemma from the proof.

Lemma 10.2.2. *Suppose that X is a compact metric space and that we have a nested sequence $S_1 \supseteq S_2 \supseteq S_3 \supseteq \cdots$ of nonempty, closed subsets of X. Then the intersection $\bigcap_{n=1}^{\infty} S_n$ is nonempty.*

Remark. You might be interested in comparing this with Lemma 7.2.3, where a similar conclusion was reached assuming that X is complete and that the diameters of S_i tend to 0.

Proof. Suppose the intersection is empty. Then the complements S_i^c (which are open sets) are an open cover of X. By compactness, there is a finite subcover. In particular, for some n the sets $S_1^c, \dots, S_n^c$ cover X. However, we have $S_1^c \subseteq S_2^c \subseteq \cdots \subseteq S_n^c$, and therefore S_n^c covers (is equal to) X. But this is a contradiction, since S_n is nonempty. □

Proof. (Proof of Proposition 10.2.1.) Let X be the space in question, and suppose that $(x_n)_{n=1}^{\infty}$ is a sequence of elements of X. We wish to find a convergent subsequence of this sequence.

For each natural number n, set $A_n := \{x_n, x_{n+1}, x_{n+2}, \dots\}$. Obviously, $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, and so $\bar{A}_1 \supseteq \bar{A}_2 \supseteq \bar{A}_3 \supseteq \cdots$. Applying Lemma 10.2.2, we see that $\bigcap_{n=1}^{\infty} \bar{A}_n$ is nonempty.

Let a be a point in this intersection. We inductively construct a subsequence $(x_{n_k})_{k=1}^{\infty}$ such that $d(x_{n_k}, a) < 1/k$ for all k; it is then clear that this subsequence converges (to a) and the proof will be complete. Suppose that $n_1, \dots, n_k$ have already been constructed. Now a lies in $\bar{A}_{n_k+1}$, that is to say the closure of the set $\{x_{n_k+1}, x_{n_k+2}, \dots\}$. In particular, there is some element of this sequence at distance less than $1/(k+1)$ from a, and we can take this to be our $x_{n_{k+1}}$. □

10.3. Sequential compactness implies compactness

The converse of Proposition 10.2.1 is also true.

Proposition 10.3.1. *A sequentially compact metric space is compact.*

As a consequence of this, Proposition 10.2.1 and Theorem 9.5.3, we have the following substantial and important theorem.

Theorem 10.3.2. *Let X be a metric space. Then the following are equivalent:*

(i) *X is compact;*
(ii) *X is sequentially compact;*
(iii) *X is complete and totally bounded.*

Lemma 9.2.1, Corollary 9.4.3, Proposition 10.2.1 and Proposition 10.3.1 also imply the following important result.

Theorem 10.3.3 (Heine-Borel). *A subset of $\mathbf{R}^n$ is compact if and only if it is closed and bounded.*

We turn now to the proof of Proposition 10.3.1, which is nonexaminable.

Proof. (Proof of Proposition 10.3.1, non-examinable.) Let X be a sequentially compact metric space. By (the easy direction of) Proposition 9.5.3, X is complete and totally bounded. For $m = 1, 2, 3, \dots$, fix some collection of balls $B_1^{(m)}, \dots, B_{k_m}^{(m)}$ of radius 2^{-m} which cover X.

Suppose we have an open cover of X by sets U_i, $i \in I$, which has no finite subcover. Then one of the balls $B_j^{(1)}$ is not covered by finitely many of the U_i; let us write B_1 for this ball.

Now consider the balls $B_j^{(2)}$ which intersect B_1. One of these is not covered by finitely many of the U_i (otherwise B_1 would be). Write B_2 for this ball.

Now consider the balls $B_j^{(3)}$ which intersect B_2, and so on.

Continuing in this fashion, we obtain a sequence $B_1, B_2, \dots$ of open balls, with B_m having radius 2^{-m}, $B_m \cap B_{m+1} \neq \emptyset$ for all m, and with none of the B_j covered by finitely many of the U_i. Let x_m be the centre of B_m. Then, since B_m and B_{m+1} intersect in some point t, we have

$$d(x_m, x_{m+1}) \leqslant d(x_m, t) + d(x_{m+1}, t) < 2^{-m} + 2^{-(m+1)} < 2 \cdot 2^{-m}.$$

By the triangle inequality and summing the geometric series, it follows that for any $n \geqslant m$ we have

$$\begin{aligned} d(x_m, x_n) &\leqslant d(x_m, x_{m+1}) + \cdots + d(x_{n-1}, x_n) \\ &< 2(2^{-m} + 2^{-(m+1)} + \cdots) = 4 \cdot 2^{-m}. \end{aligned}$$

Therefore $(x_n)_{n=1}^{\infty}$ is a Cauchy sequence. Since X is complete, we have $\lim_{n\to\infty} x_n = x$ for some $x \in X$. Since the sets U_i cover X, one of them must contain x. Let us suppose U_1 contains x. Then, since U_1 is open, some ball $B(x, \varepsilon)$ is contained in U_1.

Choose n large enough that $d(x_n, x) < \varepsilon/2$, and also that $2^{-n} < \varepsilon/2$. Recalling that B_n is the ball of radius 2^{-n} centred on x_n, it follows that $B_n \subseteq B(x, \varepsilon)$. But then $B_n \subseteq U_1$, contrary to the assumption that B_n is not covered by finitely many of the U_i.

We were wrong to assume the existence of an open cover of X with no finite subcover, and so X is indeed compact. □

CHAPTER 11

Conformal maps

11.1. The extended complex plane $\mathbf{C}_\infty$

It is a remarkable fact that it is possible to "add the point at infinity" to $\mathbf{C}$ in such a way that the resulting space $\mathbf{C}_\infty$ has pleasant analytic properties. For instance, one can extend the function $f(z) = 1/z$ to a continuous bijection on this space, by setting $f(0) = \infty$ and $f(\infty) = 0$, and one can make rigorous sense of such statements as $\infty + 1 = \infty$. The aim of this section is to study $\mathbf{C}_\infty$, which is known as the extended complex plane. Note that $\mathbf{C}_\infty$ is a very basic example of a *Riemann surface*, one of the main objects of study in the Geometry of Surfaces course in Part B.

11.1.1. Stereographic projection.

Let

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

be the unit sphere of radius 1 centred at the origin in $\mathbf{R}^3$. View the complex plane $\mathbf{C}$ as the copy of $\mathbf{R}^2$ inside $\mathbf{R}^3$ given by the plane $\{(x, y, 0) \in \mathbf{R} : x, y \in \mathbf{R}\}$. Thus $z = x + iy$ corresponds to the point $(x, y, 0)$. Let N be the "north pole" $N = (0, 0, 1)$ of $\mathbb{S}$.

We can define a bijective map $S : \mathbf{C} \to \mathbb{S} \setminus \{N\}$ as follows. To determine $S(z)$, join z to N by a straight line, and let $S(z)$ be the point where this line meets the sphere $\mathbb{S}$. This map (or more accurately its inverse) is called stereographic projection.

It is not too hard to give an explicit formula for $S(z)$.

Lemma 11.1.1. *Suppose that $z = x + iy$. Then*

$$S(z) = \left(\frac{2x}{x^2 + y^2 + 1}, \frac{2y}{x^2 + y^2 + 1}, \frac{x^2 + y^2 - 1}{x^2 + y^2 + 1}\right).$$

Proof. The general point on the line joining z and N is $t(0, 0, 1) + (1 - t)(x, y, 0)$. There is a unique value of t with $t \neq 1$ for which this point lies on the sphere, namely $t = (x^2 + y^2 - 1)/(x^2 + y^2 + 1)$, as can be easily checked. □

We remark that the same formula can be written in the alternative form

$$S(z) = \frac{1}{1 + |z|^2}\left(2\Re(z), 2\Im(z), |z|^2 - 1\right).$$

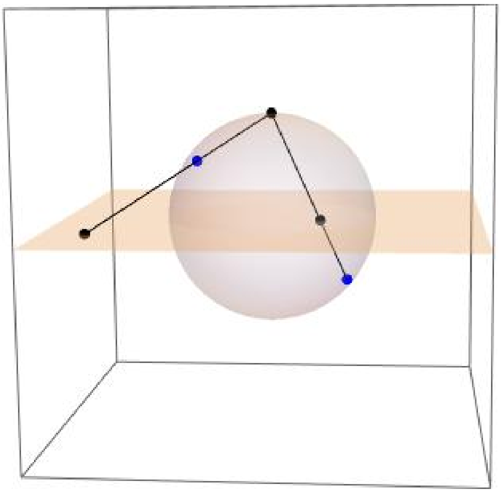

FIGURE 1. The stereographic projection map.

As we have seen, $\mathbf{C}$ may be identified with $\mathbb{S} \setminus \{N\}$ by stereographic projection. The set $\mathbb{S} \setminus \{N\}$ has a natural metric, namely the one induced from the Euclidean metric on $\mathbf{R}^3$. This induces a metric d on $\mathbf{C}$, the unique metric on $\mathbf{C}$ such that S is an isometry. To spell it out,

$$d(z,w) := \|S(z) - S(w)\|.$$

Here is a formula for this metric.

LEMMA 11.1.2. *For any $z, w \in \mathbf{C}$ we have*

$$d(z,w) = \frac{2|z-w|}{\sqrt{1+|z|^2}\sqrt{1+|w|^2}}$$

Proof. Since $\|S(z)\| = \|S(w)\| = 1$ we have $\|S(z)-S(w)\|^2 = 2-2\langle S(z), S(w)\rangle$, where $\langle , \rangle$ is the usual Euclidean inner product on $\mathbf{R}^3$. Using the formulæ (and after a little computation),

$$\langle S(z), S(w)\rangle = 1 - \frac{2|z-w|^2}{(1+|z|^2)(1+|w|^2)}.$$

Therefore

$$\|S(z) - S(w)\|^2 = \frac{4|z-w|^2}{(1+|z|^2)(1+|w|^2)}$$

as required. □

This d is not the same as the usual metric. However, it is very similar to it. For instance, on any bounded set $\{z \in \mathbf{C} : |z| \leqslant K\}$ we have

$$c_1|z-w| \leqslant d(z,w) \leqslant c_2|z-w|$$

for some $c_1, c_2 > 0$ depending on K. In fact, we could take $c_2 = 2$ and $c_1 = \frac{1}{K^2}$ for $K \geqslant 1$ (exercise). That is, d is *strongly equivalent* to the usual metric on any such set, in the sense described in Section 4.3 of the metric spaces part of the course. Therefore d is equivalent (but not strongly equivalent) to the usual metric on all of $\mathbf{C}$. Recall what this means: any ball $B(z,\varepsilon)$ in the usual metric is contained in some ball $B_d(z,\varepsilon')$ in the metric d, and vice versa.

Therefore, as remarked in the metric spaces notes, notions such as limit and continuity are the same whether we work with the usual metric or with d.

11.1.2. Adding in ∞. Now it is time to add in the point at infinity, which we will call ∞ (note this is just a symbol).

Now (exercise) as $|z| \to \infty$, $S(z) \to N$. Therefore, once we have identified $\mathbf{C}$ with $\mathbb{S}\setminus\{N\}$, it is natural to identify ∞ with N, and hence $\mathbf{C}_\infty = \mathbf{C} \cup \{\infty\}$ with the whole sphere $\mathbb{S}$. We extend the map S to a map $S : \mathbf{C}_\infty \to \mathbb{S}$ by defining $S(\infty) = N$.

Using, once again, the Euclidean metric on $\mathbb{S}$, we can extend d to a metric on $\mathbf{C}_\infty$, the unique metric for which the map S is an isometry.

Lemma 11.1.3. *For any $z \in \mathbf{C}$ we have*

$$d(z,\infty) = \frac{2}{\sqrt{1+|z|^2}}.$$

Proof. By definition, $d(z,\infty) = \|S(z) - S(\infty)\| = \|S(z) - N\|$, where N is the north pole on the sphere. We may now proceed in much the same way as before, except the calculation is easier this time. The details are left as an exercise. □

We turn now to a few examples, which show that adding ∞ to $\mathbf{C}$ in this way leads to a space with nice analytic properties.

Example 11.1.4 (Translations). Let $a \in \mathbf{C}$. Define $f : \mathbf{C}_\infty \to \mathbf{C}_\infty$ by $f(z) = z + a$ for $z \in \mathbf{C}$ and $f(\infty) = \infty$. Then f is a continuous bijection.

Proof. Clearly f is continuous with respect to the usual metric on $\mathbf{C}$. Therefore, restricted to $\mathbf{C}$, it is also continuous with respect to d, since d is equivalent to the usual metric.

It remains to check continuity at ∞. Let $\varepsilon > 0$. Now if $\delta > 0$ and if $d(z,\infty) < \delta$ then $|z| > \sqrt{\frac{4}{\delta^2} - 1}$ and so $|f(z)| > \sqrt{\frac{4}{\delta^2} - 1} - |a|$. This tends to ∞ as $\delta \to 0$, so by choosing δ small enough in terms of ε it will follows that $d(f(z),\infty) = \frac{2}{\sqrt{1+|f(z)|^2}} < \varepsilon$. □

Example 11.1.5 (Dilations). Let $b \in \mathbf{C}$ with $b \neq 0$. Define $f : \mathbf{C}_\infty \to \mathbf{C}_\infty$ by $f(z) = bz$ for $z \in \mathbf{C}$ and $f(\infty) = \infty$. Then f is a continuous bijection.

Proof. This is very similar to the argument for translations and we leave the details as an exercise. □

The final example is the most interesting one.

Example 11.1.6 (Inversion). Define $f : \mathbf{C}_\infty \to \mathbf{C}_\infty$ by $f(z) = 1/z$ for $z \in \mathbf{C} \setminus \{0\}$, $f(0) = \infty$ and $f(\infty) = 0$. Then f is a continuous bijection.

Proof. As before, the equivalence of d and the usual metric on $\mathbf{C}$ means that f is continuous except possibly at 0 and ∞.

We prove that f is continuous at 0, leaving the continuity at ∞ as an exercise (similar to Example 11.1.4).

Let $\varepsilon > 0$ be small. Then there is δ such that $\frac{2t}{\sqrt{1+t^2}} \leqslant \varepsilon$ for all $t \in [0, \delta]$. If $|z| < \delta$, then

$$d(f(z), f(0)) = d(1/z, \infty) = \frac{2}{\sqrt{1 + \frac{1}{|z|^2}}} = \frac{2|z|}{\sqrt{1+|z|^2}} \leqslant \varepsilon.$$

This indeed shows that f is continuous at 0. □

There is a nice way to analyse Example 11.1.6, by considering what f looks like under the identification of $\mathbf{C}_\infty$ with the unit sphere $\mathbb{S}$. One can easily check using Lemma 11.1.1 that if $S(z) = (s, u, v) \in \mathbb{S}$ then $S(f(z)) = (s, -u, -v)$. That is, under the identification $S : \mathbf{C}_\infty \to \mathbb{S}$, f corresponds to the (obviously continuous) map $(s, u, v) \mapsto (s, -u, -v)$, that is to say rotation by π about the x-axis.

11.1.3. Möbius maps.

In this subsection and subsequent ones we look at an important class of maps from $\mathbf{C}_\infty$ to itself, the Möbius maps.

Definition 11.1.7. The general linear group $\mathrm{GL}_2(\mathbf{C})$ consists of all nonsingular 2×2 matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $a, b, c, d \in \mathbf{C}$, with the group operation being matrix multiplication.

Each element $g \in \mathrm{GL}_2(\mathbf{C})$ gives a Möbius map $\Psi_g : \mathbf{C}_\infty \to \mathbf{C}_\infty$. Roughly, this is given by the formula

$$\Psi_g(z) := \frac{az+b}{cz+d},$$

but one needs to be careful about ∞, as follows:

- If $c \neq 0$ then we define $\Psi_g(-d/c) = \infty$ and $\Psi_g(\infty) = a/c$;
- If $c = 0$ then we define $\Psi_g(\infty) = \infty$.

We remark that two elements $g, g' \in \mathrm{GL}_2(\mathbf{C})$ give the same Möbius map if $g = \lambda g'$ for some $\lambda \neq 0$.

An important fact about Möbius maps is that composing them corresponds to multiplying the relevant matrices. That is to say, we have the following.

Proposition 11.1.8 (Composition of Möbius maps). *We have* $\Psi_{g_1 g_2} = \Psi_{g_1} \circ \Psi_{g_2}$. *That is,* $\mathrm{GL}_2(\mathbf{C})$ *acts on* $\mathbf{C}_\infty$ *via Möbius maps.*

One can prove this by a direct calculation (which we leave as an exercise).

11.1.4. Decomposing Möbius maps. Earlier in this section, we looked at translations, dilations and inversion from $\mathbf{C}_\infty$ to $\mathbf{C}_\infty$. It turns out that these are all Möbius maps, and moreover that an arbitrary Möbius map can be built from maps of these types.

That they are all Möbius maps is straightforward. Indeed,

- The translation $z \mapsto z + a$ is the Möbius map $\Psi_{T(a)}$, where $T(a) = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$;
- The dilation $z \mapsto bz$ is the Möbius map $\Psi_{D(b)}$, where $D(b) = \begin{pmatrix} b & 0 \\ 0 & 1 \end{pmatrix}$;
- The inversion $z \mapsto 1/z$ is the Möbius map Ψ_J, where $J = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

Lemma 11.1.9. *Every Möbius map can be written as a composition of translations, dilations and inversions.*

Proof. Let Ψ_g, $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, be the Möbius map we are interested in. Suppose first that $c \neq 0$. Then, putting aside any worries about ∞, we have the following chain of compositions:

$$z \xrightarrow{\Psi_{D(c)}} cz \xrightarrow{\Psi_{T(d)}} cz + d \xrightarrow{\Psi_J} \frac{1}{cz+d} \xrightarrow{\Psi_{D(\frac{bc-ad}{c})}} \frac{b - \frac{ad}{c}}{cz+d} \xrightarrow{\Psi_{T(\frac{a}{c})}} \frac{az+b}{cz+d}.$$

This certainly suggests (very strongly!) that

$$\Psi_g = \Psi_{T(\frac{a}{c})} \circ \Psi_{D(\frac{bc-ad}{c})} \circ \Psi_J \circ \Psi_{T(d)} \circ \Psi_{D(c)}.$$

A rigorous proof follows from Proposition 11.1.8 and the following identity of matrices (which is of course an easy check):

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = T\left(\frac{a}{c}\right) \cdot D\left(\frac{bc-ad}{c}\right) \cdot J \cdot T(d) \cdot D(c).$$

The case $c = 0$ is much easier and we leave this as an exercise (in this case only a dilation and a translation are required). □

11.1.5. Basic geometry of Möbius maps. In this final subsection, we look at one key example of how Möbius maps transform $\mathbf{C}_\infty$.

Definition 11.1.10. A *circline* is either

- A circle in $\mathbf{C}$ (considered as a subset of $\mathbf{C}_\infty$) or
- A line in $\mathbf{C}$ (considered as a subset of $\mathbf{C}_\infty$) together with the point $\{\infty\}$.

Note that lines in $\mathbf{C}$ are given by equations of the form $|z-a| = |z-b|$ for a and b distinct in $\mathbf{C}$ (think about why).

Proposition 11.1.11. *Möbius maps take circlines to circlines.*

Proof. (sketch: non-examinable) By Lemma 11.1.9 it is enough to check this for translations, dilations and inversions. So it is a straightforward case-by-case analysis. (The only interesting cases being circles through the origin which map under inversion to lines not through the origin, and vice-versa.) □

11.2. Conformal transformations

Another important feature of the stereographic projection map is that it is *conformal*, meaning that it preserves angles. The following definition helps us to formalize what this means:

Definition 11.2.1. If $\gamma\colon [-1,1] \to \mathbb{C}$ is a C^1 path which has $\gamma'(t) \neq 0$ for all t, then we say that the line $\{\gamma(t) + s\gamma'(t) : s \in \mathbb{R}\}$ is the *tangent line* to γ at $\gamma(t)$, and the vector $\gamma'(t)$ is a tangent vector at $\gamma(t) \in \mathbb{C}$.

Note that this definition gives us a notion of tangent vectors at points on paths in subsets of $\mathbb{R}^n$, since the notion of a C^1 path extends readily to such paths (we just require all n component functions are continuously differentiable). In particular, if $\gamma\colon [-1,1] \to \mathbb{S} \subset \mathbb{R}^3$ is a C^1 path, it is easy to check that the tangent vectors at a point $\gamma(t) \in \mathbb{S}$ all lie in the plane perpendicular to $\gamma(t)$ – simply differentiate the identity $f(\gamma(t)) = 1$ where $f(x,y,z) = x^2 + y^2 + z^2$ using the chain rule.

We can now state what we mean by a conformal map:

Definition 11.2.2. Let U be an open subset of $\mathbb{C}$ and suppose that $T\colon U \to \mathbb{C}$ (or $\mathbb{S}$) is continuously differentiable in the real sense (so all its partial derivatives exist and are continuous). If $\gamma_1, \gamma_2\colon [-1,1] \to U$ are two paths with $z_0 = \gamma_1(0) = \gamma_2(0)$ then $\gamma_1'(0)$ and $\gamma_2'(0)$ are two tangent vectors at z_0, and we may consider the angle between them (formally speaking this is the difference of their arguments). By our assumption on T, the compositions $T \circ \gamma_1$ and $T \circ \gamma_2$ are C^1-paths through $T(z_0)$, thus we obtain a pair of tangent vectors at $T(z_0)$. We say that T is *conformal* at z_0 if for every pair of C^1 paths γ_1, γ_2 through z_0, the angle between their tangent vectors at z_0 is equal to the angle between the tangent vectors at $T(z_0)$ given by the C^1 paths $T \circ \gamma_1$ and $T \circ \gamma_2$. We say that T is conformal on U if it is conformal at every $z \in U$.

One of the main reasons we focus on conformal maps here is because holomorphic functions give us a way of producing many examples of them, as the following result shows.

Proposition 11.2.3. *Let $f\colon U \to \mathbb{C}$ be a holomorphic map and let $z_0 \in U$ be such that $f'(z_0) \neq 0$. Then f is conformal at z_0. In particular, if $f\colon U \to \mathbb{C}$ has nonvanishing derivative on all of U, it is conformal on all of U.*

Proof. We need to show that f preserves angles at z_0. Let γ_1 and γ_2 be C^1-paths with $\gamma_1(0) = \gamma_2(0) = z_0$. Then we obtain paths η_1, η_2 through $f(z_0)$ where $\eta_1(t) = f(\gamma_1(t))$ and $\eta_2(t) = f(\gamma_2(t))$. We show that a version of the chain rule applies to these compositions. For $i = 1, 2$ we have

$$\eta_i'(0) = \lim_{h\to 0} \frac{f(\gamma_i(h)) - f(\gamma(0))}{h} = \lim_{h\to 0} \frac{f(\gamma_i(h)) - f(z_0)}{\gamma_i(h) - z_0} \cdot \frac{\gamma_i(h) - z_0}{h}$$

For suitably small h, $\gamma_i(h) \neq z_0$ as $\gamma_i'(0) \neq 0$. (If there was arbitrarily small $h > 0$ with $\gamma_i(h) = z_0$, then the limit $\gamma_i'(0)$ could only be zero.) Also $\lim_{h\to 0} \dfrac{f(\gamma_i(h)) - f(z_0)}{\gamma_i(h) - z_0} = f'(z_0)$. So if we set $f'(z_0) = \rho e^{i\theta}$ we have

$$\eta_i'(0) = f'(z_0)\gamma_i'(0) = \rho e^{i\theta}\gamma_i'(0), \quad i = 1, 2.$$

Hence if ϕ_1 and ϕ_2 are the arguments of $\gamma_1'(0)$ and $\gamma_2'(0)$, then the arguments of $\eta_1'(0)$ and $\eta_2'(0)$ are $\phi_1 + \theta$ and $\phi_2 + \theta$ respectively. It follows that the difference between the two pairs of arguments, that is, the angles between the curves at z_0 and $f(z_0)$, are the same. □

Example 11.2.4. The function $f(z) = z^2$ has $f'(z)$ nonzero everywhere except the origin. It follows f is a conformal map from $\mathbb{C}^\times$ to itself. Note that the condition that $f'(z)$ is nonzero is necessary – if we consider the function $f(z) = z^2$ at $z = 0$, $f'(z) = 2z$ which vanishes precisely at $z = 0$, and it is easy to check that at the origin f in fact doubles the angles between tangent vectors.

Lemma 11.2.5. *The sterographic projection map $S\colon \mathbb{C} \to \mathbb{S}$ is conformal.*

Proof. (Non-examinable sketch) Let z_0 be a point in $\mathbb{C}$, and suppose that $\gamma_1(t) = z_0 + tv_1$ and $\gamma_2(t) = z_0 + tv_2$ are two paths having tangents v_1 and v_2 at $z_0 = \gamma_1(0) = \gamma_2(0)$. Then the lines L_1 and L_2 they describe, together with the point N, determine planes H_1 and H_2 in $\mathbb{R}^3$, and moreover the image of the lines under stereographic projection is the intersection of these planes with $\mathbb{S}$. Since the intersection of $\mathbb{S}$ with any plane is either empty or a circle, it follows that the paths γ_1 and γ_2 get sent to two circles C_1 and C_2 passing through $P = S(z_0)$ and N. Now by symmetry, these circles meet at the same angle at N as they do at P. Now the tangent lines of C_1 and C_2 at N are just the intersections of H_1 and H_2 with the plane tangent to $\mathbb{S}$ at N. But this means the angle between them will be the same as that between the intersection of H_1 and H_2 with the complex plane, since it is parallel to the tangent plane of $\mathbb{S}$ at N. Thus the angles between C_1 and C_2 at P and L_1 and L_2 at z_0 coincide as required. □

Although it follows easily from what we have already done, it is worth highlighting the following:

Lemma 11.2.6. *Möbius transformations are conformal.*

Proof. As we have already shown, any holomorphic map is conformal wherever its derivative is nonzero.

For a Möbius map we have $f(z) = \frac{az+b}{cz+d}$ and

$$f'(z) = \frac{ad-bc}{(cz+d)^2} \neq 0,$$

for all $z \neq -d/c$, thus f is conformal at each $z \in \mathbb{C}\backslash\{-d/c\}$. □

Since a Möbius map is given by the four entries of a 2×2 matrix, up to simultaneous rescaling, the following result is perhaps not too surprising.

Proposition 11.2.7. *If z_1, z_2, z_3 and w_1, w_2, w_3 are triples of pairwise distinct complex numbers, then there is a unique Möbius transformation f such that $f(z_i) = w_i$ for each $i = 1, 2, 3$.*

Proof. It is enough to show that, given any triple (z_1, z_2, z_3) of complex numbers, we can find a Möbius transformations which takes z_1, z_2, z_3 to $0, 1, \infty$ respectively. Indeed if f_1 is such a transformation, and f_2 takes $0, 1, \infty$ to w_1, w_2, w_3 respectively, then clearly $f_2 \circ f_1^{-1}$ is a Möbius transformation which takes z_i to w_i for each i.

Now consider

$$f(z) = \frac{(z-z_1)(z_2-z_3)}{(z-z_3)(z_2-z_1)}$$

It is easy to check that $f(z_1) = 0, f(z_2) = 1, f(z_3) = \infty$, and clearly f is a Möbius transformation as required. If any of z_1, z_2 or z_3 is ∞, then one can find a similar transformation (for example by letting $z_i \to \infty$ in the above formula). Indeed if $z_1 = \infty$ then we set $f(z) = \frac{z_2-z_3}{z-z_3}$; if $z_2 = \infty$, we take $f(z) = \frac{z-z_1}{z-z_3}$; and finally if $z_3 = \infty$ take $f(z) = \frac{z-z_1}{z_2-z_1}$.

To see the f is unique, suppose f_1 and f_2 both took z_1, z_2, z_3 to w_1, w_2, w_3. Then taking Möbius transformations g, h sending z_1, z_2, z_3 and w_1, w_2, w_3 to $0, 1, \infty$ the transformations hf_1g^{-1} and hf_2g^{-1} both take $(0, 1, \infty)$ to $(0, 1, \infty)$. But suppose $T(z) = \frac{az+b}{cz+d}$ is any Möbius transformation with $T(0) = 0$, $T(1) = 1$ and $T(\infty) = \infty$. Since T fixes ∞ it follows $c = 0$. Since $T(0) = 0$ it follows that $b/d = 0$ hence $b = 0$, thus $T(z) = a/d \cdot z$, and since $T(1) = 1$

it follows $a/d = 1$ and hence $T(z) = z$. Thus we see that

$$hf_1g^{-1} = hf_2g^{-1} = \mathrm{id},$$

and so $f_1 = f_2$ as required. □

Example 11.2.8. The above lemma shows that we can use Möbius transformations as a source of conformal maps. For example, suppose we wish to find a conformal transformation which takes the upper half plane $\mathbb{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$ to the unit disk $B(0,1)$. The boundary of $\mathbb{H}$ is the real line, and we know Möbius transformations take lines to lines or circles, and in the latter case this means the point $\infty \in \mathbb{C}_\infty$ is sent to a finite complex number. Now any circle is uniquely determined by three points lying on it, and we know Möbius transformations allow us to take any three points to any other three points. Thus if we take f the Möbius map which sends $0 \mapsto -i$, and $1 \mapsto 1$, $\infty \mapsto i$ the real axis will be sent to the unit circle. Now we have

$$f(z) = \frac{iz+1}{z+i}$$

(one can find f in a similar fashion to the proof of Proposition 11.2.7).

So far, we have found a Möbius transformation which takes the real line to the unit circle. Since $\mathbb{C}\backslash\mathbb{R}$ has two connected components, the upper and lower half planes, $\mathbb{H}$ and $i\mathbb{H}$, and similarly $\mathbb{C}\backslash\mathbb{S}^1$ has two connected components, $B(0,1)$ and $\mathbb{C}\backslash\bar{B}(0,1)$. Since a Möbius transformation is continuous, it maps connected sets to connected sets, thus to check whether $f(\mathbb{H}) = B(0,1)$ it is enough to know which component of $\mathbb{C}\backslash\mathbb{S}^1$ a single point in $\mathbb{H}$ is sent to. But $f(i) = 0 \in B(0,1)$, so we must have $f(\mathbb{H}) = B(0,1)$ as required.