METRIC SPACES AND COMPLEX ANALYSIS.

KEVIN MCGERTY.

1. Introduction

In Prelims you studied Analysis, the rigorous theory of calculus for (real-valued) functions of a single real variable. This term we will largely focus on the study of functions of a complex variable, but we will begin by seeing how much of the theory developed last year can in fact can be made to work, with relatively little extra effort, in a significantly more general context.

Recall the trajectory of the Prelims Analysis course – initially it focused on sequences and developed the notion of the limit of a sequence which was crucial for essentially everything which followed[1]. Then it moved to the study of continuity and differentiability, and finally it developed a theory of integration. This term's course will follow approximately the same pattern, but the contexts we work in will vary a bit more. To begin with we will focus on limits and continuity, and attempt to gain a better understanding of what is needed in order for make sense of these notions.

Example 1.1. Consider for example one of the key definitions of Prelims analysis, that of the *continuity* of a function. Recall that if $f\colon \mathbb{R} \to \mathbb{R}$ is a function, we say that f is continuous at $a \in \mathbb{R}$ if, for any $\epsilon > 0$, we can find a $\delta > 0$ such that if $|x - a| < \delta$ then $|f(x) - f(a)| < \epsilon$. Stated somewhat more informally, this means that no matter how small an ϵ we are given, we can ensure $f(x)$ is within distance ϵ of $f(a)$ provided we demand x is sufficiently close to – that is, within distance δ of – the point a.

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. (Note that we needed this notion of distance in the above definition of continuity for both the pairs (x, a) and $(f(x), f(a))$.) 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

Date: February 8, 2020.

[1]Although continuity is introduced via ϵs and δs, the notion can be expressed in terms of convergent sequences. Similarly one can define the integral in terms of convergent sequences.

precise, we will therefore need to give a mathematically rigorous definition of what a "notion of distance" on a set should be.

As a first step, consider as an example the case of $\mathbb{R}^n$. The dot product on vectors in $\mathbb{R}^n$ gives us a notion of distance between vectors in $\mathbb{R}^n$: Recall that if $v = (v_1, \dots, v_n), w = (w_1, \dots, w_n)$ are vectors in $\mathbb{R}^n$ then we set

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

and we define the length of a vector to be[2] $\|v\| = \langle v, v \rangle^{1/2}$. Recall that the Cauchy-Schwarz inequality then says that $|\langle v, w \rangle| \leq \|v\| \|w\|$. It has the following important consequence for the length function:

Lemma 1.2. *If $x, y \in \mathbb{R}^n$ then $\|x + y\| \leq \|x\| + \|y\|$.*

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

$$\|x + y\|^2 \leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2.$$

But since $\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + 2\langle x, y \rangle + \|y\|^2$, this inequality is immediate from the Cauchy-Schwarz inequality. $\square$

Once we have a notion of length for vectors, we also immediately have a way of defining the distance between them – we simply take the length of the vector $v - w$. Explicitly, this is:

$$\|v - w\| = \big(\sum_{i=1}^{n} (v_i - w_i)^2\big)^{1/2}.$$

Now that we have defined the distance between any two vectors in $\mathbb{R}^n$, we can immediately make sense both of what it means for a function $f \colon \mathbb{R}^n \to \mathbb{R}$ to be continuous[3] as above and also what it means for a sequence to converge.

Definition 1.3. If $(v^k)_{k \in \mathbb{N}}$ is a sequence of vectors in $\mathbb{R}^n$ (so $v^k = (v_1^k, \dots, v_n^k)$) we say $(v^k)_{k \in \mathbb{N}}$ *converges* to $w \in \mathbb{R}^n$ if for any $\epsilon > 0$ there is an $N > 0$ such that for all $k \geq N$ we have $\|v^k - w\| < \epsilon$.

If $f \colon \mathbb{R}^n \to \mathbb{R}$ and $a \in \mathbb{R}^n$ then we say that f is *continuous at a* if for any $\epsilon > 0$ there is a $\delta > 0$ such that $|f(a) - f(x)| < \epsilon$ whenever $\|x - a\| < \delta$. (As usual, we say that f is continuous on $\mathbb{R}^n$ if it is continuous at every $a \in \mathbb{R}^n$.)

Many of the results about convergence for sequences of real or complex numbers which were established last year readily extend to sequences in $\mathbb{R}^n$, with almost identical proofs. As an example, just as for sequences of real or complex numbers, we have the following:

[2]Sometimes the notation $\|v\|_2$ is used for this length function – we will see later there are other natural choices for the length of a vector in $\mathbb{R}^n$.

[3]More ambitiously, using the notions of distance we have for $\mathbb{R}^n$ and $\mathbb{R}^m$ you can readily make sense of the notion of continuity for a function $g \colon \mathbb{R}^n \to \mathbb{R}^m$.

Lemma 1.4. *Suppose that $(v^k)_{k\geq 1}$ is a sequence in $\mathbb{R}^n$ which converges to $w \in \mathbb{R}^n$ and also to $u \in \mathbb{R}^n$. Then $w = u$, that is, limits are unique.*

Proof. We prove this by contradiction: suppose $w \neq u$. Then $d = \|w-u\| > 0$, so since (v^k) converges to w we can find an $N_1 \in \mathbb{N}$ such that for all $k \geq N$ we have $\|w - v^k\| < d/2$. Similarly, since (v^k) converges to u we can find an N_2 such that for all $k \geq N_2$ we have $\|v^k - u\| < d/2$. But then if $k \geq \max\{N_1, N_2\}$ we have

$$d = \|w-u\| = \|(w-v^k)+(v^k-u)\| \leq \|w-v^k\|+\|v^k-u\| < d/2+d/2 = d,$$

where in the first inequality we use Lemma 1.2. Thus we have a contradiction as required. □

2. Metric Spaces

We now come to the definition of a metric space. To motivate it, let's consider what a notion of distance on a set X should mean: Given any two points in X, we should have a non-negative real number – the distance between them. Thus a distance on a set X should be a function $d\colon X \times X \to \mathbb{R}_{\geq 0}$, but we must also decide what properties such a function should have in order to capture our intuition of distance. A couple of properties suggest themselves immediately – the distance between two points $x, y \in X$ should be symmetric, that is, the distance from x to y should[4] be the same as the distance from y to x, and the distance between two points should be 0 precisely when they are equal. Note that this latter property was one of the crucial ingredients in the proof of the uniqueness of limits as we just saw. The only other requirement we will make of a distance function is known as the "triangle inequality", a version of which we established in Lemma 1.2 and which was also essential in the above uniqueness proof. Thus altogether our requirements yield in the following definition:

Definition 2.1. Let X be a set and suppose that $d\colon X \times X \to \mathbb{R}$. Then we say that d is a *distance function* on X if it has the following properties: For all $x, y, z \in X$:

(1) (*Positivity*): $d(x, y) \geq 0$ and $d(x, y) = 0$ if and only if $x = y$.
(2) (*Symmetry*): $d(x, y) = d(y, x)$.
(3) (*Triangle inequality*): If $x, y, z \in X$ then we have

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

Note that for the normal distance function in the plane $\mathbb{R}^2$, the third property expresses the fact that the length of a side of a triangle is at most the sum of the lengths of the other two sides (hence the name!). We will write a metric space as a pair (X, d) of a set and a distance function $d\colon X\times X \to \mathbb{R}_{\geq 0}$

[4] In fact it's possible to think of contexts where this assumption doesn't hold – consider *e.g.* swimming in a river – going upstream is harder work than going downstream, so if your notion of distance took this into account it would fail to be symmetric.

satisfying the axioms above. If the distance function is clear from context, we may, for convenience, simply write X rather than (X, d).

Example 2.2. The vector space $\mathbb{R}^n$ equipped with the distance function $d_2(v, w) = \|v - w\| = \langle v - w, v - w\rangle^{1/2}$ is a metric space: The first two properties of the metric d_2 are immediate from the definition, while the triangle inequality follows from Lemma 1.2.

Example 2.3. In Prelims Linear Algebra, you met the notion of an inner product space $(V, \langle -, -\rangle)$ (over the real or complex numbers). For any two vectors $v, w \in V$ setting $d(v, w) = \|v - w\|$, where $\|v\| = \langle v, v\rangle^{1/2}$, gives V a notion of distance. Since the Cauchy-Schwarz inequality holds in any inner product space, Lemma 1.2 holds in any inner product space (the proof is word for word the same), it follows that d is also a metric in this more general setting.

Definition 2.4. If (X, d_X) is a metric space and $A \subseteq X$ then we set

$$\text{diam}(A) = \sup\{d(a_1, a_2) : a_1, a_2 \in X\} \in \mathbb{R}_{\geq 0} \cup \{\infty\},$$

(where we take $\text{diam}(A) = \infty$ if the $\{d(a_1, a_2) : a_1, a_2 \in A\}$ is not bounded above. If $\text{diam}(A)$ is finite then we say that A is a *bounded* subset of X.

To make good our earlier assertion, we now define the notions of continuity and convergence in a metric space.

Definition 2.5. Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f \colon X \to Y$ is said to be continuous at $a \in X$ if for any $\epsilon > 0$ there is a $\delta > 0$ such that for any $x \in X$ with $d_X(a, x) < \delta$ we have $d_Y(f(x), f(a)) < \epsilon$. We say f is *continuous* if it is continuous at every $a \in X$.

If $(x_n)_{n\geq 1}$ is a sequence in X, and $a \in X$, then we say $(x_n)_{n\geq 1}$ *converges to* a if, for any $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $d_X(x_n, a) < \epsilon$.

In fact it is clear that the notion of uniform continuity also extends to functions between metric spaces: A function $f \colon X \to Y$ is said to be *uniformly continuous* if, for any $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x_1, x_2 \in X$ with $d_X(x_1, x_2) < \delta$ we have $d_Y(f(x_1), f(x_2)) < \epsilon$.

For later use, we also note that a function $f \colon X \to Y$ is said to be *bounded* if its image $f(X)$ is a bounded subset of Y in the sense of Definition 2.4, that is, if

$$\{d_Y(f(x), f(y)) : x, y \in X\} \subseteq \mathbb{R}$$

is a bounded subset of $\mathbb{R}$. Note that, unlike continuity or uniform continuity, the condition that a function is bounded only requires that Y has a metric (X need not).

Example 2.6. Consider the case of $\mathbb{R}^n$ again. The distance function d_2 coming from the dot product makes $\mathbb{R}^n$ into a metric space, as we have already seen. On the other hand it is not the only reasonable notion of distance one can take. We can define for $v, w \in \mathbb{R}^n$

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

Each of these functions clearly satisfies the positivity and symmetry properties of a metric. We have already checked the triangle inequality for d_2, while for d_1 and d_∞ it follows readily from the triangle inequality for $\mathbb{R}$.

Example 2.7. 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*[5] of X.

Example 2.8. 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.

Example 2.9. A slightly more interesting example is the *Hamming distance* on words: if A is a finite set which we think of as an "alphabet", then a word of length n in just an element of A^n, that is, a sequence of n elements in the alphabet. The Hamming distance between two such words $\mathbf{a} = (a_1, \ldots, a_n), \mathbf{b} = (b_1, \ldots, b_n)$ is

$$d_H(\mathbf{a}, \mathbf{b}) = |\{i \in \{1, 2, \ldots, n\} : a_i \neq b_i\}.$$

An important special case of this is the space of binary sequences of length n, that is, where the alphabet A is just $\{0, 1\}$. In this case one can view set of words of length n in this alphabet as a subset of $\mathbb{R}^n$, and moreover you can check that the Hamming distance function is the same as the subspace metric induced by the d_1 metric on $\mathbb{R}^n$ given above.

Example 2.10. If (X, d) is a metric space, then we can consider the space $X^{\mathbb{N}}$ of all sequences in X. That is, the elements of $X^{\mathbb{N}}$ are sequences $(x_n)_{n\geq 1}$ in X. While there is no obvious metric on the whole space of sequences, if we take $X_b^{\mathbb{N}}$ to be the space of *bounded* sequences, that is, sequences such that the set $\{d_\infty(x_n, x_m) : n, m \in \mathbb{N}\} \subset \mathbb{R}$ is bounded, then the function[6]

$$d_\infty((x_n)_{n\geq 1}, (y_n)_{n\geq 1}) = \sup_{n \in \mathbb{N}} d(x_n, y_n),$$

[5] This is completely standard terminology, though it's a little unfortunate if X is a vector space, where we use the word subspace to mean *linear* subspace also. Context (usually) makes it clear which meaning is intended, and I'll try and be as clear about this as possible!

[6] The fact that the sequences are bounded ensure the right-hand side is finite.

is a metric on $X_b^{\mathbb{N}}$. It clearly satisfies positivity and symmetry, and the triangle inequality follows from the inequality

$$d(x_n, z_n) \leq d(x_n, y_n) + d(y_n, z_n) \leq d_\infty((x_n), (y_n)) + d_\infty((y_n), (z_n)),$$

by taking the supremum of the left-hand side over $n \in \mathbb{N}$.

Example 2.11. 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. In fact this can be done in a number of ways – we will return to this issue later. One method is to set $d_{X\times Y} = \max\{d_X, d_Y\}$, that is if $x_1, x_2 \in X$ and $y_1, y_2 \in Y$ then we set

$$d_{X\times Y}((x_1, y_1), (x_2, y_2)) = \max\{d_X(x_1, x_2), d_Y(y_1, y_2)\}.$$

It is straight-forward to check that this is indeed a metric on $X \times Y$. It is also easy to see that if $f \colon Z \to X \times Y$ is a function from a metric space Z to $X \times Y$, so that we may write $f(z) = (f_X(z), f_Y(z))$ with $f_X(z) \in X$ and $f_Y(z) \in Y$, then f is continuous if and only if f_X and f_Y are both continuous. Problem set 1 asks you to check this is also true when you use the metric on $X \times Y$ given by

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

Example 2.12. If (X, d_X) and (Y, d_Y) are metric spaces, then we can also consider the set $\mathcal{B}(X, Y)$ of *bounded* functions from X to Y. This set has a natural metric given by

$$d(f, g) = \sup_{x\in X} d_Y(f(x), g(x)).$$

Indeed one can check that $d(f, g)$ is finite for any $f, g \in \mathcal{B}(X, Y)$, so that since d_Y is non-negatively valued, so is d. This space has a natural subspace consisting of the continuous bounded function $\mathcal{C}_b(X, Y)$.

Example 2.13. Consider the set $\mathbb{P}(\mathbb{R}^n)$ of lines in $\mathbb{R}^n$ (that is, one-dimensional subspace of $\mathbb{R}^n$, or lines through the origin). A natural 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 $\mathbb{R}^2$. The distance between two such lines given by the above formula is then $\sin(\theta)$ where θ is the angle between the two lines.

The next exercise is the natural generalization of the result you saw last year which showed that continuity could be expressed in terms of convergent sequences. It show it one uses exactly the same argument, just phrased in the language of metric spaces.

Exercise 2.14. Let $f\colon X \to Y$ be a function. Show that f is continuous at $a \in X$ if and only if for every sequence $(x_k)_{k\geq 0}$ converging to a we have $f(x_k) \to f(a)$ as $k \to \infty$.

Solution: Suppose that f is continuous at a. Then given $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in X$ with $d(x,a) < \delta$ we have $d(f(x), f(a)) < \epsilon$. Now if $(x_k)_{k\geq 0}$ is a sequence tending to a then there is an $N > 0$ such that $d(a, x_k) < \delta$ for all $k \geq N$. But then for all $k \geq N$ we see that $d(f(a), f(x_k)) < \epsilon$, so that $f(x_k) \to f(a)$ as $k \to \infty$ as required.

For the converse, we use the contrapositive, hence we suppose that f is not continuous at a. Then there is an $\epsilon > 0$ such that for all $\delta > 0$ there is some $x \in X$ with $d(x,a) < \delta$ and $d(f(x), f(a)) \geq \epsilon$. Chose for each $k \in \mathbb{Z}_{>0}$ a point $x_k \in X$ with $d(x_k, a) < 1/k$ but $d(f(x_k), f(a)) \geq \epsilon$. Then $d(x_k, a) < 1/k \to 0$ as $k \to \infty$ so that $x_k \to a$ as $k \to \infty$, but since $d(f(x_k), f(a))) \geq \epsilon$ for all k clearly $(f(x_k))_{k\geq 0}$ does not tend to $f(a)$.

We now review some of the algebra of limits-type results from last year in our more general context:

Definition 2.15. If X is a metric space we write $\mathcal{C}(X) = \{f\colon X \to \mathbb{R} : f \text{ is continuous}\}$ for the set of continuous real-valued functions on X. (Here the real line is viewed as a metric space equipped with the metric coming from the absolute value).

Lemma 2.16. *The set $\mathcal{C}(X)$ is a vector space. Moreover if $f, g \in \mathcal{C}(X)$ then so is $f.g$.*

Proof. $\mathcal{C}(X)$ is a subset of the vector space of all real-valued functions on X, so we just need to check it is closed under addition and multiplication (since we can view scalars as constant functions, the latter clearly being continuous).

To see that $\mathcal{C}(X)$ is closed under multiplication, suppose that $f, g \in \mathcal{C}(X)$ and $a \in X$. To see that $f.g$ is continuous at a, note that if $\epsilon > 0$ is given, then since both f and g are continuous at a, we may find a δ_1 such that $|f(x) - f(a)| < \min\{1, \epsilon/2(|g(a)|+1)\}$ for all $x \in X$ with $d(x,a) < \delta_1$ and a $\delta_2 > 0$ such that $|g(x) - g(a)| < \epsilon/2(|f(a)|+1)$ for all $x \in X$ with $d(x,a) < \delta_2$. Setting $\delta = \min\{\delta_1, \delta_2\}$ it follows that for all $x \in X$ with $d(x,a) < \delta$ we have

$$\begin{aligned} |f(x)g(x) - f(a)g(a)| &= |f(x)g(x) - f(x)g(a) + f(x)g(a) - f(a)g(a)| \\ &\leq |f(x)||g(x) - g(a)| + |f(x) - f(a)||g(a)| \\ &\leq (|f(a)|+1)|g(x) - g(a)| + |f(x) - f(a)||g(a)| \\ &< \epsilon/2 + \epsilon/2 = \epsilon \end{aligned}$$

where in the third line we use the fact that $|f(x)| < |f(a)| + 1$ for all $x \in X$ such that $d(x,a) < \delta_1$. Since a was arbitrary, this shows that $f.g$ lies in $\mathcal{C}(X)$. Checking that $\mathcal{C}(X)$ is closed under addition is similar but easier, and we leave it as an exercise for the reader to check the details. $\square$

Exercise 2.17. One can also check that if $f\colon X \to \mathbb{R}$ is continuous at a and $f(a) \neq 0$ then $1/f$ is continuous at a. Again this is proved just as in the single-variable case. Problem set 1 asks you to provide the details for this.

3. Normed vector spaces.

If we start with a vector space V, for example the set of solutions to a homogeneous linear differential equation, then it is natural to consider metrics which interact with the linear structures – addition and scalar multiplication– of the space.

Two natural conditions to consider are the following: for any vectors $x, y, z \in \mathbb{R}^n$ and any scalar λ we have

(1) $d(x+z, y+z) = d(x,y)$,
(2) $d(\lambda x, \lambda y) = |\lambda| d(x,y)$.

The first of these is known as *translation invariance* and the second is a kind of *homogeneity*.

A vector space V with a distance function compatible with the vector space structure in the above sense is then clearly determined by the function from V to the non-negative real numbers given by $v \mapsto d(v,0)$. The following definition and Lemma formalize this discussion.

Definition 3.1. Let V be a (real or complex) vector space. A *norm* on V is a function $\|.\|\colon V \to \mathbb{R}$ which satisfies the following properties:

(1) (*Positivity*): $\|x\| \geq 0$ for all $x \in V$ and $\|x\| = 0$ if and only if $x = 0$.
(2) (*Homogeneity*): if $x \in V$ and λ is a scalar then
$$\|\lambda.x\| = |\lambda|\|x\|.$$
(3) (*Triangle inequality*): If $x, y \in V$ then $\|x+y\| \leq \|x\| + \|y\|$.

Note that in the second property $|\lambda|$ denotes the absolute value of λ if V is a real vector space, and the modulus of λ if V is a complex vector space.

Remark 3.2. If there is the potential for ambiguity, we will write the norm on a vector space V as $\|.\|_V$, but usually this is clear from the context, and so just as for metric spaces we will write $\|.\|$ for the norm on all vector spaces we consider.

Lemma 3.3. *If V is a vector space with a norm $\|.\|$ then the function $d\colon V \times V \to \mathbb{R}_{\geq 0}$ given by $d(x,y) = \|x - y\|$ is a metric which is compatible with the vector space structure in that:*

(1) *For all $x, y \in V$ we have*
$$d(\lambda.x, \lambda.y) = |\lambda| d(x,y).$$

(2) $d(x+z, y+z) = d(x,y)$.

Conversely, if d *is a metric satisfying the above conditions then* $\|v\| = d(v,0)$ *is a norm on* V*.*

Proof. This follows immediately from the definitions. □

Example 3.4. As discussed above, if $V = \mathbb{R}^n$ then the metrics d_1, d_2, d_∞ all come from the norms. We denote these by $\|x\|_1 = \sum_{i=1}^m |x_i|$ and $\|x\|_2 = (\sum_{i=1}^m x_i^2)^{1/2}$ and $\|x\|_\infty = \max_{1\leq i\leq m} |x_i|$.

Since the most natural maps between vector spaces are linear maps, it is natural to ask when a linear map between normed vector spaces is continuous. The following lemma gives an answer to this question:

Lemma 3.5. *Let* $f\colon V \to W$ *be a linear map between normed vector spaces. Then* f *is continuous if and only if* $\{\|f(x)\| : \|x\| \leq 1\}$ *is bounded.*

Proof. If f is continuous, then it is continuous at $0 \in V$ and so there is a $\delta > 0$ such that for all $v \in V$ with $\|v\| < \delta$ we have $\|f(v) - f(0)\| = \|f(v)\| < \epsilon$. But then if $\|v\| \leq 1$ we have $\frac{\delta}{2}\|f(v)\| = \|f(\frac{\delta}{2}.v))\| < \epsilon$, and hence $\|f(v)\| \leq \frac{2\epsilon}{\delta}$.

For the converse, if we have $\|f(v)\| < M$ for all v with $\|v\| \leq 1$, then if $\epsilon > 0$ is given we may pick $\delta > 0$ so that $\delta.M < \epsilon$ and hence if $\|v - w\| < \delta$ we have

$$\|f(v) - f(w)\| = \|f(v-w)\| = \delta\|f(\delta^{-1}(v-w))\| \leq \delta.M < \epsilon,$$

so that f is in fact uniformly continuous on V. □

Remark 3.6. The boundedness condition above can be rephrased as saying there is a constant $K > 0$ such that $\|f(v)\| \leq K.\|v\|$, since any non-zero vector v can be rescaled to a vector of unit length, $v/\|v\|$.

An important source of (normed) vector spaces for us will be the space of functions on a set X (usually a metric space). Indeed if X is any set, the space of all real-valued functions on X is a vector space – addition and scalar multiplication are defined "pointwise" just as for functions on the real line. It is not obvious how to make this into a normed vector space, but if we restrict to the subspace $\mathcal{B}(X)$ of *bounded* functions there is an reasonably natural choice of norm.

Definition 3.7. If X is any set we define

$$\mathcal{B}(X) = \{f\colon X \to \mathbb{R} : f(X) \subset \mathbb{R} \text{ bounded}\},$$

to be the space of bounded functions on X, that is $f \in \mathcal{B}(X)$ if and only if there is some $K > 0$ such that $|f(x)| < K$ for all $x \in X$. For $f \in \mathcal{B}(X)$ we set $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

Lemma 3.8. *Let* X *be any set, then* $(\mathcal{B}(X), \|.\|_\infty)$ *is a normed vector space.*

Proof. To see that $\mathcal{B}(X)$ is a vector space, note that if $f, g \in \mathcal{B}(X)$ then we may find $N_1, N_2 \in \mathbb{R}_{>0}$ such that $f(X) \subseteq [-N_1, N_1]$ and $g(X) \subseteq [-N_2, N_2]$. But then clearly $(f+g)(X) \subseteq [-N_1 - N_2, N_1 + N_2]$ and if $\lambda \in \mathbb{R}$ then $(\lambda.f)(X) \subseteq [-|\lambda|N_1, |\lambda|N_1]$, so that $\lambda.f \in \mathcal{B}(X)$ and $f + g \in \mathcal{B}(X)$.

Next we check that $\|f\|_\infty$ is a norm: it is clear from the definition that $\|f\|_\infty \geq 0$ with equality if and only if f is identically zero. Compatibility with scalar multiplication is also immediate, while for the triangle inequality note that if $f, g \in \mathcal{B}(X)$, then for all $x \in X$ we have

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

Taking the supremum over $x \in X$ then yields the result. □

We will write d_∞ for the metric associated to the norm $\|.\|_\infty$.

If X is itself a metric space, we also have the space $\mathcal{C}(X)$ of continuous real-valued functions on X. While $\mathcal{C}(X)$ does not automatically have a norm, the subspace $\mathcal{C}_b(X) = \mathcal{C}(X) \cap \mathcal{B}(X)$ of *bounded* continuous functions clearly inherits a norm from $\mathcal{B}(X)$.

Example 3.9. One can check that if $X = [a, b]$ then if $(f_n)_{n\geq 1}$ is a sequence in[7] $\mathcal{C}([a, b]) = \mathcal{C}_b([a, b])$ then $f_n \to f$ in $(\mathcal{C}_b(X), d_\infty)$ (where d_∞ is the metric given by the norm $\|.\|_\infty$) if and only if f_n tends to f uniformly.

Example 3.10. For certain spaces X, we can define other natural metrics on the space of continuous functions: Let $X = [a, b] \subset \mathbb{R}$ be a closed interval. Then we know that in fact all continuous functions on X are bounded, so that $\|.\|_\infty$ defines a norm on $\mathcal{C}([a, b])$. We can also define analogues of the norms $\|.\|_1$ and $\|.\|_2$ on $\mathbb{R}^n$ using the integral in place of summation: Let

$$\|f\|_1 = \int_a^b |f(t)|dt,$$
$$\|f\|_2 = \Big(\int_a^b f(t)^2 dt\Big)^{1/2}$$

Lemma 3.11. *Suppose that $a < b$ so that the interval $[a, b]$ has positive length. Then the functions $\|.\|_1$ and $\|.\|_2$ are norms on $\mathcal{C}([a, b])$.*

Proof. The compatibility with scalars and the triangle inequality both follow from standard properties of the integral. The interesting point to check here is that both $\|.\|_1$ and $\|.\|_2$ satisfy postitivity – continuity[8] is crucial for this! Indeed if $f = 0$ clearly $\|f\|_1 = \|f\|_2 = 0$. On the other hand if $f \neq 0$ then there is some $x_0 \in [a, b]$ such that $f(x_0) \neq 0$, and so $|f(x_0)| > 0$. Since f is continuous at x_0, there is a $\delta > 0$ such that $|f(x) - f(x_0)| < |f(x_0)|/2$

[7] The result from Prelims Analysis showing any continuous function on a closed bounded interval is bounded implies the equality $\mathcal{C}([a, b]) = \mathcal{C}_b([a, b])$.

[8] So in particular, $\|.\|_1$ and $\|.\|_2$ are *not* norms on the space of Riemann integrable functions on $[a, b]$.

for all $x \in [a,b]$ with $|x - x_0| < \delta$. But the it follows that for $x \in [a,b]$ with $|x - x_0| < \delta$ we have $|f(x)| \geq |f(x_0)| - |f(x) - f(x_0)| > |f(x_0)|/2$. Now set

$$s(x) = \begin{cases} |f(x_0)|/2, & \text{if } x \in [a,b] \cap (x_0 - \delta, x_0 + \delta) \\ 0, & \text{otherwise} \end{cases}$$

Since the interval $[a,b] \cap (x_0 - \delta, x_0 + \delta)$ has length at least $d = \min\{\delta, (b-a)\}$ we see that $\int_a^b s(x)dx \geq d.|f(x_0)|/2 > 0$. Since $s(x) \leq |f(x)|$ for all $x \in [a,b]$ it follows from the positivity of the integral that $0 < d|f(x_0)|/2 \leq \|f\|_1$. Similarly we see that $\|f\|_2 \geq f\sqrt{d}|f(x_0)|/2$, so that both $\|.\|_1$ and $\|.\|_2$ satisfy the positivity property. □

There are very similar metrics on certain sequence spaces:

Example 3.12. Let

$$\ell_1 = \{(x_n)_{n \geq 1} : \sum_{n \geq 1} |x_n| < \infty\}$$
$$\ell_2 = \{(x_n)_{n \geq 1} : \sum_{n \geq 1} x_n^2 < \infty\}$$
$$\ell_\infty = \{(x_n)_{n \geq 1} : \sup_{n \in \mathbb{N}} |x_n| < \infty\}.$$

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

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

(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 problem sets investigate the example of ℓ_2 in some detail.

4. Metrics and convergence

Recall that if (X,d) is a metric space, then a sequence (x_n) in X converges to a point $a \in X$ if for any $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that for all $n \geq N$ we have $d(x_n, a) < \epsilon$. In the case of $\mathbb{R}^m$, although d_1, d_2, d_∞ are all different distance functions, they in fact give the same notion of convergence. To see this we need the following:

Lemma 4.1. *Let $x, y \in \mathbb{R}^m$. Then we have*

$$d_2(x,y) \leq d_1(x,y) \leq \sqrt{m} d_2(x,y); \quad d_\infty(x,y) \leq d_2(x,y) \leq \sqrt{m} d_\infty(x,y).$$

Proof. It is enough to check the corresponding inequalities for the norms $\|x\|_i$ (where $i \in \{1, 2, \infty\}$) that is, we may assume $y = 0$. For the first inequality, note that

$$\|x\|_1^2 = (\sum_{i=1}^m |x_i|)^2 = \sum_{i=1}^m x_i^2 + \sum_{1\leq i<j\leq m} 2|x_i x_j| \geq \sum_{i=1}^m x_i^2 = \|x\|_2^2,$$

so that $\|x\|_2 \leq \|x\|_1$. On the other hand, if $x = (x_1, \dots, x_m)$, set $a = (|x_1|, |x_2|, \dots, |x_m|)$ and $\mathbf{1} = (1, 1, \dots, 1)$. Then by the Cauchy-Schwarz inequality we have

$$\|x\|_1 = \langle \mathbf{1}, a\rangle \leq \sqrt{m}.\|a\|_2 = \sqrt{m}.\|x\|_2$$

The second pair of inequalities is simpler. Note that clearly

$$\max_{1\leq i\leq m} |x_i| = \max_{1\leq i\leq m} (x_i^2)^{1/2} \leq (\sum_{i=1}^m x_i^2)^{1/2},$$

yielding one inequality. On the other hand, since for each i we have $|x_i| \leq \|x\|_\infty$ by definition, clearly

$$\|x\|_2^2 = \sum_{i=1}^m |x_i|^2 \leq m\|x\|_\infty^2,$$

giving $\|x\|_2/\sqrt{m} \leq \|x\|_\infty$ as required. $\square$

Lemma 4.2. *If $(x^n) \subset \mathbb{R}^m$ is a sequence then (x^n) converges to $a \in \mathbb{R}^m$ with respect to the metric d_2, if and only if it does with respect to the metric d_1, if and only if it does so with respect to the metric d_∞. Thus the three metrics all yield the same notion of convergence.*

Proof. Suppose (x^n) converges to a with respect to the metric d_2. Then for any $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d_2(x^n, a) < \epsilon/\sqrt{m}$ for all $n \geq N$. It follows from the previous Lemma that for $n \geq N$ we have

$$d_1(x^n, a) \leq \sqrt{m}.d_2(x^n, a) < \sqrt{m}.(\epsilon/\sqrt{m}) = \epsilon,$$

and so (x^n) converges to a with respect to d_1 also. Similarly we see that convergence with respect to d_1 implies convergence with respect to d_2 using $\|x\|_2 \leq \|x\|_1$. In the same fashion, the inequalities $d_\infty(x, y) \leq d_2(x, y) \leq \sqrt{m}d_\infty(x, y)$ yield the equivalence of the notions of convergence for d_2 and d_∞. $\square$

Remark 4.3. *(Non-examinable): If X is any set and d_1, d_2 are two metrics on X, we say they are equivalent if there are positive constants K, L such that*

$$d_1(x, y) \leq K d_2(x, y); \quad d_2(x, y) \leq L d_1(x, y).$$

The proof of the previous Lemma extends to show that if two metrics are equivalent, then a sequence converges with respect to one metric if and only if it does with respect to the other.

5. Open and closed sets

In this section we will define two special classes of subsets of a metric space – the open and closed subsets. To motivate their definition, recall that we have two ways of characterizing continuity in a metric space: the "ϵ-δ" definition, and the description in terms of convergent sequences. Examining the former will lead us to the notion of an open set, while examining the latter will lead us to the notion of a limit point and hence that of a closed set.

The definitions of continuity and convergence can be made somewhat more geometric if we introduce the notion of a ball in a metric space:

Definition 5.1. Let (X, d) is a metric space. If $x_0 \in X$ and $\epsilon > 0$ then we define the *open ball of radius* ϵ to be the set

$$B(x_0, \epsilon) = \{x \in X : d(x, x_0) < \epsilon\}.$$

Similarly we defined the *closed ball* of radius ϵ about x_0 to be the set

$$\bar{B}(x_0, \epsilon) = \{x \in X : d(x, x_0) \leq \epsilon\}.$$

The term "ball" comes from the case where $X = \mathbb{R}^3$ equipped with the usual Euclidean notion of distance. When $X = \mathbb{R}$ an open/closed ball is just an open/closed interval.

Recall that if $f\colon X \to Y$ is a function between any two sets, then given any subset $Z \subseteq Y$ we let[9] $f^{-1}(Z) = \{x \in X : f(x) \in Z\}$. The set $f^{-1}(Z)$ is called the *pre-image* of Z under the function f.

Lemma 5.2. *Let (X, d) and (Y, d) be metric spaces. Then $f\colon X \to Y$ is continuous at $a \in X$ if and only if, for any open ball $B(f(a), \epsilon)$ centred at $f(a)$ there is an open ball $B(a, \delta)$ centred at a such that $f(B(a, \delta)) \subseteq B(f(a), \epsilon)$, or equivalently $B(a, \delta) \subseteq f^{-1}(B(f(a), \epsilon))$.*

Proof. This follows directly from the definitions. (*Check this!*) □

We have seen in the last section that different metrics on a set X can give the same notions of continuity. The next definition is motivated by this – it turns out that we can attach to a metric a certain class of subsets of X known as *open sets* and knowing these open sets suffices to determine which functions on X are continuous. Informally, a subset $U \subseteq X$ is open if, for any point $y \in U$, every point sufficiently close to y in X is also in U. Thus, if $y \in U$, it has some "wiggle room" – we may move slightly away from y while still remaining in U. The rigorous definition is as follows:

Definition 5.3. If (X, d) is a metric space then we say a subset $U \subset X$ is *open* (or *open in* X) if for each $y \in U$ there is some $\delta > 0$ such that $B(y, \delta) \subseteq U$. More generally, if $Z \subseteq X$ and $z \in Z$ then we say Z is a *neighbourhood* of z if

[9]The notion is not meant to suggest that f is invertible, though when it is, the preimage of any point in Y is a single point in X, so the notation is in this sense consistent. Note that formally, f^{-1} as defined here is a function from the power set of Y to the power set of X.

there is a $\delta > 0$ such that $B(z, \delta) \subseteq Z$. Thus a subset $U \subseteq X$ is open exactly when it is a neighbourhood of all of its elements.

The collection $\mathcal{T} = \{U \subset X : U \text{ open in } X\}$ of open sets in a metric space (X, d) is called the *topology* of X.

We first note an easy lemma, which can be viewed as a consistency check on our terminology.

Lemma 5.4. *Let (X, d) be a metric space. If $a \in X$ and $\epsilon > 0$ then $B(a, \epsilon)$ is an open set.*

Proof. We need to show that $B(a, \epsilon)$ is a neighbourhood of each of its points. If $x \in B(a, \epsilon)$ then by definition $r = \epsilon - d(a, x) > 0$. We claim that $B(x, r) \subseteq B(a, \epsilon)$. Indeed by the triangle inequality we have for $z \in B(x, r)$

$$d(z, a) \leq d(z, x) + d(x, a) < r + d(x, a) = \epsilon,$$

as required. □

Remark 5.5. While reading the above proof, please draw a picture of the case where $X = \mathbb{R}^2$ with the standard metric d_2.

Next let us observe some basic properties of open sets.

Lemma 5.6. *Let (X, d) be metric space and let $\mathcal{T}$ be the associated topology on X. Then we have*

(1) *The subsets X and $\emptyset$ are open.*
(2) *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.*
(3) *If I is finite and $\{U_i : i \in I\}$ are open sets then $\bigcap_{i \in I} U_i$ is open in X.*

Proof. The first claim is trivial. For the second claim, if $x \in \bigcup_{i \in} U_i$ then there is some $i \in I$ with $x \in U_i$. Since U_i is open, there is an $\epsilon > 0$ such that

$$B(x, \epsilon) \subset U_i \subseteq \bigcup_{i \in I} U_i,$$

so that $\bigcup_{i \in I} U_i$ is a neighbourhood of each of its points as required. Applying this to the case $I = \emptyset$ shows that $\emptyset \subseteq X$ is open (or simply note that the empty set satisfies the condition to be an open set vacuously).

For the final claim, if I is finite and $x \in \bigcap_{i \in I} U_i$, then for each i there is an $\epsilon_i > 0$ such that $B(x, \epsilon_i) \subseteq U_i$. But then since I is finite, $\epsilon = \min(\{\epsilon_i : i \in I\} \cup \{1\}) > 0$, and

$$B(x, \epsilon) \subseteq \bigcap_{i \in I} B(x, \epsilon_i) \subseteq \bigcap_{i \in I} U_i,$$

so that $\bigcap_{i \in I} U_i$ is an open subset as required. Applying this to the case $I = \emptyset$ shows that X is open (or simply note that if $U = X$ and $x \in X$ then $B(x, \epsilon) \subseteq X$ for *any* positive ϵ so that X is open). □

Remark 5.7. Apart from being trivial, the first part of the above lemma is also redundant in that it follows from the second and third: If I is an indexing set, then a collection $\{U_i : i \in I\}$ of subsets of X is just a function $u\colon I \to \mathcal{P}(X)$ where $\mathcal{P}(X)$ denotes the power set of X, where by convention[10] we write $U_i \subseteq X$ for $u(i)$. Then union $\bigcup_{i \in I} U_i$ of the collection of subsets $\{U_i : i \in I\}$ is then $\{x \in X : \exists i \in I, x \in U_i\}$, while the intersection of the collection $\{U_i : i \in I\}$ is just $\{x \in X : \forall i \in I, x \in U_i\}$. Using this, one readily sees that if $I = \emptyset$ then the intersection of the collection is X and the union is the empty set $\emptyset$.

Exercise 5.8. Using Lemma 4.1, show that the topologies $\mathcal{T}_i$ on $\mathbb{R}^n$ given by the norms d_i $(i = 1, 2, \infty)$ coincide.

Example 5.9. A subset U of $\mathbb{R}$ is open if for any $x \in U$ there is an open interval centred at x contained in U. Thus we can readily see that the finiteness condition for intersections is necessary: if $U_i = (-1/i, 1)$ for $i \in \mathbb{N}$ then each U_i is open but $\bigcap_{i \in \mathbb{N}} U_i = [0, 1)$ and $[0, 1)$ is not open because it is not a neighbourhood of 0.

One important consequence of the fact that arbitrary unions of open sets are open is the following:

Definition 5.10. Let (X, d) be a metric space and let $S \subseteq X$. The *interior* of S is defined to be

$$\text{int}(S) = \bigcup_{\substack{U \subseteq S \\ U \text{ open}}} U.$$

Since the union of open subsets is always open, int(S) is an open subset of X and it is the largest open subset of X which is contained in S in the sense that any open subset of X which is contained in S is in fact contained in int(S). If $x \in \text{int}(S)$ we say that x is an *interior point* of S. One can also phrase this in terms of neighborhoods: the interior of S is the set of all points in S for which S is a neighbourhood.

Example 5.11. If $S = [a, b]$ is a closed interval in $\mathbb{R}$ then its interior is just the open interval (a, b). If we take $S = \mathbb{Q} \subset \mathbb{R}$ then int$(\mathbb{Q}) = \emptyset$.

We now show that the topology given by a metric is sufficient to characterize continuity.

Proposition 5.12. *Let X and Y be metric spaces and let $f\colon X \to Y$ be a function. If $a \in X$ then f is continuous at a if and only if for every neighbourhood $N \subseteq Y$ of $f(a)$, the preimage $f^{-1}(N)$ is a neighbourhood of $a \in X$. Moreover, 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.*

[10]This is similar to how a sequence in a space X is actually a function $a\colon \mathbb{N} \to X$, but we usually write a_n rather than $a(n)$.

Proof. First suppose that f is continuous at a, and let N be a neighhourhood of $f(a)$. Then we may find an $\epsilon > 0$ such that $B(f(a), \epsilon) \subseteq N$. Since f is continuous at a, there is a $\delta > 0$ such that $B(x, \delta) \subseteq f^{-1}(B(f(a), \epsilon)) \subseteq f^{-1}(U)$. It follows $f^{-1}(N)$ is a neighbourhood of a as required. Conversely, if $\epsilon > 0$ is given, then certainly $B(f(a), \epsilon)$ is a neighbourhood of $f(a)$, so that $f^{-1}(B(f(a), \epsilon))$ is a neighbourhood of a, hence there is a $\delta > 0$ such that $B(a, \delta) \subseteq f^{-1}(B(f(a), \epsilon))$, and thus f is continuous at a as required.

Now if f is continuous on all of X, since a set is open if and only if it is a neighbourhood of each of its points, it follows from the above that $f^{-1}(U)$ is an open subset of X for any open subset U of Y. For the converse, note that if $a \in X$ is any point of X and $\epsilon > 0$ is given then the open ball $B(f(a), \epsilon)$ is an open subset of Y, hence $f^{-1}(B(f(a), \epsilon))$ is open in X, and in particular is a neighbourhood of $a \in X$. But then there is a $\delta > 0$ such that $B(a, \delta) \subseteq f^{-1}(B(f(a), \epsilon))$, hence f is continuous at a as required. □

Example 5.13. Notice that this Proposition gives us a way of producing many examples of open sets: if $f\colon \mathbb{R}^n \to \mathbb{R}$ is any continuous function and $a, b \in \mathbb{R}$ are real numbers with $a < b$ then $\{v \in \mathbb{R}^n : a < f(x) < b\} = f^{-1}((a, b))$ is open in $\mathbb{R}^n$. Thus for example $\{(x, y) \in \mathbb{R}^2 : 1 < 2x^2 + 3xy < 2\}$ is an open subset of the plane.

Exercise 5.14. Use the characterization of continuity in terms of open sets to show that the composition of continuous functions is continuous[11].

Remark 5.15. The previous Proposition 5.12 shows, perhaps surprisingly, that we actually need somewhat less than a metric on a set X to understand what continuity means: we only need the topology induced by the metric on the set X. In particular any two metrics which give the same topology give the same notion of continuity. This motivates the following, perhaps rather abstract-seeming, definition.

Definition 5.16. If X is a set, a *topology* on X is a collection of subsets $\mathcal{T}$ of X, known as the *open subsets* which satisfy the conclusion of Lemma 5.6. That is,

(1) If $\{U_i : i \in I\}$ are in $\mathcal{T}$ then $\bigcup_{i \in I} U_i$ is in $\mathcal{T}$. In particular $\emptyset$ is an open subset.
(2) If I is finite and $\{U_i : i \in I\}$ are in $\mathcal{T}$, then $\bigcap_{i \in I} U_i$ is in $\mathcal{T}$. In particular X is an open subset of X.

A *topological space* is a pair $(X, \mathcal{T}_X)$ consisting of a set X and a choice of topology $\mathcal{T}_X$ on X.

Motivated by Proposition 5.12, if $f\colon X \to Y$ is a function between two topological spaces $(X, \mathcal{T}_X)$ and $(Y, \mathcal{T}_Y)$ we say that f is *continuous* if for every open subset $U \in \mathcal{T}_Y$ we have $f^{-1}(U) \in \mathcal{T}_X$, that is, $f^{-1}(U)$ is an open subset of X.

[11] This is easy, the point is just to check you see how easy it is!

Remark 5.17. There are a variety of ways of stating the axioms for a topology. They are often phrased by stating separately that X and $\emptyset$ are open. For example the Topology course choses the axioms:

(1) The sets X and $\emptyset$ are open.
(2) If U and V are open, then $U \cap V$ is open.
(3) If I is any indexing set and $\{U_i : i \in I\}$ are a collection of open sets in X then $\bigcup_{i\in I} U_i$ is open.

In this articulation of the axioms, the the condition that $\emptyset$ is open is redundant[12], while the condition that $\bigcap_{i\in I} U_i$ is open for finite indexing sets I follows from axioms 1) and 2) using induction.

The properties of a metric space which we can express in terms of open sets can equally be expressed in terms of their complements, which are known as *closed sets*. It is useful to have both formulations (as we will show, the formulation of continuity in terms of closed sets is closer to that given by convergence of sequences rather than the ϵ-δ definition).

Definition 5.18. If (X, d) is a metric space, then a subset $F \subseteq X$ is said to be a *closed* subset of X if its complement $F^c = X\backslash F$ is an open subset.

Remark 5.19. 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 $\mathbb{R}$ the set $\mathbb{R}$ 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.

The following lemma follows easily from Lemma 5.6 by using DeMorgan's Laws.

Lemma 5.20. *Let (X, d) be a metric space and let $\{F_i : i \in I\}$ be a collection of closed subsets.*

(1) *The intersection $\bigcap_{i\in I} F_i$ is a closed subset. In particular X is a closed subset of X.*
(2) *If I is finite then $\bigcup_{i\in I} F_i$ is closed. In particular the empty set $\emptyset$ is a closed subset of X.*

Moreover, if $f: X \to Y$ is a function between two metric spaces X and Y then f is continuous if and only if $f^{-1}(G)$ is closed for every closed subset $G \subseteq Y$.

Proof. The properties of closed sets follow immediately from DeMorgan's law, while the characteriszation of continuity follows from the fact that if $G \subset Y$ is any subset of Y we have $f^{-1}(G^c) = (f^{-1}(G))^c$, that is, $X\backslash f^{-1}(G) = f^{-1}(Y\backslash G)$. □

[12] This is not necessarily a terrible thing, for example in giving the axioms for a group, one can require the existence of a two-sided identity and of two-sided inverses, or just the existence of a left-identity and left-inverses. Although the two-sided version is contains redundant stipulations it is nevertheless the most commonly used one.

Lemma 5.21. *If (X,d) is a metric space then any closed ball $\bar{B}(a,r)$ for $r \geq 0$ is a closed set. In particular, singleton sets are closed.*

Proof. We must show that $X \backslash \bar{B}(a,r)$ is open. If $y \in X \backslash \bar{B}(a,r)$ then $d(a,y) > r$, so that $\epsilon = d(a,y) - r > 0$. But then if $z \in B(y,\epsilon)$ we have

$$d(a,z) \geq d(a,y) - d(z,y) > d(a,y) - \epsilon = r,$$

so that $z \notin \bar{B}(a,r)$. It follows that $B(y,\epsilon) \subseteq X \backslash \bar{B}(a,r)$ and so $X \backslash \bar{B}(a,r)$ is open as required. $\square$

The relation between closed sets and convergent sequences mentioned at the beginning of this section arises through the notion of a limit point which we now define.

Definition 5.22. If (X,d) is a metric space and $Z \subseteq X$ is any subset, then we say a point $a \in X$ is a *limit point* if for any $\epsilon > 0$ we have $\big(B(a,\epsilon) \backslash \{a\}\big) \cap Z \neq \emptyset$. If $a \in Z$ and a is *not* a limit point of Z we say that a is an *isolated point* of Z. The set of limit points of Z is denoted Z'. Notice that if $Z_1 \subseteq Z_2$ are subsets of X then it follows immediately from the definition that $Z_1' \subseteq Z_2'$.

Example 5.23. If $Z = (0,1] \cup \{2\} \subset \mathbb{R}$ then 0 is a limit point of Z which does not lie in Z, while 2 is an isolated point of Z because $B(2,1/2) \cap Z = (1.5, 2.5) \cap Z = \{2\}$.

If (x_n) is a sequence in (X,d) which converges to $\ell \in X$ then $\{x_n : n \in \mathbb{N}\}$ is either empty or equal to $\{\ell\}$. (See the problem set.)

The term "limit point" is motivated by the following easy result:

Lemma 5.24. *If S is a subset of a metric space (X,d) then $x \in S'$ if and only if there is a sequence in $S \backslash \{x\}$ converging to x.*

Proof. If x is a limit point then for each $n \in \mathbb{N}$ we may pick $z_n \in B(x,1/n) \cap (S \backslash \{x\})$. Then clearly $z_n \to x$ as $n \to \infty$ as required. Conversely if (z_n) is a sequence in $S \backslash \{x\}$ converging to x and $\delta > 0$ is given, there is an $N \in \mathbb{N}$ such that $z_n \in B(x,\delta)$ for all $n \geq N$. It follows that $B(x,\delta) \cap (S \backslash \{x\})$ is nonempty as required. $\square$

The fact that a subset of a metric space is closed can be characterized in terms of limit points (and hence in terms of convergent seqeunces):

The fact that any intersection of closed subsets is closed has an important consequence – given any subset S of a metric space (X,d) there is a unique smallest closed set which contains S.

Definition 5.25. Let (X,d) be a metric space and let $S \subseteq X$. Then the set

$$\bar{S} = \bigcap_{\substack{G \supseteq S \\ G \text{ closed}}} G,$$

is the *closure* of S. It is closed because it is the intersection of closed subsets of X and is the smallest closed set containing S in the sense that if G is any closed set containing S then G contains $\bar{S}$. If $S \subseteq Y \subseteq X$ we say that S is *dense* in Y if $Y \subseteq \bar{S}$. (Thus every point of Y lies in S or is a limit point of S.)

Example 5.26. The rationals $\mathbb{Q}$ are a dense subset of $\mathbb{R}$, as is the set $\{\frac{a}{2^n} : a \in \mathbb{Z}, n \in \mathbb{N}\}$.

Definition 5.27. The notions of closure and interior also allow us to define the *boundary* ∂S of a subset S of a metric space to be $\bar{S}\backslash\text{int}(S)$.

Proposition 5.28. *Let (X, d) be a metric space and let $S \subseteq X$. Then*

$$S \cup S' = \bar{S}.$$

In particular, a subset S is closed if and only if $S' \subseteq S$, i.e. if and only if S contains all of its limit points.

Proof. Let $Y = S \cup S'$. Since $S \subseteq \bar{S}$, certainly $S' \subseteq (\bar{S})'$, and as $\bar{S}$ is closed, by Lemma **??**, $(\bar{S})' \subseteq \bar{S}$. Hence $Y \subseteq \bar{S}$. To see the opposite inclusion, suppose that $a \notin Y$. Then there is a $\delta > 0$ such that $B(a, \delta) \cap S = \emptyset$. It follows that $S \subseteq B(a, \delta)^c$ and thus since $B(a, \delta)^c$ is closed, $\bar{S} \subseteq B(a, \delta)^c$, and so certainly $a \notin \bar{S}$. It follows $\bar{S} \subseteq Y$ and hence $\bar{S} = Y$ are required. □

Remark 5.29. If $Z \subseteq X$ is an arbitrary subset you can check that $(Z')' \subseteq Z'$, that is, the limit points of Z' are limit points of Z. It then follows from Proposition 5.28 that Z' is closed, since it contains its limit points.

Exercise 5.30. Show that if $S \subseteq X$ and $a \in X$, then $a \in \bar{S}$ if and only if there is a sequence (x_n) in S with $x_n \to a$.

Solution: First suppose that (x_n) is a sequence in S and $x_n \to y$ as $n \to \infty$. Let $M = \{n \in \mathbb{N} : x_n \neq y\}$. If M is infinite then the corresponding subsequence $(x_n)_{n \in M}$ lies in $S\backslash\{y\}$ and clearly converges to y, so that $y \in S'$ by Lemma 5.24. If M is finite, then $x_n = y$ for infinitely many n so certainly $y \in S$. Conversely, if $y \in \bar{S}$ then by Proposition 5.28, either $y \in S$ or $y \in S'$. If $y \in S$ we may take the constant sequence $x_n = y$ while if $y_n \in S'\backslash S$ then we are again done by Lemma 5.24.

Example 5.31. In general, it need *not* be the case that $\bar{B}(a, r)$ is the closure of $B(a, r)$. Since we have seen that $\bar{B}(a, r)$ is closed, it is always true that $\overline{B(a, r)} \subseteq \bar{B}(a, r)$ but the containment can be proper. As a (perhaps silly-seeming) example take any set X with at least two elements equipped with the discrete metric. Then if $x \in X$ we have $\{x\} = B(x, 1)$ is an open set consisting of the single point $\{x\}$. Since singletons are always closed we see that $\overline{B(x, 1)} = B(x, 1) = \{x\}$. On the other hand $\bar{B}(x, 1) = X$ the entire set, which is strictly larger than $\{x\}$ by assumption.

Remark 5.32. Combining the above characterization of closed sets in terms of limit points and the characterization of continuity in terms of closed sets we can give yet another description of continuity for a function $f \colon X \to Y$ between metric spaces: If $Z \subset Y$ is a subset of Y which contains all its limit points then so does $f^{-1}(Z)$. Yet another characterization can be given using the notion of the closure of a set, namely that a function $f\colon X \to Y$ is continuous if and only if for any subset $Z \subseteq X$ we have $f(\overline{Z}) \subseteq \overline{f(Z)}$. It is easy to relate this to the definition of continuity in terms of convergent sequences.

6. Subspaces of metric spaces

If (X, d) is a metric space, then as we noted before, any subset $Y \subseteq X$ is automatically also a metric space since the distance function $d\colon X \times X \to \mathbb{R}_{\geq 0}$ restricts to a distance function on Y. The set Y thus has a topology given by this metric. In this section we show that this topology is easy to describe in terms of the topology on X. The key to this description is the simple observation that the open balls in Y are just the intersection of the open balls in X with Y. For clarity, for $y \in Y \subseteq X$ 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. Thus $B_Y(y, r) = Y \cap B_X(y, r)$.

Lemma 6.1. *If (X, d) is a metric space and $Y \subseteq X$ 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 $U = V \cap Y$. 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. If $U = Y \cap V$ where V is open in X and $y \in U$ then there is a $\delta > 0$ such that $B_X(y, \delta) \subseteq V$. But then $B_Y(y, \delta) = B_X(y, \delta) \cap Y \subseteq V \cap Y = U$ and so U is a neighhourbood of each of its points as required. On the other hand, if U is an open subset of Y then for each $y \in U$ we may pick an open ball $B_Y(y, \delta_y) \subseteq U$. It follows that $U = \bigcup_{y \in U} B_Y(y, \delta_y)$. But then if we set $V = \bigcup_{y \in U} B_X(y, \delta_y)$ it is immediate that V is open in X and $V \cap Y = U$ as required.

The corresponding result for closed sets follows readily: F is closed in Y if and only if $Y \backslash F$ is open in Y which by the above happens if and only if it equals $Y \cap V$ for some open set in X. But this is equivalent to $F = Y \cap V^c$, the intersection of Y with the closed set V^c. $\square$

Remark 6.2. The lemma shows that the topology on X determines the topology on the subspace $Y \subseteq X$ directly. It is easy to see that if $(X, \mathcal{T})$ is an abstract topological space and $Y \subseteq X$ then the collection $\mathcal{T}_Y = \{U \cap Y : U \in \mathcal{T}\}$ is a topology on Y which is called the *subspace topology*.

Remark 6.3. It is important here to note that the property of being open or closed is a relative one – it depends on which metric space you are working in. Thus for example if (X, d) is a metric space and $Y \subseteq X$ then Y is always open viewed as a subset of itself (since the whole space is always an open subset) but it of course need not be an open subset of X! For example, $[0,1]$ is not open in $\mathbb{R}$ but it is an open subset of itself.

Example 6.4. Let's consider a more interesting example: Let $X = \mathbb{R}$ and let $Y = [0,1] \cup [2,3]$. As a subset of Y the set $[0,1]$ is both open and closed. To see that it is open, note that if $x \in [0,1]$ then

$$\begin{aligned} B_Y(x, 1/2) = B_{\mathbb{R}}(x, 1/2) \cap Y &= (x - \frac{1}{2}, x + \frac{1}{2}) \cap ([0,1] \cup [2,3]) \\ &= (x - \frac{1}{2}, x + \frac{1}{2}) \cap [0,1] \subset [0,1], \end{aligned}$$

Similarly we see that $B_Y(x, 1/2) \subseteq [2,3]$ if $x \in [2,3]$ so that $[2,3]$ is also open in Y. It follows $[0,1]$ is both open and closed in Y (as is $[2,3]$).

7. Homeomorphisms and isometries

If (X,d) and (Y,d) are metric spaces it is natural to ask when we wish to consider X and Y equivalent. There is more than one way to answer this question – the first, perhaps most obvious one, is the following:

Definition 7.1. A function $f \colon X \to Y$ between metric spaces (X, d_X) and (Y, d_Y) is said to be an *isometry* if

$$d_Y(f(x), f(y)) = d_X(x,y) \quad \forall x, y \in X$$

An isometry is automatically injective. If there is a surjective (and hence bijective) isometry between two metric spaces X and Y we say that X and Y are *isometric*.

Example 7.2. Let $X = \mathbb{R}^2$ (equipped with the Euclidean metric[13] d_2). The collection of all bijective isometries from X to itself forms a group, the *isometry group* of the plane. Clearly the translations $t_v \colon \mathbb{R}^2 \to \mathbb{R}^2$ are isometries, where $v \in \mathbb{R}^2$ and $t_v(x) = x + v$. Similarly, if $A \in \mathrm{Mat}_2(\mathbb{R})$ is an orthogonal matrix, so that $A^t A = I$, then $x \mapsto Ax$ is an isometry: since $d_2(Ax, Ay) = \|A(x) - A(y)\| = \|A(x-y)\|$ it is enough to check that $\|Ax\| = \|x\|$, but this is clear since $\|Ax\|^2 = (Ax).(Ax) = xA^t Ax = x^t Ix = \|x\|$.

In fact these two kinds of isometries generate the full group of isometries. If $T \colon \mathbb{R}^2 \to \mathbb{R}^2$ is any isometry, let $v = T(0)$. Then $T_1 = t_{-v} \circ T$ is an isometry which fixes the origin. Thus it remains to show that any isometry which fixes the origin is in fact linear. But you showed in Prelims Geometry that any such isometry of $\mathbb{R}^n$ must preserve the inner product (because it preserves the norm and you can express the inner product in terms of the norm). It follows such an isometry takes an orthonormal basis to an orthonormal basis, from which linearity readily follows. (Note that this argument works just as well in $\mathbb{R}^n$.)

[13] Unless it is explicitly stated otherwise, we will always take $\mathbb{R}^n$ to be a metric space equipped with the d_2 metric.

Example 7.3. Let $S^n = \{x \in \mathbb{R}^{n+1} : \|x\|_2 = 1\}$ be the n-sphere (so S^1 is a circle and S^2 is the usual sphere). Clearly $O_{n+1}(\mathbb{R})$ acts by isometries on S^n. In fact you can show that $\mathrm{Isom}(S^n) = O_{n+1}(\mathbb{R})$. To prove this one needs to show that any isometry of S^n extends to an isometry of $\mathbb{R}^{n+1}$ which fixes the origin.

We have already seen that on $\mathbb{R}^n$ the metrics d_1, d_2, d_∞, although different, induce the same notion of convergence and continuity[14] . The notion of isometry is thus in some sense too rigid a notion of equivalence if these are the notions we are primarily interested in. A weaker, but often more useful, notion of equivalence is the following:

Definition 7.4. Let $f\colon X \to Y$ be a continuous function between metric spaces X and Y. We say that f is a *homeomorphism* if there is a continuous function $g\colon Y \to X$ such that $f \circ g = \mathrm{id}_Y$ and $g \circ f = \mathrm{id}_X$. If there is a homeomorphism between two metric spaces X and Y we say they are *homeomorphic*.

Remark 7.5. Note that the definition implies that f is bijective as a map of sets but it is *not* true in general[15] that a continuous bijection is necessarily a homeomorphism. To see this, 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. Its inverse $g\colon Y \to X$ is however not continuous at 1 – the one-sided limits of g as x tends to 1 from above and below are 1 and 2 respectively.

Example 7.6. The closed disk $\bar{B}(0,1)$ of radius 1 in $\mathbb{R}^2$ is homoemorphic to the square $[-1,1] \times [-1,1]$. The easiest way to see this is 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.

The open interval $(-1,1)$ is homeomorphic to $\mathbb{R}$: 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 $\mathbb{R}$. 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 $\mathbb{R}$, (a,b), $(-\infty,a)$ and (a,∞) are all homeomorphic for any $a,b \in \mathbb{R}$ with $a < b$.

---

[14] There is actually a slightly subtle point here – to know that $(\mathbb{R}^n, d_1)$ and $(\mathbb{R}^n, d_2)$ are not isometric we would need to show that there is no bijective map $\alpha\colon \mathbb{R}^n \to \mathbb{R}^n$ such that $d_2(\alpha(x), \alpha(y)) = d_1(x,y)$ for all $x, y \in \mathbb{R}^n$.

[15] This is unlike the examples you have seen in algebra – the inverse of a bijective linear map is automatically linear, and the inverse of a bijective group homomorphism is automatically a homomorphism. Similarly, the inverse of a bijective isometry is also an isometry.

8. Completeness

One of the important notions in Prelims analysis was that of a Cauchy sequence. This is a notion, like convergence, which makes sense in any metric space.

Definition 8.1. Let (X, d) be a metric space. A sequence (x_n) in X is said to be a *Cauchy sequence* if, for any $\epsilon > 0$, there is an $N \in \mathbb{N}$ such that $d(x_n, x_m) < \epsilon$ for all $n, m \geq N$.

The following lemma establishes basic properties of Cauchy sequences in an arbitrary metric space which you saw before for real sequences.

Lemma 8.2. *Let (X, d) be a metric space.*

1. *If (x_n) is a convergent sequence then it is Cauchy.*
2. *Any Cauchy sequence is bounded.*

Proof. Suppose that $x_n \to \ell$ as $n \to \infty$ and $\epsilon > 0$ is given. Then there is an $N \in \mathbb{N}$ such that $d(x_n, \ell) < \epsilon/2$ for all $n \geq N$. It follows that if $n, m \geq N$ we have

$$d(x_n, x_m) \leq d(x_n, \ell) + d(\ell, x_m) < \epsilon/2 + \epsilon/2 = \epsilon,$$

so that (x_n) is a Cauchy sequence as required.

If (x_n) is a Cauchy sequence, then taking $\epsilon = 1$ in the definition, we see that there is an $N \in \mathbb{N}$ such that $d(x_n, x_m) < 1$ for all $n, m \geq N$. It follows that if we set

$$M = \max\{1, d(x_1, x_N), d(x_2, x_N), \ldots, d(x_{N-1}, x_N)\}$$

then for all $n \in \mathbb{N}$ we have $x_n \in B(x_N, M)$ so that (x_n) is bounded as required. $\square$

Part (1) of the lemma motivates the following definition:

Definition 8.3. A metric space (X, d) is said to be *complete* if every Cauchy sequence in X converges.

Example 8.4. One of the main results in Analysis I was that $\mathbb{R}$ is complete, and it is easy to deduce from this that $\mathbb{R}^n$ is complete also (since a sequence in $\mathbb{R}^n$ converges if and only if each of its coordinates converge).

On the other hand, consider the metric space $(0, 1]$: The sequence $(1/n)$ converges in $\mathbb{R}$ (to 0) so the sequence is Cauchy in $\mathbb{R}$ and hence also in $(0, 1]$, however it does not converge in $(0, 1]$.

The previous example suggests a connection between completeness and closed sets. One precise statement of this form is the following:

Lemma 8.5. *Let (X, d) be a complete metric space and let $Y \subseteq X$. Then Y is complete if and only if Y is a closed subset of X.*

Proof. Note that if (x_n) is a Cauchy sequence in Y then it is certainly a Cauchy sequence in X. Since X is complete, (x_n) converges in X, say

$x_n \to a$ as $n \to \infty$. Thus (x_n) converges in Y precisely when $a \in Y$. It follows that Y is complete if and only if it contains the limits of all sequences (x_n) in Y which converge in X. But Lemma 5.30 shows that the set of limits of all sequences in Y is exactly $\bar{Y}$, hence Y is complete if and only if $\bar{Y} \subseteq Y$, that is, if and only if Y is closed.

□

Another useful consequence of completeness is that it guarantees certain intersections of closed sets are non-empty:

Lemma 8.6. *Let (X, d) be a complete metric space and suppose that $D_1 \supseteq D_2 \supseteq \ldots$ form a nested sequence of non-empty closed sets in X with the property that diam$(D_k) \to 0$ as $k \to \infty$. Then there is a unique point $w \in X$ such that $w \in D_k$ for all $k \geq 1$.*

Proof. For each k pick $z_k \in D_k$. Then since the D_k are nested, $z_k \in D_l$ for all $k \geq l$, and hence the assumption on the diameters ensures that (z_k) is a Cauchy sequence. Let $w \in X$ be its limit. Since D_k is closed and contains the subsequence $(z_{n+k})_{n \geq 0}$ it follows $w \in D_k$ for each $k \geq 1$. To see that w is unique, suppose that $w' \in D_k$ for all k. Then $d(w, w') \leq \text{diam}(D_k)$ and since $\text{diam}(D_k) \to 0$ as $k \to \infty$ it follows $d(w, w') = 0$ and hence $w = w'$. □

Remark 8.7. Notice that the property of a metric space being complete is *not* preserved by homeomorphism – we have seen that $(0, 1)$ is homeomorphic to $\mathbb{R}$ but the former is not complete, while the latter is. This is because a homeomorphism does not have to take Cauchy sequences to Cauchy sequences.

Example 8.8. Let $Y = \{z \in \mathbb{C} : |z| = 1\} \backslash \{1\}$. Then Y is homeomorphic to $(0, 1)$ via the map $t \mapsto e^{2\pi i t}$, but their respective closures $\bar{Y}$ and $[0, 1]$ however are not homeomorphic. (We will seem a rigorous proof of this later using the notion of connectedness.) The metric spaces Y and $(0, 1)$ contain information about their closures in $\mathbb{R}^2$ which is lost when we only consider the topologies the metrics give: the space Y has Cauchy sequences which don't converge in Y, but these all converge to $1 \in \mathbb{C}$, whereas in $(0, 1)$ there are two kinds of Cauchy sequences which do not converge in $(0, 1)$ – the ones converging to 0 and the ones converging to 1. The point here is that given two Cauchy sequences we can detect if they converge to the same limit without knowing what that the limit actually is: (x_n) and (y_n) converge to the same limit if for all $\epsilon > 0$ there is an $N \in \mathbb{N}$ such that $d(x_n, y_n) < \epsilon$ for all $n \geq N$. Using this idea one can define what is called the *completion* of a metric space (X, d): this is a complete metric space (Y, d) such which X embeds isometrically into as a dense[16] subset. For example, the real numbers $\mathbb{R}$ are the completion of $\mathbb{Q}$.

Many naturally arising metric spaces are complete. We now give a important family of such: recall that if X is any set, the space $\mathcal{B}(X)$ of bounded

[16]that is, Y is the closure of X.

real-valued functions on X is normed vector space where if $f \in \mathcal{B}(X)$ we define its norm to be $\|f\|_\infty = \sup_{x\in X}|f(x)|$.

Theorem 8.9. *Let X be a set. The normed vector space $(\mathcal{B}(X), \|.\|_\infty)$ is complete.*

Proof. Let $(f_n)_{n\geq 1}$ be a Cauchy sequence in $\mathcal{B}(X)$. Then we have for each $x \in X$
$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\|_\infty \to 0,$$
as $n, m \to \infty$. It follows that the sequence $(f_n(x))$ is a Cauchy sequence of real numbers and hence since $\mathbb{R}$ is complete it converges to a real number. Thus we may define a function $f\colon X \to \mathbb{R}$ by setting $f(x) = \lim_{n\to\infty} f_n(x)$.

We claim $f_n \to f$ in $\mathcal{B}(X)$. Note that this requires us to show both that $f \in \mathcal{B}(X)$ and $f_n \to f$ with respect to the norm $\|.\|_\infty$. To check these both hold, fix $\epsilon > 0$. Since (f_n) is Cauchy, we may find an $N \in \mathbb{N}$ such that $\|f_n - f_m\|_\infty < \epsilon$ for all $n, m \geq N$. Thus we have for all $x \in X$ and $n, m \geq N$
$$|f_n(x) - f_m(x)| \leq \|f_n - f_m\| < \epsilon.$$
But now letting $n \to \infty$ we see that for any $m \geq N$ we have $|f(x) - f_m(x)| \leq \epsilon$ for all $x \in X$. But then for any such m we certainly have $f - f_m \in \mathcal{B}(X)$ so that[17] $f = f_m + (f - f_m) \in \mathcal{B}(X)$, and since $\|f - f_m\|_\infty \leq \epsilon$ for all $m \geq N$ it follows $f_m \to f$ as $m \to \infty$ as required.

□

As we already observed, if X is also a metric space then we can also consider the space of bounded continuous functions $\mathcal{C}_b(X)$ on X. This is a normed subspace of $\mathcal{B}(X)$, and the following theorem is a generalization of the result you saw last year showing that a uniform limit of continuous functions is continuous (the proof is essentially the same also).

Theorem 8.10. *Let (X, d) be a metric space. The space $\mathcal{C}_b(X)$ is a complete normed vector space.*

Proof. Since we have shown in Theorem 8.9 that $\mathcal{B}(X)$ is complete, by Lemma 8.5 we must show that $\mathcal{C}_b(X)$ is a closed subset of $\mathcal{B}(X)$. Let (f_n) be a Cauchy sequence of bounded continuous functions on X. By Theorem 8.9 this sequence converges to a bounded function $f\colon X \to \mathbb{R}$. We must show that f is continuous. Suppose that $a \in X$ and let $\epsilon > 0$. Then since $f_n \to f$ there is an $N \in \mathbb{N}$ such that $\|f - f_n\|_\infty < \epsilon/3$ for all $n \geq N$. Moreover, if we fix $n \geq N$ then since f_n is continuous, there is a $\delta > 0$ such that $|f_n(x) - f_n(a)| < \epsilon/3$ for all $x \in B(a, \delta)$. But then for $x \in B(a, \delta)$ we have
$$\begin{aligned}|f(x) - f(a)| &\leq |f(x) - f_n(x)| + |f_n(x) - f_n(a)| + |f_n(a) - f(a)| \\ &< \epsilon/3 + \epsilon/3 + \epsilon/3 = \epsilon.\end{aligned}$$
It follows that f is continuous at a, and since a was arbitrary, f is a continuous function as required.

□

[17] Recall from Lemma 3.8 that $\mathcal{B}(X)$ is a vector space!

Remark 8.11. If X and Y are metric spaces, as we saw in Example 2.12, one can also consider the space $\mathcal{B}(X,Y)$ of bounded functions from X to Y, that is, functions $f\colon X \to Y$ such that $f(X)$ is a bounded subset of Y, along with its subspace $\mathcal{C}_b(X,Y)$ of bounded continuous functions. These are no longer normed vector spaces, but they are both complete metric spaces provided Y is, as you are asked to show in the second problem sheet.

Lemma 8.12. *("Weierstrass M-test"): Let X be a metric space. Suppose that (f_n) is a sequence in $\mathcal{C}_b(X)$ and $(M_n)_{n\geq 0}$ is a sequence of non-negative real numbers such that $\|f_n\|_\infty \leq M_n$ for all $n \in \mathbb{Z}_{\geq 0}$ and $\sum_{n\geq 0} M_n$ exists. Then the series $\sum_{n\geq 0} f_n$ converges in $\mathcal{C}_b(X)$.*

Proof. Let $S_n = \sum_{k=0}^{N} f_k$ be the sequence of partial sums. Since we know $\mathcal{C}_b(X)$ is complete, it suffices to prove that the sequence $(S_n)_{m\geq 0}$ is Cauchy. But if $n \leq m$ then we have

$$\|S_m - S_n\| \leq \sum_{k=n+1}^{m} \|f_k\| \leq \sum_{k=n+1}^{m} M_k,$$

and since $\sum_{k\geq 0} M_k$ converges, the sum $\sum_{k=n+1}^{m} M_k$ tends to zero as $m,n \to \infty$ as required. □

Finally, we conclude this section with a theorem which is extremely useful, and is a natural generalization of a result you saw last year in constructive mathematics. We first need some terminology:

Definition 8.13. Let (X,d) and (Y,d) 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 $K \geq 0$ such that

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

If $Y = X$ and $K \in [0,1)$ then we say that f is a *contraction mapping* (or simply a *contraction*). Any Lipschitz map is continuous, and in fact uniformly continuous, as is easy to check.

The reason for the restriction of the term contraction to maps from a space to itself is the following theorem. The result is a broad generalization of a result you saw before in the Constructive Mathematics course in Prelims, which you will also see put to good use in the Differential Equations course this term.

Theorem 8.14. *Let (X,d) 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 $z \in X$ such that $f(z) = z$.*

Proof. If $y_1, y_2 \in X$ are such that $f(y_1) = y_1$ and $f(y_2) = y_2$ we have $d(y_1,y_2) = d(f(y_1), f(y_2)) \leq Kd(y_1,y_2)$ so that $(1-K)d(y_1,y_2) \leq 0$. Since $K \in [0,1)$ and the function d is nonnegative this is possible only if $d(y_1,y_2) = 0$ and hence $y_1 = y_2$. It follows that f has at most one fixed point.

To see that f has a fixed point, fix $a \in X$ and consider the sequence defined by $x_0 = a$ and $x_n = f(x_{n-1})$ for $n \geq 1$. We claim that (x_n) converges and that its limit z is the unique fixed point of f. Indeed if $x_n \to z$ as $n \to \infty$ then since f is continuous we have

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

Thus z is indeed a fixed point. Thus it remains to show that (x_n) is convergent. Since (X, d) is complete, we need only show that (x_n) is Cauchy. To see this this note first that for $n \geq 1$ we have $d(x_n, x_{n-1}) \leq K^{n-1} d(f(a), a)$ (by induction). But then if $n \geq m$ by the triangle inequality we have

$$\begin{aligned} d(x_n, x_m) &\leq \sum_{k=1}^{n-m} d(x_{m+k}, x_{m+k-1}) \leq d(a, f(a))K^m \sum_{k=1}^{n-m} K^{k-1} \\ &\leq \frac{d(a, f(a))}{1-K} K^m, \end{aligned}$$

which clearly tends to 0 as $n, m \to \infty$. It follows (x_n) is a Cauchy sequence as required. □

Remark 8.15. This theorem is important not just for the statement, but because the proof shows us how to find the fixed point! (Or rather, at least how to approximate it). The sequence (x_n) in the proof converges to the fixed point, and in fact does so quickly – if we start with an initial guess a, and z is the actual fixed point, then $d(x_n, z) \leq K^n.d(a, z)$.

Remark 8.16. It is worth checking to what extent the hypotheses of the theorem are necessary. One might think of a weaker notion of contraction, for example: if $f\colon X \to X$ has the property that $d(f(x), f(y)) < d(x, y)$ for all $x, y \in X$ then it is easy to see that f has at most one fixed point, but the example $f\colon [1, \infty) \to [1, \infty)$ where $f(x) = x + 1/x$ shows that such a map need not have any fixed points.

The requirement that X is complete is also clearly necessary: if $f\colon (0, 1) \to (0, 1)$ is given by $f(x) = x/2$ clearly f is a contraction, but f has no fixed points in $(0, 1)$.

9. Connected sets

In this section we try to understand what makes a space "connected". There are in fact more than one approaches one can take to this question. We will consider two, and show that for reasonably nice spaces the two notions in fact coincide[18].

The first definition we make tries to capture the fact that the space should not "fall apart" into separate pieces. Since we can always write a space with more than one element as a disjoint union of two subsets, we must take

[18]In particular, for the open subsets of the complex plane which are the sets we will be most interested in for second part of the course, the two notions will coincide, but both characterizations of connectedness will be useful.

into account the metric, or at least the topology, of our space in making a definition.

Example 9.1. Let $X = [0,1]$ and let $A = [0,1/2)$ and $B = [1/2,1]$. Then clearly $X = A \cup B$ so that X can be divided into two disjoint subsets. However, the point $1/2 \in B$ has points in A arbitrarily close to it, which, intuitively speaking, is why it is "glued" to A.

This suggests that we might say that a decomposition of metric space X into two subsets A and B might legitimately show X to be disconnected if no point of A was a limit point of B and vice versa. This is precisely the content of our definition.

Definition 9.2. Suppose that (X,d) is a metric space. We say that X is *disconnected* if we can write $X = U \cup V$ where U and V are nonempty open subsets of X and $U \cap V = \emptyset$. We say that X is *connected* if it is not disconnected.

Note that if $X = U \cup V$ and U and V are both open and disjoint, then $U = V^c$ is also closed, as is V. Thus U and V also contain all of their limit points, so that no limit point of A lies in B and vice versa.

Remark 9.3. Note that if (X,d) is a metric space and $A \subseteq X$, then the condition that A is connected can be rewritten as follows: if U, V are open in X and $U \cap V \cap A = \emptyset$ then whenever $A \subseteq U \cup V$, either $A \subseteq U$ or $A \subseteq V$.

As the previous remark shows, there are a few ways of expressing the above definition which are all readily seen to be equivalent. We record the most common in the following lemma.

Lemma 9.4. *Let (X,d) be a metric space. The following are equivalent.*

(1) *X is connected.*
(2) *If $f\colon X \to \{0,1\}$ is a continuous function then f is constant.*
(3) *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 $\mathbb{R}$, or equivalently with the discrete metric.)

Proof. (1) $\iff$ (2): Let $f\colon X \to \{0,1\}$ be a continuous function. Then since the singleton sets $\{0\}$ and $\{1\}$ are both open in $\{0,1\}$ each of $f^{-1}(0)$ and $f^{-1}(1)$ are open subsets of X which are clearly disjoint. It follows if X is connected then one must be the empty set, and hence f is constant as required. Conversely, if X is not connected then we may write $X = A \cup B$ where A and B are nonempty disjoint open sets. But then the function $f\colon X \to \{0,1\}$ which is 1 on A and 0 on B is non-constant and by the characterization of continuity in terms of open sets, f is clearly continuous.

(1) $\iff$ (3): If X is disconnected then we may write $X = A \cup B$ where A and B are disjoint nonempty open sets. But then $A^c = B$ so that A is closed (as is $B = A^c$) so that A and B proper sets of X which are both open and closed. Conversely, if A is a proper subset of X which is closed and

open then A^c is also a proper subset which is both closed and open so that the decomposition $X = A \cup A^c$ shows that X is disconnected. $\square$

Example 9.5. If $X = [0,1] \cup [2,3] \subset \mathbb{R}$ then we have seen that both $[0,1]$ and $[2,3]$ are open in X, hence since X is their disjoint union, X is not connected.

Lemma 9.6. *Let (X,d) be a metric space.*

$i)$ 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.
$ii)$ If $A \subseteq X$ is connected then if B is such that $A \subseteq B \subseteq \bar{A}$, the set B is also connected.
$iii)$ If $f\colon X \to Y$ is continuous and $A \subseteq X$ is connected then $f(A) \subseteq Y$ is connected.

Proof. For the first part, 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.

For the second part, consider B such that $A \subseteq B \subseteq \bar{A}$, and 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 $A \cap U \cap V = \emptyset$, so that $A \subseteq U$ or $A \subseteq V$. By symmetry we may assume $A \subseteq U$. But then $A \subseteq V^c$ since $A \cap U \cap V = \emptyset$ and since V^c is closed $B \subseteq \bar{A} \subseteq V^c$, and hence $B \subseteq U$, hence B is connected.

For the final part, note that since f is continuous, if $f(A) \subseteq U \cup V$ for U and V open in Y with $U \cap V \cap f(A) = \emptyset$, then $A \subset f^{-1}(U) \cup f^{-1}(V)$, $f^{-1}(U) \cap f^{-1}(V) \cap A = \emptyset$ and $f^{-1}(U), f^{-1}(V)$ are open in X. Since A is connected it must lie entirely in one of $f^{-1}(U)$ or $f^{-1}(V)$ and hence $f(A)$ must lie entirely in U or V as required. $\square$

Remark 9.7. Notice that $iii)$ in the previous Lemma implies that if X and Y are homeomorphic, they if X is connected so is Y, and vice versa. Note also that $iii)$ allows us to generalize the characterization of connectedness in terms of functions to the set $\{0,1\}$. We say that a metric (or topological) space is *discrete* if every point is an open set. It is easy to see that the connected subsets of a discrete metric space are precisely the singleton sets, thus any continuous function from a connected set to a discrete set must be constant. This applies for example to sets such as $\mathbb{N}$ and $\mathbb{Z}$, which will be very useful for us later in the course.

Definition 9.8. Part $i)$ of Lemma 9.6 has an important consequence: if (X,d) is a metric space and $x_0 \in X$, then the set of connected subsets of X which contain x_0 is closed under unions, that is, if $\{C_i : i \in I\}$ is any collection of connected subsets containing x_0 then $\bigcup_{i\in I} C_i$ is again a connected subset containing x_0. This means that

$$C_{x_0} = \bigcup_{\substack{C \subseteq X \text{ connected}, \\ x_0 \in C}} C,$$

is the largest[19] connected subset of X which contains x_0, in the sense that any connected subset of X which contains x_0 lies in C_{x_0}. It is called the *connected component* of X containing x_0. The space X is the disjoint union of its connected components.

9.1. **Connected sets in $\mathbb{R}$.**

Proposition 9.9. *The real line $\mathbb{R}$ is connected.*

Proof. Let U and V be open subsets of $\mathbb{R}$ such that $\mathbb{R} = U \cup V$ and $U \cap V = \emptyset$. Suppose for the sake of a contradiction that both U and V are non-empty so that we may pick $x \in U$ and $y \in V$. By symmetry we may assume that $x < y$ (since $U \cap V = \emptyset$ we cannot have $x = y$). Since $[x, y]$ is bounded and $x \in U$, if we let $S = \{z \in [x, y] : z \in U\}$, then $c = \sup(S)$ exists, and certainly $c \in [x, y]$. If $c \in U$ then $c \neq y$ and as U is open there is some $\epsilon_1 > 0$ such that $B(c, \epsilon_1) \subseteq U$. Thus if we set $\delta = \min\{\epsilon_1/2, (y - c)/2\} > 0$ we have $c + \delta \in U \cap [x, y]$ contradicting the fact that c is an upper bound for S. Similarly if $c \in V$ then there is an $\epsilon_2 > 0$ such that $B(c, \epsilon_2) \subseteq V$. But then $\emptyset = (c - \epsilon_2, c] \cap U \supseteq (c - \epsilon_2, c] \cap S$, so that $c - \epsilon_2$ is an upper bound for S, contradiction the fact that c is the least upper bound of S. It follows that one of U or V is the empty set as required. $\square$

Corollary 9.10. *The real line $\mathbb{R}$, every half-line (a, ∞), $(-\infty, a)$, $[a, \infty)$ or $(-\infty, a]$ and any interval are all connected subsets of $\mathbb{R}$.*

Proof. The previous proposition establishes that $\mathbb{R}$ is connected, and since we say in Example 7.6 that every open interval (a, b) or open half-line (a, ∞), $(-\infty, a)$ is homeomorphic to $\mathbb{R}$ they are also connected. The remaining cases the follow from part $ii)$ of Lemma 9.6. $\square$

Exercise 9.11. Show that any interval or half-line is homeomorphic to one of $[0, 1]$, $[0, 1)$ or $(0, 1)$.

Lemma 9.12. *Suppose that $A \subset \mathbb{R}$ is a connected set. Then A is either $\mathbb{R}$, an interval, or a half-line. Thus these are precisely the connected subsets of $\mathbb{R}$.*

Proof. Suppose that $x, y \in A$ and $x < y$. We claim that $[x, y] \subseteq A$. Indeed if this is not the case then there is some c with $x < c < y$ and $c \notin A$. But then $A = \big(A \cap (-\infty, c)\big) \cup \big((A \cap (c, \infty)\big)$ so that A is not connected.

[19] This is the analogous to the definition of the interior of a subset S of X, which is the largest open subset of X contained in S.

If we let $\sup(A) = +\infty$ if A is not bounded above and $\inf(A) = -\infty$ if A is not bounded below, then by the approximation property it follows that

$$(\inf(A), \sup(A)) = \bigcup_{\substack{x,y \in A \\ x \leq y}} [x, y] \subseteq A,$$

so that A is an interval or half-line as required. (The $\inf(A)$ and $\sup(A)$ may or may not lie in A, leading to open, closed, or half-open intervals and open or closed half-lines.) □

Proposition 9.13. *(Intermediate Value Theorem.) Let $f\colon [a, b] \to \mathbb{R}$ be a continuous function. Then the image of f is an interval in $\mathbb{R}$. In particular, f takes every value between $f(a)$ and $f(b)$.*

Proof. Since $[a, b]$ is connected, its image must be connected, and hence by the above it is an interval. The in particular claim follows. □

Remark 9.14. Note that for the Intermediate Value Theorem we only needed to know that $[a, b]$ was connected and that a connected subset A of $\mathbb{R}$ has the property that if $x \leq y$ lie in A then $[x, y] \subseteq A$.

9.2. **Path connectedness.** A quite different approach to connectedness might start assuming that, whatever a connected set should be, the closed interval should be one[20].

Definition 9.15. Let (X, d) be a metric space. A *path* in X is a continuous function $\gamma\colon [a, b] \to X$ where $[a, b]$ is any non-empty closed interval. If $x, y \in X$ then we say there is a path between x and y if there is a path $\gamma\colon [a, b] \to X$ such that $\gamma(a) = x$ and $\gamma(b) = y$. We say that the metric space X is *path-connected* if there is a path between any two points in X. Note that since every closed interval $[a, b]$ is homeomorphic to $[0, 1]$ one can equivalently require that paths are continuous functions $\gamma\colon [0, 1] \to X$. In the subsequent discussion we will, for convenience, impose this condition.

There are a number of useful 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

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

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

Definition 9.16. There is a notion of *path-component* for a metric space: Let us define a relation on points in X as follows: Say $x \sim y$ if there is a path from x to y in X. The constant path $\gamma(t) = x$ (for all $t \in [0, 1]$) shows that

[20]Since we've seen that the closed interval is connected according to our previous definition, it shouldn't be too surprising that we will readily be able to see our second notion of connectedness implies the first. The subtle point will be that it is actually in general a strictly *stronger* condition.

this relation is reflexive. If γ is a path from x to y then γ^- is a path from y to x, so the relation is symmetric. Finally if γ_1 is a path from x to y and γ_2 is a path from y to z then $\gamma_1 \star \gamma_2$ is a path from x to z, so the relation is transitive. It follows that $\sim$ is an equivalence relation and its equivalence classes, which partition X, are known as the *path components* of X.

We now relate the two notions of connectedness.

Proposition 9.17. *Let (X, d) be a metric space. If X is path-connected then it is connected. If X is an open subset of V where V is a normed vector space, then X is path-connected if it is connected.*

Proof. Suppose that X is path-connected. To see X is connected we use the characterization of connectedness in terms of functions to $\{0, 1\}$. Consider such a function $f \colon X \to \{0, 1\}$. We wish to show that f is constant, that is, we need to show that if $x, y \in X$ then $f(x) = f(y)$. But Z is path-connected, so there is a path $\gamma \colon [0, 1] \to X$ such that $\gamma(0) = x$ and $\gamma(1) = y$. But then $f \circ \gamma$ is a continuous function from the connected set $[0, 1]$ to $\{0, 1\}$ so that $f \circ \gamma$ must be constant. But then $f(x) = f \circ \gamma(0) = f \circ \gamma(1) = f(y)$ as required.

Now suppose that X is open in V where V is a normed vector space. Let x_0 be a point in X and let P be its path component. Then if $v \in P$, since X is open, there is an open ball $B(v, r) \subseteq Z$. Given any point w in $B(v, r)$ we have the path $\gamma_w(t) = tw + (1 - t)v$ from v to w, and hence concatenating a path from x_0 to v with γ_v we see that w lies in P. It follows that $B(v, r) \subseteq P$ so that P is open in V. But since X is the disjoint union of its path components, it follows that if Z is connected it must have at most one path-component and so is path-connected as required. □

Remark 9.18. Note that it is easy to see that if (X, d) is path-connected and $f \colon X \to Y$ is continuous, then the image of X under f is a path-connected subset of Y: if $y_1 = f(x_1)$ and $y_2 = f(x_2)$ are in the image of f, then if we pick a path $\gamma \colon [0, 1] \to X$ from x_1 to x_2 in X, clearly $f \circ \gamma$ is a path from y_1 to y_2 in $f(X)$.

Example 9.19. In general it is not true that a connected set need be path-connected. One reason the two notions differ is because, as well as being connected, the closed interval is what is known as *compact*, a notion we will examine shortly. One consequence of this is that if (X, d) is a metric space and $A \subset X$ is a path-connected subspace then $\bar{A}$, the closure of A need *not* be path-connected, despite the fact that we have already seen that it must be connected.

Consider the subset $A \subseteq \mathbb{R}^2$ given by

$$A = \{(t, \sin(1/t) : t \in (0, 1]\}.$$

Since A is clearly the image of $(0, 1]$ under a continuous map, it is a connected subset of $\mathbb{R}^2$, and hence its closure $\bar{A} = A \cup (\{0\} \times [-1, 1])$ is also connected. We claim however that $\bar{A}$ is *not* path-connected. To see informally why this is the case, suppose $\gamma\colon [0,1] \to \mathbb{R}^2$ has a path from $(1, \sin(1))$ to $(0, 1)$. Then the first and second coordinates $x(t)$ and $y(t)$ of γ are continuous functions on a closed interval, so they are uniformly continuous. By the intermediate value theorem $x(t)$ must take every value between 1 and 0, but then $y(t)$ must oscillate between -1 and 1 infinitely often which violates uniform continuity.

10. Compactness

One of the most fundamental theorems in Prelims Analysis was the Bolzano-Weierstrass theorem on bounded sequences of real numbers. It is the key technical ingredient in a number of the main theorems in the whole sequence – the completeness of the reals, the fact that a continuous function on a closed interval is bounded and attains its bounds, the equivalence of continuity and uniform continuity for functions on a closed interval all rely on it.

In this section we study metric spaces in which the conclusion of the Bolzano-Weierstrass theorem holds, and show that not only do many of the results from Prelims which relied on the Bolanzo-Weierstrass theorem extend to these metric spaces (which is perhaps unsurprising) but also that the class of such spaces is quite rich – it includes for example all closed bounded subsets of $\mathbb{R}^n$ for any n.

Definition 10.1. Let (X, d) be a metric space. We say that X is (*sequentially*[21]) *compact* if any sequence $(x_n)_{n\geq 1}$ in X contains a subsequence $(x_{n_k})_{k\geq 1}$ for which there exists an $\ell \in X$ with $x_{n_k} \to \ell$ as $k \to \infty$.

Example 10.2. You saw last year that any bounded sequence of real numbers contains a convergent subsequence. This readily implies that any closed interval $[a, b] \subset \mathbb{R}$ is compact: Indeed if (x_n) is a sequence in $[a, b]$ then clearly it is bounded, so it contains a convergent subsequence (x_{n_k}), say $x_{n_k} \to \ell$ as $k \to \infty$. But since limits preserve weak inequalities (or in the language we have now developed, $[a, b]$ is a closed subset of $\mathbb{R}$ and so contains its limit points) we must have $\ell \in [a, b]$ and hence $[a, b]$ is compact.

It is also easy to see that $(a, b]$, $[a, b)$ and (a, b) are *not compact* when $b > a$: Take $(a, b]$ for example: a tail of the sequence $(a + 1/n)_{n\geq 1}$ will lie in $(a, b]$ and any subsequence of it will converge to $a \notin (a, b]$ since $(a + 1/n)_{n\geq 1}$ does, thus $(a + 1/n)_{n\geq 1}$ has no subsequence which converges in $(a, b]$.

We now establish some basic properties of compact metric spaces:

Lemma 10.3. *Let (X, d) be a metric space and suppose $Z \subseteq X$ is a subspace.*

(1) *If Z is compact then Z is closed and bounded.*

---

[21]The word "compact" is in general used for a notion which is discussed in Section 11. For metric spaces the two notions are equivalent. [Aside: the two notions make sense for arbitrary topological spaces, where they turn out *not* to be equivalent.]

(2) *If X is compact and Z is closed in X then Z is compact.*

Proof. Suppose that Z is compact in X. If $a \in X$ is a limit point of Z then there is a sequence (z_n) in Z which converges to a. Since Z is compact, the sequence (z_n) has a subsequence (z_{n_k}) which converges in Z. But since the limit of a subsequence of a convergent sequence is just the limit of the original sequence we have

$$a = \lim_{n\to\infty} z_n = \lim_{k\to\infty} z_{n_k} \in Z.$$

Thus Z contains all its limit points and hence Z is closed. Next suppose that Z is unbounded in X. Then picking $z_0 \in Z$ we may find $z_n \in Z$ with $d(z_0, z_n) \geq n$ for each $n \in \mathbb{N}$. But then if (z_n) had a convergent subsequence (z_{n_k}) say $z_{n_k} \to b \in Z$ then we would have $d(z_{n_k}, z_0) \geq n_k \geq k$ and also $d(z_{n_k}, z_0) \to d(b, z_0)$, which is a contradiction, since a convergent sequence of real numbers must be bounded.

Now suppose that X is compact and Z is closed in X. Then if (z_n) is a sequence in Z, since X is compact it has a convergent subsequence (z_{n_k}) tending to $c \in X$ say. But then c is a limit point of Z and since Z is closed $c \in Z$, so that (z_n) has a convergent subsequence in Z as required.

□

The next Lemma essentially shows that compactness, like connectedness, is a topological property:

Lemma 10.4. *Let (X, d) and (Y, d) be metric spaces and suppose that $f\colon X \to Y$ is continuous. Then if X is compact, $f(X)$ is a compact subspace of Y. In particular, if X is compact and $f\colon X \to \mathbb{R}$ is continuous, then f is bounded and attains its bounds.*

Proof. Suppose that (y_n) is a sequence in $f(X) \subseteq Y$. Then for each n we may pick an $x_n \in X$ such that $f(x_n) = y_n$. Since X is compact 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 $y_{n_k} = f(x_{n_k}) \to f(a) \in f(X) \subseteq Y$, so that (y_n) has a convergent subsequence whose limit lies in $f(X)$ as required.

For the final sentence, note that $f(X)$ is a compact subset of $\mathbb{R}$ and hence by Lemma 10.3 it is closed and bounded. But this precisely means that the image of f is bounded and attains its bounds as required. □

Remark 10.5. The previous Lemma also shows that compactness is a property which is preserved by homoeomorphisms: If $f\colon X \to Y$ is a continuous bijection with $g\colon Y \to X$ its continuous inverse, then if X is compact $f(X) = Y$ must be compact, while conversely if Y is compact then $X = g(Y)$ must be compact.

Theorem 10.6. *Let $f\colon X \to Y$ be a continuous function and suppose that X is a compact metric space. Then f is uniformly continuous.*

Proof. Suppose for the sake of a contradiction that f is not uniformly continuous. Then there exists some $\epsilon > 0$ such that for each $n \in \mathbb{N}$ we may find $a_n, b_n \in X$ such that $d(a_n, b_n) < 1/n$ but $d(f(a_n), f(b_n)) \geq \epsilon$. Now since X is compact, (a_n) contains a convergent subsequence, (a_{n_k}) say, and since $d(a_{n_k}, b_{n_k}) \leq 1/n_k \leq 1/k$ it follows $\lim_{k\to\infty} a_{n_k} = \lim_{k\to\infty} b_{n_k} = c$ say. But since f is continuous at c there is a $\delta > 0$ such that for all $x \in X$ with $d(c,x) < \delta$, we have $d(f(c), f(x)) < \epsilon/2$. As both (a_{n_k}) and (b_{n_k}) tend to c, for all sufficiently large k we will have $d(c, a_{n_k}), d(c, b_{n_k}) < \delta$ and hence

$$\epsilon \leq d(f(a_{n_k}), f(b_{n_k})) \leq d(f(a_{n_k}), f(c)) + d(f(c), f(b_{n_k})) < \epsilon/2 + \epsilon/2 < \epsilon,$$

which is a contradiction. Thus f must be uniformly continuous as required. □

10.1. **Compactness and products: a generalization of the Bolzano-Weierstrass theorem.** Recall from Example 2.11 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 by setting

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

Example 10.7. Writing $\mathbb{R}^n = \mathbb{R}^{n-1} \times \mathbb{R}$ this gives us an inductive definition of a metric on $\mathbb{R}^n$. Check that the metric one obtains is the metric d_∞. Since we know this metric is equivalent to the metrics d_1 and d_2 if we can characterize the compact subsets of $\mathbb{R}^n$ equipped with the metric $d = d_\infty$ then we also characterize the compact subsets of $\mathbb{R}^n$ with respect to either d_1 and d_2.

Using the above definition of a metric on products of metric spaces makes the following result easy to check:

Lemma 10.8. *Let X and Y be metric spaces. A sequence $((x_n, y_n))_{n\geq 1}$ in $X \times Y$ converges if and only if (x_n) converges in X and (y_n) converges in Y.*

Proof. It is clear from the definitions that the projection maps $p_X \colon X \times Y \to X$ and $p_Y \colon X \times Y \to Y$ are continuous (in fact they are Lipschitz continuous with Lipschitz constant 1). It follows that if (x_n, y_n) converges in $X \times Y$ then (x_n) and (y_n) must converge.

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

$$d((x_n, y_n), (a, b)) = \max\{d(x_n, a), d(y_n, b)\} \to 0$$

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

Proposition 10.9. *Let X and Y be compact metric spaces. Then $X \times Y$ is compact.*

Proof. Let (x_n, y_n) be a sequence in $X \times Y$. As X is compact, the sequence (x_n) in X has a convergent subsequence (x_{n_k}), say $x_{n_k} \to a \in X$ as $k \to \infty$. But then consider the sequence (y_{n_k}) in Y. Since Y is compact this in turn has a convergent subsequence $(y_{n_{k_r}})_{r\geq 1}$, say $y_{n_{k_r}} \to b \in Y$. But since $(x_{n_{k_r}})$ is a subsequence of x_{n_k} is also converges to a and hence by the previous Lemma $(x_{n_{k_r}}, y_{n_{k_r}}) \to (a, b)$ and (x_n, y_n) has a convergent subsequence as required. □

It is now easy to give a generalisation of the Bolzano-Weierstrass theorem to $\mathbb{R}^n$.

Theorem 10.10. *(Bolzano-Weierstrass in $\mathbb{R}^n$). A subset $X \subseteq \mathbb{R}^n$ is compact if and only if it is closed and bounded.*

Proof. We have already seen in Lemma 10.3 that if X is compact in $\mathbb{R}^n$ then it must be closed and bounded, thus it remains to show that any such set is compact. But if X is bounded then there is an $R > 0$ such that[22] $X \subseteq B(0, R) = [-R, R]^n$. Now by the Bolzano-Weierstrass theorem for $\mathbb{R}$, any closed interval such as $[-R, R]$ is compact. But then using Proposition 10.9 and induction it follows readily that $[-R, R]^n$ is compact, but then again by Lemma 10.3 it follows that X, being a closed subset of a compact metric space, is compact as required. □

Remark 10.11. Note that in a general metric space X, a closed bounded subset of X need *not* be compact. An example of this is given by taking $\mathcal{C}_b(\mathbb{R})$ the normed space of continuous bounded functions on the real line equipped with $\|.\|_\infty$ the supremum metric. If we let

$$f(t) = \begin{cases} 2t, & 0 \leq t \leq 1/2; \\ 2(1-t), & 1/2 \leq t \leq 1 \end{cases}$$

and set $f_n(t) = f(t+n)$ the each f_n is bounded and in fact has $\|f_n\|_\infty = 1$, so that they all lie in $\bar{B}(0, 1)$. However, if $n \neq m$ it is easy to see that $\|f_n - f_m\|_\infty = 1$, so that (f_n) has no convergent subsequence and thus $\bar{B}(0, 1)$ is not compact, despite clearly being closed and bounded in $\mathcal{C}_b(\mathbb{R})$.

10.2. **Boundedness, completeness and compactness.** In a general metric space the property of being bounded is much weaker than one's instincts initially imagine. One can show for example that any metric space is homeomorphic to a metric space which is bounded. There is however a property stronger than boundedness which is often more useful:

Definition 10.12. A metric space X is said to be *totally bounded* if, given any $\epsilon > 0$ there is a finite set $\{x_1, x_2, \ldots, x_n\}$ in X such that $X = \bigcup_{i=1}^n B(x_i, \epsilon)$.

Lemma 10.13. *Let X be a compact metric space. Then X is totally bounded.*

Proof. Suppose that $r > 0$ is given and that, for the sake of a contradiction, no such set S exists. We claim there exists a sequence (a_i) in X such that $d(a_i, a_j) \geq r$ for every $i \neq j$. Indeed suppose we have $\{a_1, \ldots, a_n\}$ such that $d(a_i, a_j) \geq r$ whenever $1 \leq i \neq j \leq n$ (one can begin with the empty set). Our assumption that the union of any finite collection of open r-balls cannot cover X, implies that there must exist an a_{n+1} such that $d(a_{n+1}, a_i) \geq r$ for all i, ($1 \leq i \leq n$), and hence we may construct the sequence (a_i) inductively as required. But any such sequence clearly cannot contain a convergent subsequence, and hence we have a contradiction.

□

[22] Recall that the "open balls" in the d_∞ metric are hypercubes.

Proposition 10.14. *Let X be a compact metric space. Then X is complete.*

Proof. Suppose that (x_n) is a Cauchy sequence in X. Since X is compact, (x_n) has a convergent subsequence (x_{n_k}) say, so that $x_{n_k} \to a \in X$ as $k \to \infty$. We claim that $x_n \to a$ as $n \to \infty$. Indeed given $\epsilon > 0$ there is some $N \in \mathbb{N}$ such that for all $n, m \geq N$ we have $d(x_n, x_m) < \epsilon/2$. Now since $x_{n_k} \to a$ as $k \to \infty$ we may find a K such that $d(x_{n_k}, a) < \epsilon/2$ for all $k \geq K$ and $n_K > N$. But then if $n \geq N$ we have

$$d(x_n, a) \leq d(x_n, x_{n_K}) + d(x_{n_K}, a) < \epsilon/2 + \epsilon/2 = \epsilon,$$

as required. □

Remark 10.15. We have shown that if X is a compact metric space then it is complete and totally bounded. In fact any complete and totally bounded metric space is compact as we will now show.

Lemma 10.16. *Let X be a totally bounded metric space and suppose that (x_n) is a sequence in X. Then (x_n) has a subsequence which is a Cauchy sequence.*

Proof. Since X is totally bounded, for every $n \in \mathbb{Z}_{\geq 0}$ there is a finite collection of open balls $\{B_i^n : i \in M_n\}$ each with radius 2^{-n} whose union is all of X (thus the indexing set M_n is finite). Since M_0 is finite, there is some $i_0 \in M_0$ such that $S_0 = \{n \in \mathbb{N} : x_n \in B_{i_0}^0\}$ is infinite. Now suppose inductively that $S_0 \supseteq S_1 \supseteq \ldots \supseteq S_{k-1}$ have been chosen, each an infinite subset of $\mathbb{N}$ with the property that for each $j = 0, 1, \ldots, k-1$ there is an $i_j \in M_j$ with $x_n \in B_{i_j}^j$ for all $n \in S_j$. Thus all the x_ns with $n \in S_j$ lie in an open ball of radius 2^{-j}. Then since S_{k-1} is infinite and M_k is finite there is an $i_k \in N_k$ such that

$$S_k = \{n \in S_{k-1} : x_n \in B_{i_k}^k\}.$$

is infinite. Proceeding in this way[23] we get an infinite nested collection of sequences of integers $S_k = \{n_1^k < n_2^k < \ldots\}$ such that for each k, $(x_{n_i^k})_{i \geq 1}$ is a subsequence of (x_n) which lies in $B_{i_k}^k$, and hence the terms of this subsequence are at distance at most 2^{-n+1} from each other. But then the subsequence (y_k) where $y_k = x_{n_k^k}$ must be a Cauchy subsequence of (x_n): If $m \geq k$ then by construction all the terms $y_m = x_{n_m^m}$ are such that $n_m^m \in S_m \subseteq S_k$ and hence they are at distance at most 2^{-k+1} apart from each other and hence since $2^{-k+1} \to 0$ as $k \to \infty$ it follows that (y_k) is Cauchy as required. □

Remark 10.17. The same "divide and conquer" proof strategy can be used to prove that $[-R, R]^n$ is sequentially compact in $\mathbb{R}^n$, as you can find in many textbooks. The additional subtlety of this proof is that we need an

[23]This part of the proof is similar to the argument we used to prove that a product of compact metric spaces $X \times Y$ is compact. We need a new trick here however – the diagonal argument – to deal with the fact that now we obtain an infinite number of nested subsequences.

infinite nested sequence of subsequences, and hence have to use a version of Cantor's diagonal argument to finish the proof.

Corollary 10.18. *A complete and totally bounded metric space X is compact.*

Proof. By Lemma 10.16, any sequence (x_n) in X has a Cauchy subsequence. Since X is complete, this subsequence converges, and hence X is compact as required. $\square$

11. Compactness and open sets

We have already noted that compactness is a "topological property" of metric spaces, in the sense that two metric spaces which are homeomorphic have to either both be compact or both be non-compact. This might lead one to consider if the notion of compactness can be expressed in terms of open sets. In fact this is possible, though we wont quite prove the equivalence of the definition we give in terms of open sets to the one we began with in terms of convergence of subsequences[24]. For clarity in this section we will refer to the notion of compactness given by the existence of convergent subsequences as *sequential compactness*. The key definition is the following:

Definition 11.1. 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 $X = \bigcup_{i \in J} U_i = X$ 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*. Recall that if Z is a subspace of a metric space X, then the open sets of Z are of the form $Z \cap U$ where U is an open subset of X. In this situation it is often convenient to think of an open cover of Z as a collection $\mathcal{U} = \{U_i : i \in I\}$ of open subsets of X whose union contains (but need not be equal to) the subspace Z.

We can now give the definition of compactness in terms of open covers:

Definition 11.2. A metric space (X, d) is *compact* if every open cover $\mathcal{U} = \{U_i : i \in I\}$ has a finite subcover.

For example, any finite subset of a metric space is compact. To have some more non-trivial examples, we prove the following:

Proposition 11.3. *(Heine-Borel.) The interval $[a, b]$ is compact.*

Proof. Let $\mathcal{U} = \{U_i : i \in I\}$ be an open cover of $[a, b]$ (where we view the U_i as open subsets of $\mathbb{R}$). Then set $S = \{x \in [a, b] : [a, x] \text{ lies in a finite union of } U_i\text{s}\}$. Then S is a non-empty subset of $[a, b]$ (because $a \in S$). Let $c = \sup(S)$. We may find a $U_{i_0} \in \mathcal{U}$ such that $c \in U_{i_0}$ and hence a $\delta > 0$ with $(c - \delta, c + \delta) \subseteq U_{i_0}$. Now by the approximation property there is a $d \in S$ with

[24] One should be a little careful here – the two notions are equivalent for metric spaces, but for general topological spaces they are distinct.

$c - \delta < d \leq c$, and so there is a finite subset of I, say $i_1, \ldots, i_n$, such that $[a,d] \subseteq U_{i_1} \cup \ldots \cup U_{i_n}$. But then clearly $[a, c+\delta) \subseteq (U_{i_1} \cup \ldots \cup U_{i_n}) \cup U_{i_0}$ so that $[a,b] \cap [a, c+\delta) \subseteq S$, which contradicts the definition of c unless $c = b \in S$. But then $\mathcal{U}$ has a finite subcover as required.

□

It is easy to prove that a closed subset of a compact metric space is compact, which combined with the previous proposition shows that any closed bounded subset of $\mathbb{R}$ is compact (note we have already see this for sequentially compact subsets of $\mathbb{R}$). The next Proposition shows compactness implies sequential compactness, hence all the results we have shown for such metric spaces also apply to compact metric space. We first need a technical lemma.

Lemma 11.4. *Let (x_n) be a sequence in a metric space X, and let $A_n = \{x_k : k \geq n\}$. Then (x_n) has a convergent subsequence if and only if $\bigcap_{n\geq 1} \bar{A}_n \neq \emptyset$.*

Proof. Suppose (x_n) has a convergent subsequence (x_{n_k}), so that $x_{n_k} \to \ell \in X$ as $k \to \infty$. Then since for any $m \in \mathbb{N}$ all terms of the subsequence $(x_{n_{k+m}})_{k\geq 1}$ lie in A_m, it follows that $\ell \in \bar{A}_m$ for all m, so that the intersection $\bigcap_{n\geq 1} \bar{A}_n$ is non-empty.

Conversely, suppose that $\ell \in \bigcap_{n\geq 1} \bar{A}_n$. Then we claim there is a subsequence of (x_n) tending to ℓ: Certainly since $\ell \in \bar{A}_1$, we may find an x_{n_1} such that $d(x_{n_1}, a) < 1$. Now suppose that $n_1 < n_2 < \ldots < n_k$ have been found such that $d(x_{n_j}, \ell) < 1/j$ for each j with $1 \leq j \leq k$. Then since $\ell \in \bar{A}_{n_k+1}$ we may find an $n_{k+1} > n_k$ with $d(x_{n_{k+1}}, \ell) < 1/(k+1)$. This subsequence (x_{n_k}) clearly converges to ℓ so we are done. □

Proposition 11.5. *Let (X,d) be a compact metric spaces. Then every sequence in X has a convergent subsequence, that is, X is sequentially compact.*

Proof. Suppose that (x_n) is a sequence in X. For each $n \in \mathbb{N}$ let $A_n = \{x_k : k \geq n\}$. Then $\bar{A}_1 \supseteq \bar{A}_2 \supseteq \ldots$ form a nested sequence of non-empty closed subsets of X. Now by Lemma 11.4 we know that (x_n) has a convergent subsequence if and only if $\bigcap_{n\geq 1} \bar{A}_n$ is non-empty. Thus if we suppose for the sake of contradiction that the sequence (x_n) has no convergent subsequence it follows that $\bigcap_{n\geq 1} \bar{A}_n = \emptyset$. But then if we let $U_n = X \backslash \bar{A}_n$ we have $X = \bigcup_{n\geq 1} U_n$, so that $\{U_n : n \geq 1\}$ is an open cover of X. However $U_1 \subseteq U_2 \subseteq \ldots$ and each is a proper subset of X, thus this cover clearly has no finite subcover, contradicting the assumption that X is compact. □

We end this section with a simple Lemma on compact sets which are contained in an open subset of a metric space, which will be useful later in the course:

Lemma 11.6. *Let (X,d) be a metric space and suppose $K \subseteq U \subseteq X$ where K is compact and U is open. Then there is an $\epsilon > 0$ such that for any $z \in K$ we have $B(z, \epsilon) \subseteq U$.*

Proof. Suppose for the sake of contradiction that no such ϵ exists. Then for each $n \in \mathbb{N}$ we may find sequences $x_n \in K$ and $y_n \in U^c$ with $|x_n - y_n| < 1/n$. But since K is sequentially compact we can find a convergent subsequence of (x_n), say (x_{n_k}) which converges to $p \in K$. But then it follows (y_{n_k}) also converges to p, which is impossible since $p \in K \subseteq U$ while (y_{n_k}) is a sequence in the U^c and as U^c is closed it must contain all its limit points. □

Exercise 11.7. Use the technique of the proof of the previous Lemma to show that if Ω is an open subset of $\mathbb{R}^n$ then it can be written as a countable union of compact subsets, $\Omega = \bigcup_{n=1}^{\infty} K_n$.

11.1. **Compactness and function spaces.**

Definition 11.8. If X is a metric spaces and $\mathcal{F}$ is collection of real-valued function on X, we say that $\mathcal{F}$ is *equicontinuous* if, for any $\epsilon > 0$ there is a δ (which *only* depends on ϵ) such that whenever $d(x, y) < \delta$ we have $|f(x) - f(y)| < \epsilon$ for *every* $f \in \mathcal{F}$. A collection of continuous functions $\mathcal{F}$ on X is *uniformly bounded* if it is bounded as a subset of the normed vector space $(\mathcal{C}_b(X), \|.\|_\infty)$.

Theorem 11.9. *(Arzela-Ascoli): Let X be a compact metric space and let $\mathcal{F} \subseteq \mathcal{C}(X)$ be a collection of continuous functions on X which are equicontinuous and uniformly bounded. Then any sequence (f_n) in $\mathcal{F}$ contains a subsequence (f_{n_k}) which converges uniformly on X.*

Proof. To prove the theorem it suffices to check that $\mathcal{F}$ is totally bounded in $\mathcal{C}(X)$, since then the completeness of $\mathcal{C}(X)$ implies that $\bar{\mathcal{F}}$ is complete and totally bounded[25] and hence compact.

Thus we must show that $\mathcal{F}$ is totally bounded. Suppose that $\epsilon > 0$ is given. Then since $\mathcal{F}$ is equicontinuous we know that there is a $\delta > 0$ such that if $x, y \in X$ are such that $d(x, y) < \delta$ then $|f(x) - f(y)| < \epsilon/6$. Now X is compact and hence totally bounded, so that we may find a finite set $\{x_1, x_2, \ldots, x_n\} \subseteq X$ such that $X = \bigcup_{i=1}^{n} B(x_i, \delta)$. Now since $\mathcal{F}$ is uniformly bounded, there is some $N > 0$ such that $f(X) \subseteq [-N, N]$ for each $f \in \mathcal{F}$. Pick an integer $M > 0$ so that $2N/M < \epsilon/6$ and divide $[-N, N]$ into M equal parts I_j, $1 \leq j \leq M$. Let A denote the set of n^M functions $\alpha\colon \{1, \ldots, n\} \to \{1, \ldots, M\}$ and for each such α, pick a function $f_\alpha \in \mathcal{F}$ (if it exists) such that $f(x_i) \in I_{\alpha(i)}$. We claim that the open balls $B(f_\alpha, \epsilon)$ cover $\mathcal{F}$ as α runs over those functions α for which f_α exists.[26]

Indeed suppose that $f \in \mathcal{F}$. Then for each $i \in \{1, 2, \ldots, n\}$ we must have $f(x_i) \in I_{\alpha(i)}$ for some $\alpha\colon A$. Consider $d(f, f_\alpha)$ (which exists by assumption). For each $x \in X$ then there is some $i \in \{1, 2, \ldots, n\}$ such that $x \in B(x_i, \delta)$.

[25] It is a straight-forward exercise to check that if A is a totally bounded subspace of a metric space X then $\bar{A}$ is also totally bounded.

[26] It may be helpful to draw a picture in the case $X = [a, b]$.

Thus

$$\begin{aligned} d(f(x), f_\alpha(x)) &\leq d(f(x), f(x_i)) + d(f(x_i), f_\alpha(x_i)) + d(f_\alpha(x_i), f_\alpha(x) \\ &\leq \epsilon/6 + |I_{\alpha(i)}| + \epsilon/6 < \epsilon/2. \end{aligned}$$

Since this holds for all $x \in X$ it follows that $\|f - f_\alpha\|_\infty \leq \epsilon/2 < \epsilon$ and hence $f \in B(f_\alpha, \epsilon)$. Thus $\mathcal{F}$ is totally bounded as required.

□

Remark 11.10. The previous theorem implies closed bounded equicontinuous subsets of $\mathcal{C}(X)$ are compact. In fact the converse is also true. Since a compact subspace $\mathcal{F}$ of any metric space is automatically closed and bounded, one only needs to show that $\mathcal{F}$ is equicontinuous. To prove this one uses the that if $\mathcal{F}$ is compact subset then it is totally bounded, combined with the fact that since X is compact any $f \in \mathcal{C}(X)$ is uniformly continuous.

Remark 11.11. The are various ways to generalise the above theorem to spaces X which are not compact. For example, if Ω is an open subset of $\mathbb{R}^n$, one can show that Ω can be written as a countable union $\Omega = \bigcup_{n=1}^\infty K_n$ where each K_n is a closed bounded subset of Ω and then deduce that if (f_n) is a sequence in an equicontinuous uniformly bounded family of functions $\mathcal{F} \subseteq \mathcal{C}_n(\Omega)$, there is a subsequence (f_{n_k}) which converges uniformly on any compact subset of Ω.

12. THE COMPLEX PLANE: TOPOLOGY AND GEOMETRY.

For the rest of the course we will study functions on $\mathbb{C}$ the complex plane, focusing on those which satisfy the complex analogue of differentiability. We will thus need the notions of convergence and limits which $\mathbb{C}$ possesses because it is a metric space (in fact normed vector space).

In this regard, the complex plane is just $\mathbb{R}^2$ and we have seen that there are a number of norms on $\mathbb{R}^2$ which give us the same notion of convergence (and open sets). The additional structure of multiplication which we equip $\mathbb{R}^2$ with when we view it as the complex plane however, makes it natural to prefer the Euclidean one $|z| = \sqrt{(Re(z)^2 + Im(z)^2}$. More explicitly, if $z = (a, b)$ and $w = (c, d)$ are vectors in $\mathbb{R}^2$, then we define their product to be

$$z.w = (ac - bd, ad + bc).$$

It is straight-forward, though a bit tedious, to check that this defines an associative, commutative multiplication on $\mathbb{R}^2$ such that every non-zero element has a multiplicative inverse: if $z = (a, b) \neq (0, 0)$ has $z^{-1} = (a, -b)/(a^2 + b^2)$. The number $(1, 0)$ is the multiplicative identity (and so is denoted 1) while $(0, 1)$ is denoted i (or j if you're an engineer) and satisfies $i^2 = -1$. Since $(1, 0)$ and $(0, 1)$ form a basis for $\mathbb{R}^2$ we may write any complex number z uniquely in the form $a + ib$ where $a, b \in \mathbb{R}$. We refer to a and b as the *real* and *imaginary* parts of z, and denote them by $\Re(z)$ and $\Im(z)$ or $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ respectively.

Definition 12.1. If $z = (a, b)$ we write $\bar{z} = (a, -b)$ for the *complex conjugate* of z. It is easy to check that $\overline{zw} = \bar{z}.\bar{w}$ and $\overline{z + w} = \bar{z} + \bar{w}$. The Euclidean norm on $\mathbb{R}^2$ is related to the multiplication of complex numbers by the formula $|z| = \sqrt{z\bar{z}}$, which moreover makes it clear that $|zw| = |z||w|$. (We call such a norm *multiplicative*). If $z \neq 0$ then we will also write $\arg(z) \in \mathbb{R}/2\pi\mathbb{Z}$ for the angle z makes with the positive half of the real axis.

Because subsets of the complex plane can have a much richer structure than subsets of the real line, the topological material we developped in the first half of the course will be indespensible in understanding complex differentiable functions. We will need the notions of completeness, compactness, and connectedness, along with the basic notions of open and closed sets.

Definition 12.2. A connected open subset D of the complex plane will be called a *domain*. As we have already seen, an open set in $\mathbb{C}$ is connected if and only if it is path-connected.

We will also use the notations of closure, interior and boundary of a subset of the complex plane. The *diameter* $\mathrm{diam}(X)$ of a set X is $\sup\{|z - w| : z, w \in X\}$. A set is bounded if and only if it has finite diameter. Recall that the Heine-Borel theorem in the case of $\mathbb{R}^2$ ensures that a subset $X \subseteq \mathbb{C}$ is compact (that is, every open covering has a finite subcover) if and only if it is closed and bounded.

Definition 12.3. Because the complex numbers form a field, we can, for a function $f\colon U \to \mathbb{C}$ defined on some subset $U \subseteq \mathbb{C}$ which is a neighbourhood of $a \in U$, define the (complex) derivative of f at a to be

$$\lim_{z\to a} \frac{f(z)-f(a)}{z-a},$$

exactly as in the real variable case. We say that f is *complex differentiable* at a, and if f is complex differentiable at every $a \in U$ then we say that f is *holomorphic* on U.

It is straight-forward to check from this definition that the basic results about real derivatives, such as the product rule and quotient rule, carry over to the complex setting – the proofs are identical to the real case (except $|.|$ means the modulus of a complex number rather than the absolute value of a real number).

Proposition 12.4. *Let U be an open subset of $\mathbb{C}$ and let f, g be complex-valued functions on U.*

(1) *If f, g are differentiable at $z_0 \in U$ then $f+g$ and fg are differentiable at z_0 with*

$$(f+g)'(z_0) = f'(z_0) + g'(z_0); \quad (f.g)'(z_0) = f'(z_0).g(z_0) + f(z_0).g'(z_0).$$

(2) *If f, g are differentiable at z_0 and $g(z_0) \neq 0$ and $g'(z_0) \neq 0$ then f/g is differentiable at z_0 with*

$$(f/g)'(z_0) = \frac{f'(z_0)g(z_0) - f(z_0)g'(z_0)}{g'(z_0)^2}.$$

(3) *If U and V are open subsets of $\mathbb{C}$ and $f\colon V \to U$ and $g\colon U \to \mathbb{C}$ where f is complex differentiable at $z_0 \in V$ and g is complex differentiable at $f(z_0) \in U$ then $g \circ f$ is complex differentiable at z_0 with*

$$(g \circ f)'(z_0) = g'(f(z_0)).f'(z_0).$$

Proof. These are proved in exactly the same way as they are for a function of a single real variable. □

Remark 12.5. Just as for a single real variable, the basic rules of differentiation stated above allow one to check that polynomial functions are differentiable: Using the product rule and induction one sees that z^n has derivative nz^{n-1} for all $n \geq 0$ (as a constant obviously has derivative 0). Then by linearity it follows every polynomial is differentiable.

13. The extended complex plane

In this section we introduce the extended complex plane. As a set, the extended complex plane $\mathbb{C}_\infty$ is simply the complex plane union a single additional point denoted ∞. Although we cannot extend the algebraic properties of the complex plane[27] to $\mathbb{C}_\infty$, we will be able to extend its topological and analytic properties. To understand the metric/topological structure of $\mathbb{C}_\infty$ we will use a construction from real geometry, while to understand what it should mean for a function on $\mathbb{C}_\infty$ to be differentiable, we will use complex geometry.

Example 13.1. We start with a simpler example which is the real analogue of the above approaches to construct an "extended real line": We wish to build a natural space which added a point at infinity to the real line $\mathbb{R}$. If we embed the real line into the plane as the set R of points $\{(1,t) : t \in \mathbb{R}\}$, then clearly every line through the origin $(0,0)$ intersects R in a unique point, except for the y-axis, which is parallel to R. Thus the set of lines in the plane $\mathbb{R}^2$ naturally adds a "point at infinity" to the real line. Now any line L through the origin is spanned by any of its nonzero elements, and we can use this to give ourselves parametrizations of part of the space of all lines: So long as L is not the y-axis, it has a unique element of the form $(1,t)$, and so long as it is not the x-axis it has a unique point with coordinates $(s,1)$. This gives us two systems of parametrizations (both defined almost everywhere) attaching L to t or s, and the two parametrizations are related (where they are both defined) by $s = 1/t$.

Alternatively, if one draws the circle tangent to the y-axis and the line R, one sees that each line through the origin intersects that circle in two points, the origin and one other, except for the y-axis. Thus we can naturally identify the lines in the plane (and so our extended real line) with a circle.

[*Alternatively, another slightly more abstract way to see that the space of lines through the origin is a circle, is to note that any line intersects the unit circle in two opposite points, thus we can identify the space of lines in $\mathbb{R}^2$ with the space we obtain by identifying opposite points. This might sound abstract, but if you consider the restriction to the unit circle of the map $z \mapsto z^2$ on $\mathbb{R}^2$ (identified as $\mathbb{C}$), it sends opposite points on the circle to the same point, so this shows the space we get is just a circle again!*]

Let us now examine how similar ideas will let us construct the extended complex plane $\mathbb{C}_\infty$. We begin with the analogue of the circle construction, which is known as the Riemann sphere.

13.1. **Stereographic projection.** Let $\mathbb{S}^2 = \{(x,y,z) \in \mathbb{R}^3 : x^2+y^2+z^2 = 1\}$ be the unit sphere of radius 1 centred at the origin in $\mathbb{R}^3$, and view the complex plane as the copy of $\mathbb{R}^2$ inside $\mathbb{R}^3$ given by the plane $\{(x,y,0) \in$

[27] Though it is sometimes useful to have conventions such as $z + \infty = \infty$

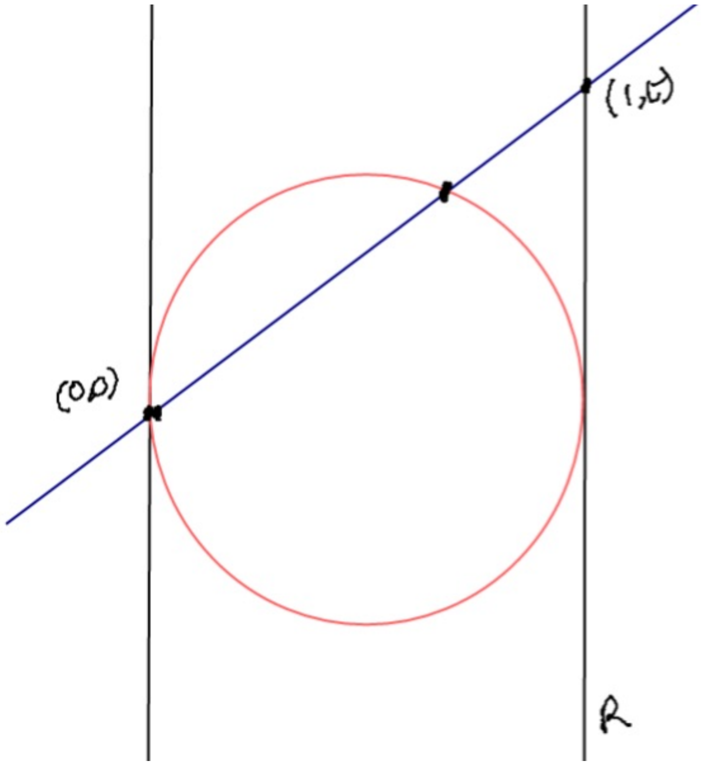

FIGURE 1. The extended real line.

$\mathbb{R}^3 : x, y \in \mathbb{R}\}$. Let N be the "north pole" $N = (0,0,1)$ of the sphere $\mathbb{S}^2$. Given a point $z \in \mathbb{C}$, there is a unique line passing through N and z, which intersects $\mathbb{S}\backslash\{N\}$ in a point $S(z)$. This map gives a bijection between $\mathbb{C}$ and $\mathbb{S}\backslash\{N\}$. Indeed, explicitly, if $(X, Y, Z) \in \mathbb{S}\backslash\{N\}$ then it corresponds to[28] $z \in \mathbb{C}$ where $z = x + iy$ with $x = X/(1-Z)$ and $y = Y/(1-Z)$. Correspondingly, given $z = x + iy \in \mathbb{C}$ we have

$$\begin{aligned} S(z) &= \Big(\frac{2x}{x^2+y^2+1}, \frac{2y}{x^2+y^2+1}, \frac{x^2+y^2-1}{x^2+y^2+1}\Big) \\ &= \frac{1}{1+|z|^2}\big(2\Re(z), 2\Im(z), |z|^2-1\big). \end{aligned} \tag{13.1}$$

Thus if we set $S(\infty) = N$, then we get a bijection between $\mathbb{C}_\infty$ and $\mathbb{S}^2$, and we use this identification to make $\mathbb{C}_\infty$ into a metric space (and thus we obtain a notion of continuity for $\mathbb{C}_\infty$): As a subset of $\mathbb{R}^3$ equipped with the Euclidean metric $\mathbb{S}^2$ is naturally a metric space.

[28] Any point on the line between N and (X, Y, Z) can be written as $t(0,0,1) + (1 - t)(X, Y, Z)$ for some $t \in \mathbb{R}$. It is then easy to calculate where this line intersects the plane given by the equation $z = 0$.

Lemma 13.2. *The metric induced on $\mathbb{C}_\infty$ by S is given by*

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

for any $z, w \in \mathbb{C}$.

Proof. First consider the case where $z, w \in \mathbb{C}$. Since $S(z), S(w) \in \mathbb{S}^2$ we see that $\|S(z) - S(w)\|^2 = 2 - 2S(z).S(w)$. But using (13.1) we see that

$$\begin{aligned} S(z).S(w) &= \frac{2(z\bar{w} + \bar{z}w) + (|z|^2-1)(|w|^2-1)}{(1+|z|^2)(1+|w|^2)} \\ &= \frac{2(z\bar{w} + \bar{z}w) + z\bar{z}w\bar{w} - z\bar{z} - w\bar{w} + 1}{(1+|z|^2)(1+|w|^2)} \\ &= 1 - \frac{2|z-w|^2}{(1+|z|^2)(1+|w|^2)} \end{aligned}$$

so that

$$d_2(S(z), S(w))^2 = \frac{4|z-w|^2}{(1+|z|^2)(1+|w|^2)}$$

as required. The case where one or both of z, w is equal to ∞ is similar but easier. $\square$

Remark 13.3. Note that in particular, $S(z)$ tends to $N = (0,0,1)$ if and only if $|z| \to \infty$, thus our notation $z \to \infty$ now takes on a literal meaning, consistent with its previous definition. One way we can use this is as follows: If we take $f(z) = 1/z$ defined on $\mathbb{C}\backslash\{0\}$ and extend it to a map $\tilde{f}\colon \mathbb{C} \to \mathbb{C}_\infty$ by setting $\tilde{f}(0) = \infty$, then $\tilde{f}$ is a continuous function on the entire complex plane.

The geometry of the sphere nicely unites lines and circles in the plane as the following Lemma shows:

Lemma 13.4. *The map $S\colon \mathbb{C} \to \mathbb{S}$ induces a bijection between lines in $\mathbb{C}$ and circles in $\mathbb{S}$ which contain N, and a bijection between circles in $\mathbb{C}$ and circles in $\mathbb{S}$ not containing N.*

Proof. A circle in $\mathbb{S}$ is given by the intersection of $\mathbb{S}$ with a plane H. Any plane H in $\mathbb{R}^3$ is given by an equation of the form $aX + bY + cZ = d$, and H intersects $\mathbb{S}$ provided $a^2 + b^2 + c^2 > d^2$. Indeed to see this note that H intersects the sphere in a circle if and only if its distance to the origin is less than 1. Since the closest vector to the origin on H is perpendicular to the plane it is a scalar multiple of (a,b,c), so it must be $\frac{d}{a^2+b^2+c^2}(a,b,c)$, hence H is at distance $d^2/(a^2+b^2+c^2)$ from the origin and the result follows. Moreover, clearly H contains N if and only if $c = d$.

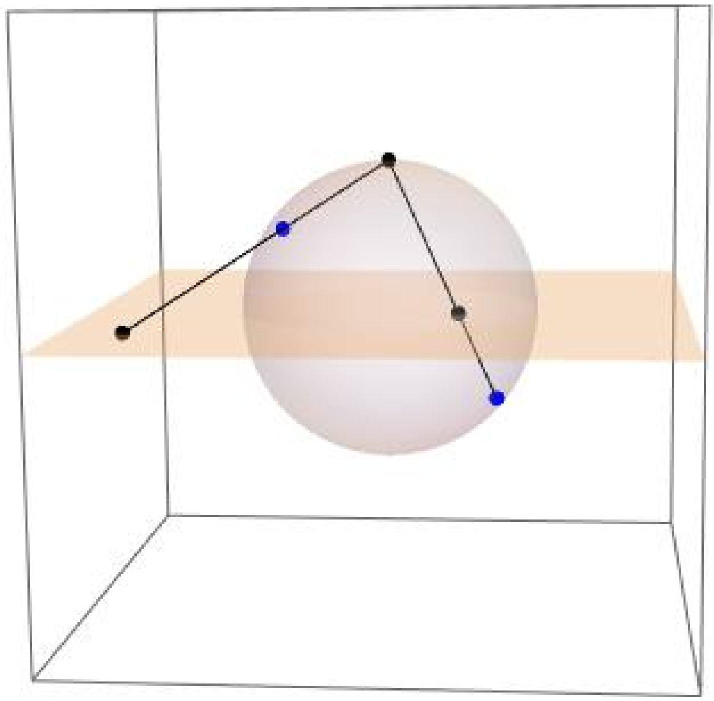

FIGURE 2. The stereographic projection map.

Now from the explicit formulas for S we see that if $z = x + iy$ then $S(z)$ lies on this plane if and only if

$$\begin{aligned} 2ax + 2by + c(x^2 + y^2 - 1) &= d(x^2 + y^2 + 1) \\ \iff (c - d)(x^2 + y^2) + 2ax + 2by - (c + d) &= 0 \end{aligned}$$

Clearly if $c = d$ this is the equation of a line, while conversely if $c \neq d$ it is the equation of a circle in the plane. Indeed if $c \neq d$, we can normalize and insist that $c - d = 1$, whence our equation becomes

$$(x + a)^2 + (y + b)^2 = (a^2 + b^2 + c + d) \tag{13.2}$$

that is, the circle with centre $(-a, -b)$ and radius $\sqrt{a^2 + b^2 + c + d}$. Note that the condition the plane intersected $\mathbb{S}$ becomes the condition that $a^2 + b^2 + c + d > 0$, that is, exactly the condition that Equation (13.2) has a non-empty solution set.

To complete the proof, we need to show that all circles and lines in $\mathbb{C}$ are given by the form of the above equation. When $c = d$ we get $2(ax+by-c) = 0$, and clearly the equation of every line can be put into this form. When $c \neq d$ as before assume $c - d = 1$, then letting $a, b, c + d$ vary freely we see that we can obtain circle in the plane as required. $\square$

13.2. **The projective line.** Our second approach to the extended complex plane is via the projective line $\mathbb{P}^1$: this is, as a set, simply the collection of one-dimensional subspaces of $\mathbb{C}^2$. Although we cannot readily draw a picture of these as we could in the real case, the same analysis we did in that setting extends to the complex one: If e_1, e_2 denote the standard basis of $\mathbb{C}^2$ then we have two subsets of $\mathbb{P}^1$, each naturally in bijection with $\mathbb{C}$. If we set $U_0 = \mathbb{P}^1 \backslash \mathbb{C}.e_1$ and $U_1 = \mathbb{P}^1 \backslash \mathbb{C}e_2$, then we have maps $i_0, i_\infty \colon \mathbb{C} \to \mathbb{P}^1$ given by $i_0(z) = \mathbb{C}.(ze_1 + e_2)$ and $i_\infty(z) = \mathbb{C}.(e_1 + ze_2)$ whose images are U_0 and U_1 respectively. Given a nonzero vector $(z, w) \in \mathbb{C}^2$ we will write $[z, w] \in \mathbb{P}^1$ for the line it spans. (The numbers z, w are often called the *homogeneous coordinates* of $[z, w]$. They are only defined up to simultaneous rescaling.)

Thus $\mathbb{P}^1$ is covered by two pieces U_0 and U_∞ whose union is all of $\mathbb{P}^1$. We can use this to make $\mathbb{P}^1$ a topological space: we say that V is an open subset of $\mathbb{P}^1$ if and only if $V \cap U_0$ and $V \cap U_\infty$ are identified with open subsets of $\mathbb{C}$ via the bijections i_0 and i_1 respectively. It is a good exercise to check that this does indeed define a topology on $\mathbb{P}^1$ (in which both U_0 and U_∞ are open, since $\mathbb{C}$ and $\mathbb{C}^\backslash\{0\}$ are open in $\mathbb{C}$. We however will take a more direct approach: Note that we can identify $\mathbb{P}^1$ with $\mathbb{C}_\infty$ using the map $i_0 \colon \mathbb{C} \to \mathbb{P}^1$ extending it to $\mathbb{C}_\infty$ by sending ∞ to $\mathbb{C}e_1$ and we can thus transport the metric on $\mathbb{C}_\infty$ (which of course we obtained in turn from our identification on $\mathbb{C}_\infty$ with $\mathbb{S}^2$) to that on $\mathbb{P}^1$. Perhaps surprisingly, this metric has a natural expression in terms of the Hermitian form $\langle \cdot, \cdot \rangle$ on $\mathbb{C}^2$ as the next Lemma shows:

Lemma 13.5. *The metric induced on $\mathbb{P}^1$ by its identification with $\mathbb{C}_\infty$ is given by*

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

where $v \in L_1 \backslash \{0\}$ and $w \in L_2 \backslash \{0\}$.

Proof. Suppose $L_1 = [z, 1]$ and $L_2 = [w, 1]$. Then the formula in the statement of the Lemma gives

$$\begin{aligned}
d(L_1, L_2) &= 2\sqrt{1 - \frac{|z\bar{w} + 1|^2}{(1 + |z|^2)(1 + |w^2)}} \\
&= 2\sqrt{\frac{1 + |z|^2 + |w|^2 + |z|^2|w|^2 - |z|^2|w|^2 - z\bar{w} - \bar{z}w - 1}{(1 + |z|^2)(1 + |w|^2)}} \\
&= 2\sqrt{\frac{|z - w|^2}{(1 + |z|^2)(1 + |w|^2}} = \frac{2|z - w|}{\sqrt{1 + |z|^2}\sqrt{1 + |w|^2}}
\end{aligned}$$

The case when $L_2 = \infty = \mathbb{C}e_1$ is similar but easier. ∎

One advantage of thinking of $\mathbb{C}_\infty$ as the projective line is that we can use the charts U_0 and U_∞ to define what it means for a function f on $\mathbb{C}_\infty$ to be holomorphic:

Definition 13.6. Suppose that $f\colon W \to \mathbb{P}^1$ is a continuous function on an open subset W of $\mathbb{P}^1$, and let $L \in W$. Suppose that $L \in U_p$ and $f(L) \in U_q$ where $p, q \in \{0, \infty\}$. Then $f^{-1}(U_q) \cap U_p$ is an open set in $U_p \subset \mathbb{P}^1$, which via i_p (or rather its inverse) we can identify with an open subset V of $\mathbb{C}$, and its image under f lies in U_q which we can identify with $\mathbb{C}$ via i_q^{-1}. Thus f yields a continuous function $\tilde{f}\colon V \to \mathbb{C}$, where $\tilde{f} = i_q^{-1} \circ f \circ i_p$ and we say f is holomorphic at L if $\tilde{f}$ is holomorphic at $i_p(z) = L$.

$$\begin{tikzcd}
f^{-1}(U_q) \cap U_p \arrow[r, "f"] & U_p \arrow[d, "i_q^{-1}"] \\
V \subseteq C \arrow[u, "i_p"] \arrow[r, "\tilde{f}"] & \mathbb{C}
\end{tikzcd}$$

Since most points in $\mathbb{P}^1$ lie in both U_0 *and* U_∞ the above definition seems ambiguous. In fact, where there is a choice, it does not matter what which of U_0 or U_∞ you pick. This is because $i_0^{-1} \circ i_\infty(z) = i_\infty^{-1} \circ i_0(z) = 1/z$ for all $z \in \mathbb{C}\backslash\{0\}$ and the function $1/z$ is complex differentiable with complex differentiable inverse (itself!) on $\mathbb{C}\backslash\{0\}$. This fact and the chain rule combine to show that the definition is independent of any choices. The essential point is that if $f(z)$ is complex differentiable, then so are $f(1/z), 1/f(z)$ and $1/f(1/z)$ wherever they are defined.

Example 13.7. Consider the example of $f(z) = 1/(z^2+1)$ viewed as a function $f\colon \mathbb{C} = U_0 \to \mathbb{P}^1$, where we extend it to a function on all of $\mathbb{C}$ by continuity, so that $f(0) = \infty$. We claim that f is in fact complex differentiable. To check this near 0 we must write $f(z)$ in the form $[1 : f_\infty(z)]$ and check if f_∞ is complex diffentiable. For $z \neq 0$, by definition $f(z) = [1/(z^2+1) : 1]$, thus since $[1/(z^2+1) : 1] = [1 : z^2+1]$ we see that function $f_\infty(z) = z^2+1$ which is clearly complex differentiable at $z = 0$ as required.

You can check using this definition that a holomorphic function $f\colon \mathbb{C} \to \mathbb{P}^1$ are precisely the meromorphic functions, and with a bit more work show that the holomorphic functions f which are defined on all of $\mathbb{P}^1$ are exactly the set of rational functions.

Recall that we have identified $\mathbb{C}_\infty$ with the projective line $\mathbb{P}^1$. The general linear group $\mathrm{GL}_2(\mathbb{C})$ acts on $\mathbb{C}^2$ in the natural way, and this induces an action on the set of lines in $\mathbb{C}$. We thus get an action of $\mathrm{GL}_2(\mathbb{C})$ on $\mathbb{P}^1$, and so on the extended complex plane. Explicitly, if $v = (z_1, z_2)^t$ spans a line $L = \mathbb{C}.v$ then if $g \in \mathrm{GL}_2(\mathbb{C})$ is given by a matrix

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

we see that

$$g(L) = \mathbb{C}.g(v) = \mathbb{C}\begin{pmatrix} az_1 + bz_2 \\ cz_1 + dz_2 \end{pmatrix}.$$

In particular, using our embedding $i_0\colon \mathbb{C} \to \mathbb{P}^1$ we see that

$$g(i_0(z)) = \mathbb{C}.g\begin{pmatrix} z \\ 1 \end{pmatrix} = \mathbb{C}.\begin{pmatrix} az+b \\ cz+d \end{pmatrix} = \mathbb{C}.\begin{pmatrix} \frac{az+b}{cz+d} \\ 1 \end{pmatrix} = i_0(\frac{az+b}{cz+d}).$$

Note that $f(-d/c) = \infty$ and $f(\infty) = a/c$, as is easily checked using the fact that $\infty = [1:0] \in \mathbb{P}^1$.

Definition 13.8. The induced maps $z \mapsto \frac{az+b}{cz+d}$ from the extended complex plane to itself are known as *Mobius maps* or *Mobius transformations*. Since they come from the action of $\mathrm{GL}_2(\mathbb{C})$ on $\mathbb{P}^1$ they automatically form a group. Note this means that every Mobius transformation is a bijection of the extended complex plane to itself, and moreover its inverse is also a Mobius transformation. In particular, since rational functions on $\mathbb{C}$ yield holomorphic functions on $\mathbb{C}_\infty$, every Mobius transformation gives an invertible holomorphic function on $\mathbb{C}_\infty$.

$$\mathrm{Mob} = \{f(z) = \frac{az+b}{cz+d} : ad-bc \neq 0\}.$$

Note that if we rescale a, b, c, d by the same (nonzero) scalar, then we get the same transformation. In group theoretic terms, the map from $\mathrm{GL}_2(\mathbb{C})$ to Mob has a kernel, the scalar matrices, thus Mob is a *quotient group* of $\mathrm{GL}_2(\mathbb{C})$. As a quotient group it is usually denoted $\mathrm{PGL}_2(\mathbb{C})$ the *projective general linear group*.

Any Mobius transformation can be understood as a composition of a small collection of simpler transformations, as we will now show. This can be useful because it allows us to prove certain results about all Mobius transformations by checking them for the simple transformations.

Definition 13.9. A transformation of the form $z \mapsto az$ where $a \neq 0$ is called a *dilation*. A transformation of the form $z \mapsto z+b$ is called a *translation*. The transformation $z \mapsto 1/z$ is called *inversion*. Note that these are all Mobius transformations, and the inverse of a dilation is a dilation, the inverse of a translation is a translation, while inversion is an involution and so is its own inverse.

Lemma 13.10. *Any Mobius transformation can be written as a composition of dilations, translations and an inversion.*

Proof. Let G denote the set of all Mobius transformations which can be obtained as compositions of dilations, translations and inversions. The set G is a subgroup of Mob. We wish to show it is the full group of Mobius transformations.

First note that any transformation of the form $z \mapsto az+b$ is a composition of the dilation $z \mapsto az$ and the translation $z \mapsto z+b$. Moreover, if $f(z) = \frac{az+b}{cz+d}$ is a Mobius transformation and $c = 0$ then $f(z) = (a/d)z + (b/d)$ (note if $c = 0$ then $ad - bc \neq 0$ implies $d \neq 0$) and so is a composition of a dilation and a translation. If $c \neq 0$ then we have

$$\frac{az+b}{cz+d} = \frac{(a/c)(cz+d) + (b - da/c)}{cz+d} = \frac{a}{c} + (b - d/a)\frac{1}{cz+d}. \tag{13.3}$$

Now $z \mapsto \frac{1}{cz+d}$ is the composition of an inversion with the map $z \mapsto cz + d$, and so lies in G. But then by equation (13.3) we have $f(z)$ is a composition of this map with a dilation and a translation, and so f lies in G. Since f was an arbitrary transformation with $c \neq 0$ it follows $G = \text{Mob}$ as required. □

Remark 13.11. The subgroup of Mob generated by translations and dilations is the group of $\mathbb{C}$-linear affine transformations $\text{Aff}(\mathbb{C}) = \{f(z) = az + b : a \neq 0\}$ of the complex plane. It is the stablizer of ∞ in Mob.

Remark 13.12. One should compare the statement of the previous Lemma with the theory or reduced row echelon form in Linear Algebra: any invertible 2×2 matrix will have the identity matrix as its reduced row echelon form, and the elementary row operations correspond essentially to the simple transformations which generate the Mobius group. This can be used to give an alternative proof of the Lemma.

As an example of how we can use this result to study Mobius transformations, we prove the following:

Lemma 13.13. *Let $f \colon \mathbb{C}_\infty \to \mathbb{C}_\infty$ be a Mobius transformation. Then f takes circles to circles. (Here we view $\mathbb{C}_\infty$ as $\mathbb{S}^2$ so that by Lemma 13.4 a circle in $\mathbb{C}_\infty$ is a line or a circle in $\mathbb{C}$).*

Proof. Since a line in $\mathbb{C}$ is given by the equation $\Im(az) = s$ where $s \in \mathbb{R}$ and $|a| = 1$, while a circle is given by the equation $|z - a| = r$ for $a \in \mathbb{C}$, $r \in \mathbb{R}_{>0}$, it is easy to check that any dilation or translation takes a line to a line and a circle to a circle.

The case of $z \mapsto 1/z$ is more interesting. One way to show it preserves lines and circles is to use the fact that these are both just circles viewed on the Riemann sphere. A direct calculation shows that the map $z \mapsto 1/z = \bar{z}/|z|^2$ corresponds to the map $(x, y, z) \mapsto (x, -y, -z)$, which is just the rotation by π about the x-axis, which is an isometry and so certainly preserves circles on unit sphere.

□

Exercise 13.14. Let $a, b \in \mathbb{C}$ be distinct complex numbers and let $k \in (0, 1]$. Then the locus of complex numbers satisfying $|z - a| = k.|z - b|$ is a line if $k = 1$ and is a circle otherwise.

Solution: Let $f(z) = (z - a)/(z - b)$. Since $a \neq b$ this is a Mobius map. The condition that $|z - a| = k|z - b|$ is just that $|f(z)| = k$, thus the locus of points satisfying this condition is the image of the circle of radius k centred at the origin under the Mobius map $f^{-1}(z) = (az - b)/(z - 1)$. Since we have seen Mobius maps take lines and circles to lines and circles, this image must be a line or a circle. Since $f^{-1}(1) = \infty$, the image is a circle if $k < 1$ and a line if $k = 1$.

14. Complex differentiability and the Cauchy-Riemann equations

We begin by recalling one way of defining the derivative of a real-valued function:

Definition 14.1. Suppose that $E \subseteq \mathbb{R}$ and $f\colon E \to \mathbb{R}$ is a function. If E is a neighbourhood of $x_0 \in \mathbb{R}$ then we say that f is differentiable at x_0 if there is a real number α such that for all $x \in E$ we have

$$f(x) = f(x_0) + \alpha(x - x_0) + \epsilon(x)|x - x_0|,$$

where $\epsilon(x) \to \epsilon(x_0) = 0$ as $x \to x_0$. If α exists it is unique and we write $\alpha = f'(x_0)$.

Remark 14.2. Note that rearranging the above equation we have, for $x \neq a$, $|\epsilon(x)| = |\frac{f(x)-f(a)}{x-a} - \alpha|$, thus the condition that $\epsilon(x) \to 0$ as $x \to a$ is equivalent to $\lim_{x\to a} \frac{f(x)-f(a)}{x-a} = \alpha$. This also shows the uniqueness of α.

Note also that if E is not a neighbourhood of a, then the above definition still makes sense, but more precise terminology is often used. For example if $E = [a, b]$ with $a < b$ and we take $x_0 = a$ then we say f has a right-hand derivative at x_0 if $\lim_{x\to a}(f(x) - f(a))/(x - a)$ exists as $x \to a$ with $x \in [a, b]$.

The above formulation of the definition of the derivative is a precise formulation of the statement that a function is differentiable at a point a if there is a "best linear approximation", or tangent line, to f near x_0 – that is, the function $x \mapsto f(x_0) + f'(x_0).(x - x_0)$. (The condition that the error term $\epsilon(x)|x - x_0|$ goes to zero faster than x tends to x_0 since $\epsilon(x)$ also tends to zero as x tends to x_0 is the rigorous meaning given to the adjective "best".) This has the advantage that it generalizes immediately to many variables:

Definition 14.3. Suppose that $E \subseteq \mathbb{R}^2$ is an open set, and $f\colon E \to \mathbb{R}^2$. Then we say that f is differentiable at $a \in E$ if there is a linear map $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that

$$f(z) = f(a) + T(z - a) + \epsilon(x)\|z - a\|$$

where $\epsilon(z) \to \epsilon(a) = 0$ as $z \to a$. If such a map T exists it is unique, and we denote it as $Df(a)$ (or sometimes Df_a. It is known as the *total derivative*[29] of f at a.

One can prove the uniqueness of Df_a directly, but it is more illuminating to understand the relation of α to the partial derivatives: If $v \in \mathbb{R}^2$ we define the *directional derivative* of f at a in the direction v to be

$$\partial_v f(a) = \lim_{t\to 0} \frac{f(a + t.v) - f(a)}{t},$$

(if this limit exists). When f is differentiable at a with derivative T, then it follows from the definitions that $t^{-1}(f(a+t.v) - f(a)) = T(v) \pm \epsilon(t.v)\|v\| \to$

[29] As opposed to the partial derivatives.

$T(v)$ as $t \to 0$, so the directional derivative of f at a all exist. In particular if $z = (x, y)$ and we write $f(z) = (u(x, y), v(x, y)))$ the directional derivatives in the direction of the standard basis vectors e_1 and e_2 are just $(\partial_x u, \partial_x v)$ and $(\partial_y u, \partial_y v)$. Thus we see that if T exists then its matrix with respect to the standard basis is just given by

$$\begin{pmatrix} \partial_x u & \partial_y u \\ \partial_x v & \partial_y v \end{pmatrix}$$

that is the matrix of T is just the *Jacobian matrix* of the partial derivatives of f (and hence the total derivative is uniquely determined, as asserted above).

We are now ready to define what it means for $f : U \to \mathbb{C}$ a function on an open subset U of $\mathbb{C}$, to be *complex differentiable*: We simply require that the linear map T is complex linear, or in other words, that T is given by multiplication by a complex number $f'(a)$:

Definition 14.4. A function $f : U \to \mathbb{C}$ on an open subset U of $\mathbb{C}$ is complex differentiable at $a \in U$ if there exists a complex number $f'(a)$ such that

$$f(z) = f(a) + f'(a).(z - a) + \epsilon(z).|z - a|,$$

where as before $\epsilon(z) \to \epsilon(a) = 0$ as $z \to a$.

Remark 14.5. If a function $f : U \to \mathbb{C}$ on an open subset U of $\mathbb{C}$ is everywhere complex differentiable on U we say it is *holomorphic* on U. We will use the terms "complex differentiable" and "holomorphic" interchangeably. (The term "analytic" is also commonly used, we will come back to that term later.)

Since the standard basis corresponds to $\{1, i\}$, since $(r + is)(x + iy) = (rx - sy) + i(sx + ry)$, the matrix of the linear map given by multiplication by $w = r + is$ is just

$$\begin{pmatrix} r & s \\ -s & r \end{pmatrix}$$

This gives us our first important result about complex differentiability:

Lemma 14.6. *(Cauchy-Riemann equations): If U is an open subset of $\mathbb{C}$ and $f : U \to \mathbb{C}$, then f is complex differentiable at $a \in U$ if and only if it is real-differentiable and the partial derivatives satisfy the equations:*

$$\partial_x u = \partial_y v, \quad \partial_x v = -\partial_y u.$$

Proof. This follows immediately from the definitions above. Note that it also shows that the complex derivative satisfies $f'(a) = \partial_x f = \partial_x u + i\partial_x v$ and $f'(a) = \frac{1}{i}\partial_y f = \frac{1}{i}(\partial_y u + i\partial_y v)$. □

Remark 14.7. Since the operation of multiplication by a complex number w is a composition of a rotation (by the argument of w) and a dilation (by the modulus of w) the matrix of the corresponding linear map is, up to scalar, a rotation matrix. The Cauchy-Riemann equations just capture this fact for the matrix of the total (real) derivative of a complex differentiable function.

A subtlety of real-differentiability in many variables is that it is possible for the partial derivatives of a function to exist without the function being differentiable in the sense of Definition 14.3. In most reasonable situations however, the following theorem shows that this does not happen:

Theorem 14.8. *Let U be an open subset of $\mathbb{R}^2$ and $f\colon U \to \mathbb{R}^2$. Let f have components f_1, f_2 so that $f = (f_1, f_2)^t$. If, for $i = 1, 2$, the partial derivatives $\partial_x f_i, \partial_y f_i$ exist and are continuous at $z_0 \in U$ then f is differentiable at z_0.*

The proof of this (although it is not hard – one only needs the definitions and the single-variable mean-value theorem) is not part of this course. For completeness, a proof is given in the Appendix. Combining this theorem with the Cauchy-Riemann equations gives a criterion for complex-differentiability:

Theorem 14.9. *Suppose that U is an open subset of $\mathbb{C}$ and let $f\colon U \to \mathbb{C}$ be a function. If f is differentiable as a function of two real variables with continuous partial derivatives satisfying the Cauchy-Riemann equations on U, then f is complex differentiable on U.*

Proof. Since the partial derivatives are continuous, Theorem 14.8 shows that f is differentiable as a function of two real variables, with total derivative given by the matrix of partial derivatives. If f also satisfies the Cauchy-Riemann equations, then by Lemma 14.6 it follows it is complex differentiable as required. □

Example 14.10. The previous theorem allows us to show that the complex logarithm is a holomorphic function – up to the issue that we cannot define it continuously on the whole complex plane! The function $z \mapsto e^z$ is not injective, since $e^{z+2n\pi i} = e^z$ for all $n \in \mathbb{Z}$ thus it cannot have an inverse defined on all of $\mathbb{C}$. However, since $e^{x+iy} = e^x(\cos(y) + i\sin(y))$, it follows that if we pick a ray through the origin, say $B = \{z \in \mathbb{C} : \Im(z) = 0, \Re(z) \leq 0\}$, then we may define $\mathrm{Log}\colon \mathbb{C}\backslash B \to \mathbb{C}$ by setting $\mathrm{Log}(z) = log(|z|) + i\theta$ where $\theta \in (-\pi, \pi]$ is the argument of z. Clearly $e^{\mathrm{Log}(z)} = z$, while $\mathrm{Log}(e^z)$ differs from z by an integer multiple of $2\pi i$.

We claim that Log is complex differentiable: To show this we use Theorem 14.9. Indeed the function $L(x,y) = (\log(\sqrt{x^2+y^2}), \theta) = (L_1, L_2)$ has

$$\begin{aligned} \partial_x L_1 &= \frac{x}{x^2+y^2}, \quad \partial_y L_1 = \frac{y}{x^2+y^2}, \\ \partial_x L_2 &= -\frac{y}{x^2+y^2}, \quad \partial_y L_2 = \frac{x}{x^2+y^2}. \end{aligned}$$

where in calculating the partial derivatives of L_2 we used that it is equal to $\arctan(y/x)$ in $(-\pi/2, \pi/2)$ (and one can similarly use other inverse trigonometric functions in the rest of the complex plane). Examining the formulae we see that the partial derivatives are all continuous, and obey the Cauchy-Riemann equations, so that Log is indeed complex differentiable.

14.1. **Harmonic functions.** Recall that the two-dimensional Laplace operator Δ is the differential operator $\partial_x^2 + \partial_y^2$ (defined on functions $f\colon \mathbb{R}^2 \to \mathbb{R}$ which are twice differentiable in the sense that their partial derivatives are again differentiable). A function which is in the kernel of the Laplace operator is said to be *harmonic*, that is, a function $u\colon D \to \mathbb{R}$ defined on an open subset D of $\mathbb{R}^2$ is harmonic if $\Delta(u) = \partial_x^2 u + \partial_y^2 u = 0$.

If we work over the complex numbers, then the Laplacian can be factorized[30] as

$$\Delta = (\partial_x + i\partial_y)(\partial_x - i\partial_y) = (\partial_x - i\partial_y)(\partial_x + i\partial_y).$$

The two first-order differential operators $\partial_x + i\partial_y$ and $\partial_x - i\partial_y$ are closely related to the Cauchy-Riemann equations, as we now show, which yields an important connection between complex-differentiable functions and harmonic functions.

Definition 14.11. The *Wirtinger* (partial) derivatives are defined to be $\partial_z = \frac{1}{2}(\partial_x - i\partial_y)$ and $\partial_{\bar{z}} = \frac{1}{2}(\partial_x + i\partial_y)$. By the equation above, we have $\Delta = 4\partial_z\partial_{\bar{z}} = 4\partial_{\bar{z}}\partial_z$ (as operators on twice continuously differentiable functions).

Remark 14.12. Notice that, as you study in Differential Equations, to obtain D'Alembert's solution to the one-dimensional wave equation, one factors $\partial_x^2 - \partial_y^2 = (\partial_x - \partial_y)(\partial_x + \partial_y)$, and then performs the change of coordinates $\eta = x + y, \xi = x - y$. Over the complex numbers, the above factorization of Δ shows that we can analyze the Laplacian in a similar way.

Exercise 14.13. Show that if $T\colon \mathbb{C} \to \mathbb{C}$ is any *real linear* map (that is, viewing $\mathbb{C}$ as $\mathbb{R}^2$ we have $T\colon \mathbb{R}^2 \to \mathbb{R}^2$ is a linear map) then there are unique $a, b \in \mathbb{C}$ such that $T(z) = az + b\bar{z}$. (*Hint: note that the map $z \mapsto az + b\bar{z}$ is $\mathbb{R}$-linear. What matrix does it correspond to as a map from $\mathbb{R}^2$ to itself?*)

Lemma 14.14. *Let U be an open subset of $\mathbb{C}$ and let $f: U \to \mathbb{C}$. Then f satisfies the Cauchy-Riemann equations if and only if $\partial_{\bar{z}} f = 0$.*

Proof. Let $f(z) = u(z) + iv(z)$ where u and v are real-valued. Then we have

$$\partial_{\bar{z}} f = (\partial_x + i\partial_y)(u + iv) = (\partial_x u - \partial_y v) + i(\partial_x v + \partial_y u),$$

thus the result follows by taking real and imaginary parts. $\square$

Corollary 14.15. *Suppose that U is an open subset of $\mathbb{C}$ and $f: U \to \mathbb{C}$ is complex differentiable and $f(z) = u(z) + iv(z)$ are its real and imaginary parts. If u and v are twice continuously*[31] *differentiable then they are harmonic on U. Moreover any function $g\colon U \to \mathbb{R}$ is harmonic if it is twice continuously differentiable and $\partial_z(g)$ is complex differentiable.*

[30] Acting on functions which are twice continuously differentiable, the two first order factors commute.

[31] That is, all of their second partial deriviatives exist and are continuous.

Proof. The previous Lemma shows that if f is complex differentiable then $\partial_{\bar{z}} f = 0$. Since the Laplacian Δ is equal to $4\partial_z \partial_{\bar{z}}$ it follows that

$$\Delta(\Re(f)) = \Re(\Delta(f)) = \Re(4\partial_z \partial_{\bar{z}}(f)) = 0,$$

so that $\Re(f)$ is harmonic. Similarly we find $\Im(f)$ is harmonic. The final part is also immediate from the fact that $\Delta = 4\partial_{\bar{z}}\partial_z$. $\square$

Remark 14.16. We will shortly see that if $f = u + iv$ is complex differentiable then it is in fact infinitely complex differentiable. Since we have seen that $f' = \partial_x f = \frac{1}{i}\partial_y f$ it follows that u and v are in fact infinitely differentiable so the condition in the previous lemma on the existence and continuity of their second derivatives holds automatically. For a proof of the fact that the mixed partial derivatives of a twice continuously differentiable function are equal, see the Appendix.

Corollary 14.15 motivates the following definition:

Definition 14.17. If $u\colon \mathbb{R}^2 \to \mathbb{R}$ is a harmonic function, we say that $v\colon \mathbb{R}^2 \to \mathbb{R}$ is a *harmonic conjugate* of u if $f(z) = u + iv$ is holomorphic.

Notice that if u is harmonic, it is twice differentiable so that its partial derivatives are continuously differentiable. It follows that a function v is a harmonic conjugate precisely if the pair (u, v) satisfy the Cauchy-Riemann equations. Thus provided we can integrate these equations to find v, a harmonic conjugate will exist. We will show later that, at least when the second partial derivatives are continuous, this can always been done locally in the plane.

14.2. **Power series.** Another important family of examples are the functions which arise from power series. We review here the main results about complex power series which were proved in Analysis II last year:

Definition 14.18. Let $(a_n)_{n\geq 0}$ be a sequence of complex numbers. Then we have an associated sequence of polynomials $s_n(z) = \sum_{k=0}^{n} a_k z^k$. Let S be the set on which this sequence converges pointwise, that is

$$S = \{z \in \mathbb{C} : \lim_{n\to\infty} s_n(z) \text{ exists}\}.$$

Note that since $s_n(0) = a_0$ we have $0 \in S$ so in particular S is nonempty. On the set S, we can define a function $s(z) = \lim_n s_n(z) = \sum_{k=0}^{\infty} a_k z^k$ which we call a *power series*. We define the *radius of convergence* R of the power series $\sum_{k\geq 0} a_k z^k$ to be $\sup\{|z| : z \in S\}$ (or ∞ if S is unbounded).

By convention, given any sequence of complex numbers $(c_n)_{n\geq 0}$ we write $\sum_{k=0}^{\infty} c_k z^k$ for the corresponding power series (even though it may be that it converges only for $z = 0$).

We can give an explicit formula for the radius of convergence using the notion of $\limsup$ which we now recall:

Definition 14.19. If $(a_n)_{n\geq 0}$ is a sequence of real numbers, set $s_n = \sup\{a_k : k \geq n\} \in \mathbb{R} \cup \{\infty\}$ (where we take $s_n = \infty$ if $\{a_k : k \geq n\}$ is not bounded above). Then the sequence (s_n) is either constant and equal to ∞ or eventually becomes a decreasing sequence of real numbers. In the first case we set $\limsup_n a_n = \infty$, whereas in the second case we set $\limsup_n a_n = \lim_n s_n$ (which is finite if (s_n) is bounded below, and equal to $-\infty$ otherwise).

Lemma 14.20. *Let $\sum_{k\geq 0} a_k z^k$ be a power series, let S be the subset of $\mathbb{C}$ on which it converges and let R be its radius of convergence. Then we have*

$$B(0,R) \subseteq S \subseteq \bar{B}(0,R).$$

The series converges absolutely on $B(0,R)$ and if $0 \leq r < R$ then it converges uniformly on $\bar{B}(0,r)$. Moreover, we have

$$1/R = \limsup_n |a_n|^{1/n}.$$

Proof. Let $L = \limsup_n |a_n|^{1/n} \in [0,\infty]$. If $L = 0$ then the statement should be understood to say that the radius of convergence R is ∞, while if $L = \infty$ we take $R = 0$. These two cases are in fact similar but easier than the case where $L \in (0,\infty)$, so we will only give the details for the case where L is finite and positive. Let $s_n = \sup\{|a_k|^{1/k} : k \geq n\}$ so that $L = \lim_{n\to\infty} s_n$.

If $0 < s < 1/L$ we can find an $\epsilon > 0$ such that $(L+\epsilon).s = r < 1$. Thus by definition, for sufficiently large n we have $|a_n|^{1/n} \leq s_n < L + \epsilon$ so that if $|z| \leq s$ we have

$$|a_n||z|^n \leq [(L+\epsilon)|z|]^n \leq r^n,$$

and hence by the comparison test, $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely and uniformly on $\bar{B}(0,s)$. It follows the power series converges everywhere in $B(0,1/L)$.

On the other hand, if $|z| > 1/L$ we can find an $\epsilon_1 > 0$ such that $|z|(L - \epsilon_1) = r > 1$. But then for all k we have $s_k \geq L$ since (s_n) is decreasing, and hence by the approximation property for each k we can find an $n_k \geq k$ with $|a_{n_k}|^{1/n_k} > s_k - \epsilon_1 \geq L - \epsilon$ and hence $|a_{n_k} z^{n_k}| > r^k$. Thus $|a_n z^n|$ has a subsequence which does not tend to zero, so the series cannot converge. It follows the radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$ is $1/L$ as claimed.

□

The next lemma is a relatively straight-forward consequence of standard algebra of limits style results:

Lemma 14.21. *Let $s(z) = \sum_{k=0}^{\infty} a_k z^k$ and $t(z) = \sum_{k=0}^{\infty} b_k z^k$ be power series with radii of convergence R_1 and R_2 respectively and let $T = min\{R_1, R_2\}$.*

(1) *Let $c_n = \sum_{k+l=n} a_k b_l$, then the power series $\sum_{n=0}^{\infty} c_n z^n$ has radius of convergence at least T and if $|z| < T$ we have*

$$\sum_{n=0}^{\infty} c_n z^n = s(z)t(z).$$

Thus the product of power series is a power series.

(2) *If $s(z)$ and $t(z)$ are as above, then $\sum_{k=0}^{\infty}(a_k+b_k)z^k$ is a power series which converges to $s(z)+t(z)$ in $B(0,T)$, thus the sum of power series is again a power series.*

Proof. This was established in Prelims Analysis II. Note that T is only a lower bound for the radius of convergence in each case – it is easy to find examples where the actual radius of convergence of the sum or product is strictly larger than T. □

The behaviour of a power series at its radius of convergence is in general a rather complicated phenomenon. The following result, which we shall not prove, gives some information however. Some of the ideas involved in its proof are investigated in Problem Set 4.

Theorem 14.22. *(Abel's theorem:) Suppose that (a_n) is a sequence of complex numbers and $\sum_{n=0}^{\infty} a_n$ exists. Then the series $\sum_{n=0}^{\infty} a_n z^n$ converges for $|z|<1$ and*

$$\lim_{\substack{r\in(-1,1)\\ r\uparrow 1}} \big(\sum_{n=0}^{\infty} a_n r^n\big) = \sum_{n=0}^{\infty} a_n.$$

Proof. Note that since the series $\sum_{n=0}^{\infty} a_n z^n$ converges at $z=1$ by assumption, its radius of convergence is at least 1, so that the first statement holds. For some idea of what goes into the proof of the second part, see the Problem sets. □

Proposition 14.23. *Let $s(z)=\sum_{k\geq 0} a_k z^k$ be a power series, let S be the domain on which it converges, and let R be its radius of convergence. Then power series $t(z)=\sum_{k=1}^{\infty} k a_k z^{k-1}$ also has radius of convergence R and on $B(0,R)$ the power series s is complex differentiable with $s'(z)=t(z)$. In particular, it follows that a power series is infinitely complex differentiable within its radius of convergence.*

Proof. This is proved in Prelims Analysis II. An alternative proof is given in Appendix II. □

Example 14.24. The previous Proposition gives us a large supply of complex differentiable functions. For example,

$$\exp(z)=\sum_{n=0}^{\infty}\frac{z^n}{n!},\quad \cos(z)=\sum_{n=0}^{\infty}(-1)^n\frac{z^{2n}}{(2n)!},\quad \sin(z)=\sum_{n=0}^{\infty}(-1)^n\frac{z^{2n+1}}{(2n+1)!},$$

are all complex differentiable on the whole complex plane (since $R=\infty$ in each case). Note that one can use the above theorem to show that $\cos(z)^2+\sin(z)^2=1$ for all $z\in\mathbb{C}$, but since $\sin(z)$ and $\cos(z)$ are not in general real, this does not imply that $|\sin(z)|$ or $|\cos(z)|$ at most 1. (In fact it is easy to check that they are both unbounded on $\mathbb{C}$). Using what we have already established about power series it is also easy to check that the complex

$\sin$ function encompases both the real trigonometric and real hyperbolic functions, indeed:

$$\sin(a + ib) = \sin(a)\cosh(b) + i\cos(a)\sinh(b).$$

Example 14.25. Let $s(z) = \sum_{n=1}^{\infty} \frac{z^n}{n}$. Then $s(z)$ has radius of convergence 1, and in $B(0,1)$ we have $s'(z) = \sum_{n=0}^{\infty} z^n = 1/(1-z)$, thus this power series is a complex differentiable function which extends the function $-\log(1-z)$ on the interval $(-1,1)$ to the open disc $B(0,1) \subset \mathbb{C}$. We will see later that we will not be able to extend the function $\log$ to a complex differentiable function on $\mathbb{C}\backslash\{0\}$ – we will only be able to construct a "multi-valued" extension.

Note that, slightly more generally, we can work with power series centred at an arbitrary point $z_0 \in \mathbb{C}$. Such power series are functions given by an expression of the form

$$f(z) = \sum_{n \geq 0} a_n (z - z_0)^n.$$

All the results we have shown above immediately extend to these more general power series, since if

$$g(z) = \sum_{n \geq 0} a_n z^n,$$

then the function f is obtained from g simply by composing with the translation $z \mapsto z - z_0$. In particular, the chain rule shows that

$$f'(z) = \sum_{n \geq 1} n a_n (z - z_0)^{n-1}.$$

15. BRANCH CUTS

It is often the case that we study a holomorphic function on a domain $D \subseteq \mathbb{C}$ which does not extend to a function on the whole complex plane.

Example 15.1. Consider the square root "function" $f(z) = z^{1/2}$. Unlike the case of real numbers, every complex number has a square root, but just as for the real numbers, there are two possiblities unless $z = 0$. Indeed if $z = x + iy$ and $w = u + iv$ has $w^2 = z$ we see that

$$u^2 - v^2 = x; \quad 2uv = y,$$

and so

$$u^2 = \frac{x + \sqrt{x^2 + y^2}}{2}, v^2 = \frac{y + \sqrt{x^2 + y^2}}{2}.$$

where the requirement that u^2, v^2 are nonnegative determines the signs. Hence taking square roots we obtain the two possible solutions for w satifying $w^2 = z$. (Note it looks like there are four possible sign combinations in the above, however the requirement that $2uv = y$ means the sign of u determines that of v.) In polars it looks simpler: if $z = re^{i\theta}$ then $w = \pm r^{1/2}e^{i\theta/2}$. Indeed this expression gives us a continuous choice of square root except at the positive real axis: for any $z \in \mathbb{C}$ we may write z uniquely as $re^{i\theta}$ where $\theta \in [0, 2\pi)$, and then set $f(z) = r^{1/2}e^{i\theta/2}$. But now for θ small and positive, $f(z) = r^{1/2}e^{i\theta}$ has small positive argument, but if $z = re^{(2\pi-\epsilon)i}$ we find $f(z) = r^{1/2}e^{(\pi-\epsilon/2)i}$, thus $f(z)$ in the first case is just above the positive real axis, while in the second case $f(z)$ is just below the negative real axis. Thus the function f is only continuous on $\mathbb{C}\backslash\{z \in \mathbb{C} : \Im(z) = 0, \Re(z) > 0\}$. Using Theorem 14.9 you can check f is also holomorphic on this domain. The positive real axis is called a *branch cut* for the *multi-valued function* $z^{1/2}$. By chosing different intervals for the argument (such as $(-\pi, \pi]$ say) we can take different cuts in the plane and obtain different *branches* of the function $z^{1/2}$ defined on their complements.

We formalize these concepts as follows:

Definition 15.2. A *multi-valued function* or *multifunction* on a subset $U \subseteq \mathbb{C}$ is a map $f : U \to \mathcal{P}(\mathbb{C})$ assigning to each point in U a subset[32] of the complex numbers. A *branch* of f on a subset $V \subseteq U$ is a function $g : V \to \mathbb{C}$ such that $g(z) \in f(z)$, for all $z \in V$. If g is continuous (or holomorphic) on V we refer to it as a continuous, (respectively holomorphic) branch of f. We will primarily be interested in branches of multifunctions which are holomorphic.

Remark 15.3. In order to distinguish between multifunctions and functions, it is sometimes useful to introduce some notation: if we wish to consider $z \mapsto z^{1/2}$ as a multifunction, then to emphasize that we mean a multifunction we will write $[z^{1/2}]$. Thus $[z^{1/2}] = \{w \in \mathbb{C} : w^2 = z\}$. Similarly we write $[\mathrm{Log}(z)] = \{w \in \mathbb{C} : e^w = z\}$. This is not a uniform convention in the subject, but is used, for example, in the text of Priestley.

Thus the square root $z \mapsto [z^{1/2}]$ is a multifunction, and we saw above that we can obtain holomorphic branches of it on a cut plane $\mathbb{C}\backslash R$ where $R = \{te^{i\theta} : t \in \mathbb{R}_{\geq 0}\}$. The point here is that both the origin and infinity as "branch points" for the multifunction $[z^{1/2}]$.

Definition 15.4. Suppose that $f : U \to \mathcal{P}(\mathbb{C})$ is a multi-valued function defined on an open subset U of $\mathbb{C}$. We say that $z_0 \in U$ is not a branch point of f if there is an open disk[33] $D \subseteq U$ containing z_0 such that there is a holomorphic branch of f defined on $D\backslash\{z_0\}$. We say z_0 is a *branch point* otherwise. When $\mathbb{C}\backslash U$ is bounded, we say that f does not have a branch

[32] We use the notation $\mathcal{P}(X)$ to denote the *power set* of X, that is, the set of all subsets of X.

[33] In fact any simply-connected domain – see our discussion of the homotopy form of Cauchy's theorem.

point at ∞ if there is a branch of f defined on $\mathbb{C}\backslash B(0,R) \subseteq U$ for some $R>0$. Otherwise we say that ∞ is a branch point of f.

A *branch cut* for a multifunction f is a curve in the plane on whose complement we can pick a holomorphic branch of f. Thus a branch cut must contain all the branch points.

Example 15.5. Another important example of a multi-valued function which we have already discussed is the complex logarithm: as a multifunction we have $\text{Log}(z) = \{\log(|z|) + i(\theta + 2n\pi) : n \in \mathbb{Z}\}$ where $z = |z|e^{i\theta}$. To obtain a branch of the multifunction we must make a choice of argument function $\arg\colon \mathbb{C} \to \mathbb{R}$ we may define

$$\text{Log}(z) = \log(|z|) + i\arg(z),$$

which is a continuous function away from the branch cut we chose. By convention, the *principal branch* of Log is defined by taking $\arg(z) \in (-\pi, \pi]$.

Another important class of examples of multifunctions are the *fractional power* multifunctions $z \mapsto [z^\alpha]$ where $\alpha \in \mathbb{C}$: These are given by

$$z \mapsto \exp(\alpha.[\text{Log}(z)]) = \{\exp(\alpha.w) : w \in \mathbb{C}, e^w = z\}$$

Note this is includes the square root multifunction we discussed above, which can be defined without the use of exponential function. Indeed if $\alpha = m/n$ is rational, $m \in \mathbb{Z}, n \in \mathbb{Z}_{>0}$, then $[z^\alpha] = \{w \in \mathbb{C} : w^m = z^n\}$. For $\alpha \in \mathbb{C}\backslash\mathbb{Q}$ however we can only define $[z^\alpha]$ using the exponential function. Clearly from its definition, anytime we choose a branch $L(z)$ of $[Log(z)]$ we obtain a corresponding branch $\exp(\alpha.L(z))$ of $[z^\alpha]$. If $L(z)$ is the principal branch of $[\text{Log}(z)]$ then the corresponding branch of $[z^\alpha]$ is called the *principal branch* of $[z^\alpha]$.

Example 15.6. Let $F(z)$ be the multi-function

$$[(1+z)^\alpha] = \{\exp(\alpha.w) : w \in \mathbb{C}, \exp(w) = 1+z\}.$$

Using $L(z)$ the principal branch of $[\text{Log}(z)]$ we obtain a branch $f(z)$ of $[(1+z)^\alpha]$ given by $f(z) = \exp(\alpha.L(1+z))$. Let $\binom{\alpha}{k} = \frac{1}{k!}\alpha.(\alpha-1)\dots(\alpha-k+1)$. We want to show that a version of the binomial theorem holds for this branch of the multifunction $[(1+z)^\alpha]$. Let

$$s(z) = \sum_{k=0}^{\infty} \binom{\alpha}{k} z^k,$$

By the ratio test, $s(z)$ has radius of convergence equal to 1, so that $s(z)$ defines a holomorphic function in $B(0,1)$. Moreover, you can check using the properties of power series established in the previous section that, within $B(0,1)$, $s(z)$ satisfies $(1+z)s'(z) = a.s(z)$.

Now $f(z)$ is defined on $\mathbb{C}\backslash(-\infty,-1)$, and hence on all of $B(0,1)$. Moreover[34] We claim that within the open ball $B(0,1)$ the power series $s(z) =$

[34]Any continuous branch $L(z)$ of $[\text{Log}(z)]$ is holomorphic where it is defined and satisfies $\exp(L(z)) = z$, hence by the chain rule one obtains $L'(z) = 1/z$.

$\sum_{n=0}^{\infty} \binom{\alpha}{k} z^k$ coincides with $f(z)$. Indeed we have

$$\frac{d}{dz}(L(s(z))) = s'(z)/s(z) = \alpha/(1+z) = \frac{d}{dz}(\alpha L(1+z))$$

so that $L(s(z)) = \alpha.L(1+z)+c$ for some constant c (as $B(0,1)$ is connected) which by evaluating at $z=0$ we find is zero. Finally, it follows that $s(z) = \exp(\alpha L(1+z))$ so that $s(z) \in [(1+z)^\alpha]$ as required.

Example 15.7. A more interesting example is the function $f(z) = [(z^2-1)^{1/2}]$. Using the principal branch of the square root function, we obtain a branch f_1 of f on the complement of $E = \{z \in \mathbb{C} : z^2-1 \in (-\infty, 0]\}$, which one calculates is equal to $(-1,1) \cup i\mathbb{R}$. If we cross either the segment $(-1,1)$ or the imaginary axis, this branch of f is discontinuous.

To find another branch, note that we may write $f(z) = \sqrt{z-1}\sqrt{z+1}$, thus we can take the principal branch of the square root for each of these factors. More explicity, if we write $z = 1 + re^{i\theta_1} = -1 + se^{i\theta_2}$ where $\theta_1, \theta_2 \in (-\pi, \pi]$ then we get a branch of f given by $f_2(z) = \sqrt{rs}.e^{i(\theta_1+\theta_2)/2}$. Now the factors are discontinuous on $(-\infty, 1]$ and $(\infty, -1]$ respectively, however let us examine the behaviour of their product: If z crosses the negative real axis at $\Im(z) < -1$ then θ_1 and θ_2 both jumps by 2π, so that $(\theta_1+\theta_2)/2$ jumps by 2π, and hence $\exp((\theta_1+\theta_2)/2)$ is in fact continuous. On the other hand, if we cross the segment $(-1,1)$ then only the factor $\sqrt{z-1}$ switches sign, so our branch is discontinuous there. Thus our second branch of f is defined away from the cut $[-1,1]$.

Example 15.8. The branch points of the complex logarithm are 0 and infinity: indeed if $z_0 \neq 0$ then we can find a half-plane $H = \{z \in \mathbb{C} : \Im(az) > 0\}$, for some $a \in \mathbb{C}$, $|a| = 1$, such that $z_0 \in H$. We can chose a continuous choice of argument function on H, and this gives a holomorphic branch of Log defined on H and hence on the disk $B(z_0, r)$ for r sufficiently small. The logarithm also has a branch point at infinity, since we cannot chose a continous argument function on $\mathbb{C} \backslash B(0,R)$ for any $R > 0$. (We will return to this point when discussing the winding number later in the course.)

Note that if $f(z) = [\sqrt{z^2-1}]$ then the second of our branches f_2 discussed above shows that f does not have a branch point at infinity, whereas both 1 and -1 are branch points – as we move in a sufficiently small circle around we cannot make a continuous choice of branch. One can given a rigorous proof of this using the branch f_2: given any branch g of $[\sqrt{z^2-1}]$ defined on $B(1,r)$ for $r < 1$ one proves that $g = \pm f_2$ so that g is not continuous on $B(0,r) \cap (-1,1)$. See Problem Sheet 4, question 5, for more details.

Example 15.9. A more sophisticated point of view on branch points and cuts uses the theory of Riemann surfaces. As a first look at this theory, consider the multifunction $f(z) = [\sqrt{z^2-1}]$ again. Let $\Sigma = \{(z,w) \in \mathbb{C}^2 : w^2 = z^2 - 1\}$ (this is an example of a Riemann surface). Then we have two maps from Σ to $\mathbb{C}$, projecting along the first and second factor:

$p_1(z,w) = z$ and $p_2(z,w) = w$. Now if $g(z)$ is a branch of f, it gives us a map $G\colon \mathbb{C} \to \Sigma$ where $G(z) = (z, g(z))$. If we take $f_2(z) = \sqrt{z-1}\sqrt{z+1}$ (using the principal branch of the square root function in each case, then let $\Sigma_+\{(z, f_2(z)) : z \notin [-1,1]\}$ and $\Sigma_- = \{(z, -f_2(z)) : z \notin [-1,1]\}$, then $\Sigma_+ \cup \Sigma_-$ covers all of Σ apart from the pairs (z,w) where $z \in [-1,1]$. For such z we have $w = \pm i\sqrt{1-z^2}$, and Σ is obtained by gluing together the two copies Σ_+ and Σ_- of the cut plane $\mathbb{C}\backslash[-1,1]$ along the cut locus $[-1,1]$. However, we must examine the discontinuity of g in order to see how this gluing works: the upper side of the cut in Σ_+ is glued to the lower side of the cut in Σ_- and similarly the lower side of the cut in Σ_+ is glued to the upper side of Σ_-.

Notice that on Σ we have the (single-valued) function $p_2(z,w) = w$, and any map $q\colon U \to \Sigma$ from an open subset U of $\mathbb{C}$ to Σ such that $p_1 \circ q(z) = z$ gives a branch of $f(z) = \sqrt{z^2-1}$ given by $p_2 \circ q$. Such a function is called a *section* of p_1. Thus the multi-valued function on $\mathbb{C}$ becomes a single-valued function on Σ, and a branch of the multifunction corresponds to a section of the map $p_1\colon \Sigma \to \mathbb{C}$. In general, given a multi-valued function f one can construct a Riemann surface Σ by gluing together copies of the cut complex plane to obtain a surface on which our multifunction becomes a single-valued function.

16. Paths and Integration

Paths will play a crucial role in our development of the theory of complex differentiable functions. In this section we review the notion of a path and define the integral of a continuous function along a path.

16.1. **Paths.** Recall that a *path* in the complex plane is a continuous function $\gamma\colon [a,b] \to \mathbb{C}$. A path is said to be *closed* if $\gamma(a) = \gamma(b)$. If γ is a path, we will write γ^* for its image, that is

$$\gamma^* = \{z \in \mathbb{C} : z = \gamma(t), \text{ some } t \in [a,b]\}.$$

Although for some purposes it suffices to assume that γ is continuous, in order to make sense of the integral along a path we will require our paths to be (at least piecewise) differentiable. We thus need to define what we mean for a path to be differentiable:

Definition 16.1. We will say that a path $\gamma\colon [a,b] \to \mathbb{C}$ is *differentiable* if its real and imaginary parts are differentiable as real-valued functions. Equivalently, γ is differentiable at $t_0 \in [a,b]$ if

$$\lim_{t \to t_0} \frac{\gamma(t) - \gamma(t_0)}{t - t_0}$$

exists, and then we denote this limit as $\gamma'(t_0)$. (If $t = a$ or b then we interpret the above as a one-sided limit.) We say that a path is C^1 if it is differentiable and its derivative $\gamma'(t)$ is continuous.

We will say a path is *piecewise* C^1 if it is continuous on $[a,b]$ and the interval $[a,b]$ can be divided into subintervals on each of which γ is C^1. That is, there is a finite sequence $a = a_0 < a_1 < \ldots < a_m = b$ such that $\gamma_{|[a_i,a_{i+1}]}$ is C^1. Thus in particular, the left-hand and right-hand derivatives of γ at a_i $(1 \leq i \leq m-1)$ may not be equal.

Remark 16.2. Note that a C^1 path may not have a well-defined tangent at every point: if $\gamma\colon [a,b] \to \mathbb{C}$ is a path and $\gamma'(t) \neq 0$, then the line $\{\gamma(t) + s\gamma'(t) : s \in \mathbb{R}\}$ is tangent to γ^*, however if $\gamma'(t) = 0$, the image of γ may have no tangent line there. Indeed consider the example of $\gamma\colon [-1,1] \to \mathbb{C}$ given by

$$\gamma(t) = \begin{cases} t^2 & -1 \leq t \leq 0 \\ it^2 & 0 \leq t \leq 1. \end{cases}$$

Since $\gamma'(0) = 0$ the path is C^1, even though it is clear there is no tangent line to the image of γ at 0.

If $s\colon [a,b] \to [c,d]$ is a differentiable map, then we have the following version of the chain rule, which is proved in exactly the same way as the real-valued case. It will be crucial in our definition of the integral of functions $f\colon \mathbb{C} \to \mathbb{C}$ along paths.

Lemma 16.3. *Let $\gamma\colon [c,d] \to \mathbb{C}$ and $s\colon [a,b] \to [c,d]$ and suppose that s is differentiable at t_0 and γ is differentiable at $s_0 = s(t_0)$. Then $\gamma \circ s$ is differentiable at t_0 with derivative*

$$(\gamma \circ s)'(t_0) = s'(t_0).\gamma'(s(t_0)).$$

Proof. Let $\epsilon\colon [c,d] \to \mathbb{C}$ be given by $\epsilon(s_0) = 0$ and

$$\gamma(x) = \gamma(s_0) + \gamma'(s_0)(x - s_0) + (x - s_0)\epsilon(x),$$

(so that this equation holds for all $x \in [c,d]$), then $\epsilon(x) \to 0$ as $x \to s_0$ by the definition of $\gamma'(s_0)$, *i.e.* ϵ is continuous at t_0. Substituting $x = s(t)$ into this we see that for all $t \neq t_0$ we have

$$\frac{\gamma(s(t)) - \gamma(s_0)}{t - t_0} = \frac{s(t) - s(t_0)}{t - t_0}\big(\gamma'(s(t)) + \epsilon(s(t))\big).$$

Now $s(t)$ is continuous at t_0 since it is differentiable there hence $\epsilon(s(t)) \to 0$ as $t \to t_0$, thus taking the limit as $t \to t_0$ we see that

$$(\gamma \circ s)'(t_0) = s'(t_0)(\gamma'(s_0) + 0) = s'(t_0)\gamma'(s(t_0)),$$

as required. □

Definition 16.4. If $\phi\colon [a,b] \to [c,d]$ is continuously differentiable with $\phi(a) = c$ and $\phi(b) = d$, and $\gamma\colon [c,d] \to \mathbb{C}$ is a C^1-path, then setting $\tilde{\gamma} = \gamma \circ \phi$, by Lemma 16.3 we see that $\tilde{\gamma}\colon [a,b] \to \mathbb{C}$ is again a C^1-path with the same image as γ and we say that $\tilde{\gamma}$ is a *reparametrization* of γ.

Definition 16.5. We will say two parametrized paths $\gamma_1 \colon [a,b] \to \mathbb{C}$ and $\gamma_2 \colon [c,d] \to \mathbb{C}$ are *equivalent* if there is a continuously differentiable bijective function $s \colon [a,b] \to [c,d]$ such that $s'(t) > 0$ for all $t \in [a,b]$ and $\gamma_1 = \gamma_2 \circ s$. It is straight-forward to check that equivalence is indeed an equivalence relation on parametrized paths, and we will call the equivalence classes *oriented curves* in the complex plane. We denote the equivalence class of γ by $[\gamma]$. The condition that $s'(t) > 0$ ensures that the path is traversed in the same direction for each of the parametrizations γ_1 and γ_2. Moreover γ_1 is piecewise C^1 if and only if γ_2 is.

Recall that we saw before (in a general metric space) that any path $\gamma \colon [a,b] \to \mathbb{C}$ has an *opposite* path γ^- and that two paths $\gamma_1 \colon [a,b] \to \mathbb{C}$ and $\gamma_2 \colon [c,d] \to \mathbb{C}$ with $\gamma_1(b) = \gamma_2(c)$ can be *concatenated* to give a path $\gamma_1 \star \gamma_2$. If $\gamma, \gamma_1, \gamma_2$ are piecewise C^1 then so are γ^- and $\gamma_1 \star \gamma_2$. (Indeed a piecewise C^1 path is precisely a finite concatenation of C^1 paths).

Remark 16.6. Note that if $\gamma \colon [a,b] \to \mathbb{C}$ is piecewise C^1, then by choosing a reparametrization by a function $\psi \colon [a,b] \to [a,b]$ which is strictly increasing and has vanishing derivative at the points where γ fails to be C^1, we can replace γ by $\tilde{\gamma} = \gamma \circ \psi$ to obtain a C^1 path with the same image. For this reason, some texts insist that C^1 paths have everywhere non-vanishing derivative. In this course we will not insist on this. Indeed sometimes it is convenient to consider a *constant* path, that is a path $\gamma \colon [a,b] \to \mathbb{C}$ such that $\gamma(t) = z_0$ for all $t \in [a,b]$ (and hence $\gamma'(t) = 0$ for all $t \in [a,b]$).

Example 16.7. The most basic example of a closed curve is a circle: If $z_0 \in \mathbb{C}$ and $r > 0$ then the path $z(t) = z_0 + re^{2\pi i t}$ (for $t \in [0,1]$) is the simple closed path with *positive orientation* encircling z_0 with radius r. The path $\tilde{z}(t) = z_0 + re^{-2\pi i t}$ is the simple closed path encircling z_0 with radius r and *negative orientation*.

Another useful path is a line segment: if $a, b \in \mathbb{C}$ then the path $\gamma_{[a,b]} \colon [0,1] \to \mathbb{C}$ given by $t \mapsto a + t(b-a) = (1-t)a + tb$ traverses the line segment from a to b. We denote the corresponding oriented curve by $[a,b]$ (which is consistent with the notation for an interval in the real line). One of the simplest classes of closed paths are triangles: given three points a, b, c, we define the triangle, or triangular path, associated to them, to be the concatenation of the associated line segments, that is $T_{a,b,c} = \gamma_{a,b} \star \gamma_{b,c} \star \gamma_{c,a}$.

16.2. **Integration along a path.** To define the integral of a complex-valued function along a path, we first need to be able to integrate functions $F \colon [a,b] \to \mathbb{C}$ on a closed interval $[a,b]$ taking values in $\mathbb{C}$. Last year in Analysis III the Riemann integral was defined for a function on a closed interval $[a,b]$ taking values in $\mathbb{R}$, but it is easy to extend this to functions taking values in $\mathbb{C}$: Indeed we may write $F(t) = G(t) + iH(t)$ where G, H are functions on $[a,b]$ taking real values. Then we say that F is Riemann integrable if both

G and H are, and we define:

$$\int_a^b F(t)dt = \int_a^b G(t)dt + i\int_a^b H(t)dt$$

It is easy to check that the integral is then complex linear, that is, if F_1, F_2 are complex-valued Riemann integrable functions on $[a,b]$, and $\alpha, \beta \in \mathbb{C}$, then $\alpha F_1 + \beta F_2$ is Riemann integrable and

$$\int_a^b (\alpha.F_1 + \beta.F_2)dt = \alpha.\int_a^b F_1 dt + \beta.\int_a^b F_2 dt.$$

Note that if F is continuous, then its real and imaginary parts are also continuous, and so in particular Riemann integrable[35]. The class of Riemann integrable (real or complex valued) functions on a closed interval is however slightly larger than the class of continuous functions, and this will be useful to us at certain points. In particular, we have the following:

Lemma 16.8. *Let $[a,b]$ be a closed interval and $S \subset [a,b]$ a finite set. If f is a bounded continuous function (taking real or complex values) on $[a,b]\backslash S$ then it is Riemann integrable on $[a,b]$.*

Proof. The case of complex-valued functions follows from the real case by taking real and imaginary parts. For the case of a function $f\colon [a,b]\backslash S \to \mathbb{R}$, let $a = x_0 < x_1 < x_2 < \ldots < x_k = b$ be any partition of $[a,b]$ which includes the elements of S. Then on each open interval (x_i, x_{i+1}) the function f is bounded and continuous, and hence integrable. We may therefore set

$$\int_a^b f(t)dt = \int_{x_0}^{x_1} f(t)dt + \int_{x_1}^{x_2} f(t)dt + \ldots + \int_{x_{k-1}}^{x_k} f(t)dt$$

The standard additivity properties of the integral then show that $\int_a^b f(t)dt$ is independent of any choices. □

Remark 16.9. Note that normally when one speaks of a function f being integrable on an interval $[a,b]$ one assumes that f is defined on all of $[a,b]$. However, if we change the value of a Riemann integrable function f at a finite set of points, then the resulting function is still Riemann integrable and its integral is the same. Thus if one prefers the function f in the previous lemma to be defined on all of $[a,b]$ one can define f to take any values at all on the finite set S.

It is easy to check that the Riemann integral of complex-valued functions is complex linear. We also note a version of the triangle inequality for complex-valued functions:

[35] It is clear this definition extends to give a notion of the integral of a function $f\colon [a,b] \to \mathbb{R}^n$ – we say f is integrable if each of its components is, and then define the integral to be the vector given by the integrals of each component function.

Lemma 16.10. *Suppose that $F\colon [a,b] \to \mathbb{C}$ is a complex-valued function. Then we have*

$$|\int_a^b F(t)dt| \leq \int_a^b |F(t)|dt.$$

Proof. First note that if $F(t) = x(t) + iy(t)$ then $|F(t)| = \sqrt{x^2+y^2}$ so that if F is integrable $|F(t)|$ is also[36]. We may write $\int_a^b F(t)dt = re^{i\theta}$, where $r \in [0,\infty)$ and $\theta \in [0, 2\pi)$. Now taking the components of F in the direction of $e^{i\theta}$ and $e^{i(\theta+\pi/2)} = ie^{i\theta}$, we may write $F(t) = u(t)e^{i\theta} + iv(t)e^{i\theta}$. Then by our choice of θ we have $\int_a^b F(t)dt = e^{i\theta}\int_a^b u(t)dt$, and so

$$|\int_a^b F(t)dt| = |\int_a^b u(t)dt| \leq \int_a^b |u(t)|dt \leq \int_a^b |F(t)|dt,$$

where in the first inequality we used the triangle inequality for the Riemann integral of real-valued functions. □

We are now ready to define the integral of a function $f\colon \mathbb{C} \to \mathbb{C}$ along a piecewise-C^1 curve.

Definition 16.11. If $\gamma\colon [a,b] \to \mathbb{C}$ is a piecewise-C^1 path and $f\colon \mathbb{C} \to \mathbb{C}$, then we define the integral of f along γ to be

$$\int_\gamma f(z)dz = \int_a^b f(\gamma(t))\gamma'(t)dt.$$

In order for this integral to exist in the sense we have defined, we have seen that it suffices for the functions $f(\gamma(t))$ and $\gamma'(t)$ to be bounded and continuous at all but finitely many t. Our definition of a piecewise C^1-path ensures that $\gamma'(t)$ is bounded and continuous away from finitely many points (the boundedness follows from the existence of the left and right hand limits at points of discontinuity of $\gamma'(t)$). For most of our applications, the function f will be continuous on the whole image γ^* of γ, but it will occasionally be useful to weaken this to allow $f(\gamma(t))$ finitely many (bounded) discontinuities.

Lemma 16.12. *If $\gamma\colon [a,b] \to \mathbb{C}$ be a piecewise C^1 path and $\tilde{\gamma}\colon [c,d] \to \mathbb{C}$ is an equivalent path, then for any continuous function $f\colon \mathbb{C} \to \mathbb{C}$ we have*

$$\int_\gamma f(z)dz = \int_{\tilde{\gamma}} f(z)dz.$$

In particular, the integral only depends on the oriented curve $[\gamma]$.

Proof. Since $\tilde{\gamma}$ is equivalent to γ there is a continuously differentiable function $s\colon [c,d] \to [a,b]$ with $s(c) = a$, $s(d) = b$ and $s'(t) > 0$ for all $t \in [c,d]$.

[36] The simplest way to see this is to use that fact that if ϕ is continuous and f is Riemann integrable, then $\phi \circ f$ is Riemann integrable.

Suppose first that γ is C^1. Then by the chain rule we have

$$\begin{aligned}\int_{\tilde{\gamma}} f(z)dz &= \int_c^d f(\gamma(s(t)))(\gamma \circ s)'(t)dt \\ &= \int_c^d f(\gamma(s(t))\gamma'(s(t))s'(t)dt \\ &= \int_a^b f(\gamma(s))\gamma'(s)ds \\ &= \int_\gamma f(z)dz.\end{aligned}$$

where in the second last equality we used the change of variables formula. If $a = x_0 < x_1 < \ldots < x_n = b$ is a decomposition of $[a, b]$ into subintervals such that γ is C^1 on $[x_i, x_{i+1}]$ for $1 \leq i \leq n-1$ then since s is a continuous increasing bijection, we have a corresponding decomposition of $[c, d]$ given by the points $s^{-1}(x_0) < \ldots < s^{-1}(x_n)$, and we have

$$\begin{aligned}\int_{\tilde{\gamma}} f(z)dz &= \int_c^d f(\gamma(s(t))\gamma'(s(t))s'(t)dt \\ &= \sum_{i=0}^{n-1} \int_{s^{-1}(x_i)}^{s^{-1}(x_{i+1})} f(\gamma(s(t))\gamma'(s(t))s'(t)dt \\ &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(\gamma(x))\gamma'(x)dx \\ &= \int_a^b f(\gamma(x))\gamma'(x)dx = \int_\gamma f(z)dz.\end{aligned}$$

where the third equality follows from the case of C^1 paths established above. $\square$

Definition 16.13. If $\gamma \colon [a, b] \to \mathbb{C}$ is a C^1 path then we define the *length* of γ to be

$$\ell(\gamma) = \int_a^b |\gamma'(t)|dt.$$

Using the chain rule as we did to show that the integrals of a function $f \colon \mathbb{C} \to \mathbb{C}$ along equivalent paths are equal, one can check that the length of a parametrized path is also constant on equivalence classes of paths, so in fact the above defines a length function for oriented curves. The definition extends in the obvious way to give a notion of length for piecewise C^1-paths. More generally, one can define the integral *with respect to arc-length* of a function $f \colon U \to \mathbb{C}$ such that $\gamma^* \subseteq U$ to be

$$\int_\gamma f(z)|dz| = \int_a^b f(\gamma(t))|\gamma'(t)|dt.$$

This integral is invariant with respect to C^1 reparametrizations $s\colon [c,d] \to [a,b]$ if we require $s'(t) \neq 0$ for all $t \in [c,d]$ (the condition $s'(t) > 0$ is not necessary because of this integral takes the modulus of $\gamma'(t)$). In particular $\ell(\gamma) = \ell(\gamma^-)$.

The integration of functions along piecewise smooth paths has many of the properties that the integral of real-valued functions along an interval possess. We record some of the most standard of these:

Proposition 16.14. *Let $f, g\colon U \to \mathbb{C}$ be continuous functions on an open subset $U \subseteq \mathbb{C}$ and $\gamma, \eta\colon [a,b] \to \mathbb{C}$ be piecewise-C^1 paths whose images lie in U. Then we have the following:*

(1) *(Linearity): For $\alpha, \beta \in \mathbb{C}$,*

$$\int_\gamma (\alpha f(z) + \beta g(z)) dz = \alpha \int_\gamma f(z) dz + \beta \int_\gamma g(z) dz.$$

(2) *If γ^- denotes the opposite path to γ then*

$$\int_\gamma f(z) dz = -\int_{\gamma^-} f(z) dz.$$

(3) *(Additivity): If $\gamma \star \eta$ is the concatenation of the paths γ, η in U, we have*

$$\int_{\gamma\star\eta} f(z) dz = \int_\gamma f(z) dz + \int_\eta f(z) dz.$$

(4) *(Estimation Lemma.) We have*

$$\left|\int_\gamma f(z) dz\right| \leq \sup_{z\in\gamma^*} |f(z)|.\ell(\gamma).$$

Proof. Since f, g are continous, and γ, η are piecewise C^1, all the integrals in the statement are well-defined: the functions $f(\gamma(t))\gamma'(t)$, $f(\eta(t))\eta'(t)$, $g(\gamma(t))\gamma'(t)$ and $g(\eta(t))\eta'(t)$ are all Riemann integrable. It is easy to see that one can reduce these claims to the case where γ is smooth. The first claim is immediate from the linearity of the Riemann integral, while the second claim follows from the definitions and the fact that $(\gamma^-)'(t) = -\gamma'(a+b-t)$. The third follows immediately for the corresponding additivity property of Riemann integrable functions.

For the fourth part, first note that $\gamma([a,b])$ is compact in $\mathbb{C}$ since it is the image of the compact set $[a,b]$ under a continuous map. It follows that the function $|f|$ is bounded on this set so that $\sup_{z\in\gamma([a,b])} |f(z)|$ exists. Thus we have

$$\begin{aligned}
\Big|\int_\gamma f(z)dz\Big| &= \Big|\int_a^b f(\gamma(t))\gamma'(t)dt\Big| \\
&\leq \int_a^b |f(\gamma(t))||\gamma'(t)|dt \\
&\leq \sup_{z\in\gamma^*}|f(z)|\int_a^b |\gamma'(t)|dt \\
&= \sup_{z\in\gamma^*}|f(z)|.\ell(\gamma).
\end{aligned}$$

where for the first inequality we use the triangle inequality for complex-valued functions as in Lemma 16.10 and the positivity of the Riemann integral for the second inequality. □

Remark 16.15. We give part (4) of the above proposition a name (the "estimation lemma") because it will be very useful later in the course. We will give one important application of it now:

Proposition 16.16. *Let $f_n\colon U\to\mathbb{C}$ be a sequence of continuous functions on an open subset U of the complex plane. Suppose that $\gamma\colon[a,b]\to\mathbb{C}$ is a path whose image is contained in U. If (f_n) converges uniformly to a function f on the image of γ then*

$$\int_\gamma f_n(z)dz \to \int_\gamma f(z)dz.$$

Proof. We have

$$\begin{aligned}
\left|\int_\gamma f(z)dz - \int_\gamma f_n(z)dz\right| &= \left|\int_\gamma (f(z)-f_n(z))dz\right| \\
&\leq \sup_{z\in\gamma^*}\{|f(z)-f_n(z)|\}.\ell(\gamma),
\end{aligned}$$

by the estimation lemma. Since we are assuming that f_n tends to f uniformly on γ^* we have $\sup\{|f(z)-f_n(z)| : z\in\gamma^*\}\to 0$ as $n\to\infty$ which implies the result. □

Definition 16.17. Let $U\subseteq\mathbb{C}$ be an open set and let $f\colon U\to\mathbb{C}$ be a continuous function. If there exists a differentiable function $F\colon U\to\mathbb{C}$ with $F'(z)=f(z)$ then we say F is a *primitive* for f on U.

The fundamental theorem of calculus has the following important consequence[37]:

Theorem 16.18. *(Fundamental theorem of Calculus): Let $U\subseteq\mathbb{C}$ be a open and let $f\colon U\to\mathbb{C}$ be a continuous function. If $F\colon U\to\mathbb{C}$ is a primitive for f and $\gamma\colon[a,b]\to U$ is a piecewise C^1 path in U then we have*

$$\int_\gamma f(z)dz = F(\gamma(b)) - F(\gamma(a)).$$

[37] You should compare this to the existence of a potential in vector calculus.

In particular the integral of such a function f around any closed path is zero.

Proof. First suppose that γ is C^1. Then we have

$$\begin{aligned}\int_\gamma f(z)dz = \int_\gamma F'(z)dz &= \int_a^b F'(\gamma(t))\gamma'(t)dt \\ &= \int_a^b \frac{d}{dt}(F\circ\gamma)(t)dt \\ &= F(\gamma(b)) - F(\gamma(a)),\end{aligned}$$

where in second line we used a version of the chain rule[38] and in the last line we used the Fundamental theorem of Calculus from Prelims analysis on the real and imaginary parts of $F \circ \gamma$.

If γ is only[39] piecewise C^1, then take a partition $a = a_0 < a_1 < \ldots < a_k = b$ such that γ is C^1 on $[a_i, a_{i+1}]$ for each $i \in \{0, 1, \ldots, k-1\}$. Then we obtain a telescoping sum:

$$\begin{aligned}\int_\gamma f(z) &= \int_a^b f(\gamma(t))\gamma'(t)dt \\ &= \sum_{i=0}^{k-1}\int_{a_i}^{a_{i+1}} f(\gamma(t))\gamma'(t)dt \\ &= \sum_{i=0}^{k-1}(F(\gamma(a_{i+1})) - F(\gamma(a_i))) \\ &= F(\gamma(b)) - F(\gamma(a)),\end{aligned}$$

Finally, since γ is closed precisely when $\gamma(a) = \gamma(b)$ it follows immediately that the integral of f along a closed path is zero. □

Remark 16.19. If $f(z)$ has finitely many point of discontinuity $S \subset U$ but is bounded near them, and $\gamma(t) \in S$ for only finitely many t, then provided F is continuous and $F' = f$ on $U \backslash S$, the same proof shows that the fundamental theorem still holds – one just needs to take a partition of $[a, b]$ to take account of those singularities along with the singularities of $\gamma'(t)$.

Theorem 16.18 already has an important consequence:

Corollary 16.20. *Let U be a domain and let $f : U \to \mathbb{C}$ be a function with $f'(z) = 0$ for all $z \in U$. Then f is constant.*

Proof. Pick $z_0 \in U$. Since U is path-connected, if $w \in U$, we may find[40] a piecewise C^1-path $\gamma : [0, 1] \to U$ such that $\gamma(a) = z_0$ and $\gamma(b) = w$. Then by

[38] See the appendix for a discussion of this – we need a version of the chain rule for a composition of real-differentiable functions $f : \mathbb{R}^2 \to \mathbb{R}^2$ and $g : \mathbb{R} \to \mathbb{R}^2$.

[39] The reason we must be careful about this case is that the Fundamental Theorem of Calculus only holds when the integrand is continuous.

[40] Check that you see that if U is an open subset of $\mathbb{C}$ which is path-connected then any two points can be joined by a piecewise C^1-path.

Theorem 16.18 we see that

$$f(w) - f(z_0) = \int_\gamma f'(z)dz = 0,$$

so that f is constant as required. □

The following theorem is a kind of converse to the fundamental theorem:

Theorem 16.21. *If U is a domain (i.e. it is open and path connected) and $f: U \to \mathbb{C}$ is a continuous function such that for any closed path in U we have $\int_\gamma f(z)dz = 0$, then f has a primitive.*

Proof. Fix z_0 in U, and for any $z \in U$ set

$$F(z) = \int_\gamma f(z)dz.$$

where $\gamma\colon [a,b] \to U$ with $\gamma(a) = z_0$ and $\gamma(b) = z$.

We claim that $F(z)$ is independent of the choice of γ. Indeed if γ_1, γ_2 are two such paths, let $\gamma = \gamma_1 \star \gamma_2^-$ be the path obtained by concatenating γ_1 and the opposite γ_2^- of γ_2 (that is, γ traverses the path γ_1 and then goes backward along γ_2). Then γ is a closed path and so, using Proposition 16.14 we have

$$0 = \int_\gamma f(z)dz = \int_{\gamma_1} f(z)dz + \int_{\gamma_2^-} f(z)dz,$$

hence since $\int_{\gamma_2^-} f(z)dz = -\int_{\gamma_2} f(z)dz$ we see that $\int_{\gamma_1} f(z)dz = \int_{\gamma_2} f(z)dz$.

Next we claim that F is differentiable with $F'(z) = f(z)$. To see this, fix $w \in U$ and $\epsilon > 0$ such that $B(w,\epsilon) \subseteq U$ and choose a path $\gamma\colon [a,b] \to U$ from z_0 to w. If $z_1 \in B(w,\epsilon) \subseteq U$, then the concatenation of γ with the straight-line path $s\colon [0,1] \to U$ given by $s(t) = w + t(z-w)$from w to z is a path γ_1 from z_0 to z. It follows that

$$\begin{aligned} F(z_1) - F(w) &= \int_{\gamma_1} f(z)dz - \int_\gamma f(z)dz \\ &= (\int_\gamma f(z)dz + \int_s f(z)dz) - \int_\gamma f(z)dz \\ &= \int_s f(z)dz. \end{aligned}$$

But then we have for $z_1 \neq w$

$$\begin{aligned} \left|\frac{F(z_1) - F(w)}{z_1 - w} - f(w)\right| &= \left|\frac{1}{z_1 - w}\left(\int_0^1 f(w + t(z_1 - w)(z_1 - w)dt\right) - f(w)\right| \\ &= \left|\int_0^1 (f(w + t(z_1 - w)) - f(w))dt\right| \\ &\leq \sup_{t\in[0,1]} |f(w + t(z_1 - w)) - f(w)| \\ &\to 0 \text{ as } z_1 \to w \end{aligned}$$

as f is continuous at w. Thus F is differentiable at w with derivative $F'(w) = f(w)$ as claimed. $\square$

Remark 16.22. Note that any two primitives for a function f differ by a constant: This follows immediately from Corollary 16.20, since if F_1 and F_2 are two primitives, their difference $(F_1 - F_2)$ has zero derivative.

17. WINDING NUMBERS

The previous section on the fundamental theorem of calculus in the complex plane shows that not every holomorphic function can have a primitive. The most fundamental example of this is the function $f(z) = 1/z$ on the domain $\mathbb{C}^\times$.

Example 17.1. Let $f\colon \mathbb{C}^\times \to \mathbb{C}^\times$ be the function $f(z) = 1/z$. Then f does not have a primitive on $\mathbb{C}^\times$. Indeed if $\gamma\colon [0,1] \to \mathbb{C}$ is the path $\gamma(t) = \exp(2\pi i t)$ then

$$\int_\gamma f(z)dz = \int_0^1 f(\gamma(t))\gamma'(t)dt = \int_0^1 \frac{1}{\exp(2\pi i t)}.(2\pi i \exp(2\pi i t))dt = 2\pi i.$$

Since the path γ is closed, this integral would have to be zero if $f(z)$ has a primitive in an open set containing γ^*, thus $f(z)$ has no primitive on $\mathbb{C}^\times$ as claimed.

Note that $1/z$ *does* have a primitive on any domain in $\mathbb{C}^\times$ where we can chose a branch of the argument function (or equivalently a branch of [Log(z)]): Indeed if $l(z)$ is a branch of [Log(z)] on a domain $D \subset \mathbb{C}^\times$ then since $\exp(l(z)) = z$ the chain rule shows that $\exp(l(z)).l'(z) = 1$ and hence $l'(z) = 1/z$.

In the present section we investigate the change in argument as we move along a path. It will turn out to be a basic ingredient in computing integrals around closed paths. In more detail, suppose that $\gamma\colon [0,1] \to \mathbb{C}$ is a closed path which does not pass through 0. We would like to give a rigorous definition of the number of times γ "goes around the origin". Roughly speaking, this will be the change in argument $\arg(\gamma(t))$, and therein lies the difficulty, since $\arg(z)$ cannot be defined continuously on all of $\mathbb{C}\backslash\{0\}$. The next Proposition shows that we *can* however always define the argument as a continuous function *of the parameter* $t \in [0,1]$:

Proposition 17.2. *Let* $\gamma\colon [0,1] \to \mathbb{C}\backslash\{0\}$ *be a path. Then there is continuous function* $a\colon [0,1] \to \mathbb{R}$ *such that*

$$\gamma(t) = |\gamma(t)|e^{2\pi i a(t)}.$$

Moreover, if a *and* b *are two such functions, then there exists* $n \in \mathbb{Z}$ *such that* $a(t) = b(t) + n$ *for all* $t \in [0,1]$.

Proof. By replacing $\gamma(t)$ with $\gamma(t)/|\gamma(t)|$ we may assume that $|\gamma(t)| = 1$ for all t. Since γ is continuous on a compact set, it is uniformly continuous, so that there is a $\delta > 0$ such that $|\gamma(s) - \gamma(t)| < \sqrt{3}$ for any s, t with $|s - t| < \delta$.

Choose an integer $n > 0$ such that $n > 1/\delta$ so that on each subinterval $[i/n, (i+1)/n]$ we have $|\gamma(s) - \gamma(t)| < \sqrt{3}/2$. Now on any half-plane in $\mathbb{C}$ we may certainly define a holomorphic branch of $[\mathrm{Log}(z)]$ (simply pick a branch cut along a ray in the opposite half-plane) and hence a continuous argument function, and if $|z_1| = |z_2| = 1$ and $|z_1 - z_2| < \sqrt{3}$, then the angle between z_1 and z_2 is at most $\pi/3$. It follows there exists a continuous functions $a_i\colon [j/n, (j+1)/n] \to \mathbb{R}$ such that $\gamma(t) = e^{2\pi i a_j(t)}$ for $t \in [j/n, (j+1)/n]$ (since $\gamma([j/n, (j+1)/n])$ must lie in an arc of length at most $2\pi/3$). Now since $e^{2\pi i a_j(j/n)} = e^{2\pi i a_{j-1}(j/n)}$ $a_{j-1}(j/n)$ and $a_i(j/n)$ differ by an integer. Thus we can successively adjust the a_j for $j > 1$ by an integer (as if $\gamma(t) = e^{2\pi i a_j(t)}$ then $\gamma(t) = e^{2\pi i (a(t)+n)}$ for any $n \in \mathbb{Z}$) to obtain a continuous function $a\colon [0,1] \to \mathbb{C}$ such that $\gamma(t) = e^{2\pi i a(t)}$ as required. Finally, the uniqueness statement follows because $e^{2\pi i (a(t)-b(t))} = 1$, hence $a(t) - b(t) \in \mathbb{Z}$, and since $[0,1]$ is connected it follows $a(t) - b(t)$ is constant as required. $\square$

Definition 17.3. If $\gamma\colon [0,1] \to \mathbb{C}\backslash\{0\}$ is a closed path and $\gamma(t) = |\gamma(t)|e^{2\pi i a(t)}$ as in the previous lemma, then since $\gamma(0) = \gamma(1)$ we must have $a(1) - a(0) \in \mathbb{Z}$. This integer is called the *winding number* $I(\gamma, 0)$ of γ around 0. It is uniquely determined by the path γ because the function a is unique up to an integer. By translation, if γ is any closed path and z_0 is not in the image of γ, we may define the winding number $I(\gamma, z_0)$ of γ about z_0 in the same fashion. Explicitly, if γ is a closed path with $z_0 \notin \gamma^*$ then let $t\colon \mathbb{C} \to \mathbb{C}$ be given by $t(z) = z - z_0$ and define $I(\gamma, z_0) = I(t \circ \gamma, 0)$.

Remark 17.4. Note that if $\gamma\colon [0,1] \to U$ where $0 \notin U$ and there exists a holomorphic branch $L\colon U \to \mathbb{C}$ of $[\mathrm{Log}(z)]$ on U, then $I(\gamma, 0) = 0$. Indeed in this case we may define $a(t) = \Im(L(\gamma(t)))$, and since $\gamma(0) = \gamma(1)$ it follows $a(1) - a(0) = 0$ as claimed. Note also that the definition of the winding number only requires the closed path γ to be continuous, not piecewise C^1. Of course as usual, we will mostly only be interested in piecewise C^1 paths, as these are the ones along which we can integrate functions.

We now see that the winding number has a natural interpretation in term of path integrals: Note that if γ is piecewise C^1 then the function $a(t)$ is also piecewise C^1, since any branch of the logarithm function is in fact differentiable where it is defined, and $a(t)$ is locally given as $\Im(\log(\gamma(t))$ for a suitable branch.

Lemma 17.5. *Let γ be a piecewise C^1 closed path and $z_0 \in \mathbb{C}$ a point not in the image of γ. Then the winding number $I(\gamma, z_0)$ of γ around z_0 is given by*

$$I(\gamma, z_0) = \frac{1}{2\pi i}\int_\gamma \frac{dz}{z - z_0}.$$

Proof. If $\gamma\colon [0,1] \to \mathbb{C}$ we may write $\gamma(t) = z_0 + r(t)e^{2\pi i a(t)}$ (where $r(t) = |\gamma(t) - z_0| > 0$ is continuous and the existence of $a(t)$ is guaranteed by

Proposition 17.2). Then we have

$$\begin{aligned}\int_\gamma \frac{dz}{z-z_0} &= \int_0^1 \frac{1}{r(t)e^{2\pi i a(t)}}.\left(r'(t)+2\pi i r(t)a'(t)\right)e^{2\pi i a(t)}dt \\ &= \int_0^1 r'(t)/r(t) + 2\pi i a'(t)dt = [\log(r(t)) + 2\pi i a(t)]_0^1 \\ &= 2\pi i(a(1)-a(0)),\end{aligned}$$

since $r(1) = r(0) = |\gamma(0) - z_0|$. □

The next Proposition will be useful not only for the study of winding numbers. We first need a definition:

Definition 17.6. If $f\colon U \to \mathbb{C}$ is a function on an open subset U of $\mathbb{C}$, then we say that f is *analytic* on U if for every $z_0 \in \mathbb{C}$ there is an $r > 0$ with $B(z_0, r) \subseteq U$ such that there is a power series $\sum_{k=0}^\infty a_k(z-z_0)^k$ with radius of convergence at least r and $f(z) = \sum_{k=0}^\infty a_k(z-z_0)^k$. An analytic function is holomorphic, as any power series is (infinitely) complex differentiable.

Proposition 17.7. *Let U be an open set in $\mathbb{C}$ and let $\gamma\colon [0,1] \to U$ be a closed path. If $f(z)$ is a continuous function on γ^* then the function*

$$I_f(\gamma, w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-w}dz,$$

is analytic. in w.

In particular, if $f(z) = 1$ this shows that the function $w \mapsto I(\gamma, w)$ is a continuous function on $\mathbb{C}\backslash\gamma^*$, and hence, since it is integer-valued, it is constant on the connected components of $\mathbb{C}\backslash\gamma^*$.

Proof. We wish to show that $I_\gamma(f(w)$ is holomorphic at each $z_0 \in \mathbb{C}\backslash\gamma^*$. Translating if necessary we may assume $z_0 = 0$.

Now since $\mathbb{C}\backslash\gamma^*$ is open, there is some $r > 0$ such that $B(0, 2r) \cap \gamma^* = \emptyset$. We claim that $I_f(\gamma, w)$ is holomorphic in $B(0.r)$. Indeed if $w \in B(0, r)$ and $z \in \gamma^*$ it follows that $|w/z| < 1/2$. Moreover, since γ^* is compact, $M = \sup\{|f(z)| : z \in \gamma^*\}$ is finite, and hence

$$|f(z).w^n/z^{n+1}| = |f(z)||z|^{-1}|w/z|^n < \frac{M}{2r}(1/2)^n, \quad \forall z \in \gamma^*.$$

It follows from the Weierstrass M-test that the series

$$\sum_{n=0}^\infty \frac{f(z).w^n}{z^{n+1}} = \sum_{n=0}^\infty \frac{f(z)}{z}(w/z)^n = \frac{f(z)}{z}(1-w/z)^{-1} = \frac{f(z)}{z-w}$$

viewed as a function of z, converges uniformly on γ^* to $f(z)/(z-w)$. Thus for all $w \in B(0, r)$ we have

$$I_f(\gamma, w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)dz}{z-w} = \sum_{n=0}^\infty \left(\frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z^{n+1}}dz\right)w^n,$$

hence $I_f(\gamma, w)$ is given by a power series in $B(0, r)$ (and hence is also holomorphic there) as required.

Finally, if $f = 1$, then since $I_1(\gamma, z) = I(\gamma, z)$ is integer-valued, it follows it must be constant on any connected component of $\mathbb{C}\backslash\gamma^*$ as required. $\square$

Remark 17.8. Note that since the coefficients of a power series centred at a point z_0 are given by its derivatives at that point, the proof above actually also gives formulae for the derivatives of $g(w) = I_f(\gamma, w)$ at z_0:

$$g^{(n)}(z_0) = \frac{n!}{2\pi i}\int_\gamma \frac{f(z)dz}{(z-z_0)^{n+1}}.$$

Remark 17.9. If γ is a closed path then γ^* is compact and hence bounded. Thus there is an $R > 0$ such that the connected set $\mathbb{C}\backslash B(0, R) \cap \gamma^* = \emptyset$. It follows that $\mathbb{C}\backslash\gamma^*$ has exactly one unbounded connected component. Since

$$\left|\int_\gamma \frac{d\zeta}{\zeta - z}\right| \leq \ell(\gamma).\sup_{\zeta\in\gamma^*}|1/(\zeta - z)| \to 0$$

as $z \to \infty$ it follows that $I(\gamma, z) = 0$ on the unbounded component of $\mathbb{C}\backslash\gamma^*$.

Definition 17.10. Let $\gamma\colon [0, 1] \to \mathbb{C}$ be a closed path. We say that a point z is in the *inside*[41] of γ if $z \notin \gamma^*$ and $I(\gamma, z) \neq 0$. The previous remark shows that the inside of γ is a union of bounded connected components of $\mathbb{C}\backslash\gamma^*$. (We don't, however, know that the inside of γ is necessarily non-empty.)

Example 17.11. Suppose that $\gamma_1\colon [-\pi, \pi] \to \mathbb{C}$ is given by $\gamma_1 = 1 + e^{it}$ and $\gamma_2\colon [0, 2\pi] \to \mathbb{C}$ is given by $\gamma_2(t) = -1 + e^{-it}$. Then if $\gamma = \gamma_1 \star \gamma_2$, γ traverses a figure-of-eight and it is easy to check that the inside of γ is $B(1, 1) \cup B(-1, 1)$ where $I(\gamma, z) = 1$ for $z \in B(1, 1)$ while $I(\gamma, z) = -1$ for $z \in B(-1, 1)$.

Remark 17.12. It is a theorem, known as the *Jordan Curve Theorem*, that if $\gamma\colon [0, 1] \to \mathbb{C}$ is a simple closed curve, so that $\gamma(t) = \gamma(s)$ if and only if $s = t$ or $s, t \in \{0, 1\}$, then $\mathbb{C}\backslash\gamma^*$ is the union of precisely one bounded and one unbounded component, and on the bounded component $I(\gamma, z)$ is either 1 or -1. If $I(\gamma, z) = 1$ for z on the inside of γ we say γ is postively oriented and we say it is negatively oriented if $I(\gamma, z) = -1$ for z on the inside.

18. Cauchy's theorem

The key insight into the study of holomorphic functions is Cauchy's theorem, which (somewhat informally) states that if $f\colon U \to \mathbb{C}$ is holomorphic and γ is a path in U whose interior lies entirely in U then $\int_\gamma f(z)dz = 0$. It will follow from this and Theorem 16.21 that, at least locally, every holomorphic function has a primitive. The strategy to prove Cauchy's theorem goes as follows: first show it for the simplest closed contours – triangles.

[41]The term *interior* of γ might be more natural, but we have already used this in the first part of the course to mean something quite different.

Then use this to deduce the existence of a primitive (at least for certain kinds of sufficiently nice open sets U which are called "star-like") and then use Theorem 16.18 to deduce the result for arbitrary paths in such open subsets. We will discuss more general versions of the theorem later, after we have applied Cauchy's theorem for star-like domains to obtain important theorems on the nature of holomorphic functions. First we recall the definition of a triangular path:

Definition 18.1. A *triangle* or *triangular path* T is a path of the form $\gamma_1 \star \gamma_2 \star \gamma_3$ where $\gamma_1(t) = a + t(b-a)$, $\gamma_2(t) = b + t(c-b)$ and $\gamma_3(t) = c + t(a-c)$ where $t \in [0,1]$ and $a, b, c \in \mathbb{C}$. (Note that if $\{a, b, c\}$ are collinear, then T is a degenerate triangle.) That is, T traverses the boundary of the triangle with vertices $a, b, c \in \mathbb{C}$. The solid triangle $\mathcal{T}$ bounded by T is the region

$$\mathcal{T} = \{t_1 a + t_2 b + t_3 c : t_i \in [0,1], \sum_{i=1}^{3} t_i = 1\},$$

with the points in the interior of $\mathcal{T}$ corresponding to the points with $t_i > 0$ for each $i \in \{1, 2, 3\}$. We will denote by $[a, b]$ the line segment $\{a + t(b-a) : t \in [0,1]\}$, the side of T joining vertex a to vertex b. Whenever it is not evident what the vertices of the triangle T are, we will write $T_{a,b,c}$.

Theorem 18.2. *(Cauchy's theorem for a triangle): Suppose that $U \subseteq \mathbb{C}$ is an open subset and let $T \subseteq U$ be a triangle whose interior is entirely contained in U. Then if $f \colon U \to \mathbb{C}$ is holomorphic we have*

$$\int_T f(z) dz = 0$$

Proof. The proof proceeds using a version of the "divide and conquer" strategy one uses to prove the Bolzano-Weierstrass theorem. Suppose for the sake of contradiction that $\int_T f(z)dz \neq 0$, and let $I = |\int_T f(z)dz| > 0$. We build a sequence of smaller and smaller triangles T^n around which the integral of f is not too small, as follows: Let $T^0 = T$, and suppose that we have constructed T^i for $0 \leq i < k$. Then take the triangle T^{k-1} and join the midpoints of the edges to form four smaller triangles, which we will denote S_i $(1 \leq i \leq 4)$.

Then we have $\int_{T^{k-1}} f(z)dz = \sum_{i=1}^{4} \int_{S_i} f(z)dz$, since the integrals around the interior edges cancel (see Figure 3). In particular, we must have

$$I_k = |\int_{T^{k-1}} f(z)dz| \leq \sum_{i=1}^{4} |\int_{S_i} f(z)dz|,$$

so that for some i we must have $|\int_{S_i} f(z)dz| \geq I_{k-1}/4$. Set T^k to be this triangle S_i. Then by induction we see that $\ell(T^k) = 2^{-k}\ell(T)$ while $I_k \geq 4^{-k} I$.

Now let $\mathcal{T}$ be the solid triangle with boundary T and similarly let $\mathcal{T}^k$ be the solid triangle with boundary T^k. Then we see that $\operatorname{diam}(\mathcal{T}^k) =$

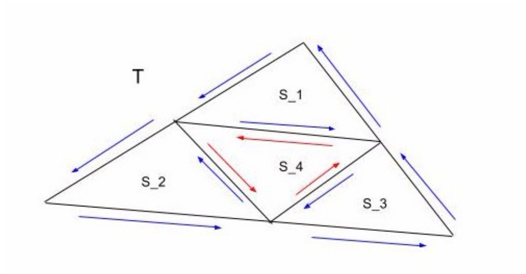

FIGURE 3. Subdivision of a triangle

$2^{-k}\text{diam}(\mathcal{T}) \to 0$, and the sets $\mathcal{T}^k$ are clearly nested. It follows from Lemma 8.6 that there is a unique point z_0 which lies in every $\mathcal{T}^k$. Now by assumption f is holomorphic at z_0, so we have

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + (z - z_0)\psi(z),$$

where $\psi(z) \to 0 = \psi(z_0)$ as $z \to z_0$. Note that ψ is continuous and hence integrable on all of U. Now since the linear function $z \mapsto f'(z_0)z + f(z_0)$ clearly has a primitive it follows from Theorem 16.18

$$\int_{T^k} f(z)dz = \int_{T^k} (z - z_0)\psi(z)dz$$

Now since z_0 lies in $\mathcal{T}^k$ and z is on the boundary T^k of $\mathcal{T}^k$, we see that $|z - z_0| \leq \text{diam}(\mathcal{T}^k) = 2^{-k}\text{diam}(T)$. Thus if we set $\eta_k = \sup_{z \in T^k} |\psi(z)|$, it follows by the estimation lemma that

$$\begin{aligned} I_k = \Big| \int_{T^k} (z - z_0)\psi(z)dz \Big| &\leq \eta_k.\text{diam}(T^k)\ell(T^k) \\ &= 4^{-k}\eta_k.\text{diam}(T).\ell(T). \end{aligned}$$

But since $\psi(z) \to 0$ as $z \to z_0$, it follows $\eta_k \to 0$ as $k \to \infty$, and hence $4^k I_k \to 0$ as $k \to \infty$. On the other hand, by construction we have $4^k I_k \geq I > 0$, thus we have a contradiction as required. □

Definition 18.3. Let X be a subset in $\mathbb{C}$. We say that X is *convex* if for each $z, w \in U$ the line segment between z and w is contained in X. We say that X is *star-like* if there is a point $z_0 \in X$ such that for every $w \in X$ the line segment $[z_0, w]$ joining z_0 and w lies in X. We will say that X is star-like with respect to z_0 in this case. Thus a convex subset is thus starlike with respect to every point it contains.

Example 18.4. A disk (open or closed) is convex, as is a solid triangle or rectangle. On the other hand a cross, such as $\{0\} \times [-1, 1] \cup [-1, 1] \times \{0\}$ is star-like with respect to the origin, but is not convex.

Theorem 18.5. *(Cauchy's theorem for a star-like domain): Let U be a star-like domain. Then if $f: U \to \mathbb{C}$ is holomorphic and $\gamma: [a, b] \to U$ is a closed path in*

U we have

$$\int_\gamma f(z)dz = 0.$$

Proof. The proof proceeds similarly to the proof of Theorem 16.21: by Theorem 16.18 it suffices to show that f has a primitive in U. To show this, let $z_0 \in U$ be a point for which the line segment from z_0 to every $z \in U$ lies in U. Let $\gamma_z = z_0 + t(z - z_0)$ be a parametrization of this curve, and define

$$F(z) = \int_{\gamma_z} f(\zeta)d\zeta.$$

We claim that F is a primitive for f on U. Indeed pick $\epsilon > 0$ such that $B(z,\epsilon) \subseteq U$. Then if $w \in B(z,\epsilon)$ note that the triangle T with vertices z_0, z, w lies entirely in U by the assumption that U is star-like with respect to z_0. It follows from Theorem 18.2 that $\int_T f(\zeta)d\zeta = 0$, and hence if $\eta(t) = w + t(z - w)$ is the straight-line path going from w to z (so that T is the concatenation of γ_w, η and γ_z^-) we have

$$\begin{aligned}
\Big|\frac{F(z)-F(w)}{z-w} - f(z)\Big| &= \Big|\int_\eta \frac{f(\zeta)}{z-w}d\zeta - f(z)\Big| \\
&= \Big|\int_0^1 f(w+t(z-w))dt - f(z)\Big| \\
&= \Big|\int_0^1 (f(w+t(z-w)) - f(z)dt\Big| \\
&\leq \sup_{t\in[0,1]} |f(w+t(z-w)) - f(z)|,
\end{aligned}$$

which, since f is continuous at w, tends to zero as $w \to z$ so that $F'(z) = f(z)$ as required.

□

Note that our proof of Cauchy's theorem for a star-like domain D proceeded by showing that any holomorphic function on D has a primitive, and hence by the fundamental theorem of calculus its integral around a closed path is zero. This motivates the following definition:

Definition 18.6. We say that a domain $D \subseteq \mathbb{C}$ is *primitive*[42] if any holomorphic function $f\colon D \to \mathbb{C}$ has a primitive in D.

Thus, for example, our proof of Theorem 18.5 shows that all star-like domains are primitive. The following Lemma shows however that we can build many primitive domains which are not star-like.

Lemma 18.7. *Suppose that D_1 and D_2 are primitive domains and $D_1 \cap D_2$ is connected. Then $D_1 \cup D_2$ is primitive.*

[42] This is *not* standard terminology. The reason for this will become clear later.

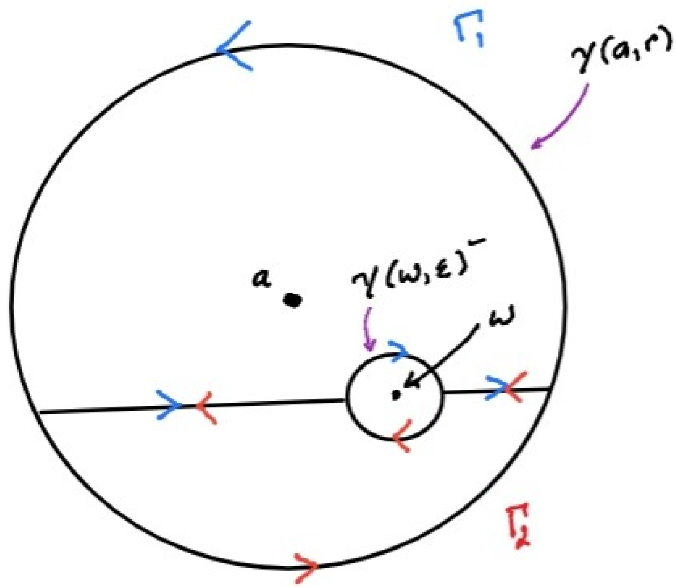

FIGURE 4. Contours for the proof of Theorem 18.8.

Proof. Let $f\colon D_1 \cup D_2 \to \mathbb{C}$ be a holomorphic function. Then $f_{|D_1}$ is a holomorphic function on D_1, and thus it has a primitive $F_1\colon D_1 \to \mathbb{C}$. Similarly $f_{|D_2}$ has a primitive, F_2 say. But then $F_1 - F_2$ has zero derivative on $D_1 \cap D_2$, and since by assumption $D_1 \cap D_2$ is connected (and thus path-connected) it follows $F_1 - F_2$ is constant, c say, on $D_1 \cap D_2$. But then if $F\colon D_1 \cup D_2 \to \mathbb{C}$ is a defined to be F_1 on D_1 and $F_2 + c$ on D_2 it follows that F is a primitive for f on $D_1 \cup D_2$ as required. $\square$

18.1. **Cauchy's Integral Formula.** We are now almost ready to prove one of the most important consequences of Cauchy's theorem – the integral formula. This formula will have incredibly powerful consequences.

Theorem 18.8. *(Cauchy's Integral Formula.) Suppose that $f\colon U \to \mathbb{C}$ is a holomorphic function on an open set U which contains the disc $\bar{B}(a,r)$. Then for all $w \in B(a,r)$ we have*

$$f(w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)}{z-w}dz,$$

where γ is the path $t \mapsto a + re^{2\pi i t}$.

Proof. Fix $w \in B(a,r)$. We use the contours Γ_1 and Γ_2 as shown in Diagram 4 (where Γ_1 follows the direction of the blue arrows, and Γ_2 the directions of the red arrows). These paths join the circular contours $\gamma(a,r)$ and $\gamma(w,\epsilon)^-$

where ϵ is small enough to lie in the interior of $B(a, r)$. By the additivity properties of path integrals, the contributions of the line segments cancel so that

$$\int_{\Gamma_1} \frac{f(z)}{z-w} dz + \int_{\Gamma_2} \frac{f(z)}{z-w} dz = \int_{\gamma(a,r)} \frac{f(z)}{z-w} dz - \int_{\gamma(w,\epsilon)} \frac{f(z)}{z-w} dz.$$

On the other hand, each of Γ_1, Γ_2 lies in a primitive domain in which $f/(z-w)$ is holomorphic – indeed by the quotient rule, $f(z)/(z-w)$ is holomophic on $U\backslash\{w\}$ – so each of the integrals on the left-hand side vanish, and hence

$$\frac{1}{2\pi i} \int_{\gamma(a,r)} \frac{f(z)}{z-w} dz = \frac{1}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{f(z)}{z-w} dz.$$

Thus we can replace the integral over the circle $\gamma(a, r)$ with an integral over an arbtirary small circle centred at w itself. But for such a small circle,

$$\begin{aligned}
\frac{1}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{f(z)}{z-w} dz &= \frac{1}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{f(z)-f(w)}{z-w} dz + \frac{f(w)}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{dz}{z-w}. \\
&= \frac{1}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{f(z)-f(w)}{z-w} dz + f(w) I(\gamma(w,\epsilon), w) \\
&= \frac{1}{2\pi i} \int_{\gamma(w,\epsilon)} \frac{f(z)-f(w)}{z-w} dz + f(w)
\end{aligned}$$

But since f is complex differentiable at $z = w$, the term $(f(z)-f(w))/(z-w)$ is bounded as $\epsilon \to 0$, so that by the estimation lemma its integral over $\gamma(w, \epsilon)$ tends to zero. Thus as $\epsilon \to 0$ the path integral around $\gamma(w, \epsilon)$ tends to $f(w)$. But since it is also equal to $(2\pi i)^{-1} \int_{\gamma(a,r)} f(z)/(z-w)dz$, which is independent of ϵ, we conclude that it must in fact be equal to $f(w)$. The result follows.

□

Remark 18.9. The same result holds for any oriented curve γ once we weight the left-hand side by the winding number[43] of a path around the point $w \notin \gamma^*$, provided that f is holomorphic on the inside of γ.

18.2. **Applications of the Integral Formula.**

Remark 18.10. Note that Cauchy's integral formula can be interpreted as saying the value of $f(w)$ for w inside the circle is obtained as the "convolution" of f and the function $1/(z-w)$ on the boundary circle. Since the function $1/(z-w)$ is infinitely differentiable one can use this to show that f itself is infinitely differentiable as we will shortly show. If you take the Integral Transforms, you will see convolution play a crucial role in the theory of transforms. In particular, the convolution of two functions often inherits the "good" properties of either. We next show that in fact the formula implies a strong version of Taylor's Theorem.

[43]Which, as we used in the proof above, is 1 in the case of a point inside a positively oriented circular path.

Corollary 18.11. *If $f: U \to \mathbb{C}$ is holomorphic on an open set U, then for any $z_0 \in U$, the $f(z)$ is equal to its Taylor series at z_0 and the Taylor series converges on any open disk centred at z_0 lying in U. Moreover the derivatives of f at z_0 are given by*

$$f^{(n)}(z_0) = \frac{n!}{2\pi i}\int_{\gamma(a,r)} \frac{f(z)}{(z-z_0)^{n+1}}dz. \tag{18.1}$$

For any $a \in \mathbb{C}$, $r \in \mathbb{R}_{>0}$ with $z_0 \in B(a,r)$.

Proof. This follows immediately from the proof of Proposition 17.7, and Remark 17.8. The integral formulae of Equation 18.1 for the derivatives of f are also referred to as *Cauchy's Integral Formulae*. □

Definition 18.12. Recall that a function which is locally given by a power series is said to be *analytic*. We have thus shown that any holomorphic function is actually analytic, and from now on we may use the terms interchangeably (as you may notice is common practice in many textbooks).

One famous application of the Integral formula is known as Liouville's theorem, which will give an easy proof of the Fundamental Theorem of Algebra[44]. We say that a function $f: \mathbb{C} \to \mathbb{C}$ is *entire* if it is complex differentiable on the whole complex plane.

Theorem 18.13. *Let $f: \mathbb{C} \to \mathbb{C}$ be an entire function. If f is bounded then it is constant.*

Proof. Suppose that $|f(z)| \leq M$ for all $z \in \mathbb{C}$. Let $\gamma_R(t) = Re^{2\pi i t}$ be the circular path centred at the origin with radius R. The for $R > |w|$ the integral formula shows

$$\begin{aligned} |f(w) - f(0)| &= \Big|\frac{1}{2\pi i}\int_{\gamma_R} f(z)\Big(\frac{1}{z-w} - \frac{1}{z}\Big)dz\Big| \\ &= \frac{1}{2\pi}\Big|\int_{\gamma_R} \frac{w.f(z)}{z(z-w)}dz\Big| \\ &\leq \frac{2\pi R}{2\pi}\sup_{z:|z|=R}\Big|\frac{w.f(z)}{z(z-w)}\Big| \\ &\leq R.\frac{M|w|}{R.(R-|w|)} = \frac{M|w|}{R-|w|}, \end{aligned}$$

Thus letting $R \to \infty$ we see that $|f(w) - f(0)| = 0$, so that f is constant an required.

□

Theorem 18.14. *Suppose that $p(z) = \sum_{k=0}^{n} a_k z^k$ is a non-constant polynomial where $a_k \in \mathbb{C}$ and $a_n \neq 0$. Then there is a $z_0 \in \mathbb{C}$ for which $p(z_0) = 0$.*

[44] Which, when it comes down to it, isn't really a theorem in algebra. The most "algebraic" proof of that I know uses Galois theory, which you can learn about in Part B.

Proof. By rescaling p we may assume that $a_n = 1$. If $p(z) \neq 0$ for all $z \in \mathbb{C}$ it follows that $f(z) = 1/p(z)$ is an entire function (since p is clearly entire). We claim that f is bounded. Indeed since it is continuous it is bounded on any disc $\bar{B}(0, R)$, so it suffices to show that $|f(z)| \to 0$ as $z \to \infty$, that is, to show that $|p(z)| \to \infty$ as $z \to \infty$. But we have

$$|p(z)| = |z^n + \sum_{k=0}^{n-1} a_k z^k| = |z^n|\{|1 + \sum_{k=0}^{n-1} \frac{a_k}{z^n - k}|\} \geq |z^n|.(1 - \sum_{k=0}^{n-1} \frac{|a_k|}{|z|^{n-k}}).$$

Since $\frac{1}{|z|^m} \to 0$ as $|z| \to \infty$ for any $m \geq 1$ it follows that for sufficiently large $|z|$, say $|z| \geq R$, we will have $1 - \sum_{k=0}^{n-1} \frac{|a_k|}{|z|^{n-k}} \geq 1/2$. Thus for $|z| \geq R$ we have $|p(z)| \geq \frac{1}{2}|z|^n$. Since $|z|^n$ clearly tends to infinity as $|z|$ does it follows $|p(z)| \to \infty$ as required. □

Remark 18.15. The crucial point of the above proof is that one term of the polynomial – the leading term in this case– dominates the behaviour of the polynomial for large values of z. All proofs of the fundamental theorem hinge on essentially this point. Note that $p(z_0) = 0$ if and only if $p(z) = (z - z_0)q(z)$ for a polynomial $q(z)$, thus by induction on degree we see that the theorem implies that a polynomial over $\mathbb{C}$ factors into a product of degree one polynomials.

Corollary 18.16. *(Riemann's removable singularity theorem): Suppose that U is an open subset of $\mathbb{C}$ and $z_0 \in U$. If $f \colon U \backslash \{z_0\} \to \mathbb{C}$ is holomorphic and bounded near z_0, then f extends to a holomorphic function on all of U.*

Proof. Define $h(z)$ by

$$h(z) = \begin{cases} (z - z_0)^2 f(z), & z \neq 0; \\ 0, & z = z_0 \end{cases}$$

The clearly $h(z)$ is holomorphic on $U \backslash \{z_0\}$, using the fact that f and standard rules for complex differentiablility. On the other hand, at $z = z_0$ we see directly that

$$\frac{h(z) - h(z_0)}{z - z_0} = (z - z_0) f(z) \to 0$$

as $z \to z_0$ since f is bounded near z_0 by assumption. It follows that h is in fact holomorphic everywhere in U. But then if we chose $r > 0$ is such that $\bar{B}(z_0, r) \subset U$, then by Corollary 18.11 $h(z)$ is equal to its Taylor series centred at z_0, thus

$$h(z) = \sum_{k=0}^{\infty} a_k (z - z_0)^k.$$

But since we have $h(z_0) = h'(z_0) = 0$ we see $a_0 = a_1 = 0$, and so $\sum_{k=0}^{\infty} a_{k+2}(z - z_0)^k$ defines a holomorphic function in $B(z_0, r)$. Since this clearly agrees with $f(z)$ on $B(z_0, r) \backslash \{0\}$, we see that by redefining $f(z_0) = a_2$, we can extend f to a holomorphic function on all of U as required. □

We end this section with a kind of converse to Cauchy's theorem:

Theorem 18.17. *(Morera's theorem) Suppose that $f\colon U \to \mathbb{C}$ is a continuous function on an open subset $U \subseteq \mathbb{C}$. If for any closed path $\gamma\colon [a,b] \to U$ we have $\int_\gamma f(z)dz = 0$, then f is holomorphic.*

Proof. By Theorem 16.21 we know that f has a primitive $F\colon U \to \mathbb{C}$. But then F is holomorphic on U and so infinitely differentiable on U, thus in particular $f = F'$ is also holomorphic. $\square$

Remark 18.18. One can prove variants of the above theorem: If U is a star-like domain for example, then our proof of Cauchy's theorem for such domains shows that $f\colon U \to \mathbb{C}$ has a primitive (and hence will be differentiable itself) provided $\int_T f(z)dz = 0$ for every triangle in U. In fact the assumption that $\int_T f(z)dz = 0$ for all triangles whose interior lies in U suffices to imply f is holomorphic for *any* open subset U: To show f is holomorphic on U, it suffices to show that f is holomorphic on $B(a,r)$ for each open disk $B(a,r) \subset U$. But this follows from the above as disks are star-like (in fact convex). It follows that we can characterize the fact that $f\colon U \to \mathbb{C}$ is holomorphic on U by an integral condition: $f\colon U \to \mathbb{C}$ is holomorphic if and only if for all triangles T which bound a solid triangle $\mathcal{T}$ with $\mathcal{T} \subset U$, the integral $\int_T f(z)dz = 0$.

This characterization of the property of being holomorphic has some important consequences. We first need a definition:

Definition 18.19. Let U be an open subset of $\mathbb{C}$. If (f_n) is a sequence of functions defined on U, we say $f_n \to f$ *uniformly on compacts* if for every compact subset K of U, the sequence $(f_{n|K})$ converges uniformly to $f_{|K}$. Note that in this case f is continuous if the f_n are: Indeed to see that f is continuous at $a \in U$, note that since U is open, there is some $r > 0$ with $B(a,r) \subseteq U$. But then $K = \bar{B}(a, r/2) \subseteq U$ and $f_n \to f$ uniformly on K, whence f is continuous on K, and so certainly it is continuous at a.

Example 18.20. Convergence of power series $f(z) = \sum_{k=0}^{\infty} a_n z^n$ is a basic example of convergence on compacts: if R is the radius of convergences of $f(z)$ the partial sums $s_n(z)$ of the power series $B(0,R)$ converge uniformly on compacts in $B(0,R)$. The convergence is *not* necessarily uniform on $B(0,R)$, as the example $f(z) = \sum_{n=0}^{\infty} z^n$ shows. Nevertheless, since $B(0,R) = \bigcup_{r<R} \bar{B}(0,r)$ is the union of its compact subsets, many of the good properties of the polynomial functions $s_n(z)$ are inherited by the power series because the convergence is uniform on compact subsets.

Proposition 18.21. *Suppose that U is a domain and the sequence of holomorphic functions $f_n\colon U \to \mathbb{C}$ converges to $f\colon U \to \mathbb{C}$ uniformly on compacts in U. Then f is holomorphic.*

Proof. Note by the above that f is continuous on U. Since the property of being holomorphic is local, it suffices to show for each $w \in U$ that there is a ball $B(w,r) \subseteq U$ within which f is holomorphic. Since U is open, for any such w we may certainly find $r > 0$ such that $B(w,r) \subseteq U$. Then as $B(w,r)$ is convex, Cauchy's theorem for a star-like domain shows that for every closed path $\gamma\colon [a,b] \to B(w,r)$ whose image lies in $B(w,r)$ we have $\int_\gamma f_n(z)dz = 0$ for all $n \in \mathbb{N}$.

But $\gamma^* = \gamma([a,b])$ is a compact subset of U, hence $f_n \to f$ uniformly on γ^*. It follows that

$$0 = \int_\gamma f_n(z)dz \to \int_\gamma f(z)dz,$$

so that the integral of f around any closed path in $B(w,r)$ is zero. But then Theorem 16.21 shows that f has a primitive F on $B(w,r)$. But we have seen that any holomorphic function is in fact infinitely differentiable, so it follows that F, and hence f is infinitely differentiable on $B(w,r)$ as required.

□

Often functions on the complex plane are defined in terms of integrals. It is thus useful to have a criterion by which one can check if such a function is holomorphic. The following theorem gives such a criterion.

Theorem 18.22. *Let U be an open subset of $\mathbb{C}$ and suppose that $F\colon U \times [a,b]$ is a function satisfying*

(1) *The function $z \mapsto F(z,s)$ is holomorphic in z for each $s \in [a,b]$.*
(2) *F is continuous on $U \times [a,b]$*

Then the function $f\colon U \to \mathbb{C}$ defined by

$$f(z) = \int_a^b F(z,s)ds$$

is holomorphic.

Proof. Changing variables we may assume that $[a,b] = [0,1]$ (explicitly, one replaces s by $(s-a)/(b-a)$). By Theorem 18.21 it is enough to show that we may find a sequence of holomorphic functions $f_n(z)$ which converge of $f(z)$ uniformly on compact subsets of U. To find such a sequence, recall from Prelims Analysis that the Riemann integral of a continuous function is equal to the limit of its Riemann sums as the mesh of the partition used for the sum tends to zero. Using the partition $x_i = i/n$ for $0 \leq i \leq n$ evaluating at the right-most end-point of each interval, we see that

$$f_n(z) = \frac{1}{n}\sum_{i=1}^{n} F(z, i/n),$$

is a Riemann sum for the integral $\int_0^1 F(z,s)ds$, hence as $n \to \infty$ we have $f_n(z) \to f(z)$ for each $z \in U$, *i.e.* the sequence (f_n) converges pointwise to f on all of U. To complete the proof of the theorem it thus suffices to check that $f_n \to f$ as $n \to \infty$ uniformly on compact subsets of U. But if $K \subseteq U$ is compact, then since F is clearly continuous on the compact set $K \times [0,1]$, it is uniformly continuous there, hence, given any $\epsilon > 0$, there is a $\delta > 0$ such that $|F(z,s) - F(z,t)| < \epsilon$ for all $z \in \bar{B}(a,\rho)$ and $s,t \in [0,1]$ with $|s-t| < \delta$. But then if $n > \delta^{-1}$ we have for all $z \in K$

$$\begin{aligned}
|f(z) - f_n(z)| &= \big| \int_0^1 F(z,s)dz - \frac{1}{n}\sum_{i=1}^n F(z,i/n)\big| \\
&= \left| \sum_{i=1}^n \int_{(i-1)/n}^{i/n} \big(F(z,s) - F(z,i/n)\big)ds \right| \\
&\leq \sum_{i=1}^n \int_{(i-1)/n}^{i/n} |F(z,s) - F(z,i/n)|ds \\
&< \sum_{i=1}^n \epsilon/n = \epsilon.
\end{aligned}$$

Thus $f_n(z)$ tends to $f(z)$ uniformly on K as required. □

Example 18.23. If f is any continuous function on $[0,1]$, then the previous theorem shows that the function $f(z) = \int_0^1 e^{isz} f(s)ds$ is holomorphic in z, since clearly $F(z,s) = e^{isz}f(z)$ is continuous as a function on $\mathbb{C} \times [0,1]$ and, for fixed $s \in [0,1]$, F is holomorphic as a function of z. Integrals of this nature (though perhaps over the whole real line or the positive real axis) arise frequently in many parts of mathematics, as you can learn more about in the optional course on Integral Transforms.

Remark 18.24. Another way to prove the theorem is to use Morera's theorem directly: if $\gamma\colon [0,1] \to \mathbb{C}$ is a closed path in $B(a,r)$, then we have

$$\begin{aligned}
\int_\gamma f(z)dz &= \int_\gamma \big(\int_0^1 F(z,s)ds\big)dz \\
&= \int_0^1 \big(\int_\gamma F(z,s)dz\big)ds = 0,
\end{aligned}$$

where in the first line we interchanged the order of integration, and in the second we used the fact that $F(z,s)$ is holomorphic in z and Cauchy's theorem for a disk. To make this completely rigorous however, one has to justify the interchange of the orders of integration. Next term's course on Integration proves a very general result of this form known as Fubini's theorem, but for continous functions on compact subets of $\mathbb{R}^n$ one can give more elementary arguments by showing any such function is a uniform limit of linear combinations of indicator functions of "boxes" – the higher dimensional analogues of step functions – and the elementary fact that the interchange of the order of integration for indicator functions of boxes holds trivially.

19. THE IDENTITY THEOREM, ISOLATED ZEROS AND SINGULARITIES

The fact that any complex differentiable function is in fact analytic has some very surprising consequences – the most striking of which is perhaps captured by the "Identity theorem". This says that if f, g are two holomorphic functions defined on a domain U and we let $S = \{z \in U : f(z) = g(z)\}$ be the locus on which they are equal, then if S has a limit point in U it must actually be all of U. Thus for example if there is a disk $B(a, r) \subseteq U$ on which f and g agree (not matter how small r is), then in fact they are equal on all of U! The key to the proof of the Identity theorem is the following result on the zeros of a holomorphic function:

Proposition 19.1. *Let U be an open set and suppose that $g\colon U \to \mathbb{C}$ is holomorphic on U. Let $S = \{z \in U : g(z) = 0\}$. If $z_0 \in S$ then either z_0 is isolated in S (so that g is non-zero in some disk about z_0 except at z_0 itself) or $g = 0$ on a neighbourhood of z_0. In the former case there is a unique integer $k > 0$ and holomorphic function g_1 such that $g(z) = (z - z_0)^k g_1(z)$ where $g_1(z_0) \neq 0$.*

Proof. Pick any $z_0 \in U$ with $g(z_0) = 0$. Since g is analytic at z_0, if we pick $r > 0$ such that $\bar{B}(z_0, r) \subseteq U$, then we may write

$$g(z) = \sum_{k=0}^{\infty} c_k (z - z_0)^k,$$

for all $z \in B(z_0, r) \subseteq U$, where the coeficients c_k are given as in Theorem 18.11. Now if $c_k = 0$ for all k, it follows that $g(z) = 0$ for all $z \in B(0, r)$. Otherwise, we set $k = \min\{n \in \mathbb{N} : c_n \neq 0\}$ (where since $g(z_0) = 0$ we have $c_0 = 0$ so that $k \geq 1$). Then if we let $g_1(z) = (z - z_0)^{-k} g(z)$, clearly $g_1(z)$ is holomorphic on $U \backslash \{z_0\}$, but since in $B(z_0, r)$ we have we have $g_1(z) = \sum_{n=0}^{\infty} c_{k+n}(z - z_0)^n$, it follows if we set $g_1(z_0) = c_k \neq 0$ then g_1 becomes a holomorphic function on all of U. Since g_1 is continuous at z_0 and $g_1(z_0) \neq 0$, there is an $\epsilon > 0$ such that $g_1(z) \neq 0$ for all $z \in B(z_0, \epsilon)$. But $(z - z_0)^k$ vanishes only at z_0, hence it follows that $g(z) = (z - z_0)^k g_1(z)$ is non-zero on $B(a, \epsilon) \backslash \{z_0\}$, so that z_0 is isolated.

Finally, to see that k is unique, suppose that $g(z) = (z - z_0)^k g_1(z) = (z - z_0)^l g_2(z)$ say with $g_1(z_0)$ and $g_2(z_0)$ both nonzero. If $k < l$ then $g(z)/(z - z_0)^k = (z - z_0)^{l-k} g_2(z)$ for all $z \neq z_0$, hence as $z \to z_0$ we have $g(z)/(z - z_0)^k \to 0$, which contradicts the assumption that $g_1(z) \neq 0$. By symmetry we also cannot have $k > l$ so $k = l$ as required. □

Remark 19.2. The integer k in the previous proposition is called the *multiplicity* of the zero of g at $z = z_0$ (or sometimes the *order of vanishing*).

Theorem 19.3. *(Identity theorem): Let U be a domain and suppose that f_1, f_1 are holomorphic functions defined on U. Then if $S = \{z \in U : f_1(z) = f_2(z)\}$ has a limit point in U, we must have $S = U$, that is $f_1(z) = f_2(z)$ for all $z \in U$.*

Proof. Let $g = f_1 - f_2$, so that $S = g^{-1}(\{0\})$. We must show that if S has a limit point then $S = U$. Since g is clearly holomorphic in U, by Proposition

19.1 we see that if $z_0 \in S$ then either z_0 is an isolated point of S or it lies in an open ball contained in S. It follows that $S = V \cup T$ where $T = \{z \in S : z \text{ is isolated}\}$ and $V = \text{int}(S)$ is open. But since g is continuous, $S = g^{-1}(\{0\})$ is closed in U, thus $V \cup T$ is closed, and so $\text{Cl}_U(V)$, the closure[45] of V in U, lies in $V \cup T$. However, by definition, no limit point of V can lie in T so that $\text{Cl}_U(V) = V$, and thus V is open and closed in U. Since U is connected, it follows that $V = \emptyset$ or $V = U$. In the former case, all the zeros of g are isolated so that $S' = T' = \emptyset$ and S has no limit points. In the latter case, $V = S = U$ as required.

□

Remark 19.4. The requirement in the theorem that S have a limit point *lying in* U is essential: If we take $U = \mathbb{C}\backslash\{0\}$ and $f_1 = \exp(1/z) - 1$ and $f_2 = 0$, then the set S is just the points where f_1 vanishes on U. Now the zeros of f_1 have a limit point at $0 \notin U$ since $f(1/(2\pi i n)) = 0$ for all $n \in \mathbb{N}$, but certainly f_1 is not identically zero on U!

We now wish to study singularities of holomorphic functions. The key result here is Riemann's removable singularity theorem, Corollary 18.16.

Definition 19.5. If U is an open set in $\mathbb{C}$ and $z_0 \in U$, we say that a function $f : U\backslash\{z_0\} \to \mathbb{C}$ has an *isolated singularity* at z_0 if it is holomorphic on $B(z_0, r)\backslash\{z_0\}$ for some $r > 0$.

Suppose that z_0 is an isolated singularity of f. If f is bounded near z_0 we say that f has a *removable singularity* at z_0, since by Corollary 18.16 it can be extended to a holomorphic function at z_0. If f is not bounded near z_0, but the function $1/f(z)$ has a removable singularity at z_0, that is, $1/f(z)$ extends to a holomorphic function on all of $B(z_0, r)$, then we say that f has a *pole* at z_0. By Proposition 19.1 we may write $(1/f)(z) = (z - z_0)^m g(z)$ where $g(z_0) \neq 0$ and $m \in \mathbb{Z}_{>0}$. (Note that the extension of $1/f$ to z_0 must vanish there, as otherwise f would be bounded near z_0.) We say that m is the *order* of the pole of f at z_0. In this case we have $f(z) = (z - z_0)^{-m}.(1/g)$ near z_0, where $1/g$ is holomorphic near z_0 since $g(z_0) \neq 0$. If $m = 1$ we say that f has a *simple pole* at z_0.

Finally, if f has an isolated singularity at z_0 which is not removable nor a pole, we say that z_0 is an *essential singularity*.

Lemma 19.6. *Let f be a holomorphic function with a pole of order m at z_0. Then there is an $r > 0$ such that for all $z \in B(z_0, r)\backslash\{z_0\}$ we have*

$$f(z) = \sum_{n \geq -m} c_n (z - z_0)^n$$

[45]I use the notation $\text{Cl}_U(V)$, as opposed to $\bar{V}$, to emphasize that I mean the closure of V in U, not in $\mathbb{C}$, that is, $\text{Cl}_U(V)$ is equal to the union of V with the limits points of V which lie in U.

Proof. As we have already seen, we may write $f(z) = (z-z_0)^{-m}h(z)$ where m is the order of the pole of f at z_0 and $h(z)$ is holomorphic and non-vanishing at z_0. The claim follows since, near z_0, $h(z)$ is equal to its Taylor series at z_0, and multiplying this by $(z-z_0)^{-m}$ gives a series of the required form for $f(z)$. □

Definition 19.7. The series $\sum_{n\geq -m} c_n(z-z_0)^n$ is called the *Laurent series* for f at z_0. We will show later that if f has an isolated essential singularity it still has a Laurent series expansion, but the series is then involves infinitely many positive and negative powers of $(z-z_0)$.

A function on an open set U which has only isolated singularities all of which are poles is called a *meromorphic* function on U. (Thus, strictly speaking, it is a function only defined on the complement of the poles in U.)

Lemma 19.8. *Suppose that f has an isolated singularity at a point z_0. Then z_0 is a pole if and only if $|f(z)| \to \infty$ as $z \to z_0$.*

Proof. If z_0 is a pole of f then $1/f(z) = (z-z_0)^k g(z)$ where $g(z_0) \neq 0$ and $k > 0$. But then for $z \neq z_0$ we have $f(z) = (z-z_0)^{-k}(1/g(z))$, and since $g(z_0) \neq 0$, $1/g(z)$ is bounded away from 0 near z_0, while $|(z-z_0)^{-k}| \to \infty$ as $z \to z_0$, so $|f(z)| \to \infty$ as $z \to z_0$ as required.

On the other hand, if $|f(z)| \to \infty$ as $z \to z_0$, then $1/f(z) \to 0$ as $z \to z_0$, so that $1/f(z)$ has a removable singularity and f has a pole at z_0. □

Remark 19.9. The previous Lemma can be rephrased to say that f has a pole at z_0 precisely when f extends to a continuous function $f\colon U \to \mathbb{C}_\infty$ with $f(z_0) = \infty$. Moreover, you can check from Definition 13.6 that in this case, the extension is actually holomorphic. Thus the Riemann sphere allows us to put holomorphic and meromorphic functions on the same footing.

The case where f has an essential singularity is more complicated. We prove that near an isolated singularity the values of a holomorphic function are dense:

Theorem 19.10. *(Casorati-Weierstrass): Let U be an open subset of $\mathbb{C}$ and let $a \in U$. Suppose that $f\colon U\backslash\{a\} \to \mathbb{C}$ is a holomorphic function with an isolated essential singularity at a. Then for all $\rho > 0$ with $B(a,\rho) \subseteq U$, the set $f(B(a,\rho)\backslash\{a\})$ is dense in $\mathbb{C}$, that is, the closure of $f(B(a,\rho)\backslash\{a\})$ is all of $\mathbb{C}$.*

Proof. Suppose, for the sake of a contradiction, that there is some $\rho > 0$ such that $z_0 \in \mathbb{C}$ is not a limit point of $f(B(a,\rho)\backslash\{a\})$. Then the function $g(z) = 1/(f(z)-z_0)$ is bounded and non-vanishing on $B(a,\rho)\backslash\{a\}$, and hence by Riemann's removable singularity theorem, it extends to a holomorphic function on all of $B(a,\rho)$. But then $f(z) = z_0 + 1/g(z)$ has at most a pole at a which is a contradiction. □

Remark 19.11. In fact much more is true: Picard showed that if f has an isolated essential singularity at z_0 then in any open disk about z_0 the function

f takes every complex value infinitely often with at most one exception. The example of the function $f(z) = \exp(1/z)$, which has an essential singularity at $z = 0$ shows that this result is best possible, since $f(z) \neq 0$ for all $z \neq 0$.

19.1. **Principal parts.**

Definition 19.12. Recall that by Lemma 19.6 if a function f has a pole of order k at z_0 then near z_0 we may write

$$f(z) = \sum_{n \geq -k} c_n (z - z_0)^n.$$

The function $\sum_{n=-k}^{-1} c_n (z - z_0)^n$ is called the *principal part* of f at z_0, and we will denote it by $P_{z_0}(f)$. It is a rational function which is holomorphic on $\mathbb{C}\backslash\{z_0\}$. Note that $f - P_{z_0}(f)$ is holomorphic at z_0 (and also holomorphic wherever f is). The *residue* of f at z_0 is defined to be the coefficient c_{-1} and denoted $\mathrm{Res}_{z_0}(f)$.

The reason for introducing these definitions is the following: Suppose that $f \colon U \to \mathbb{C}_\infty$ is a meromorphic function with poles at a finite set $S \subseteq U$. Then for each $z_0 \in S$ we have the principal part $P_{z_0}(f)$ of f at z_0, a rational function which is holomorphic everywhere on $\mathbb{C}\backslash\{z_0\}$. The difference

$$g(z) = f(z) - \sum_{z_0 \in S} P_{z_0}(f),$$

is holomorphic on all of U (away from S the is clear because each term is, at $z_0 \in S$ the terms $P_s(f)$ for $s \in S\backslash\{z_0\}$ are all holomorphic, while $f(z) - P_{z_0}(f)$ is holomorphic at z_0 by the definition of $P_{z_0}(f)$). Thus if U is starlike and $\gamma \colon [0,1] \to U$ is any closed path in U with $\gamma^* \cap S = \emptyset$, we have

$$\int_\gamma f(z)dz = \int_\gamma g(z)dz + \sum_{z_0 \in S} \int_\gamma P_{z_0}(f)dz = \sum_{z_0 \in S} \int_\gamma P_{z_0}(f)dz.$$

The most important term in the principal part $P_{z_0}(f)$ is the term $c_{-1}/(z - z_0)$. This is because every other term has a primitive on $\mathbb{C}\backslash\{z_0\}$, hence by the Fundamental Theorem of Calculus it is the only part which contributes to the integral of $P_{z_0}(f)$ around the closed path γ. Combining these observations we see that

$$\int_\gamma f(z)dz = \sum_{z_0 \in S} \mathrm{Res}_{z_0}(f) \int_\gamma \frac{dz}{z - z_0} = 2\pi i \sum_{z_0 \in S} \mathrm{Res}_{z_0}(f).I(\gamma, z_0),$$

where $I(\gamma, z_0)$ denotes the winding number of γ about the pole z_0. This is the *residue theorem* for meromorphic functions on a starlike domain. We will shortly generalize it.

Lemma 19.13. *Suppose that f has a pole of order m at z_0, then*

$$Res_{z_0}(f) = \lim_{z \to z_0} \frac{1}{(m-1)!} \frac{d^{m-1}}{dz^{m-1}}((z - z_0)^m f(z))$$

Proof. Since f has a pole of order m at z_0 we have $f(z) = \sum_{n\geq -m} c_n(z-z_0)^n$ for z sufficiently close to z_0. Thus

$$(z-z_0)^m f(z) = c_{-m} + c_{-m+1}(z-z_0) + \ldots + c_{-1}(z-z_0)^{m-1} + \ldots$$

and the result follows from the formula for the derivatives of a power series. ☐

Remark 19.14. The last lemma is perhaps most useful in the case where the pole is simple, since in that case no derivatives need to be computed. In fact there is a special case which is worth emphasizing: Suppose that $f = g/h$ is a ratio of two holomorphic functions defined on a domain $U \subseteq \mathbb{C}$, where h is non-constant. Then f is meromorphic with poles at the zeros[46] of h. In particular, if h has a simple zero at z_0 and g is non-vanishing there, then f correspondingly has a simple pole at z_0. Since the zero of h is simple at z_0, we must have $h'(z_0) \neq 0$, and hence by the previous result

$$\mathrm{Res}_{z_0}(f) = \lim_{z\to z_0} \frac{g(z)(z-z_0)}{h(z)} = \lim_{z\to z_0} g(z). \lim_{z\to z_0} \frac{z-z_0}{h(z)-h(z_0)} = g(z_0)/h'(z_0)$$

where the last equality holds by standard Algebra of Limits results.

20. Homotopies, simply-connected domains and Cauchy's theorem

A crucial point in our proof of Cauchy's theorem for a triangle was that the interior of the triangle was entirely contained in the open set on which our holomorphic function f was defined. In general however, given a closed curve, it is not always easy to say what we mean by the "interior" of the curve. In fact there is a famous theorem, known as the Jordan Curve Theorem, which resolves this problem, but to prove it would take us too far afield. Instead we will take a slightly different strategy: in fact we will take two different approaches: the first using the notion of homotopy and the second using the winding number. For the homotopy approach, rather than focusing only on closed curves and their "interiors" we consider arbitrary curves and study what it means to deform one to another.

Definition 20.1. Suppose that U is an open set in $\mathbb{C}$ and $a, b \in U$. If $\eta\colon [0,1] \to U$ and $\gamma\colon [0,1] \to U$ are paths in U such that $\gamma(0) = \eta(0) = a$ and $\gamma(1) = \eta(1) = b$, then we say that γ and η are *homotopic* in U if there is a continuous function $h\colon [0,1]\times[0,1] \to U$ such that

$$h(0,s) = a, \quad h(1,s) = b$$
$$h(t,0) = \gamma(t), \quad h(t,1) = \eta(t).$$

One should think of h as a family of paths in U indexed by the second variable s which continuously deform γ into η.

[46]Strictly speaking, the poles of f form a subset of the zeros of h, since if g also vanishes at a point z_0, then f may have a removable singularity at z_0.

A special case of the above definition is when $a = b$ and γ and η are closed paths. In this case there is a constant path $c_a \colon [0,1] \to U$ going from a to $b = a$ which is simply given by $c_a(t) = a$ for all $t \in [0,1]$. We say a closed path starting and ending at a point $a \in U$ is *null homotopic* if it is homotopic to the constant path c_a. One can show that the relation "γ is homotopic to η" is an equivalence relation, so that any path γ between a and b belongs to a unique equivalence class, known as its homotopy class.

Definition 20.2. Suppose that U is a domain in $\mathbb{C}$. We say that U is *simply connected* if for every $a, b \in U$, any two paths from a to b are homotopic in U.

Lemma 20.3. *Let U be a convex open set in $\mathbb{C}$. Then U is simply connected. Moreover if U_1 and U_2 are homeomorphic, then U_1 is simply connected if and only if U_2 is.*

Proof. Suppose that $\gamma \colon [0,1] \to U$ and $\eta \colon [0,1] \to U$ are paths starting and ending at a and b respectively for some $a, b \in U$. Then for $(s,t) \in [0,1] \times [0,1]$ let

$$h(t,s) = (1-s)\gamma(t) + s\eta(t)$$

It is clear that h is continuous and one readily checks that h gives the required homotopy. For the moreover part, if $f \colon U_1 \to U_2$ is a homeomorphism then it is clear that f induces a bijection between continuous paths in U_1 to those in U_2 and also homotopies in U_1 to those in U_2, so the claim follows. □

Remark 20.4. (*Non-examinable*) In fact, with a bit more work, one can show that any starlike domain D is also simply-connected. The key is to show that a domain is simply-connected if all closed paths starting and ending at a given point $z_0 \in D$ are null-homotopic. If D is star-like with respect to $z_0 \in D$, then if $\gamma \colon [0,1] \to D$ is a closed path with $\gamma(0) = \gamma(1) = z_0$, it follows $h(s,t) = z_0 + s(\gamma(t) - z_0)$ gives a homotopy between γ and the constant path c_{z_0}.

Thus we see that we already know many examples of simply connected domains in the plane, such as disks, ellipsoids, half-planes. The second part of the above lemma also allows us to produce non-convex examples:

Example 20.5. Consider the domain

$$D_{\eta,\epsilon} = \{z \in \mathbb{C} : z = re^{i\theta} : \eta < r < 1, 0 < \theta < 2\pi(1-\epsilon)\},$$

where $0 < \eta, \epsilon < 1/10$ say, then $D_{\eta,\epsilon}$ is clearly not convex, but it is the image of the convex set $(0,1) \times (0, 1-\epsilon)$ under the map $(r,\theta) \mapsto re^{2\pi i\theta}$. Since this map has a continuous (and even differentiable) inverse, it follows $D_{\eta,\epsilon}$ is simply-connected. When η and ϵ are small, the boundary of this set, oriented anti-clockwise, is a version of what is called a *key-hole contour*.

We are now ready to state our extension of Cauchy's theorem. The proof is given in the Appendices.

Theorem 20.6. *Let U be a domain in $\mathbb{C}$ and $a, b \in U$. Suppose that γ and η are paths from a to b which are homotopic in U and $f\colon U \to \mathbb{C}$ is a holomorphic function. Then*

$$\int_\gamma f(z)dz = \int_\eta f(z)dz.$$

Remark 20.7. Notice that this theorem is really more general than the previous versions of Cauchy's theorem we have seen – in the case where a holomorphic function $f\colon U \to \mathbb{C}$ has a primitive the conclusion of the previous theorem is of course obvious from the Fundamental theorem of Calculus[47], and our previous formulations of Cauchy's theorem were proved by producing a primitive for f on U. One significance of the homotopy form of Cauchy's theorem is that it applies to domains U even when there is no primitive for f on U.

Theorem 20.8. *Suppose that U is a simply-connected domain, let $a, b \in U$, and let $f\colon U \to \mathbb{C}$ be a holomorphic function on U. Then if γ_1, γ_2 are paths from a to b we have*

$$\int_{\gamma_1} f(z)dz = \int_{\gamma_2} f(z)dz.$$

In particular, if γ is a closed oriented curve we have $\int_\gamma f(z)dz = 0$, and hence any holomorphic function on U has a primitive.

Proof. Since U is simply-connected, any two paths from from a to b are homotopic, so we can apply Theorem 20.6. For the last part, in a simply-connected domain any closed path $\gamma\colon [0,1] \to U$, with $\gamma(0) = \gamma(1) = a$ say, is homotopic to the constant path $c_a(t) = a$, and hence $\int_\gamma f(z)dz = \int_{c_a} f(z)dz = 0$. The final assertion then follows from the Theorem 16.21. □

Example 20.9. If $U \subseteq \mathbb{C}\backslash\{0\}$ is simply-connected, the previous theorem shows that there is a holomorphic branch of $[\mathrm{Log}(z)]$ defined on all of U (since any primitive for $f(z) = 1/z$ will be such a branch).

Remark 20.10. Recall that in Definition 18.6 we called a domain D in the complex plane *primtive* if every holomorphic function $f\colon D \to \mathbb{C}$ on it had a primitive. Theorem 20.8 shows that any simply-connected domain is primitive. In fact the converse is also true – any primitive domain is necessarily simply-connected. Thus the term "primitive domain" is in fact another name for a simply-connected domain.

The definition of winding number allows us to give another version of Cauchy's integral formula (sometimes called the *winding number* or *homology* form of Cauchy's theorem).

[47] Indeed the hypothesis that the paths γ and η are homotopic is irrelevant when f has a primitive on U.

Theorem 20.11. *Let $f\colon U \to \mathbb{C}$ be a holomorphic function and let $\gamma\colon [0,1] \to U$ be a closed path whose inside lies entirely in U, that is $I(\gamma, z) = 0$ for all $z \notin U$. Then we have, for all $z \in U\backslash\gamma^*$,*

$$\int_\gamma f(\zeta)d\zeta = 0; \quad \int_\gamma \frac{f(\zeta)}{\zeta - z}d\zeta = 2\pi i I(\gamma, z)f(z).$$

Moreover, if U is simply-connected and $\gamma\colon [a,b] \to U$ is any closed path, then $I(\gamma, z) = 0$ for any $z \notin U$, so the above identities hold for all closed paths in such U.

Remark 20.12. The "moreover" statement in fact just uses the fact that a simply-connected domain is primitive: if D is a domain and $w \notin D$, then the function $1/(z-w)$ is holomorphic on all of D, and hence has a primitive on D. It follows $I(\gamma, w) = 0$ for any path γ with $\gamma^* \subseteq D$.

Remark 20.13. This version of Cauchy's theorem has a natural extension: instead of integrating over a single closed path, one can integrate over formal sums of closed paths, which are known as *cycles*: if $a \in \mathbb{N}$ and $\gamma_1, \ldots, \gamma_k$ are closed paths and $a_1, \ldots, a_k$ are complex numbers (we will usually only consider the case where they are integers) then we define the integral around the formal sum $\Gamma = \sum_{i=1}^k a_i\gamma_i$ of a function f to be

$$\int_\Gamma f(z)dz = \sum_{i=1}^k a_i \int_{\gamma_i} f(z)dz.$$

Since the winding number can be expressed as an integral, this also gives a natural defintion of the winding number for such Γ: explicitly $I(\Gamma, z) = \sum_{i=1}^k a_i I(\gamma_i, z)$. If we write $\Gamma^* = \gamma_1^* \cup \ldots \cup \gamma_k^*$ then $I(\Gamma, z)$ is defined for all $z \notin \Gamma^*$. The winding number version Cauchy's theorem then holds (with the same proof) for cycles in an open set U, where we define the inside of a cycle to be the set of $z \in \mathbb{C}$ for which $I(\Gamma, z) \neq 0$.

Note that if z is inside Γ then it must be the case that z is inside some γ_i, but the converse is not necessarily the case: it may be that z lies inside some of the γ_i but does not lie inside Γ. One natural way in which cycles arise are as the boundaries of an open subsets of the plane: if Ω is an domain in the plane, then $\partial\Omega$, the boundary of Ω is often a *union* of curves rather than a single curve[48]. For example if $r < R$ then $\Omega = B(0,R)\backslash\bar{B}(0,r)$ has a boundary which is a union of two concentric circles. If these circles are oriented correctly, then the "inside" of the cycle Γ which they form is precisely Ω (see the discussion of Laurent series below for more details). Thus the origin, although inside each of the circles $\gamma(0,r)$ and $\gamma(0,R)$, is not inside Γ. The cycles version of Cauchy's theorem is thus closest to Green's theorem in multivariable calculus.

[48] Of course in general the boundary of an open set need not be so nice as to be a union of curves at all.

As a first application of this new form of Cauchy's theorem, we establish the *Laurent expansion* of a function which is holomorphic in an annulus. This is a generalization of Taylor's theorem, and we already saw it in the special case of a function with a pole singularity.

Definition 20.14. Let $0 \leq r < R$ be real numbers and let $z_0 \in \mathbb{C}$. An open *annulus* is a set

$$A = A(r, R, z_0) = B(z_0, R) \backslash \bar{B}(z_0, r) = \{z \in \mathbb{C} : r < |z - z_0| < R\}.$$

If we write (for $s > 0$) $\gamma(z_0, s)$ for the closed path $t \mapsto z_0 + se^{2\pi i t}$ then notice that the inside of the cycle $\Gamma_{r,R,z_0} = \gamma(z_0, R) - \gamma(z_0, r)$ is precisely A, since for any s, $I(\gamma(z_0, s), z)$ is 1 precisely if $z \in B(z_0, s)$ and 0 otherwise.

Theorem 20.15. *Suppose that $0 < r < R$ and $A = A(r, R, z_0)$ is an annulus centred at z_0. If $f : U \to \mathbb{C}$ is holomorphic on an open set U which contains $\bar{A}$, then there exist $c_n \in \mathbb{C}$ such that*

$$f(z) = \sum_{n=-\infty}^{\infty} c_n (z - z_0)^n, \quad \forall z \in A.$$

Moreover, the c_n are unique and are given by the following formulae:

$$c_n = \frac{1}{2\pi i} \int_{\gamma_s} \frac{f(z)}{(z - z_0)^{n+1}} dz,$$

where $s \in [r, R]$ and for any $s > 0$ we set $\gamma_s(t) = z_0 + se^{2\pi i t}$.

Proof. By translation we may assume that $z_0 = 0$. Since A is the inside of the cycle Γ_{r,R,z_0} it follows from the winding number form of Cauchy's integral formula that for $w \in A$ we have

$$2\pi i f(w) = \int_{\gamma_R} \frac{f(z)}{z - w} dz - \int_{\gamma_r} \frac{f(z)}{z - w} dz$$

But now the result follows in the same way as we showed holomorphic functions were analytic: if we fix w, then, for $|w| < |z|$ we have $\frac{1}{z-w} = \sum_{n=0}^{\infty} w^n / z^{n+1}$, converging uniformly in z in $|z| > |w| + \epsilon$ for any $\epsilon > 0$. It follows that

$$\int_{\gamma_R} \frac{f(z)}{z - w} dz = \int_{\gamma_R} \sum_{n=0}^{\infty} \frac{f(z) w^n}{z^{n+1}} dz = \sum_{n \geq 0} \left(\int_{\gamma_R} \frac{f(z)}{z^{n+1}} dz \right) w^n.$$

for all $w \in A$. Similarly since for $|z| < |w|$ we have[49] $\frac{1}{w-z} = \sum_{n \geq 0} z^n / w^{n+1} = \sum_{n=-1}^{-\infty} w^n / z^{n+1}$, again converging uniformly on $|z|$ when $|z| < |w| - \epsilon$ for $\epsilon > 0$, we see that

$$\int_{\gamma_r} \frac{f(z)}{w - z} dz = \int_{\gamma_r} \sum_{n=-1}^{-\infty} f(z) w^n / z^{n+1} dz = \sum_{n=-1}^{-\infty} \left(\int_{\gamma_r} \frac{f(z)}{z^{n+1}} dz \right) w^n.$$

[49] Note the sign change.

Thus taking $(c_n)_{n\in\mathbb{Z}}$ as in the statement of the theorem, we see that

$$f(w) = \frac{1}{2\pi i}\int_{\gamma_R}\frac{f(z)}{z-w}dz - \frac{1}{2\pi i}\int_{\gamma_r}\frac{f(z)}{z-w}dz = \sum_{n\in\mathbb{Z}} c_n z^n,$$

as required. To see that the c_n are unique, one checks using uniform convergence that if $\sum_{n\in\mathbb{Z}} d_n z^n$ is any series expansion for $f(z)$ on A, then the d_n must be given by the integral formulae above.

Finally, to see that the c_n can be computed using any circular contour γ_s, note that if $r \leq s_1 < s_2 \leq R$ then $f/(z-z_0)^{n+1}$ is holomorphic on the inside of $\Gamma = \gamma_{s_2} - \gamma_{s_1}$, hence by the homology form of Cauchy's theorem $0 = \int_\Gamma f(z)/(z-z_0)^{n+1}dz = \int_{\gamma_{s_2}} f(z)/(z-z_0)^{n+1}dz - \int_{\gamma_{s_1}} f(z)/(z-z_0)^{n+1}dz$. □

Remark 20.16. Note that the above proof shows that the integral $\int_{\gamma_R}\frac{f(z)}{z-w}dz$ defines a holomorphic function of w in $B(z_0,R)$, while $\int_{\gamma_r}\frac{f(z)}{z-w}dz$ defines a holomorphic function of w on $\mathbb{C}\backslash B(z_0,r)$. Thus we have actually expressed $f(w)$ on A as the difference of two functions which are holomorphic on $B(z_0,R)$ and $\mathbb{C}\backslash\bar{B}(z_0,r)$ respectively.

Corollary 20.17. *If $f\colon U \to \mathbb{C}$ is a holomorphic function on an open set U containing an annulus $A = A(r,R,z_0)$ then f has a Laurent expansion on A. In particular, if f has an isolated singularity at z_0, then it has a Laurent expansion on a punctured disc $B(z_0,r)\backslash\{z_0\}$ for sufficiently small $r > 0$.*

Proof. This follows from the previous Theorem and the fact that for any $0 \leq r \leq R$ we have

$$A(r,R,z_0) = \bigcup_{r<r_1<R_1<R} \overline{A(r_1,R_1,z_0)}.$$

The final sentence follows from the fact that $B(z_0,r)\backslash\{z_0\} = A(0,r,z_0)$. □

Definition 20.18. Let $f\colon U\backslash S \to \mathbb{C}$ be a function which is holomorphic on a domain U except at a discrete set $S \subseteq U$. Then for any $a \in S$ Corollary 20.17 shows that for $r > 0$ sufficiently small, we have

$$f(z) = \sum_{n\in\mathbb{Z}} c_n(z-a)^n, \quad \forall z \in B(a,r)\backslash\{a\}.$$

We define

$$P_a(f) = \sum_{n=-1}^{-\infty} c_n(z-a)^n,$$

to be the *principal part* of f at a. This generalizes the previous definition we gave for the principal part of a meromorphic function. Note that the proof of Theorem 20.17 shows that the series $P_a(f)$ is uniformly convergent on $\mathbb{C}\backslash B(a,r)$ for all $r > 0$, and hence defines a holomorphic function on $\mathbb{C}\backslash\{a\}$.

21. The argument principle

Lemma 21.1. *Suppose that $f\colon U \to \mathbb{C}$ is a meromorphic and has a zero of order k or a pole of order k at $z_0 \in U$. Then $f'(z)/f(z)$ has a simple pole at z_0 with residue k or $-k$ respectively.*

Proof. If $f(z)$ has a zero of order k we have $f(z) = (z - z_0)^k g(z)$ where $g(z)$ is holomorphic near z_0 and $g(z_0) \neq 0$. It follows that

$$f'(z)/f(z) = \frac{k}{z - z_0} + g'(z)/g(z),$$

and since $g(z) \neq 0$ near z_0 it follows $g'(z)/g(z)$ is holomorphic near z_0, so that the result follows. The case where f has a pole at z_0 is similar. □

Remark 21.2. Note that if U is an open set on which one can define a holomorphic branch L of $[\mathrm{Log}(z)]$ then $g(z) = L(f(z))$ has $g'(z) = f'(z)/f(z)$. Thus integrating $f'(z)/f(z)$ along a path γ will measure the change in argument around the origin of the path $f(\gamma(t))$. The residue theorem allows us to relate this to the number of zeros and poles of f inside γ, as the next theorem shows:

Theorem 21.3. *(Argument principle): Suppose that U is an open set and $f\colon U \to \mathbb{C}$ is a meromorphic function on U. If $B(a, r) \subseteq U$ and N is the number of zeros (counted with multiplicity) and P is the number of poles (again counted with multiplicity) of f inside $B(a, r)$ and f has neither on $\partial B(a, r)$ then*

$$N - P = \frac{1}{2\pi i}\int_\gamma \frac{f'(z)}{f(z)}dz,$$

where $\gamma(t) = a + re^{2\pi i t}$ is a path with image $\partial B(a, r)$. Moreover this is the winding number of the path $\Gamma = f \circ \gamma$ about the origin.

Proof. It is easy to check that $I(\gamma, z)$ is 1 if $|z - a| \leq 1$ and is 0 otherwise. Since Lemma 21.1 shows that $f'(z)/f(z)$ has simple poles at the zeros and poles of f with residues the corresponding orders the result immediately from Theorem 22.1.

For the last part, note that the winding number of $\Gamma(t) = f(\gamma(t))$ about zero is just

$$\int_{f\circ\gamma} dw/w = \int_0^1 \frac{1}{f(\gamma(t))} f'(\gamma(t))\gamma'(t)dt = \int_\gamma \frac{f'(z)}{f(z)}dz$$

□

Remark 21.4. The argument principle also holds, with the same proof, to any closed path γ on which f is continuous and non-vanishing, provided it has winding number $+1$ around its inside. Thus for example it applies to triangles, or paths built from an arc of a circle and the line segments joining the end-points to the centre of the circle, provided they are correctly oriented.

The argument principle is very useful – we use it here to establish some important results.

Theorem 21.5. *(Rouché's theorem): Suppose that f and g are holomorphic functions on an open set U in $\mathbb{C}$ and $\bar{B}(a,r) \subset U$. If $|f(z)| > |g(z)|$ for all $z \in \partial B(a,r)$ then f and $f+g$ have the same change in argument around γ, and hence the same number of zeros in $B(a,r)$ (counted with multiplicities).*

Proof. Let $\gamma(t) = a + re^{2\pi i t}$ be a parametrization of the boundary circle of $B(a,r)$. We need to show that $(f+g)/f = 1 + g/f$ has the same number of zeros as poles (Note that $f(z) \neq 0$ on $\partial B(a,r)$ since $|f(z)| > |g(z)|$.) But by the argument principle, this number is the winding number of $\Gamma(t) = h(\gamma(t))$ about zero, where $h(z) = 1 + g(z)/f(z)$. Since, by assumption, for $z \in \gamma^*$ we have $|g(z)| < |f(z)|$ and so $|g(z)/f(z)| < 1$, the image of Γ lies entirely in $B(1,1)$ and thus in the half-plane $\{z : \Re(z) > 0\}$. Hence picking a branch of Log defined on this half-plane, we see that the integral

$$\int_\Gamma \frac{dz}{z} = \mathrm{Log}(h(\gamma(1)) - \mathrm{Log}(h(\gamma(0)) = 0$$

as required.

∎

Remark 21.6. Rouche's theorem can be useful in counting the number of zeros of a function f – one tries to find an approximation to f whose zeros are easier to count and then by Rouche's theorem obtain information about the zeros of f. Just as for the argument principle above, it also holds for closed paths which having winding number about their inside.

Example 21.7. Suppose that $P(z) = z^4 + 5z + 2$. Then on the circle $|z| = 2$, we have $|z|^4 = 16 > 5.2 + 2 \geq |5z+2|$, so that if $g(z) = 5z+2$ we see that $P - g = z^4$ and P have the same number of roots in $B(0,2)$. It follows by Rouche's theorem that the four roots of $P(z)$ all have modulus less than 2. On the other hand, if we take $|z| = 1$, then $|5z+2| \geq 5 - 2 = 3 > |z^4| = 1$, hence $P(z)$ and $5z+2$ have the same number of roots in $B(0,1)$. It follows $P(z)$ has one root of modulus less than 1, and 3 of modulus between 1 and 2.

Theorem 21.8. *(Open mapping theorem): Suppose that $f \colon U \to \mathbb{C}$ is holomorphic and non-constant on a domain U. Then for any open set $V \subset U$ the set $f(V)$ is also open.*

Proof. Suppose that $w_0 \in f(V)$, say $f(z_0) = w_0$. Then $g(z) = f(z) - w_0$ has a zero at z_0 which, since f is nonconstant, is isolated. Thus we may find an $r > 0$ such that $g(z) \neq 0$ on $\bar{B}(z_0,r)\backslash\{z_0\} \subset U$ and in particular since $\partial B(z_0,r)$ is compact, we have $|g(z)| \geq \delta > 0$ on $\partial B(z_0,r)$. But then if $|w - w_0| < \delta$ it follows $|w - w_0| < |g(z)|$ on $\partial B(z_0,r)$, hence by Rouche's theorem, since $g(z)$ has a zero in $B(z_0,r)$ it follows $h(z) = g(z) + (w_0 - w) = f(z) - w$ does also, that is, $f(z)$ takes the value w in $B(z_0,r)$. Thus $B(w_0,\delta) \subseteq f(B(z_0,r))$ and hence $f(U)$ is open as required. ∎

Remark 21.9. Note that the proof actually establishes a bit more than the statement of the theorem: if $w_0 = f(z_0)$ then the multiplicity d of the zero of the function $f(z) - w_0$ at z_0 is called the *degree* of f at z_0. The proof shows that locally the function f is d-to-1, counting multiplicities, that is, there are $r, \epsilon \in \mathbb{R}_{>0}$ such that for every $w \in B(w_0, \epsilon)$ the equation $f(z) = w$ has d solutions counted with multiplicity in the disk $B(z_0, r)$.

Theorem 21.10. *(Inverse function theorem): Suppose that $f: U \to \mathbb{C}$ is injective and holomorphic and that $f'(z) \neq 0$ for all $z \in U$. If $g\colon f(U) \to U$ is the inverse of f, then g is holomorphic with $g'(w) = 1/f'(g(w))$.*

Proof. By the open mapping theorem, the function g is continuous, indeed if V is open in $f(U)$ then $g^{-1}(V) = f(V)$ is open by that theorem. To see that g is holomorphic, fix $w_0 \in f(U)$ and let $z_0 = g(w_0)$. Note that since g and f are continuous, if $w \to w_0$ then $f(w) \to z_0$. Writing $z = f(w)$ we have

$$\lim_{w \to w_0} \frac{g(w) - g(w_0)}{w - w_0} = \lim_{z \to z_0} \frac{z - z_0}{f(z) - f(z_0)} = 1/f'(z_0)$$

as required. □

Remark 21.11. Note that the non-trivial part of the proof of the above theorem is the fact that g is continuous! In fact the condition that $f'(z) \neq 0$ follows from the fact that f is bijective – this can be seen using the degree of f: if $f'(z_0) = 0$ and f is nonconstant, we must have $f(z) - f(z_0) = (z - z_0)^k g(z)$ where $g(z_0) \neq 0$ and $k \geq 1$. Since we can chose a holomorphic branch of $g^{1/k}$ near z_0 it follows that $f(z)$ is locally k-to-1 near z_0, which contradicts the injectivity of f. For details see the Appendices. Notice that this is in contrast with the case of a single real variable, as the example $f(x) = x^3$ shows. Once again, complex analysis is "nicer" than real analysis!

22. THE RESIDUE THEOREM

We can now prove one of the most useful theorems of the course – it is extremely powerful as a method for computing integrals, as you will see this course and many others.

Theorem 22.1. *(Residue theorem): Suppose that U is an open set in $\mathbb{C}$ and γ is a path whose inside is contained in U, so that for all $z \notin U$ we have $I(\gamma, z) = 0$. Then if $S \subset U$ is a finite set such that $S \cap \gamma^* = \emptyset$ and f is a holomorphic function on $U \backslash S$ we have*

$$\frac{1}{2\pi i} \int_\gamma f(z) dz = \sum_{a \in S} I(\gamma, a) Res_a(f)$$

Proof. For each $a \in S$ let $P_a(f)(z) = \sum_{n=-1}^{-\infty} c_n(a)(z-a)^n$ be the principal part of f at a, a holomorphic function on $\mathbb{C}\backslash\{a\}$. Then by definition of $P_a(f)$, the difference $f - P_a(f)$ is holomorphic at $a \in S$, and thus $g(z) =$

$f(z) - \sum_{a \in S} P_a(f)$ is holomorphic on all of U. But then by Theorem 20.11 we see that $\int_\gamma g(z)dz = 0$, so that

$$\int_\gamma f(z)dz = \sum_{a \in S} \int_\gamma P_a(f)(z)dz$$

But by the proof of Theorem 20.17, the series $P_a(f)$ converges uniformly on γ^* so that

$$\begin{aligned}\int_\gamma P_a(f)dz &= \int_\gamma \sum_{n=-1}^{-\infty} c_n(a)(z-a)^n = \sum_{n=1}^{\infty} \int_\gamma \frac{c_{-n}(a)dz}{(z-a)^n} \\ &= \int_\gamma \frac{c_{-1}(a)dz}{z-a} = I(\gamma, a)\mathrm{Res}_a(f),\end{aligned}$$

since for $n > 1$ the function $(z-a)^{-n}$ has a primitive on $\mathbb{C}\backslash\{a\}$. The result follows. □

Remark 22.2. In practice, in applications of the residue theorem, the winding numbers $I(\gamma, a)$ will be simple to compute in terms of the argument of $(z-a)$ – in fact most often they will be 0 or ± 1 as we will usually apply the theorem to integrals around simple closed curves.

22.1. **Residue Calculus.** The Residue theorem gives us a very powerful technique for computing many kinds of integrals. In this section we give a number of examples of its application.

Example 22.3. Consider the integral $\int_0^{2\pi} \frac{dt}{1+3\cos^2(t)}$. If we let γ be the path $t \mapsto e^{it}$ and let $z = e^{it}$ then $\cos(t) = \Re(z) = \frac{1}{2}(z + \bar{z}) = \frac{1}{2}(z + 1/z)$. Thus we have

$$\frac{1}{1+3\cos^2(t)} = \frac{1}{1+3/4(z+1/z)^2} = \frac{1}{1+\frac{3}{4}z^2+\frac{3}{2}+\frac{3}{4}z^{-2}} = \frac{4z^2}{3+10z^2+3z^4},$$

Finally, since $dz = izdt$ it follows

$$\int_0^{2\pi} \frac{dt}{1+3\cos^2(t)} = \int_\gamma \frac{-4iz}{3+10z^2+3z^4}dz.$$

Thus we have turned our real integral into a contour integral, and to evaluate the contour integral we just need to calculate the residues of the meromorphic function $g(z) = \frac{-4iz}{3+10z^2+3z^4}$ at the poles it has inside the unit circle. Now the poles of $g(z)$ are the zeros of the polynomial $p(z) = 3 + 10z^2 + 3z^4$, which are at $z^2 \in \{-3, -1/3\}$. Thus the poles inside the unit circle are at $\pm i/\sqrt{3}$. In particular, since p has degree 4 and has four roots, they must all be simple zeros, and so g has simple poles at these points. The residue at a simple pole z_0 can be calculated as the limit $\lim_{z \to z_0}(z - z_0)g(z)$, thus we see (compare with Remark 19.14) that

$$\begin{aligned}\mathrm{Res}_{z=\pm i/\sqrt{3}}(g(z)) = \lim_{z\to\pm i/\sqrt{3}} \frac{-4iz(z-\pm i/\sqrt{3})}{3+10z^2+3z^4} &= (\pm 4/\sqrt{3}).\frac{1}{p'(\pm i/\sqrt{3})} \\ &= (\pm 4/\sqrt{3}).\frac{1}{20(\pm i/\sqrt{3})+12(\pm i/\sqrt{3})^3} = 1/4i.\end{aligned}$$

It now follows from the Residue theorem that

$$\int_0^{2\pi} \frac{dt}{1+3\cos^2(t)} = 2\pi i\big(\mathrm{Res}_{z=i/\sqrt{3}}((g(z)) + \mathrm{Res}_{z=-i/\sqrt{3}}(g(z))\big) = \pi.$$

Remark 22.4. Often we are interested in integrating along a path which is not closed or even finite, for example, we might wish to understand the integral of a function on the positive real axis. The residue theorem can still be a power tool in calculating these integrals, provided we complete the path to a closed one in such a way that we can control the extra contribution to the integral along the part of the path we add.

Example 22.5. If we have a function f which we wish to integrate over the whole real line (so we have to treat it as an improper Riemann integral) then we may consider the contours Γ_R given as the concatenation of the paths $\gamma_1\colon[-R,R]\to\mathbb{C}$ and $\gamma_2\colon[0,1]\to\mathbb{C}$ where

$$\gamma_1(t) = -R+t; \quad \gamma_2(t) = Re^{i\pi t}.$$

(so that $\Gamma_R = \gamma_2 \star \gamma_1$ traces out the boundary of a half-disk). In many cases one can show that $\int_{\gamma_2} f(z)dz$ tends to 0 as $R\to\infty$, and by calculating the residues inside the contours Γ_R deduce the integral of f on $(-\infty,\infty)$. To see this strategy in action, consider the integral

$$\int_0^\infty \frac{dx}{1+x^2+x^4}.$$

It is easy to check that this integral exists as an improper Riemann integral, and since the integrand is even, it is equal to

$$\frac{1}{2}\lim_{R\to\infty}\int_{-R}^{R} \frac{dx}{1+x^2+x^4}dx.$$

If $f(z) = 1/(1+z^2+z^4)$, then $\int_{\Gamma_R} f(z)dz$ is equal to $2\pi i$ times the sum of the residues inside the path Γ_R. The function $f(z) = 1/(1+z^2+z^4)$ has poles at $z^2 = \pm e^{2\pi i/3}$ and hence at $\{e^{\pi i/3}, e^{2\pi i/3}, e^{4\pi i/3}, e^{5\pi i/3}\}$. They are all simple poles and of these only $\{\omega,\omega^2\}$ are in the upper-half plane, where $\omega = e^{i\pi/3}$. Thus by the residue theorem, for all $R>1$ we have

$$\int_{\Gamma_R} f(z)dz = 2\pi i\big(\mathrm{Res}_\omega(f(z)) + \mathrm{Res}_{\omega^2}(f(z))\big),$$

and we may calculate the residues using the limit formula as above (and the fact that it evaluates to the reciprocal of the derivative of $1+z^2+z^4$): Indeed since $\omega^3 = -1$ we have $\mathrm{Res}_\omega(f(z)) = \frac{1}{2\omega+4\omega^3} = \frac{1}{2\omega-4}$, while $\mathrm{Res}_{\omega^2}(f(z)) = \frac{1}{2\omega^2+4\omega^6} = \frac{1}{4+2\omega^2}$. Thus we obtain:

$$\begin{aligned}\int_{\Gamma_R} f(z)dz &= 2\pi i\big(\frac{1}{2\omega-4}+\frac{1}{2\omega^2+4}\big)\\ &= \pi i\big(\frac{1}{\omega-2}+\frac{1}{\omega^2+2}\big)\\ &= \pi i\big(\frac{\omega^2+\omega}{2(\omega-\omega^2)-5}\big) = -\sqrt{3}\pi/(-3) = \pi/\sqrt{3},\end{aligned}$$

(where we used the fact that $\omega^2+\omega = i\sqrt{3}$ and $\omega-\omega^2 = 1$). Now clearly

$$\int_{\Gamma_R} f(z)dz = \int_{-R}^{R}\frac{dt}{1+t^2+t^4} + \int_{\gamma_2} f(z)dz,$$

and by the estimation lemma we have

$$\big|\int_{\gamma_2} f(z)dz\big| \leq \sup_{z\in\gamma_2^*}|f(z)|.\ell(\gamma_2) \leq \frac{\pi R}{R^4-R^2-1} \to 0,$$

as $R\to\infty$, it follows that

$$\pi/\sqrt{3} = \lim_{R\to\infty}\int_{\Gamma_R} f(z)dz = \int_{-\infty}^{\infty}\frac{dt}{1+t^2+t^4}.$$

22.2. **Jordan's Lemma and applications.** The following lemma is a real-variable fact which is fundamental to something known as *convexity*. Note that if x, y are vectors in any vector space then the set $\{tx+(1-t)y : t \in [0,1]\}$ describes the line segment between x and y.

Lemma 22.6. *Let $g\colon \mathbb{R}\to\mathbb{R}$ be a twice differentiable function. Then if $[a,b]$ is an interval on which $g''(x) < 0$, the function g is convex on $[a,b]$, that is, for $x < y \in [a,b]$ we have*

$$g(tx+(1-t)y) \geq tg(x)+(1-t)g(y), \quad t\in[0,1].$$

Thus informally speaking, chords between points on the graph of g lie below the graph itself.

Proof. Given $x, y \in [a,b]$ and $t\in[0,1]$ let $\xi = tx+(1-t)y$, a point in the interval between x and y. Now the slope of the chord between $(x, g(x))$ and $(\xi, g(\xi))$ is, by the Mean Value Theorem, equal to $g'(s_1)$ where s_1 lies between x and ξ, while the slope of the chord between $(\xi, g(\xi))$ and $(y, g(y))$ is equal to $g'(s_2)$ for s_2 between ξ and y. If $g(\xi) < tg(x)+(1-t)g(y)$ it follows that $g'(s_1) < 0$ and $g'(s_2) > 0$. Thus by the mean value theorem for $g'(x)$ applied to the points s_1 and s_2 it follows there is an $s\in(s_1,s_2)$ with $g''(s) = (g'(s_2)-g'(s_1))/(s_2-s_1) > 0$, contradicting the assumption that $g''(x)$ is negative on (a,b). □

The following lemma is an easy application of this convexity result.

Lemma 22.7. *(Jordan's Lemma): Let $f\colon \mathbb{H} \to \mathbb{C}_\infty$ be a meromorphic function on the upper-half plane $\mathbb{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$. Suppose that $f(z) \to 0$ as $z \to \infty$ in $\mathbb{H}$. Then if $\gamma_R(t) = Re^{it}$ for $t \in [0,\pi]$ we have*

$$\int_{\gamma_R} f(z)e^{i\alpha z}dz \to 0$$

as $R \to \infty$ for all $\alpha \in \mathbb{R}_{>0}$.

Proof. Suppose that $\epsilon > 0$ is given. Then by assumption we may find an S such that for $|z| > S$ we have $|f(z)| < \epsilon$. Thus if $R > S$ and $z = \gamma_R(t)$, it follows that

$$|f(z)e^{i\alpha z}| =\leq \epsilon e^{-\alpha R\sin(t)}.$$

But now applying Lemma 22.6 to the function $g(t) = \sin(t)$ with $x = 0$ and $y = \pi/2$ we see that $\sin(t) \geq \frac{2}{\pi}t$ for $t \in [0,\pi/2]$. Similarly we have $\sin(\pi - t) \geq 2(\pi - t)/\pi$ for $t \in [\pi/2,\pi]$. Thus we have

$$|f(z)e^{i\alpha z}| \leq \begin{cases} \epsilon.e^{-2\alpha Rt/\pi}, & t \in [0,\pi/2] \\ \epsilon.e^{-2\alpha R(\pi-t)/\pi} & t \in [\pi/2,\pi] \end{cases}$$

But then it follows that

$$\Big|\int_{\gamma_R} f(z)e^{i\alpha z}dz\Big| \leq 2\int_0^{\pi/2} \epsilon R.e^{-2\alpha Rt/\pi}dt = \epsilon.\pi\frac{1-e^{-\alpha R}}{\alpha} < \epsilon.\pi/\alpha,$$

Thus since $\pi/\alpha > 0$ is independent of R, it follows that $\int_{\gamma_R} f(z)e^{i\alpha z}dz \to 0$ as $R \to \infty$ as required. $\square$

Remark 22.8. If η_R is an arc of a semicircle in the upper half plane, say $\eta_R(t) = Re^{it}$ for $0 \leq t \leq 2\pi/3$, then the same proof shows that $\int_{\eta_R} f(z)e^{i\alpha z}dz$ tends to zero as R tends to infinity. This is sometimes useful when integrating around the boudary of a sector of disk (that is a set of the form $\{re^{i\theta} : 0 \leq r \leq R, \theta \in [\theta_1,\theta_2]\}$).

It is also useful to note that if $\alpha < 0$ then the integral of $f(z)e^{i\alpha z}$ around a semicircle in the *lower* half plane tends to zero as the radius of the semicircle tends to infinity provided $|f(z)| \to 0$ as $|z| \to \infty$ in the lower half plane. This follows immediately from the above applied to $f(-z)$.

Example 22.9. Consider the integral $\int_{-\infty}^{\infty} \frac{\sin(x)}{x}dx$. This is an improper integral of an even function, thus it exists if and only if the limit of $\int_{-R}^{R} \frac{\sin(x)}{x}dx$ exists as $R \to \infty$. To compute this consider the integral along the closed curve η_R given by the concatenation $\eta_R = \nu_R \star \gamma_R$, where $\nu_R\colon [-R,R] \to \mathbb{R}$ given by $\nu_R(t) = t$ and $\gamma_R(t) = Re^{it}$ (where $t \in [0,\pi]$). Now if we let $f(z) = \frac{e^{iz}-1}{z}$, then f has a removable singularity at $z = 0$ (as is easily seen by considering the power series expansion of e^{iz}) and so is an entire function. Thus we have $\int_{\eta_R} f(z)dz = 0$ for all $R > 0$. Thus we have

$$0 = \int_{\eta_R} f(z)dz = \int_{-R}^{R} f(t)dt + \int_{\gamma_R} \frac{e^{iz}}{z}dz - \int_{\gamma_R} \frac{dz}{z}.$$

Now Jordan's lemma ensures that the second term on the right tends to zero as $R \to \infty$, while the third term integrates to $\int_0^\pi \frac{iRe^{it}}{Re^{it}} dt = i\pi$. It follows that $\int_{-R}^R f(t)dt$ tends to $i\pi$ as $R \to \infty$. and hence taking imaginary parts we conclude the improper integral $\int_{-\infty}^\infty \frac{\sin(x)}{x} dx$ is equal to π.

Remark 22.10. The function $f(z) = \frac{e^{iz}-1}{z}$ might not have been the first meromorphic function one could have thought of when presented with the previous improper integral. A more natural candidate might have been $g(z) = \frac{e^{iz}}{z}$. There is an obvious problem with this choice however, which is that it has a pole on the contour we wish to integrate around. In the case where the pole is simple (as it is for e^{iz}/z) there is standard procedure for modifying the contour: one indents it by a small circular arc around the pole. Explicitly, we replace the ν_R with $\nu_R^- \star \gamma_\epsilon \star \nu_R^+$ where $\nu_R^\pm(t) = t$ and $t \in [-R, -\epsilon]$ for ν_R^-, and $t \in [\epsilon, R]$ for ν_R^+ (and as above $\gamma_\epsilon(t) = \epsilon e^{i(\pi - t)}$ for $t \in [0, \pi]$). Since $\frac{\sin(x)}{x}$ is bounded at $x = 0$ the sum

$$\int_{-R}^{-\epsilon} \frac{\sin(x)}{x} dx + \int_\epsilon^R \frac{\sin(x)}{x} dx \to \int_{-R}^R \frac{\sin(x)}{x} dx,$$

as $\epsilon \to 0$, while the integral along γ_ϵ can be computed explicitly: by the Taylor expansion of e^{iz} we see that $\mathrm{Res}_{z=0} \frac{e^{iz}}{z} = 1$, so that $e^{iz} - 1/z$ is bounded near 0. It follows that as $\epsilon \to 0$ we have $\int_{\gamma_\epsilon} (e^{iz}/z - 1/z)dz \to 0$. On the other hand $\int_{\gamma_\epsilon} dz/z = \int_{-\pi}^0 (-\epsilon i e^{i(\pi-t)})/(e^{i(\pi-t)} dt = -i\pi$, so that we see

$$\int_{\gamma_\epsilon} \frac{e^{iz}}{z} dz \to -i\pi$$

as $\epsilon \to 0$.

Combining all of this we conclude that if $\Gamma_\epsilon = \nu_R^- \star \gamma_\epsilon \star \nu_R^+ \star \gamma_R$ then

$$\begin{aligned} 0 = \int_{\Gamma_\epsilon} f(z)dz &= \int_{-R}^{-\epsilon} \frac{e^{ix}}{x} dx + \int_{\gamma_\epsilon} \frac{e^{iz}}{z} dz + \int_\epsilon^R \frac{e^{ix}}{x} dx + \int_{\gamma_R} \frac{e^{iz}}{z} dz. \\ &= 2i \int_\epsilon^R \frac{\sin(x)}{x} + \int_{\gamma_\epsilon} \frac{e^{iz}}{z} + \int_{\gamma_R} \frac{e^{iz}}{z} dz \\ &\to 2i \int_0^R \frac{\sin(x)}{x} dx - i\pi + \int_{\gamma_R} \frac{e^{iz}}{z} dz. \end{aligned}$$

as $\epsilon \to 0$. Then letting $R \to \infty$, it follows from Jordans Lemma that the third term tends to zero so we see that

$$\int_{-\infty}^\infty \frac{\sin(x)}{x} dx = 2 \int_0^\infty \frac{\sin(x)}{x} dx = \pi$$

as required.

We record a general version of the calculation we made for the contribution of the indentation to a contour in the following Lemma.

Lemma 22.11. *Let $f\colon U \to \mathbb{C}$ be a meromorphic function with a simple pole at $a \in U$ and let $\gamma_\epsilon\colon [\alpha,\beta] \to \mathbb{C}$ be the path $\gamma_\epsilon(t) = a + \epsilon e^{it}$, then*

$$\lim_{\epsilon\to 0}\int_{\gamma_\epsilon} f(z)dz = Res_a(f).(\beta-\alpha)i.$$

Proof. Since f has a simple pole at a, we may write

$$f(z) = \frac{c}{z-a} + g(z)$$

where $g(z)$ is holomorphic near z and $c = \mathrm{Res}_a(f)$ (indeed $c/(z-a)$ is just the principal part of f at a). But now as g is holomorphic at a, it is continuous at a, and so bounded. Let $M, r > 0$ be such that $|g(z)| < M$ for all $z \in B(a,r)$. Then if $0 < \epsilon < r$ we have

$$\Big|\int_{\gamma_\epsilon} g(z)dz\Big| \leq \ell(\gamma_\epsilon)M = (\beta-\alpha)\epsilon.M,$$

which clearly tends to zero as $\epsilon \to 0$. On the other hand, we have

$$\int_{\gamma_\epsilon} \frac{c}{z-a}dz = \int_\alpha^\beta \frac{c}{\epsilon e^{it}} i\epsilon e^{it}dt = \int_\alpha^\beta (ic)dt = ic(\beta-\alpha).$$

Since $\int_{\gamma_\epsilon} f(z)dz = \int_{\gamma_\epsilon} c/(z-a)dz + \int_{\gamma_\epsilon} g(z)dz$ the result follows. ☐

22.3. **On the computation of residues and principal parts.** The previous examples will hopefully have convinced you of the power of the residue theorem. Of course for it to be useful one needs to be able to calculate the residues of functions with isolated singularities. In practice the integral formulas we have obtained for the residue are often not the best way to do this. In this section we discuss a more direct approach which is often useful when one wishes to calculate the residue of a function which is given as the ratio of two holomorphic functions.

More precisely, suppose that we have a function $F\colon U \to \mathbb{C}$ given to us as a ratio f/g of two holomorphic functions f, g on U where g is non-constant. The singularities of the function F are therefore poles which are located precisely at the (isolated) zeros of the function g, so that F is meromorphic. For convenience, we assume that we have translated the plane so as to ensure the pole of F we are interested in is at $a = 0$. Let $g(z) = \sum_{n\geq 0} c_n z^n$ be the power series for g, which will converge to $g(z)$ on any $B(0,r)$ such that $\bar{B}(0,r) \subseteq U$. Since $g(0) = 0$, and this zero is isolated, there is a $k > 0$ minimal with $c_k \neq 0$, and hence

$$g(z) = c_k z^k (1 + \sum_{n\geq 1} a_n z^n),$$

where $a_n = c_{n+k}/c_k$. Now if we let $h(z) = \sum_{n=1}^\infty a_n z^{n-1}$ then $h(z)$ is holomorphic in $B(0,r)$ – since $h(z) = (g(z) - c_k z^k)/(c_k z^{k+1})$ – and moreover

$$\frac{1}{g(z)} = \frac{1}{c_k z^k}\big(1 + zh(z)\big)^{-1},$$

Now as h is continuous, it is bounded on $\bar{B}(0,r)$, say $|h(z)| < M$ for all $z \in \bar{B}(0,r)$. But then we have, for $|z| \leq \delta = \min\{r, 1/(2M)\}$,

$$\frac{1}{g(z)} = \frac{1}{c_k z^k}\big(\sum_{n=0}^{\infty}(-1)^n z^n h(z)^n\big),$$

where by the Weierstrass M-test, the above series converges uniformly on $\bar{B}(0,\delta)$. Moreover, for any n, the series $\sum_{m\geq n}(-1)^m z^m h(z)^m$ is a holomorphic function which vanishes to order at least n at $z = 0$, so that $\frac{1}{c_k z^k}\sum_{n\geq k}(-1)^n z^n h(z)^n$ is holmorphic. It follows that the principal part of the Laurent series of $1/g(z)$ is equal to the principal part of the function

$$\frac{1}{c_k z^k}\sum_{n=1}^{k}(-1)^{k-1} z^k h(z)^k.$$

Since we know the power series for $h(z)$, this allows us to compute the principal part of $\frac{1}{g(z)}$ as claimed. Finally, the principal part $P_0(F)$ of $F = f/g$ at $z = 0$ is just the $P_0(f.P_0(g))$, the principal part of the function $f(z).P_0(g)$, which again is straight-forward to compute if we know the power series expansion of $f(z)$ at 0 (indeed we only need the first k terms of it). The best way to digest this analysis is by means of examples. We consider one next, and will examine another in the next section on summation of series.

Example 22.12. Consider $f(z) = 1/(z^2\sinh(z)^3)$. Now $\sinh(z) = (e^z - e^{-z})/2$ vanishes on $\pi i\mathbb{Z}$, and these zeros are all simple since $\frac{d}{dz}(\sinh(z)) = \cosh(z)$ has $\cosh(n\pi i) = (-1)^n \neq 0$. Thus $f(z)$ has a pole or order 5 at zero, and poles of order 3 at $\pi i n$ for each $n \in \mathbb{Z}\backslash\{0\}$. Let us calculate the principal part of f at $z = 0$ using the above technique. We will write $O(z^k)$ for the vector space of holomorphic functions which vanish to order k at 0.

$$\begin{aligned}
z^2\sinh(z)^3 &= z^2(z + \frac{z^3}{3!} + \frac{z^5}{5!} + O(z^7))^3 = z^5(1 + \frac{z^2}{3!} + \frac{z^4}{5!} + O(z^6))^3\\
&= z^5(1 + \frac{3z^2}{3!} + \frac{3z^4}{(3!)^2} + \frac{3z^4}{5!} + O(z^6))\\
&= z^5(1 + \frac{z^2}{2} + \frac{13z^4}{120} + O(z^6))\\
&= z^5\left(1 + z(\frac{z}{2} + \frac{13z^3}{120} + O(z^5))\right)
\end{aligned}$$

Thus, in the notation of the above discussion, $h(z) = \frac{z}{2} + \frac{13z^3}{120} + O(z^5)$, and so, as h vanishes to first order at $z = 0$, in order to obtain the principal part we just need to consider the first two terms in the geometric series

$(1+zh(z))^{-1} = \sum_{n=0}^{\infty}(-1)^n z^n h(z)^n$:

$$\begin{aligned}
1/z^2\sinh(z)^3 &= z^{-5}\big(1+z(\frac{z}{2}+\frac{13z^3}{120}+O(z^5))\big)^{-1} \\
&= z^{-5}\big(1-z(\frac{z}{2}+\frac{13z^3}{120})+z^2\frac{z^2}{(2!)^2}+O(z^5)\big) \\
&= z^{-5}\big(1-\frac{z^2}{2}+(\frac{1}{4}-\frac{13}{120})z^4+O(z^5)\big) \\
&= \frac{1}{z^5}-\frac{1}{2z^3}+\frac{17}{120z}+O(z).
\end{aligned}$$

Thus the principal part of $f(z)$ at 0 is $P_0(f) = \frac{1}{z^5}-\frac{1}{2z^3}+\frac{17}{120z}$, and $\mathrm{Res}_0(f) = 17/120$.

There are other variants on the above method which we could have used: For example, by the binomial theorem for an arbitrary exponent we know that if $|z|<1$ then $(1+z)^{-3} = \sum_{n\geq 0}\binom{-3}{n}z^n = 1-3z+6z^2+\ldots$. Arguing as above, it follows that for small enough z we have

$$\begin{aligned}
\sinh(z)^{-3} &= z^{-3}.(1+\frac{z^2}{3!}+\frac{z^4}{5!}+O(z^6))^{-3} \\
&= z^{-3}\left(1+(-3)\big(\frac{z^2}{3!}+\frac{z^4}{5!}\big)+6\big(\frac{z^2}{3!}+\frac{z^4}{5!}\big)^2+O(z^6)\right) \\
&= z^{-3}\left(1-\frac{z^2}{2}+\big(\frac{-3}{5!}+\frac{6}{(3!)^2}\big)z^4+O(z^6)\right) \\
&= z^{-3}\left(1-\frac{z^2}{2}+\frac{17z^4}{120}+O(z^6)\right)
\end{aligned}$$

yielding the same result for the principal part of $1/z^2\sinh(z)^3$.

22.4. **Summation of infinite series.** Residue calculus can also be a useful tool in calculating infinite sums, as we now show. For this we use the function $f(z) = \cot(\pi z)$. Note that since $\sin(\pi z)$ vanishes precisely at the integers, $f(z)$ is meromorphic with poles at each integer $n \in \mathbb{Z}$. Moreover, since f is periodic with period 1, in order to understand the poles of f it suffices to calculate the principal part of f at $z=0$. We can use the method of the previous section to do this:

We have $\sin(z) = z - \frac{z^3}{3!}+\frac{z^5}{5!}+O(z^7)$, so that $\sin(z)$ vanishes with multiplicity 1 at $z=0$ and we may write $\sin(z) = z(1-zh(z))$ where $h(z) = z/3! - z^3/5! + O(z^5)$ is holomorphic at $z=0$. Then

$$\frac{1}{\sin(z)} = \frac{1}{z}(1-zh(z))^{-1} = \frac{1}{z}\big(1+\sum_{n\geq 1}z^n h(z)^n\big) = \frac{1}{z}+h(z)+O(z^2).$$

Multiplying by $\cos(z)$ we see that the principal part of $\cot(z)$ is the same as that of $\frac{1}{z}\cos(z)$ which, using the Taylor expansion of $\cos(z)$, is clearly $\frac{1}{z}$ again. By periodicity, it follows that $\cot(\pi z)$ has a simple pole with residue $1/\pi$ at each integer $n \in \mathbb{Z}$.

We can also use this strategyto find further terms of the Laurent series of $\cot(z)$: Since our $h(z)$ actually vanishes at $z = 0$, the terms $h(z)^n z^n$ vanish to order $2n$. It follows that we obtain all the terms of the Laurent series of $\cot(z)$ at 0 up to order 3, say, just by considering the first two terms of the series $1 + \sum_{n \geq 1} z^n h(z)^n$, that is, $1 + zh(z)$. Since $\cos(z) = 1 - z^2/2! + z^4/4!$, it follows that $\cot(z)$ has a Laurent series

$$\begin{aligned} \cot(z) &= (1 - \frac{z^2}{2!} + O(z^4)).(\frac{1}{z} + (\frac{z}{3!} - \frac{z^3}{5!} + O(z^5))) \\ &= \frac{1}{z} - \frac{z}{3} + O(z^3) \end{aligned}$$

The fact that $f(z)$ has simple poles at each integer will allow us to sum infinite series with the help of the following:

Lemma 22.13. *Let $f(z) = \cot(\pi z)$ and let Γ_N denotes the square path with vertices $(N + 1/2)(\pm 1 \pm i)$. There is a constant C independent of N such that $|f(z)| \leq C$ for all $z \in \Gamma_N^*$.*

Proof. We need to consider the horizontal and vertical sides of the square separately. Note that $\cot(\pi z) = (e^{i\pi z} + e^{-i\pi z})/(e^{i\pi z} - e^{-i\pi z})$. Thus on the horizontal sides of Γ_N where $z = x \pm (N + 1/2)i$ and $-(N + 1/2) \leq x \leq (N + 1/2)$ we have

$$\begin{aligned} |\cot(\pi z)| &= \left| \frac{e^{i\pi(x \pm (N+1/2)i)} + e^{-i\pi(x \pm (N+1/2)i)}}{e^{i\pi(x \pm (N+1/2)i} - e^{-i\pi(x \pm (N+1/2)i)}} \right| \\ &\leq \frac{e^{\pi(N+1/2)} + e^{-\pi(N+1/2)}}{e^{\pi(N+1/2)} - e^{-\pi(N+1/2)}} \\ &= \coth(\pi(N + 1/2)). \end{aligned}$$

Now since $\coth(x)$ is a decreasing function for $x \geq 0$ it follows that on the horizontal sides of Γ_N we have $|\cot(\pi z)| \leq \coth(3\pi/2)$.

On the vertical sides we have $z = \pm(N + 1/2) + iy$, where $-N - 1/2 \leq y \leq N + 1/2$. Observing that $\cot(z + N\pi) = \cot(z)$ for any integer N and that $\cot(z + \pi/2) = -\tan(z)$, we find that if $z = \pm(N + 1/2) + iy$ for any $y \in \mathbb{R}$ then

$$|\cot(\pi z)| = |-\tan(iy)| = |-\tanh(y)| \leq 1.$$

Thus we may set $C = \max\{1, \coth(3\pi/2)\}$. $\square$

We now show how this can be used to sum an infinite series:

Example 22.14. Let $g(z) = \cot(\pi z)/z^2$. By our discussion of the poles of $\cot(\pi z)$ above it follows that $g(z)$ has simple poles with residues $\frac{1}{\pi n^2}$ at each non-zero integer n and residue $-\pi/3$ at $z = 0$.

Consider now the integral of $g(z)$ around the paths Γ_N: By Lemma 22.13 we know $|g(z)| \leq C/|z|^2$ for $z \in \Gamma_N^*$, and for all $N \geq 1$. Thus by the estimation lemma we see that

$$\left(\int_{\Gamma_N} g(z)dz\right) \leq C.(4N+2)/(N+1/2)^2 \to 0,$$

as $N \to \infty$. But by the residue theorem we know that

$$\int_{\Gamma_N} g(z)dz = -\pi/3 + \sum_{\substack{n\neq 0,\\ -N\leq n\leq N}} \frac{1}{\pi n^2}.$$

It therefore follows that

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \pi^2/6$$

Remark 22.15. Notice that the contours Γ_N and the function $\cot(\pi z)$ clearly allows us to sum other infinite series in a similar way – for example if we wished to calculate the sum of the infinite series $\sum_{n\geq 1}\frac{1}{n^2+1}$ then we would consider the integrals of $g(z) = \cot(\pi z)/(1+z^2)$ over the contours Γ_N.

Remark 22.16. (*Non-examinable – for interest only!*): Note that taking $g(z) = (1/z^{2k})\cot(\pi z)$ for any positive integer k, the above strategy gives a method for computing $\sum_{n=1}^{\infty} 1/n^{2k}$ (check that you see why we need to take even powers of n). The analysis for the case $k=1$ goes through in general, we just need to compute more and more of the Laurent series of $\cot(\pi z)$ the larger we take k to be.

One can show that $\zeta(s) = \sum_{n=1}^{\infty} 1/n^s$ converges to a holomorphic function of s for any $s \in \mathbb{C}$ with $\Re(s) > 1$ (as usual, we define $n^s = \exp(s.\log(n))$ where $\log$ is the ordinary real logarithm). As $s \to 1$ it can be checked that $\zeta(s) \to \infty$, however it can be shown that $\zeta(s)$ extends to a meromorphic function on all of $\mathbb{C}\backslash\{1\}$. The identity theorem shows that this extension is unique if it exists[50]. (This uniqueness is known as the principle of "analytic continuation".) The location of the zeros of the ζ-function is the famous Riemann hypothesis: apart from the "trivial zeros" at negative even integers, they are conjectured to all lie on the line $\Re(z) = 1/2$. Its values at special points however are also of interest: Euler was the first to calculate $\zeta(2k)$ for positive integers k, but the values $\zeta(2k+1)$ (for k a positive integer) remain mysterious – it was only shown in 1978 by Roger Apéry that $\zeta(3)$ is irrational for example. Our analysis above is sufficient to determine $\zeta(2k)$ once one succeeds in computing explicitly the Laurent series for $\cot(\pi z)$ or equivalently the Taylor series of $z\cot(\pi z) = iz + 2iz/(e^{2iz}-1)$. See Appendix IV for more details.

[50]It is this uniqueness and the fact that one can readily compute that $\zeta(-1) = -1/12$ that results in the rather outrageous formula $\sum_{n=1}^{\infty} n = -1/12$.

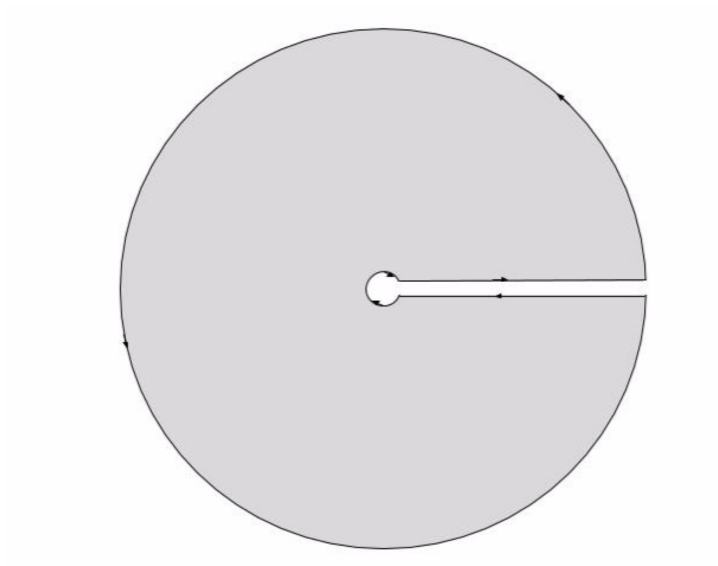

FIGURE 5. A keyhole contour.

22.5. **Keyhole contours.** There are many ingenious paths which can be used to calculate integrals via residue theory. One common contour is known (for obvious reasons) as a *keyhole contour*. It is constructed from two circular paths of radius ϵ and R, where we let R become arbitrarily large, and ϵ arbitrarily small, and we join the two circles by line segments with a narrow neck in between. Explicitly, if $0 < \epsilon < R$ are given, pick a $\delta > 0$ small, and set $\eta_+(t) = t + i\delta$, $\eta_-(t) = (R-t) - i\delta$, where in each case t runs over the closed intervals with endpoints such that the endpoints of $\eta_\pm$ lie on the circles of radius ϵ and R about the origin. Let γ_R be the positively oriented path on the circle of radius R joining the endpoints of η_+ and η_- on that circle (thus traversing the "long" arc of the circle between the two points) and similarly let γ_ϵ the path on the circle of radius ϵ which is negatively oriented and joins the endpoints of $\gamma_\pm$ on the circle of radius ϵ. Then we set $\Gamma_{R,\epsilon} = \eta_+ \star \gamma_R \star \eta_- \star \gamma_\epsilon$ (see Figure 5). The keyhole contour can sometimes be useful to evaluate real integrals where the integrand is multivalued as a function on the complex plane, as the next example shows:

Example 22.17. Consider the integral $\int_0^\infty \frac{x^{1/2}}{1+x^2}dx$. Let $f(z) = z^{1/2}/(1+z^2)$, where we use the branch of the square root function which is continuous on $\mathbb{C}\backslash\mathbb{R}_{>0}$, that is, if $z = re^{it}$ with $t \in [0, 2\pi)$ then $z^{1/2} = r^{1/2}e^{it/2}$.

We use the keyhole contour $\Gamma_{R,\epsilon}$. On the circle of radius R, we have $|f(z)| \leq R^{1/2}/(R^2-1)$, so by the estimation lemma, this contribution to the integral of f over $\Gamma_{R,\epsilon}$ tends to zero as $R \to \infty$. Similarly, $|f(z)|$ is bounded by $\epsilon^{1/2}/(1-\epsilon^2)$ on the circle of radius ϵ, thus again by the estimation lemma this contribution to the integral of f over $\Gamma_{R,\epsilon}$ tends to zero as

$\epsilon \to 0$. Finally, the discontinuity of our branch of $z^{1/2}$ on $\mathbb{R}_{>0}$ ensures that the contributions of the two line segments of the contour do not cancel but rather both tend to $\int_0^\infty \frac{x^{1/2}}{1+x^2} dx$ as δ and ϵ tend to zero.

To compute $\int_0^\infty \frac{x^{1/2}}{1+x^2} dx$ we evaluate the integral $\int_{\Gamma_{R,\epsilon}} f(z)dz$ using the residue theorem: The function $f(z)$ clearly has simple poles at $z = \pm i$, and their residues are $\frac{1}{2}e^{-\pi i/4}$ and $\frac{1}{2}e^{5\pi i/4}$ respectively. It follows that

$$\int_{\Gamma_{R,\epsilon}} f(z)dz = 2\pi i \left(\frac{1}{2}e^{-\pi i/4} + \frac{1}{2}e^{5\pi i/4} \right) = \pi\sqrt{2}.$$

Taking the limit as $R \to \infty$ and $\epsilon \to 0$ we see that $2\int_0^\infty \frac{x^{1/2}}{1+x^2} dx = \pi\sqrt{2}$, so that

$$\int_0^\infty \frac{x^{1/2}dx}{1+x^2} = \frac{\pi}{\sqrt{2}}.$$

23. 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 23.1. 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 $\gamma(t) \in \mathbb{C}$.

Remark 23.2. Note that this definition gives us a notion of tangent vectors at points on subsets of $\mathbb{R}^n$, since the notion of a C^1 path extends readily to paths in $\mathbb{R}^n$ (we just require all n component functions are continuously differentiable). In particular, if $\mathbb{S}$ is the unit sphere in $\mathbb{R}^3$ as above, a C^1 path on $\mathbb{S}$ is simply a path $\gamma\colon [a,b] \to \mathbb{R}^3$ whose image lies in $\mathbb{S}$. It is easy to check that the tangent vectors at a point $p \in \mathbb{S}$ all lie in the plane perpendicular to p – 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 23.3. 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 23.4. *Let $f: U \to \mathbb{C}$ 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 $f: U \to \mathbb{C}$ is has nonvanishing derivative on all of U, it is conformal on all of U (and locally a biholomorphism).*

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 $\eta_2(t) = f(\gamma_2(t))$. By the Chain Rule (see Lemma 26.7) we see that $\eta_1'(t) = Df_{z_0}(\gamma_1'(t))$ and $\eta_2'(t) = Df_{z_0}(\gamma_2'(t))$, and moreover if $f'(z_0) = \rho.e^{i\theta}$, then

$$Df_{z_0} = \rho. \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix},$$

(since the linear map given by multiplication by $f'(z_0)$ is precisely scaling by ρ and rotating by θ). It follows that 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.

For the final part, note that if $f'(z_0) \neq 0$ then by the definition of the degree of vanishing, the function $f(z)$ is locally biholomorphic (see the proof of the inverse function theorem). □

Example 23.5. 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 non-zero 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 23.6. *The sterographic projection map $S: \mathbb{C} \to \mathbb{S}$ is conformal.*

Proof. 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[51] 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

[51] with domain $[-1, 1]$ say – or even the whole real line, except that it is non-compact.

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. $\square$

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

Lemma 23.7. *Mobius transformations are conformal.*

Proof. As we have already shown, any holomorphic map is conformal wherever its derivative is nonzero. Since a Mobius transformation f is invertible everywhere with holomorphic inverse, its derivative must be nonzero everywhere and we are done.

One can also give a more explicit proof: If $f(z) = \frac{az+b}{cz+d}$ then it is easy to check that

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

for all $z \neq -d/c$, thus f is conformal at each $z \in \mathbb{C}\backslash\{-d/c\}$. Checking at $z = \infty, -d/c$ is similar: indeed at $\infty = [1 : 0]$ we use the map $i_\infty \colon \mathbb{C} \to \mathbb{P}^1$ given by $w \mapsto [1 : w]$ to obtain $f_\infty(w) = \frac{a+bw}{c+dw}$ and $f'_\infty(w) = \frac{bc-ad}{(c+dw)^2}$, which is certainly nonzero at $w = 0$ (and $i_\infty(0) = \infty$). $\square$

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

Proposition 23.8. *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 Mobius 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 Mobius 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 Mobius 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 Mobius 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 Mobius 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 Mobius 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 $T(z) = a/d.z$, and since $T(1) = 1$ it follows $a/d = 1$ and hence $T(z) = z$. Thus we see that $hf_1g^{-1} = hf_2g^{-1} =$ id are all the identity, and so $f_1 = f_2$ as required. □

Example 23.9. The above lemma shows that we can use Mobius 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 Mobius 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 Mobius transformations allow us to take any three points to any other three points. Thus if we take f the Mobius 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 23.8).

So far, we have found a Mobius 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 Mobius 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.

Note that if we had taken $g(z) = (z + i)/(iz + 1)$ for example, then g also maps $\mathbb{R}$ to the unit circle $\mathbb{S}^1$, but $g(-i) = 0$, so[52] g maps the lower half plane to $B(0, 1)$. If we had used this transformation, then it would be easy to "correct" it to get what we wanted: In fact there are (at least) two simple things one could do: First, one could note that the map $R(z) = -z$ (a rotation by π) sends the upper half plane to the lower half place, so that the composition $g \circ R$ is a Mobius transformation taking $\mathbb{H}$ to $B(0, 1)$. Alternatively, the inversion $j(z) = 1/z$ sends $\mathbb{C}\backslash\bar{B}(0, 1)$ to $B(0, 1)$, so that $j \circ g$ also sends $\mathbb{H}$ to $B(0, 1)$. Explicitly, we have

$$g \circ R(z) = \frac{z - i}{iz - 1} = \frac{-i(iz + 1)}{i(z + i)} = -f(z), \quad j \circ g(z) = \frac{iz + 1}{z + i} = f(z).$$

[52]A Mobius map is a continuous function on $\mathbb{C}_\infty$, and if we remove a circle from $\mathbb{C}_\infty$ the complement is a disjoint union of two connected components, just the same as when we remove a line or a circle from the plane, thus the connectedness argument works just as well when we include the point at infinity.

Note in particular that f is far from unique – indeed if f is any Mobius transformation which takes $\mathbb{H}$ to $B(0,1)$ then composing it with any Mobius transformation which preserves $B(0,1)$ will give another such map. Thus for example $e^{i\theta}.f$ will be another such transformation.

Exercise 23.10. Every Mobius transformation gives a biholomorphic map from $\mathbb{C}_\infty$ to itself, but they may not preserve the distance function d_S on $\mathbb{P}^1$. What is the subgroup of Mob which are isometries of $\mathbb{P}^1$ with respect to the distance function d_S?

Given two domains D_1, D_2 in the complex plane, one can ask if there is a conformal transformation $f\colon D_1 \to D_2$. Since a conformal transformation is in particular a homeomorphism, this is clearly not possible for completely arbitrary domains. However if we restrict to simply-connected domains (that is, domains in which any path can be continuously deformed to any other path with the same end-points), the following remarkable theorem shows that the answer to this question is yes! Since it will play a distinguished role later, we will write $\mathbb{D}$ for the unit disc $B(0,1)$.

Theorem 23.11. *(Riemann's mapping theorem): Let U be an open connected and simply-connected proper subset of $\mathbb{C}$. Then for any $z_0 \in U$ there is a unique bijective conformal transformation $f\colon U \to \mathbb{D}$ such that $f(z_0) = 0$, $f'(z_0) > 0$.*

Remark 23.12. The proof of this theorem is beyond the scope of this course, but it is a beautiful and fundamental result. The proof in fact only uses the fact that on a simply-connected domain any holomorphic function has a primitive, and hence it in fact shows that such domains are simply-connected in the topological sense (since a conformal transformation is in particular a homeomorphism, and the disc in simply-connected). It relies crucially on *Montel's theorem* on families of holomorphic functions, see for example the text of Shakarchi and Stein[53] for an exposition of the argument.

Note that it follows immediately from Liouville's theorem that there can be no bijective conformal transformation taking $\mathbb{C}$ to $B(0,1)$, so the whole complex plane is indeed an exception. The uniqueness statement of the theorem reduces to the question of understanding the conformal transformations of the disk $\mathbb{D}$ to itself.

Of course knowing that a conformal transformation between two domains D_1 and D_2 exists still leaves the challenge of constructing one. As we will see in the next section on harmonic maps, this is an important question. In simple cases one can often do so by hand, as we now show.

In addition to Mobius transformations, it is often useful to use the exponential function and branches of the multifunction $[z^\alpha]$ (away from the origin) when constructing conformal maps. We give an example of the kind of constructions one can do:

[53]*Complex Analysis*, Princeton Lecture in Analysis II, E. M. Stein & R. Shakarchi. P.U.P.

Example 23.13. Let $D_1 = B(0,1)$ and $D_2 = \{z \in \mathbb{C} : |z| < 1, \Im(z) > 0\}$. Since these domains are both convex, they are simply-connected, so Riemann's mapping theorem ensure that there is a conformal map sending D_2 to D_1. To construct such a map, note that the domain is defined by the two curves $\gamma(0,1)$ and the real axis. It can be convenient to map the two points of intersection of these curves, ± 1 to 0 and ∞. We can readily do this with a Mobius transformation:

$$f(z) = \frac{z-1}{z+1},$$

Now since f is a Mobius transformation, it follows that $f_1(\mathbb{R})$ and $f_1(\gamma(0,1))$ are lines (since they contain ∞) passing through the origin. Indeed $f(\mathbb{R}) = \mathbb{R}$, and since f had inverse $f^{-1} = \frac{z+1}{z-1}$ it follows that the image of $\gamma(0,1)$ is $\{w \in \mathbb{C} : |w-1| = |w+1|\}$, that is, the imaginary axis. Since $f(i/2) = (-3+4i)/5$ it follows by connectedness that $f(D_1)$ is the second quadrant $Q = \{w \in \mathbb{C} : \Re(z) < 0, \Im(z) > 0\}$.

Now the squaring map $s\colon \mathbb{C} \to \mathbb{C}$ given by $z \mapsto z^2$ maps Q bijectively to the half-plane $H = \{w \in \mathbb{C} : \Im(w) < 0\}$, and is conformal except at $z = 0$ (which is on the boundary, not in the interior, of Q). We may then use a Mobius map to take this half-plane to the unit disc: indeed in Example 23.9 we have already seen that the Mobius transformation $g(z) = \frac{z+i}{iz+1}$ takes the lower-half plane to the upper-half plane.

Putting everything together, we see that $F = g \circ s \circ f$ is a conformal transformation taking D_1 to D_2 as required. Calculating explicitly we find that

$$F(z) = i\left(\frac{z^2 + 2iz + 1}{z^2 - 2iz + 1}\right)$$

Remark 23.14. Note that there are couple of general principles one should keep in mind when constructing conformal transformations between two domains D_1 and D_2. Often if the boundary of D_1 has distinguished points (such as ± 1 in the above example) it is convenient to move these to "standard" points such as 0 and ∞, which one can do with a Mobius transformation. The fact that Mobius transformations are three-transitive and takes lines and circles to lines and circles and moreover act transitively on such means that we can always use Mobius transformations to match up those parts of the boundary of D_1 and D_2 given by line segments or arcs of circles. However these will not be sufficient in general: indeed in the above example, the fact that the boundary of D_1 is a union of a semicircle and a line segment, while that of D_2 is just a circle implies there is no Mobius transformation taking D_1 to D_2, as it would have to take ∂D_1 to ∂D_2, which would mean that its inverse would not take the unit circle to either a line or a circle. Branches of fractional power maps $[z^\alpha]$ are often useful as they allow us to change the angle at the points of intersection of arcs of the boundary (being conformal on the interior of the domain but not on its boundary).

23.1. **Conformal transformations and the Laplace equation.** In this section we will use the term *conformal map* or *conformal transformation* somewhat abusively to mean a holomorphic function whose derivative is nowhere vanishing on its domain of definition. (We have seen already that this implies the function is conformal in the sense of the previous section.) If there is a bijective conformal transformation between two domains U and V we say they are *conformally equivalent*.

Recall that a function $v\colon \mathbb{R}^2 \to \mathbb{R}$ is said to be *harmonic* if it is twice differentiable and $\partial_x^2 v + \partial_y^2 v = 0$. Often one seeks to find solutions to this equation on a domain $U \subset \mathbb{R}^2$ where we specify the values of v on the boundary ∂U of U. This problem is known as the *Dirichlet problem*, and makes sense in any dimension (using the appropriate Laplacian). In dimension 2, complex analysis and in particular conformal maps are a powerful tool by which one can study this problem, as the following lemma show.

Lemma 23.15. *Suppose that $U \subset \mathbb{C}$ is a simply-connected open subset of $\mathbb{C}$ and $v\colon U \to \mathbb{R}$ is twice continuously differentiable and harmonic. Then there is a holomorphic function $f\colon U \to \mathbb{C}$ such that $\Re(f) = v$. In particular, any such function v is analytic.*

Proof. (*Sketch*): Consider the function $g(z) = \partial_x v - i\partial_y v$. Then since v is twice continuously differentiable, the partial derivatives of g are continuous and

$$\partial_x^2 v = -\partial_y^2 v; \quad \partial_y \partial_x v = \partial_x \partial_y v,$$

so that g satisfies the Cauchy-Riemann equations. It follows from Theorem 14.9 that g is holomorphic. Now since U is simply-connected, it follows that g has a primitive $G\colon U \to \mathbb{C}$. But then it follows that if $G = a(z) + ib(z)$ we have $\partial_z G = \partial_x a - i\partial_y a = g(z) = \partial_x v - i\partial_y v$, hence the partial derivatives of a and v agree on all of U. But then if $z_0, z \in U$ and γ is a path between then, the chain rule[54] shows that

$$\begin{aligned}\int_\gamma (\partial_x v + i\partial_y v)dz &= \int_0^1 (\partial_x(v(\gamma(t)) + i\partial_y v(\gamma(t)))\gamma'(t)dt \\ &= \int_0^1 \frac{d}{dt}(v(\gamma(t)))dt = v(z) - v(z_0),\end{aligned}$$

Similarly, we see that the same path integral is also equal to $a(z) - a(z_0)$. It follows that $a(z) = v(z) + (a(z_0) - v(z_0))$, thus if we set $f(z) = G(z) - (G(z_0) - v(z_0))$ we obtain a holomorphic function on U whose real part is equal to v as required.

Since we know that any holomorphic function is analytic, it follows that v is analytic (and in particular, infinitely differentiable). □

[54]This uses the chain rule for a composition $g \circ f$ of real-differentiable functions $f\colon \mathbb{R} \to \mathbb{R}^2$ and $g\colon \mathbb{R}^2 \to \mathbb{R}$, applied to the real and imaginary parts of the integrand. This follows in exactly the same way as the proof of Lemma 26.7. See the remark after the proof of that lemma.

The previous Lemma shows that, at least locally (in a disk say) harmonic functions and holomorphic functions are in correspondence – given a holomorphic function f we obtain a harmonic function by taking its real part, while if u is harmonic the previous lemma shows we can associate to it a holomorphic function f whose real part equals u (and in fact examining the proof, we see that f is actually unique up to a purely imaginary constant). Thus if we are seeking a harmonic function on an open set U whose values are a given function g on ∂U, then it suffices to find a holomorphic function f on U such that $\Re(f) = g$ on the boundary ∂U.

Now if $H\colon U \to V$ was a bijective conformal transformation which extends to a homeomorphism $\bar{H}\colon \bar{U} \to \bar{V}$ which thus takes ∂U homeomorphically to ∂V, then if $f\colon V \to \mathbb{C}$ is holomorphic, so is $f \circ H$. Thus in particular $\Re(f \circ H)$ is a harmonic function on U. It follows that we can use conformal transformations to transport solutions of Laplace's equation from one domain to another: if we can use a conformal transformation H to take a domain U to a domain V where we already have a supply of holomorphic functions satisfying various boundary conditions, the conformal transformation H gives us a corresponding set of holomorphic (and hence harmonic) functions on U. We state this a bit more formally as follow:

Lemma 23.16. *If U and V are domains and $G\colon U \to V$ is a conformal transformation, then if $u\colon V \to \mathbb{R}$ is a harmonic function on V, the composition $u \circ G$ is harmonic on U.*

Proof. To see that $u \circ G$ is harmonic we need only check this in a disk $B(z_0, r) \subseteq U$ about any point $z_0 \in U$. If $w_0 = G(z_0)$, the continuity of G ensures we can find $\delta, \epsilon > 0$ such that $G(B(z_0, \delta)) \subseteq B(w_0, \epsilon) \subseteq V$. But now since $B(w_0, \epsilon)$ is simply-connected we know by Lemma 23.15 we can find a holomorphic function $f(z)$ with $u = \Re(f)$. But then on $B(z_0, \delta)$ we have $u \circ G = \Re(f \circ G)$, and by the chain rule $f \circ G$ is holomorphic, so that its real part is harmonic as required. $\square$

Remark 23.17. You can also give a more direct computational proof of the above Lemma. Note also that we only need G to be holomorphic – the fact that it is a conformal equivalence is not necessary. On the other hand if we are trying to produce harmonic functions with prescribed boundary values, then we will need to use carefully chosen conformal transformations.

This strategy for studying harmonic functions might appear over-optimistic, in that the domains one can obtain from a simple open set like $B(0, 1)$ or the upper-half plane $\mathbb{H}$ might consist of only a small subset of the open sets one might be interested in. However, the Riemann mapping theorem (Theorem 23.11) show that *every* domain which is simply connected, other than the whole complex plane itself, is in fact conformally equivalent to $B(0, 1)$. Thus a solution to the Dirichlet problem for the disk at least in principal comes close[55] to solving the same problem for any simply-connected domain! For convenience, we will write $\mathbb{D}$ for the open disk $B(0,1)$ of radius 1 centred at 0.

In the course so far, the main examples of conformal transformations we have are the following:

(1) The exponential function is conformal everywhere, since it is its own derivative and it is everywhere nonzero.
(2) Mobius transformations understood as maps on the extended complex plane are everywhere conformal.
(3) Fractional exponents: In cut planes the functions $z \mapsto z^\alpha$ for $\alpha \in \mathbb{C}$ are conformal (the cut removes the origin, where the derivative may vanish).

Let us see how to use these transformations to obtain solutions of the Laplace equation. First notice that Cauchy's integral formula suggests a way to produce solutions to Laplace's equation in the disk: Suppose that u is a harmonic function defined on $B(0,r)$ for some $r > 1$. Then by Lemma 23.15 we know there is a holomorphic function $f \colon B(0,r) \to \mathbb{C}$ such that $u = \Re(f)$. By Cauchy's integral formula, if γ is a parametrization of the positively oriented unit circle, then for all $w \in B(0,1)$ we have $f(w) = \frac{1}{2\pi i}\int_\gamma f(z)/(z-w)dz$, and so

$$u(z) = \Re\big(\frac{1}{2\pi i}\int_\gamma \frac{f(z)dz}{z-w}\big).$$

Since the integrand uses only the values of f on the boundary circle, we have almost recovered the function u from its values on the boundary. (Almost, because we appear to need the values of it harmonic conjugate). The next lemma resolves this:

Lemma 23.18. *If u is harmonic on $B(0,r)$ for $r > 1$ then for all $w \in B(0,1)$ we have*

$$u(w) = \frac{1}{2\pi}\int_0^{2\pi} f(e^{i\theta})\frac{1-|w|^2}{|e^{i\theta}-w|^2}d\theta = \frac{1}{2\pi}\int_0^{2\pi} u(e^{i\theta})\Re\big(\frac{e^{i\theta}+w}{e^{i\theta}-w}\big)d\theta.$$

Proof. (*Sketch.*) Take, as before, $f(z)$ holomorphic with $\Re(f) = u$ on $B(0,r)$. Then letting γ be a parametrization of the positively oriented unit circle we have

$$f(w) = \frac{1}{2\pi i}\int_\gamma \frac{f(z)dz}{z-w} - \frac{1}{2\pi i}\int_\gamma \frac{f(z)dz}{z-\bar{w}^{-1}}$$

where the first term is $f(w)$ by the integral formula and the second term is zero because $f(z)/(z-\bar{w}^{-1})$ is holomorphic inside all of $B(0,1)$. Gathering the terms, this becomes

$$f(w) = \frac{1}{2\pi}\int_\gamma f(z)\frac{1-|w|^2}{|z-w|^2}\frac{dz}{iz} = \frac{1}{2\pi}\int_0^{2\pi} f(e^{i\theta})\frac{1-|w|^2}{|e^{i\theta}-w|^2}d\theta.$$

[55]The issue is whether the conformal equivalence behaves well enough at the boundaries.

The advantage of this last form is that the real and imaginary parts are now easy to extract, and we see that

$$u(z) = \int_0^{2\pi} u(e^{i\theta}) \frac{1-|w|^2}{|e^{i\theta}-w|^2} d\theta.$$

Finally for the second integral expression note that if $|z| = 1$ then

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

from which one readily sees the real part agrees with the corresponding factor in our first expression. □

Now the idea to solve the Dirichlet problem for the disk $B(0,1)$ is to turn this previous result on its head: Notice that it tells us the values of u inside the disk $B(0,1)$ in terms of the values of u on the boundary. Thus if we are given the boundary values, say a (periodic) function $G(e^{i\theta})$ we might reasonably hope that the integral

$$g(w) = \frac{1}{2\pi} \int_0^{2\pi} G(e^{i\theta}) \frac{1-|w|^2}{|e^{i\theta}-w|^2} d\theta,$$

would define a harmonic function with the required boundary values. Indeed it follows from the proof of the lemma that the integral is the real part of the integral

$$\frac{1}{2\pi i} \int_\gamma G(z) \frac{1}{z-w} dz,$$

which we know from Proposition 17.7 is holomorphic in w, thus $g(w)$ is certainly harmonic. It turns out that if $w \to w_0 \in \partial B(0,1)$ then provided G is continuous at w_0 then $g(w) \to G(w_0)$, hence g is in fact a harmonic function with the required boundary value.

24. APPENDIX I: SET THEORY

In this appendix we review some elementary set theory.

If X is a set and $\{A_i : i \in I\}$ is a collection of subsets of X indexed by a set I, then we define

$$\bigcup_{i \in I} A_i = \{x \in X : \exists i \in I \text{ such that } x \in A_i\};$$
$$\bigcap_{i \in I} A_i = \{x \in X : \forall i \in I, x \in A_i\}.$$

Note that in particular this implies that if I is the empty set, then the empty intersection is X, while the empty union is the empty set.

If $A \subseteq X$ then we write A^c for the complement of A in X, that is, $A^c = \{x \in X; x \notin A\}$. *De Morgan's Laws* state that

$$\Big(\bigcup_{i \in I} A_i\Big)^c = \bigcap_{i \in I} A_i^c; \quad \Big(\bigcap_{i \in I} A_i\Big)^c = \bigcup_{i \in I} A_i^c.$$

If $: X \to Y$ is any function, then there is an induced map $f^{-1} \colon \mathcal{P}(Y) \to \mathcal{P}(X)$, where $\mathcal{P}(X)$ denotes the power set of X, that is, the set of all subsets of X. If $A \subseteq Y$ then $f^{-1}(A) = \{x \in X : f(x) \in A\}$. We call $f^{-1}(A)$ the *preimage* of A in X. The preimage map f^{-1} is compatible with unions and intersections: If $\{C_i : i \in I\}$ is a collection of subsets of Y then

$$f^{-1}\left(\bigcup_{i \in I} C_i\right) = \bigcup_{i \in I} f^{-1}(C_i); \quad f^{-1}\left(\bigcap_{i \in I} C_i\right) = \bigcap_{i \in I} f^{-1}(C_i).$$

25. Appendix II: On the connected subsets of $\mathbb{R}$

In this appendix we give an alternative approach to the classification of connected subsets of $\mathbb{R}$:

Definition 25.1. Let $E \subseteq \mathbb{R}$ be a subset of the real line. We say that E has *property I* if, whenever $x < y$ both lie in E, we have $[x, y] \subseteq I$.

Proposition 25.2. *A subset $E \subseteq \mathbb{R}$ of the real line is connected if and only if it has property I.*

Proof. First suppose that E is connected and that $x, y \in E$. By symmetry we may assume that $x < y$. If $[x, y]$ is not entirely contained in E, we may find $c \in (x, y)$ such that $c \notin E$. But then $E \subseteq (-\infty, c) \cup (c, \infty)$ and $x \in E \cap (-\infty, c)$ and $y \in E \cap (c, \infty)$ so that E is not contained entirely in one or other of the disjoint open sets $(-\infty, c)$, (c, ∞). Thus E is disconnected.

Next suppose that E has property I, and suppose $E \subseteq U \cup V$ where U and V are ope subsets of $\mathbb{R}$ with $E \cap U \cap V = \emptyset$, and, for the sake of a contradiction, that $E \cap U$ and $E \cap V$ are both non-empty. Then we may pick $x \in E \cap U$ and $y \in E \cap V$, and by symmetry assume that $x < y$. Since E has property I, the interval $[x, y]$ is entirely contained in E.

Now as $[x, y]$ is bounded and $x \in U$, if we let $S = \{z \in [x, y] : z \in U\}$, then S is non-empty and bounded and so $c = \sup(S)$ exists, and clearly $c \in [x, y]$. If $c \in U$ then $c \neq y$ and. as U is open, there is some $\epsilon_1 > 0$ such that $B(c, \epsilon_1) \subseteq U$. Thus if we set $\delta = \min\{\epsilon_1/2, (y - c)/2\} > 0$ we have $c + \delta \in U \cap [x, y]$ contradicting the fact that c is an upper bound for S. Similarly if $c \in V$ then there is an $\epsilon_2 > 0$ such that $B(c, \epsilon_2) \subseteq V$. But then $\emptyset = (c - \epsilon_2, c] \cap U \supseteq (c - \epsilon_2, c] \cap S$, so that $c - \epsilon_2$ is an upper bound for S, contradiction the fact that c is the least upper bound of S. It follows that we must have $E \subseteq U$ or $E \subseteq V$, and hence E is connected as required. $\square$

Corollary 25.3. *(The Intermediate Value Theorem): If $f : [a, b] \to \mathbb{R}$ is continuous and c lies in the closed interval with endpoints $f(a), f(b)$, then there is some $x \in [a, b]$ with $f(x) = c$.*

Proof. Clearly $[a, b]$ has property I, hence it is connected. Since f is continuous, $f([a, b])$ is connected and hence also has property I, thus since $f(a), f(b) \in f([a, b])$ the entire interval between $f(a)$ and $f(b)$ is in the image of f as required. $\square$

Lemma 25.4. *If $E \subseteq \mathbb{R}$ has property I, then E is either $\mathbb{R}$, a half-line, or an interval.*

Proof. Let us write $l = \inf(E) \in \{-\infty\} \cup \mathbb{R}$ and $u = \sup(E) \in \mathbb{R} \cup \{+\infty\}$. We claim that

$$(l, u) \subseteq E \subseteq [l, u],$$

(where if u or l is infinite, then the bracket being open or closed should be taken to mean the same thing). To establish the claim not first the right-hand inclusion is immediate from the definitions. For the left-hand inclusion, suppose that $z \in (l, u)$. The since $l < z$ the approximation property shows that there is some $y \in E$ with $l \leq y < z$, and similarly since $z < u$ there is some $y \in E$ with $z < y \leq u$. It follows $z \in [x, y] \subseteq I$ by property I, and so we are done. It is easy to see that the claim immediately implies the statement of the Lemma. $\square$

Combining the above results we obtain a classification of the connected subsets of $\mathbb{R}$.

Theorem 25.5. *The connected subsets of* $\mathbb{R}$ *are precisely* $\mathbb{R}$ *itself, all half-lines* $[a, \infty]$, (a, ∞),$(-\infty, a)$,$(-\infty, a]$ *and all bounded intervals* (a, b), $(a.b]$, $[a, b)$, (a, b) *for* $a, b \in \mathbb{R}$ *with* $a \leq b$.

26. Appendix III: some results from real analysis.

In this appendix we review some notions from multivariable calculus. While we give careful proofs, only the statements are examinable.

26.1. **Properties of the Limit Superior.** We collect here some basic facts about the $\limsup$ of a sequence of real numbers. Recall the definition:

Definition 26.1. Let (a_n) be a sequence which is bounded above (if it is not, by convention we set $\limsup_n(a_n) = +\infty$). Then for each n we may set $s_n = \sup\{a_k : k \geq n\}$. Clearly the sequence (s_n) is decreasing, and so if it is bounded below it has a limit, which we denote by $\limsup_n(a_n)$. If the sequence s_n is not bounded below, it tends to $-\infty$, and we write $\limsup_n(a_n) = -\infty$. Note that $\limsup_n(a_n) = -\infty$ if and only if $a_n \to -\infty$ as $n \to -\infty$.

The following Lemma is helpful in understanding what the properties of the $\limsup$ are.

Lemma 26.2. *Let (a_n) be a sequence of real numbers which is bounded above and let $s = \limsup_n(a_n)$. If (a_{n_k}) is any convergent subsequence of (a_n) with limit ℓ then $\ell \leq s$. Moreover, there exists a subsequence of (a_n) which converges to s, so that $\limsup_n(a_n)$ is the maximum value of the limit of a subsequence of (a_n).*

Proof. For the first part, note that by definition clearly $a_{n_k} \leq s_{n_k}$, and since (s_n) tends to s it follows the subsequence (s_{n_k}) does also, hence since limits preserve weak inequalities, $\lim_k(a_{n_k}) = l \leq s$ as required.

Let $A_n = \{a_m : m \geq n \in \mathbb{N}\}$ be the set of values of the n-th tail of the sequence (a_n). Then it is clear that s_m is in $\bar{A}_n$ for each $m \geq n$, and so $s \in \bar{A}_n$ for all n. If s is a limit point of any A_n then it is easy to see that s is a limit of a subsequence of the associated tail $(a_k)_{k \geq n}$. If, for all n, we have $s \notin A'_n$, then we must have $s \in \bar{A}_n \backslash A'_n \subseteq A_n$ for all n, hence $s = a_m$ for infinitely many m. It follows that there is a subsequence of (a_n) which is constant and equal to s, so certainly it converges to s.

□

We have the following basic property of $\limsup$, which we used in the discussion of differentiation of power series:

Lemma 26.3. *Suppose that (a_n) is a bounded sequence of real numbers. Then if (c_n) is a sequence which converges to $c \geq 0$ then $\limsup_n(c_n a_n) = c.\limsup_n a_n$.*

Proof. If (a_{n_k}) is any subsequence of (a_n) which converges to $\ell \in \mathbb{R}$, then clearly $c_{n_k}a_{n_k} \to c.\ell$ as $n \to \infty$. Since $c \geq 0$ it follows the result follows from the previous lemma which shows that $\limsup_n(c_n a_n)$ is the maximum value of the limit of a subsequence of $(c_n a_n)$. □

Remark 26.4. For sequences which are bounded below one may consider $l_n = \inf\{a_k : k \geq n\}$. Clearly (l_n) forms an increasing sequence and one sets $\liminf_n(a_n) = \lim_n l_n$. It is easy to see that $\limsup_n(a_n) = -\liminf_n(-a_n)$.

26.2. **Partial derivatives and the total derivative.**

Theorem 26.5. *Suppose that $F\colon U \to \mathbb{R}^2$ is a function defined on an open subset of $\mathbb{R}^2$, whose partial derivatives exist and are continuous on U. Then for all $z \in U$ the function F is real-differentiable, with derivative Df_z given by the matrix of partial derivative.*

Proof. Working component by component, you can check that it is in fact enough to show that a function $f\colon U \to \mathbb{R}$ with continuous partial derivatives $\partial_x f$ and $\partial_y f$ has total derivative given by $(\partial_x f, \partial_y f)$ at each $z \in U$. That is, if $z = (x, y)$ then

$$f(x+h, y+k) = f(x, y) + \partial_x f(x, y)h + \partial_y f(x, y)k + \|(h, k)\|.\epsilon(h, k),$$

where $\epsilon(h, k) \to 0$ as $(h, k) \to 0$. But now since the function $x \mapsto f(x, y)$ is differentiable at x with derivative $\partial_x f(x, y)$ we have

$$f(x+h, y) = f(x, y) + \partial_x f(x, y)h + h\epsilon_1(h)$$

where $\epsilon_1(h) \to 0$ as $h \to 0$. Now by the mean value theorem applied the function to $y \mapsto f(x+h, y)$ we have

$$f(x+h, y+k) = f(x+h, y) + \partial_y f(x+h, y+\theta_2 k)k,$$

for some $\theta_2 \in (0, 1)$. Thus using the definition of $\partial_x f(x, y)$ it follows that

$$f(x+h, y+k) = f(x, y) + \partial_x f(x, y)h + h\epsilon_1(h) + \partial_y f(x+h, y+\theta_2 k)k.$$

Thus we have

$$f(x+h, y+k) = f(x, y) + \partial_x f(x, y)h + \partial_y f(x, y)k + \|(h, k)\|\epsilon(h, k),$$

where

$$\epsilon(h, k) = \frac{h}{\sqrt{h^2+k^2}}\epsilon_1(h) + \frac{k}{\sqrt{h^2+k^2}}(\partial_y f(x+h, y+\theta_2 k) - \partial_y f(x, y)).$$

Thus since $0 \leq h/\sqrt{h^2+k^2}, k/\sqrt{h^2+k^2} \leq 1$, the fact that $\epsilon_1(h) \to 0$ as $h \to 0$ and the continuity of $\partial_y f$ at (x, y) imply that $\epsilon(h, k) \to 0$ as $(h, k) \to 0$ as required. $\square$

Remark 26.6. Note that in fact the proof didn't use the full strength of the hypothesis of the theorem – we only actually needed the existence of the partial derivatives and the continuity of one of them at (x, y) to conclude that f is real-differentiable at (x, y).

26.3. **The Chain Rule.** We establish a version of the chain rule which is needed for the proof that the existence of a primitive for a function $f\colon U \to \mathbb{C}$ implies that $\int_\gamma f(z)dz = 0$ for every closed curve γ in U. The proof requires one to use the fact that if $dF/dt = f$ on U then $f(\gamma(t))\gamma'(t)$ is the derivative of $F(\gamma(t))$. This is of course formally exactly what one would expect using the formula for the normal version of the chain rule, but one should be slightly careful: $F\colon \mathbb{C} \to \mathbb{C}$ is a function of a complex variable, while $\gamma\colon [a, b] \to \mathbb{C}$ is a function of real variable, so we are mixing real and complex differentiability.

That said, we have seen that a complex differentiable function is also differentiable in the real sense, with its derivative being the linear map given by multiplication by the complex number which is its complex derivative. Thus the result we need follows from a version of the chain rule for real-differentiable functions:

Lemma 26.7. *Let U be an open subset of $\mathbb{R}^2$ and let $F\colon U \to \mathbb{R}^2$ be a differentiable function. If $\gamma\colon [a,b] \to \mathbb{R}$ is a (piecewise) C^1-path with image in U, then $F(\gamma(t))$ is a differentiable function with*

$$\frac{d}{dt}(F(\gamma(t))) = DF_{\gamma(t)}(\gamma'(t))$$

Proof. Let $t_0 \in [a,b]$ and let $z_0 = \gamma(t_0) \in U$. Then by definition, there is a function $\epsilon(z)$ such that

$$F(z) = F(z_0) + DF_{z_0}(z - z_0) + |z - z_0|\epsilon(z),$$

where $\epsilon(z) \to 0 = \epsilon(z_0)$ as $z \to z_0$. But then

$$\frac{F(\gamma(t)) - F(\gamma(t_0))}{t - t_0} = DF_{z_0}(\frac{\gamma(t) - \gamma(t_0)}{t - t_0}) + \epsilon(\gamma(t)).\frac{|\gamma(t) - \gamma(t_0)|}{t - t_0}.$$

But now consider the two terms on the right-hand side of this expression: for the first term, note that a linear map is continuous, so since $(\gamma(t) - \gamma(t_0))/(t - t_0) \to \gamma'(t_0)$ as $t \to t_0$ we see that $DF_{z_0}(\frac{\gamma(t)-\gamma(t_0)}{t-t_0}) \to DF_{z_0}(\gamma'(t_0))$ as $t \to t_0$. On the other hand, for the second term, since $\frac{\gamma(t)-\gamma(t_0)}{t-t_0}$ tends to $\gamma'(t_0)$ as t tends to t_0, we see that $|\gamma(t) - \gamma(t_0)|/(t - t_0)$ is bounded as $t \to t_0$, while since $\gamma(t)$ is continuous at t_0 since it is differentiable there $\epsilon(\gamma(t)) \to \epsilon(\gamma(t_0)) = \epsilon(z_0) = 0$. It follows that the second term tends to zero, so that the left-hand side tends to $Df_{\gamma(t_0)}(\gamma'(t_0))$ as required. □

Remark 26.8. Notice that the proof above works in precisely the same way if F is a function from $\mathbb{R}^2$ to $\mathbb{R}$. Indeed a slight modification of the argument proves that if $F\colon \mathbb{R}^n \to \mathbb{R}^m$ and $G\colon \mathbb{R}^m \to \mathbb{R}^p$ then if F and G are differentiable, their composite $G \circ F$ is differentiable with derivative $DG_{F(x)} \circ DF_x$.

An easy application of the chain rule is the following constancy theorem. For the proof it is convenient to introduce some terminology:

Definition 26.9. We say a function $f\colon X \to Y$ between metric spaces is *locally constant* if for any $z \in X$ there is an $r > 0$ such that f is constant on $B(z,r)$.

Remark 26.10. Clearly a locally constant function is continuous, and moreover for such a function, the pre-image of any point in its image is an open set. Since for any continuous function the pre-image of a point is a closed set, it follows the pre-image of a point in the range of a locally-constant function is both open and closed. Thus if X is connected and f is locally constant, then f is in fact constant.

Proposition 26.11. *Suppose that $f: U \to \mathbb{R}^2$ is a function defined on a connected open subset of $\mathbb{R}^2$. Then if $Df_z = 0$ for all $z \in U$ the function f is constant.*

Proof. By the preceding remarks it suffices to show that f is locally constant. To see this, let $z_0 \in U$ and fix $r > 0$ such that $B(z_0, r) \subseteq U$. Then for any $z \in B(z_0, r)$ we may consider the function $F(t) = f(z_0 + t(z - z_0))$, where $t \in [0, 1]$. Note that $F = f \circ \gamma$ where $\gamma(t) = z_0 + t(z - z_0)$ is the straight line-segment from z_0 to z which lies entirely in $B(z_0, r)$ as z does. Hence applying the chain rule we have $F'(t) = Df_{z_0+t(z-z_0)}(z - z_0) = 0$ by our assumption on Df_z. It follows from the Fundamental Theorem of Calculus that

$$f(z) - f(z_0) = F(1) - F(0) = \int_0^1 F'(t)dt = 0,$$

hence f is constant on $B(z_0, r)$ as required. (The integral of the function $F'(t) = (u'(t), v'(t))$ is taken component-wise.)

$\square$

26.4. **Symmetry of mixed partial derivatives.** We used in the proof that the real and imaginary parts of a holomorphic function are harmonic the fact that partial derivatives commute on twice continuously differentiable functions. We give a proof of this for completeness. The key to the proof will be to use difference operators:

Definition 26.12. Let $f: U \to \mathbb{R}$ be a function defined on an open set $U \subset \mathbb{R}^2$. Then if $s, t \in \mathbb{R}\backslash\{0\}$ let $\Delta_1^s(f), \Delta_2^t(f)$ be the function given by

$$\Delta_1^s(f)(x, y) = \frac{f(x + s, y) - f(x, y)}{s}, \quad \Delta_2^t(f)(x, y) = \frac{f(x, y + t) - f(x, y)}{t}$$

Note that if f is differentiable at (x, y) then $\partial_x f(x, y) = \lim_{s\to 0} \Delta_1^s(f)(x, y)$ and $\partial_y f(x, y) = \lim_{t\to 0} \Delta_2^t(f)(x, y)$.

It is straight-forward to check that

$$\begin{aligned}\Delta_1^2(\Delta_2^t(f))(x, y) &= \Delta_2^t(\Delta_1^s(f))(x, y)\\ &= \frac{f(x + s, y + t) - f(x + s, y) - f(x, y + t) + f(x, y)}{st}.\end{aligned}$$

That is, the two difference operators $f \mapsto \Delta_1^s(f)$ and $f \mapsto \Delta_2^t(f)$ commute with each other. We wish to use this fact to deduce that the corresponding partial differential operators also commute, but because of the limits involved, this will not be automatic, and we will need to impose the additional hypotheses that the second partial derivatives of f are continuous functions.

Since the difference operator Δ_1^s and Δ_2^t are linear, they commute with partial differentiation so that $\partial_y \Delta_1^s(f)(x, y) = \Delta_1^s(\partial_y f)(x, y)$, and similarly for ∂_x and also for Δ_2^t and ∂_x, ∂_y.

We are now ready to prove that mixed partial derivatives are equal:

Lemma 26.13. *Suppose that $f\colon U \to \mathbb{R}$ is twice continuously differentiable, so that all its second partial derivatives exist and are continuous on U. Then*

$$\partial_x \partial_y f = \partial_y \partial_x f$$

on U.

Proof. Fix $(x,y) \in U$. Since U is open, there are $\epsilon, \delta > 0$ such that $\Delta_1^s(f)$ and $\Delta_2^t(f)$ are defined on $B((x,y),\epsilon)$ for all s,t with $|s|,|t| < \delta$. Now by definition we have

$$\partial_x \partial_y f(x,y) = \partial_x(\lim_{t\to 0} \Delta_2^t(f))(x,y) = \lim_{s\to 0}\lim_{t\to 0} \Delta_1^s \Delta_2^t(f)(x,y)$$

But now using the mean value theorem for $\Delta_2^t(f)$ in the first variable, we see that

$$\Delta_1^s \Delta_2^t(f)(x,y) = \partial_x \Delta_2^t f(x+s_1,y),$$

where s_1 lies between 0 and s. But $\partial_x \Delta_2^t(f)(x+s_1,y) = \Delta_2^t \partial_x f(x+s_1,y)$, and using the mean value theorem for $\partial_x f(x+s_1,y)$ in the second variable we see that $\Delta_2^t \partial_x f(x+s_1,y) = \partial_y \partial_x f(x+s_1, y+t_1)$ where t_1 lies between 0 and t (and note that t_1 depends both on t and s_1).

But now

$$\partial_x \partial_y f(x,y) = \lim_{s\to 0}\lim_{t\to 0} \partial_y \partial_x f(x+s_1,y+t_1) = \partial_y \partial_x f(x,y),$$

by the continuity of the second partial derivatives, so we are done. □

Example 26.14. Let $\Delta = \partial_x^2 + \partial_y^2$ be the (two-dimensional) Laplacian. Provided we are only interested in acting on twice-continuously differentiable functions $u = u(x,y)$ so that $\partial_x\partial_y(u) = \partial_y\partial_x(u)$, we can factorize Δ as

$$\Delta = (\partial_x - i\partial_y)(\partial_x + i\partial_y).$$

This is the key to the relationship between holomorphic and harmonic functions.

27. Appendix IV: Power series

In this appendix we give a proof of the following Theorem, which was established in Prelims Analysis I.

Proposition 27.1. *Let $s(z) = \sum_{k\geq 0} a_k z^k$ be a power series, let S be the domain on which it converges, and let R be its radius of convergence. Then power series $t(z) = \sum_{k=1}^{\infty} ka_k z^{k-1}$ also has radius of convergence R and on $B(0,R)$ the power series s is complex differentiable with $s'(z) = t(z)$. In particular, it follows that a power series is infinitely complex differentiable within its radius of convergence.*

Proof. First note that the power series $\sum_{k=1}^{\infty} ka_k z^{k-1}$ clearly has the same radius of convergence as $\sum_{k=1}^{\infty} ka_k z^k$, and by Lemma 14.20 this has radius of convergence[56]

$$\limsup_k |ka_k|^{1/k} = \lim_k (k^{1/k}) \limsup_k |a_k|^{1/k} = \limsup |a_k|^{1/k} = R,$$

since $\lim_{k\to\infty} k^{1/k} = 1$. Thus $s(z) = \sum_{k=0}^{\infty} a_k z^k$ and $t(z) = \sum_{k=1}^{\infty} ka_k z^{k-1}$ have the same radius of convergence. To see that $s(z)$ is complex differentiable with derivative $t(z)$, consider the sequence of polynomials f_n in two complex variables:

$$f_n(z,w) = a_n(\sum_{i=0}^{n-1} z^i w^{n-1-i}), \quad (n \geq 1).$$

Fix $\rho < R$, then for (z,w) with $|z|,|w| \leq \rho$ we have

$$|f_n(z,w)| = \big|a_n \sum_{i=0}^{n-1} z^i w^{n-i}\big| \leq |a_n| \sum_{i=0}^{n-1} |z|^i |w|^{n-i} \leq |a_n| n \rho^{n-1}$$

It therefore follows from the Weierstrass M-test with[57] $M_n = |a_n| n\rho^{n-1}$ that the series $\sum_{n\geq 0} f_n(z,w)$ converges uniformly (and absolutely) on $\{(z,w) : |z|,|w| \leq \rho\}$ to a function $F(z,w)$. In particular, it follows that $F(z,w)$ is continuous. But since $\sum_{k=1}^{n} f_k(z,z) = \sum_{k=1}^{n} ka_k z^{k-1}$, it follows that $F(z,z) = t(z)$. On the other hand, for $z \neq w$ we have $\sum_{i=0}^{k-1} z^i w^{k-i} = \frac{z^k - w^k}{z-w}$, so that

$$F(z,w) = \sum_{k=0}^{\infty} a_k \frac{z^k - w^k}{z-w} = \frac{s(z)-s(w)}{z-w},$$

hence it follows by the continuity of F that if we fix z with $|z| < \rho$ then

$$\lim_{z\to w} \frac{s(z)-s(w)}{z-w} = F(z,z) = t(z).$$

[56] This uses a standard property of $\limsup$ which is proved for completeness in Lemma 26.3 in Appendix I.

[57] We know $\sum_{n\geq 0} M_n = |a_n| n\rho^{n-1}$ converges since $\rho < R$ and $t(z)$ has radius of convergence R.

Since $\rho < R$ was arbitrary, we see that $s(z)$ is differentiable on $B(0, R)$ with derivative $t(z)$.

Finally, since we have shown that any power series is differentiable within its radius of convergence and its derivative is again a power series with the same radius of convergence, it follows by induction that any power series is in fact infinitely differentiable within its radius of convergence. $\square$

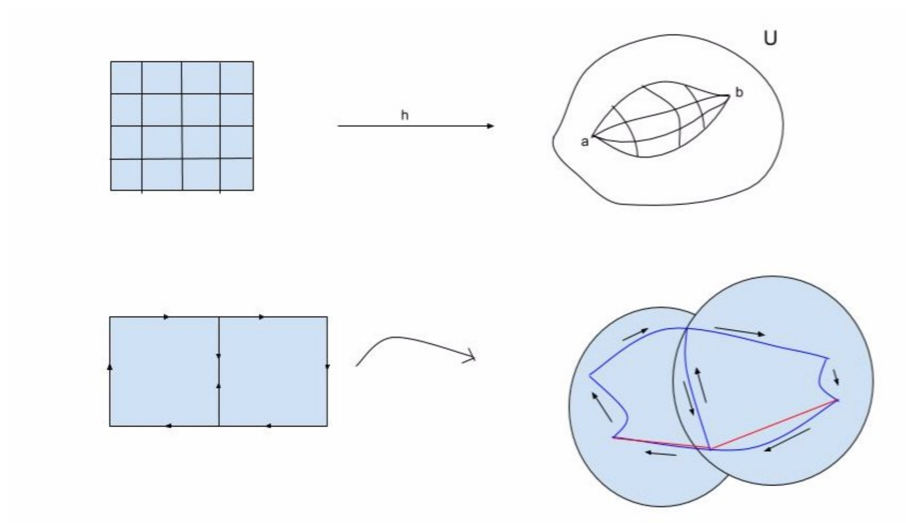

FIGURE 6. Dissecting the homotopy

28. APPENDIX V: ON THE HOMOTOPY AND HOMOLOGY VERSIONS OF CAUCHY'S THEOREM

In this appendix we give proofs of the homotopy and homology versions of Cauchy's theorem which are stated in the body of the notes. These proofs are non-examinable, but are included for the sake of completeness.

Theorem 28.1. *Let U be a domain in $\mathbb{C}$ and $a, b \in U$. Suppose that γ and η are paths from a to b which are homotopic in U and $f\colon U \to \mathbb{C}$ is a holomorphic function. Then*

$$\int_\gamma f(z)dz = \int_\eta f(z)dz.$$

Proof. The key to the proof of this theorem is to show that the integrals of f along two paths from a to b which "stay close to each other" are equal. We show this by covering both paths by finitely many open disks and using the existence of a primitive for f in each of the disks.

More precisely, suppose that $h\colon [0,1]\times[0,1]$ is a homotopy between γ and η. Let us write $K = h([0,1] \times [0,1])$ be the image of the map h, a compact subset of U. By Lemma 11.6 there is an $\epsilon > 0$ such that $B(z, \epsilon) \subseteq U$ for all $z \in K$.

Next we use the fact that, since $[0,1] \times [0,1]$ is compact, h is uniformly continuous. Thus we may find a $\delta > 0$ such that $|h(t_1, s_1) - h(t_2, w_2)| < \epsilon$ whenever $\|(t_1, s_1) - (t_2, s_2)\| < \delta$. Now pick $N \in \mathbb{N}$ such that $1/N < \delta$ and dissect the square $[0,1] \times [0,1]$ into N^2 small squares of side length $1/N$. For convenience, we will write $t_i = i/N$ for $i \in \{0, 1, \ldots, N\}$

For each $k \in \{1, 2, \ldots, N-1\}$, let ν_k be the piecewise linear path which connects the point $h(t_j, k/N)$ to $h(t_{j+1}, k/N)$ for each $j \in \{0, 1, \ldots, N)$. Explicitly, for $t \in [t_j, t_{j+1}]$, we set

$$\nu_k(t) = h(t_j, k/N)(1 - Nt - j) + h(t_{j+1}, k/N)(Nt - j)$$

We claim that

$$\int_\gamma f(z)dz = \int_{\nu_1} f(z)dz = \int_{\nu_2} f(z)dz = \ldots = \int_{\nu_{N-1}} f(z)dz = \int_\eta f(z)dz$$

which will prove the theorem. In fact, we will only show that $\int_\gamma f(z)dz = \int_{\nu_1} f(z)dz$, since the other cases are almost identical.

We may assume the numbering of our squares S_i is such that $S_1, \ldots, S_N$ list the bottom row of our N^2 squares from left to right. Let m_i be the centre of the square S_i and let $p_i = h(m_i)$. Then $h(S_i) \subseteq B(p_i, \epsilon)$ so that $\gamma([t_i, t_{i+1}]) \subseteq B(p_i, \epsilon)$ and $\nu_1([t_i, t_{i+1}]) \subseteq B(p_i, \epsilon)$ (since $B(p_i, \epsilon)$ is convex and by assumption contains $\nu_1(t_i)$ and $\nu_1(t_{i+1})$). Since $B(p_i, \epsilon)$ is convex, f has primitive F_i on each $B(p_i, \epsilon)$. Moreover, as primitives of f on a domain are unique up to a constant, it follows that F_i and F_{i+1} differ by a constant on $B(p_i, \epsilon) \cap B(p_{i+1}, \epsilon)$, where they are both defined. In particular, since $\gamma(t_i), \nu_1(t_i) \in B(p_i, \epsilon) \cap B(p_{i+1}, \epsilon)$, $(1 \leq i \leq N-1)$, we have

$$F_i(\gamma(t_i)) - F_{i+1}(\gamma(t_i)) = F_i(\nu_1(t_i)) - F_{i+1}(\nu_1(t_i)). \tag{28.1}$$

Now by the Fundamental Theorem we have

$$\int_{\gamma_{|[t_i, t_{i+1}]}} f(z)dz = F_i(\gamma(t_{i+1})) - F_i(\gamma_1(t_i)),$$
$$\int_{\nu_{1|[t_i, t_{i+1}]}} f(z)dz = F_i(\nu_1(t_{i+1})) - F_i(\nu_1(t_i))$$

Combining we find that:

$$\begin{aligned}
\int_\gamma f(z)dz &= \sum_{i=0}^{N-1} \int_{\gamma_{|[t_i,t_{i+1}]}} f(z)dz \\
&= \sum_{i=0}^{N-1} \big(F_{i+1}(\gamma(t_{i+1})) - F_{i+1}(\gamma(t_i))\big) \\
&= F_N(\gamma(t_N)) - F_1(\gamma(0)) + \sum_{i=1}^{N-1} \big(F_i(\gamma(t_i)) - F_{i+1}(\gamma(t_i))\big) \\
&= F_N(b) - F_0(a) + \big(\sum_{i=0}^{N-1} (F_i(\nu_1(t_{i+1})) - F_{i+1}(\nu_1(t_{i+1})\big) \\
&= \sum_{i=0}^{N-1} \big((F_{i+1}(\nu_1(t_{i+1})) - F_{i+1}(\nu_1(t_i))\big) \\
&= \sum_{i=0}^{N-1} \int_{\nu_{1|[t_i,t_{i+1}]}} f(z)dz = \int_{\nu_1} f(z)dz
\end{aligned}$$

where in the fourth equality we used Equation (28.1).

□

Remark 28.2. The use of the piecewise linear paths ν_k might seem unnatural – it might seem simpler to use the paths given by the homotopy, that is the paths $\gamma_k(t) = h(t, k/N)$. The reason we did not do this is because we only assume that h is continuous, so we do not know that the path γ_k is piecewise C^1 which we need in order to be able to integrate along it.

The proof of the homology form of Cauchy's theorem uses Liouville's theorem, which we proved using Cauchy's theorem for a disc.

Theorem 28.3. *Let $f\colon U \to \mathbb{C}$ be a holomorphic function and let $\gamma\colon [0,1] \to U$ be a closed path whose inside lies entirely in U, that is $I(\gamma, z) = 0$ for all $z \notin U$. Then we have, for all $z \in U\backslash\gamma^*$,*

$$\int_\gamma f(\zeta)d\zeta = 0; \quad \int_\gamma \frac{f(\zeta)}{\zeta - z}d\zeta = 2\pi i I(\gamma, z)f(z), \quad \forall z \in U\backslash\gamma^*.$$

Moreover, if U is simply-connected and $\gamma\colon [a,b] \to U$ is any closed path, then $I(\gamma, z) = 0$ for any $z \notin U$, so the above identities hold for all closed paths in such U.

Proof. We first prove the general form of the integral formula. Note that using the integral formula for the winding number and rearranging, we wish to show that

$$F(z) = \int_\gamma \frac{f(\zeta) - f(z)}{\zeta - z}d\zeta = 0$$

for all $z \in U\backslash\gamma^*$. Now if $g(\zeta, z) = (f(\zeta) - f(z))/(\zeta - z)$, then since f is complex differentiable, g extends to a continuous function on $U \times U$ if we set $g(z, z) = f'(z)$. Thus the function F is in fact defined for all $z \in U$. Moreover, if we fix ζ then, by standard properties of differentiable functions, $g(\zeta, z)$ is clearly complex differentiable as a function of z everywhere except at $z = \zeta$. But since it extends to a continuous function at ζ, it is bounded near ζ, hence by Riemann's removable singularity theorem, $z \mapsto g(\zeta, z)$ is in fact holomorphic on all of U. It follows by Theorem 18.22 that

$$F(z) = \int_0^1 g(\gamma(t), z)\gamma'(t)dt$$

is a holomorphic function of z.

Now let $\text{ins}(\gamma) = \{z \in \mathbb{C} : I(\gamma, z) \neq 0\}$ be the inside of γ, so by assumption we have $\text{ins}(\gamma) \subset U$, and let $V = \mathbb{C}\backslash(\gamma^* \cup \text{ins}(\gamma))$ be the complement of γ^* and its inside. If $z \in U \cap V$, that is, $z \in U$ but not inside γ or on γ^*, then

$$\begin{aligned} F(z) &= \int_\gamma \frac{f(\zeta)d\zeta}{\zeta - z} - f(z)\int_\gamma \frac{d\zeta}{\zeta - z} \\ &= \int_\gamma \frac{f(\zeta)d\zeta}{\zeta - z} - f(z)I(\gamma, z) \\ &= \int_\gamma \frac{f(\zeta)d\zeta}{\zeta - z} = G(z) \end{aligned}$$

since $I(\gamma, z) = 0$. Now $G(z)$ is an integral which only involves the values of f on γ^* hence it is defined for all $z \notin \gamma^*$, and by Theorem 18.22, $G(z)$ is holomorphic. In particular G defines a holomorphic function on V, which agrees with F on all of $U \cap V$, and thus gives an extension of F to a holomorphic function on all of $\mathbb{C}$. (Note that by the above, F and G will in general *not* agree on the inside of γ.) Indeed if we set $H(z) = F(z)$ for all $z \in U$ and $H(z) = G(z)$ for all $z \in V$ then H is a well-defined holomorphic function on all of $\mathbb{C}$. We claim that $|H| \to 0$ as $|z| \to \infty$, so that by Liouville's theorem, $H(z) = 0$, and so $F(z) = 0$ as required. But since $\text{ins}(\gamma)$ is bounded, there is an $R > 0$ such that $V \supseteq \mathbb{C}\backslash B(0, R)$, and so $H(z) = G(z)$ for $|z| > R$. But then setting $M = \sup_{\zeta\in\gamma^*} |f(\zeta)|$ we see

$$|H(z)| = \left|\int_\gamma \frac{f(\zeta)d\zeta}{\zeta - z}\right| \leq \frac{\ell(\gamma).M}{|z| - R}.$$

which clearly tends to zero as $|z| \to \infty$, hence $|H(z)| \to 0$ as $|z| \to \infty$ as required.

For the second formula, simply apply the integral formula to $g(z) = (z - w)f(z)$ for any $w \notin \gamma^*$. Finally, to see that if U is simply-connected the inside of γ always lies in U, note that if $w \notin U$ then $1/(z-w)$ is holomorphic on all of U, and so $I(\gamma, w) = \int_\gamma \frac{dz}{z-w} = 0$ by the homotopy form of Cauchy's theorem. $\square$

Remark 28.4. It is often easier to check a domain is simply-connected than it is to compute the interior of a path. Note that the above proof uses Liouville's theorem, whose proof depends on Cauchy's Integral Formula for a circular path, which was a consequence of Cauchy's theorem for a triangle, but apart from the final part of the proof on simply-connectd regions, we did not use the more sophisticated homotopy form of Cauchy's theorem. We have thus established the winding number and homotopy forms of Cauchy's theorem essentially independently of each other.

29. Appendix VI: Remark on the Inverse Function Theorem

In this appendix we supply[58] *the details for the claim made in the remark after the proof of the holomorphic version of the inverse function theorem.*

There is an enhancement of the Inverse Function Theorem in the holomorphic setting, which shows that the condition $f'(z) \neq 0$ is automatic (in contrast to the case of real differentiable functions, where it is essential as one sees by considering the example of the function $f(x) = x^3$ on the real line). Indeed suppose that $f \colon U \to \mathbb{C}$ is a holomorphic function on an open subset $U \subset \mathbb{C}$, and that we have $z_0 \in \mathbb{U}$ such that $f'(z_0) = 0$.
Claim: In this case, f is at least 2 to 1 near z_0, and hence is not injective.

Proof of Claim: If we let $w_0 = f(z_0)$ and $g(z) = f(z) - w_0$, it follows g has a zero at z_0, and thus it is either identically zero on the connected component of U containing z_0 (in which case it is very far from being injective!) or we may write $g(z) = (z - z_0)^k h(z)$ where $h(z)$ is holomorphic on U and $h(z_0) \neq 0$. Our assumption that $f'(z_0) = 0$ implies that k, the multiplicity of the zero of g at z_0 is at least 2.

Now since $h(z_0) \neq 0$, we have $\epsilon = |h(z_0)| > 0$ and hence by the continuity of h at z_0 we may find a $\delta > 0$ such that $h(B(z_0, \delta)) \subseteq B(h(z_0), \epsilon)$. But then by taking a cut along the ray $\{-t.h(z_0) : t \in \mathbb{R}_{>0}\}$ we can define a holomorphic branch of $z \mapsto z^{1/k}$ on the whole of $B(h(z_0), \epsilon)$. Now let $\phi \colon B(z_0, \delta) \to \mathbb{C}$ be the holomorphic function given by $\phi(z) = (z - z_0).h(z)^{1/k}$ (where by our choice of δ this is well-defined) so that $\phi'(z_0) = h(z_0)^{1/k} \neq 0$. Then clearly $f(z) = w_0 + \phi(z)^k$ on $B(z_0, \delta)$. Since $\phi(z)$ is holomorphic,the open mapping theorem ensures that $\phi(B(z_0, \delta))$ is an open set, which since it contains $0 = \phi(z_0)$, contains $B(0, r)$ for some $r > 0$. But then since $z \mapsto z^k$ is k-to-1 as a map from $B(0, r) \backslash \{0\} \to B(0, r^k) \backslash \{0\}$ it follows that f takes every value in $B(w_0, r^k) \backslash \{w_0\}$ at least k times.

[58] For interest, not examination!

30. Appendix VII: Bernoulli numbers and the ζ-function

For interest only: non-examinable.

We define the *Bernoulli numbers* via the power series expansion of $B(z) = z/(e^z - 1)$ at the origin:

$$\frac{z}{e^z-1} = \sum_{n=0}^{\infty} \frac{B_n}{n!} z^n, \tag{30.1}$$

where since $B(z)$ is defined in $B(0, 2\pi)$, by Taylor's theorem the power series has radius of convergence 2π. Since $(e^z - 1)/z = \sum_{n=0}^{\infty} z^n/(n+1)!$, we can rewrite the definition as:

$$\left(\sum_{n=0}^{\infty} \frac{z^n}{(n+1)!}\right)\left(\sum_{m=0}^{\infty} \frac{B_m}{m!} z^m\right) = 1.$$

It follows that $B_0 = 1$ and for $n \geq 1$ we have

$$\sum_{k=0}^{n} \frac{1}{k!(n-k+1)!} B_k = 0,$$

or, in terms of binomial coefficients,

$$\sum_{k=0}^{n} \binom{n+1}{k} B_k = 0.$$

Thus we can recursively compute the B_k: for example $B_0 = 1, B_1 = -1/2, B_2 = 1/6, B_3 = 0, B_4 = -1/30, B_5 = 0$. (In fact $B_{2n+1} = 0$ for all $n > 1$).

The reason we are interested in the Bernoulli numbers is that they arise when one computes the value of the ζ-function $\zeta(s) = \sum_{n=1}^{\infty} n^{-s}$ at $s = 2k$ a positive even integer. Using suitable square contours Γ_N, we showed that the value of $\zeta(2)$ is $-\frac{\pi}{2} R_1$ where R_1 is the residue of $\cot(\pi z)/z^2$ at the origin (since the residues of $\cot(\pi z)/z^2$ at the non-zero integers are $\frac{1}{\pi n^2}$). Exactly the same strategy, using the function $\cot(\pi z)/z^{2k}$, shows that $\zeta(2k)$ is equal to $-\frac{\pi}{2} R_k$ where R_k is the coefficient of z^{2k-1} in the Laurent expansion of $\cot(\pi z)$. But we have

$$\begin{aligned}
\cot(\pi z) &= \frac{\cos(\pi z)}{\sin(\pi z)} = i\frac{e^{i\pi z} + e^{-i\pi z}}{e^{i\pi z} - e^{-i\pi z}} = i\frac{e^{2i\pi z} + 1}{e^{2i\pi z} - 1} \\
&= i\left(1 + \frac{2}{e^{2\pi i z} - 1}\right) = i + \frac{1}{\pi i z} B(2\pi i z) \\
&= i + \sum_{k=0}^{\infty} \frac{B_k}{k!} (2i)^k (\pi z)^{k-1},
\end{aligned}$$

thus it follows that

$$\zeta(2k) = -\frac{\pi}{2}\frac{B_k}{k!} 2^{2k} (-1)^k (\pi)^{2k-1} = (-1)^{k+1} \frac{2^{2k-1}\pi^{2k} B_{2k}}{(2k)!}$$

MATHEMATICAL INSTITUTE, OXFORD.