Chapter 1

The real number system

1.1 The field of rational numbers

We consider the set $\mathbb{N}_0 = \{1, 2, \cdots\}$ consisting of all natural numbers equipped with the usual addition and multiplication. We introduce a new symbol 0, called zero, and consider

$$\mathbb{N} = \{0, 1, 2, \cdots\}.$$

We define $m + 0 = m,\ m \cdot 0 = 0$ for any $m \in \mathbb{N}$. The equation $x + m = 0,\ m \in \mathbb{N}, m \neq 0$ has no solution in $\mathbb{N}$. To be able to solve this equation we introduce negative integers $-1, -2, -3, \cdots$ etc., with the property that $m + (-m) = 0$ and consider

$$\mathbb{Z} = \{0, \pm 1, \pm 2, \cdots\}.$$

Elements of $\mathbb{Z}$ are called integers and we note that with respect to addition $\mathbb{Z}$ is a group in the sense that (i) $m + n \in \mathbb{Z}$ for all $m, n \in \mathbb{Z}$ (ii) $m + 0 = m$ for all $m \in \mathbb{Z}$ and (iii) for any $m \in \mathbb{Z},\ m + (-m) = 0$. We also note that $m(n + p) = mn + mp$ for $m, n, p \in \mathbb{Z}$ and $m \cdot 1 = m$. So with addition and multiplication $\mathbb{Z}$ forms what is called a ring. We refer to $\mathbb{Z}$ as the ring of integers.

In $\mathbb{Z}$ we would like to solve the equation $mx + n = 0$ but a solution exists if and only if n is a multiple of m, meaning there exists $p \in \mathbb{Z}$ such that $n = mp$ in which case $x = -p$ is the unique solution. So we look for solutions of the above equation in a number system which properly contains $\mathbb{Z}$.

To this end let us consider the cartesian product $G = \mathbb{Z} \times \mathbb{Z}$ defined to be the set of all ordered pairs $(m, n), m, n \in \mathbb{Z}$. We remark that $(m, n) = (p, q)$

if and only if $m = p$ and $n = q$. G becomes a ring if we define $(m, n)+(p, q) = (m + p, n + q)$ and $(m, n)(p, q) = (mp, nq)$. In this group $(1, 1)$ serves as the multiplicative identity as can be seen from the definition.

The equation $mx + n = 0$ now takes the form $(m, n)(x, y) + (p, q) = 0$ which amounts to solve the equation $(m, n)(x, y) = (1, 1)$. It is easy to see that this can be solved in G only when (m, n) is one of $(1, 1), (1, -1), (-1, 1)$ or $(-1, -1)$. (It is also clear that the above cannot be solved if either $m = 0$ or $n = 0$). So the passage to G from $\mathbb{Z}$ doesn't seem to help much.

From G let us remove all elements (m, n) where n is zero. Let $G_0 = \{(m, n) \in G : \ n \neq 0\}$ and consider the diagonal

$$I = \{(p, p) \in G_0\}.$$

Let us weaken the equation $(m, n)(x, y) = (1, 1)$ by demanding only that $(m, n)(x, y) \in I$. This problem clearly has a solution $(x, y) = (n, m)$ since $(m, n)(n, m) = (mn, nm) \in I$.

In G_0 let us introduce a relation $\sim$ as follows. We say that $(m, n) \sim (p, q)$ if and only if $mq = np$. Then it can be easily checked that (i) $(m, n) \sim (m, n)$ (reflexive) (ii) $(m, n) \sim (p, q)$ implies $(p, q) \sim (m, n)$(symmetric) (iii) $(m, n) \sim (p, q)$ and $(p, q) \sim (r, s)$ implies $(m, n) \sim (r, s)$ (transitive). Such a relation is called an equivalence relation.

Using $\sim$ we can decompose G_0 into a disjoint union of subsets. For $(m, n) \in G_0$ define $[m, n] = \{(p, q) \in G_0 : \ (m, n) \sim (p, q)\}$. Then $[m, n] = [p, q]$ whenever $(m, n) \sim (p, q)$ and $[m, n] \cap [p, q] = \emptyset$ if (m, n) is not related to (p, q). Further $G_0 = \bigcup[m, n]$. Note that $I = [1, 1]$ in our notation.

Suppose now that $(m, n)(p, q) \in I$. If $(m', n') \sim (m, n)$ and $(p', q') \sim (p, q)$ we claim that $(m', n')(p', q') \in I$ also. To see this we are given $mp = nq, m'n = n'm$ and $p'q = q'p$. Need to show that $m'p' = n'q'$. Look at

$$m'p'nq = m'np'q = n'mpq'$$

which gives $m'p' = n'q'$ since $mp = nq$. Hence the property $(m, n)(p, q) \in I$ is indeed a property of the equivalence classes $[m, n]$ and $[p, q]$ rather than the individual representatives.

Let us therefore try to make $\widetilde{G} = \{[m, n] : (m, n) \in \ G_0\}$ into a number system in which we can solve $gx + h = 0$. Let us define

$$[m, n][p, q] = [mp, nq].$$

It is now an easy matter to check that if $(m', n') \in [m, n]$ and $(p', q') \in [p, q]$ then

$$[mp, nq] = [m'p', n'q']$$

which means our definition is meaningful. But we have to be careful with the addition since the obvious definition

$$[m,n]+[p,q]=[m+p,n+q]$$

is not 'well defined' i.e., this definition is not independent of the representatives from $[m,n]$ and $[p,q]$. Instead if we define

$$[m,n]+[p,q]=[mq+np,nq]$$

then this becomes a well defined addition and $\widetilde{G}$ becomes a ring.

Moreover, for any $[m,n]\in\widetilde{G}$ we have $[m,n]+[0,1]=[m,n]$ and for $[m,n]\in\widetilde{G},[m,n]\neq[0,1]$, $[m,n][n,m]=[1,1]$. Thus $[0,1]$ is the additive identity and $[1,1]$ is the multiplicative identity. And in $\widetilde{G}$ every nonzero element has a multiplicative inverse. Such a system is called a field. We call $\widetilde{G}$ the field of rationals.

Note that the map $\varphi:\mathbb{Z}\to\widetilde{G}$ defined by $\varphi(m)=[m,1]$ satisfies

(i) $\varphi(m+n)=\varphi(m)+\varphi(n)$

(ii) $\varphi(mn)=\varphi(m)\varphi(n)$.

Under these conditions we say that φ is a homomorphism of $\mathbb{Z}$ into $\widetilde{G}$ and we identify $\varphi(\mathbb{Z})$ with $\mathbb{Z}$ and say that $\mathbb{Z}$ is a subring of $\widetilde{G}$.

Notation: It is customary to denote $\widetilde{G}$ by $\mathbb{Q}$ and $[m,n]$ by $\frac{m}{n}$. When there is no confusion we will use this traditional notation.

1.2 The set of rationals as a metric space

We have seen that in $\mathbb{Q}$ we can solve all the first order equations $ax+b=0$ where $a,b\in\mathbb{Q}$. However, the same is not true for second order equations, e.g. the equation $x^2=3$ has no solution in $\mathbb{Q}$. We may introduce a symbol, say $\sqrt{3}$, which is defined to be a solution of the above equation and enlarge $\mathbb{Q}$ into a bigger field, call it $\mathbb{Q}(\sqrt{3})$, in which the above equation $x^2=3$ can be solved. One way to do this to just define $\mathbb{Q}(\sqrt{3})$ to be the set of all formal symbols of the form $a+b\sqrt{3}$ where $a,b\in\mathbb{Q}$ and let

$$(a+b\sqrt{3})+(a'+b'\sqrt{3})=(a+a')+(b+b')\sqrt{3}$$

$$\text{and } (a+b\sqrt{3})(a'+b'\sqrt{3})=(aa'+3bb')+(ab'+ba')b\sqrt{3}.$$

One can check that $\mathbb{Q}\subset\mathbb{Q}(\sqrt{3})$ and $\mathbb{Q}(\sqrt{3})$ is indeed a field.

But in $\mathbb{Q}(\sqrt{3})$ there is no guarantee that the equation $x^2 = 2$ or $x^2 = 5$ has a solution. We can of course define symbols $\sqrt{2}, \sqrt{3}, \sqrt{5}$ and so on and keep on extending $\mathbb{Q}$ so that $x^2 = 2, x^2 = 3, x^2 = 5$ etc will have solutions. But then there is no end to such a procedure. Even if we achieve a field where all second order equations can be solved, there is no way we can rest assured that higher order equations also will have solutions. Instead of proceeding as above with one step at a time (which algebraists love to do) we would like to get an extension of $\mathbb{Q}$ in which all polynomial equations can be solved.

In order to do this we make $\mathbb{Q}$ into a 'metric' space(whatever it means). We have to begin by defining positive and negative integers - we call $n \in \mathbb{Z}$ positive and write $n > 0$ if $n \in \{1, 2, \cdots\}$; we call n negative and write $n < 0$ if $-n \in \{1, 2, \cdots\}$. The inequality $n \geq 0 (n \leq 0)$ stands for $n > 0$ or $n = 0$(resp. $n < 0$ or $n = 0$). We extend this notion to elements of $\mathbb{Q}$ by defining $>$ and $<$ as follows:

We say that $r \in \mathbb{Q}$ is positive, $r > 0$ if $mn > 0$ where $r = [m, n]$; $r < 0$ if $mn < 0$. For $r, s \in \mathbb{Q}$ we say $r \geq s$ whenever $r - s \geq 0$ etc. It can be verified that given $r \in \mathbb{Q}$, $r \neq 0$ either $r > 0$ or $r < 0$ holds.

We then define $|r|$ for $r \in \mathbb{Q}$ by setting $|r| = r$ when $r \geq 0$ and $|r| = -r$ when $r < 0$. In terms of the modulus function (or absolute value) $|\cdot|$ we define a distance function, also called metric, as follows: $d : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ is defined by $d(r, s) = |r - s|$. It is immediate that

(i) $d(r, s) \geq 0$, $d(r, s) = 0$ if and only if $r = s$

(ii) $d(r, s) = d(s, r)$

(iii) $d(r, s) \leq d(r, t) + d(t, s)$.

The third one is called the triangle inequality.

The set $\mathbb{Q}$ equipped with d is an example of a metric space. (Later we will deal with several metric spaces.)

Once we have a metric which measures the distance between members of $\mathbb{Q}$ we can discuss convergence. First let us recall the definition of a sequence and set up notation. By a sequence we mean a function, say, $\varphi : \mathbb{N} \to \mathbb{Q}$. Traditionally we write a sequence as (a_n) or $(a_n)_{n=0}^{\infty}$ where $a_n = \varphi(n)$. In the notation the order is given importance. The sequence (a_n) is not the set $\{a_n : n \in \mathbb{N}\}$. We can now define the convergence of a sequence (a_n) in $\mathbb{Q}$.

We say that (a_n) converges to $a \in \mathbb{Q}$ if the following happens: For every $p \in \mathbb{N}_0$, however large it is, we can find n (depending on p) such that

$d(a_j, a) < \frac{1}{p}$ for all $j \geq n$.(i.e., a_j's get arbitrarily close to a, measuring the distance in terms of d).

If the above happens we write $(a_n) \to a$. It is obvious from the definition that $(a_n) \to a$ iff the sequence $(a_n - a) \to 0$. It is equally obvious that $(a_n) \to a$ and $(b_n) \to b$ implies $(a_n + b_n) \to a + b$. With little bit of work we can show that $(a_n b_n) \to ab$. For this we need the fact that $(a_n) \to a$ implies there exists $M \in \mathbb{Q}$ such that $|a_n| \leq M$ for all n - we say that (a_n) is bounded. This means every convergent sequence is bounded - but the converse is not true as the example (a_n) with $a_n = (-1)^n$ shows.(Amuse yourself by trying to prove the convergence of this sequence.)

Convergent sequences in $\mathbb{Q}$ enjoy another important property. Suppose $(a_n) \to a$. If $p \in \mathbb{N}$ we can choose N such that $d(a_n, a) < \frac{1}{2p}$ whenever $n \geq N$. By the triangle inequality

$$d(a_n, a_m) \leq d(a_n, a) + d(a, a_m) < \frac{1}{p}$$

for all $n, m \geq N$. This warrants a definition as it is very important.

A sequence (a_n) is said to be Cauchy if given any $p \in \mathbb{N}$ there exists $N \in \mathbb{N}$ such that $d(a_n, a_m) < \frac{1}{p}$ for all $n, m \geq N$.

Roughly speaking, as we go along the sequence the distance between adjacent terms becomes smaller and smaller.

A natural question arises now: Is it true that every Cauchy sequence (a_n) in $\mathbb{Q}$ converges to some $a \in \mathbb{Q}$? (If so we would be saying $\mathbb{Q}$ is complete). Unfortunately it is not true.

Proposition 1.2.1. *$\mathbb{Q}$ is not complete, i.e., there exist Cauchy sequences which do not converge in $\mathbb{Q}$.*

We only need to produce one example of a Cauchy sequence which does not converge in $\mathbb{Q}$. Let us take

$$a_n = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{1}{n!}$$

where $n! = 1 \cdot 2 \cdots n$. Note that $n! > 2^{n-1}$ or $\frac{1}{n!} < (\frac{1}{2})^{n-1}$ for any $n \geq 1$. In order to show that (a_n) is Cauchy we need the above estimate and also the formula

$$1 + a + a^2 + \cdots a^m = (1 - a^{m+1})(1 - a)^{-1}$$

for any $a \in \mathbb{Q}, a \neq 1$. This can be proved by expanding $(1-a)(\ 1+ a+ a^2+ \cdots a^m)$. Now when $n > m \geq 1$ we have

$$0 \leq a_n - a_m = \frac{1}{(m+1)!} + \frac{1}{(m+2)!} + \cdots + \frac{1}{n!}$$

which by the estimate $n! > 2^{n-1}$ gives

$$a_n - a_m \leq \frac{1}{2^m} + \frac{1}{2^{m+1}} + \cdots + \frac{1}{2^{n-1}}$$

the right hand side of which simplifies to

$$\frac{1}{2^m} \cdot \left(1 - \left(\frac{1}{2}\right)^{n-m}\right)\left(1 - \frac{1}{2}\right)^{-1}$$

which is strictly less than 2^{-m+1}. Hence

$$a_n - a_m < 2^{-m+1},\ n > m.$$

We only need to observe that given $p \in \mathbb{N}$ we can always choose N such that $\frac{1}{p} > 2^{-m+1}$ for $m \geq N$.(In fact $N = p + 1$ will work. Why?)

It remains to show that (a_n) does not converge in $\mathbb{Q}$. We prove this by contradiction. Suppose $(a_n) \to a$ for some $a \in \mathbb{Q}$. As $a_n > 0$ is an increasing sequence, it follows that $a > 0$. Assume $a = \frac{p}{q}$ where $p, q \in \mathbb{N}, q \neq 0$. As the constant sequence $(q!)$ converges to $q!$ we see that $(q!a_n) \to p(q-1)!$, an integer. But $q!a_n = m + b_n$ where

$$m = q! + q! + \frac{q!}{2!} + \cdots + \frac{q!}{q!} \in \mathbb{N}$$

$$\text{and } \ b_n = \frac{1}{q+1} + \frac{1}{(q+1)(q+2)} + \cdots + \frac{1}{(q+1)\cdots(n)}.$$

On the one hand $(b_n) = (q!a_n - m) \to$ an integer. On the other hand $b_n < \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^{n-q}} < 1$. This contradiction proves the claim.

1.3 The system of real numbers

We have seen that the metric space $(\mathbb{Q}, d)$ is not complete - there are Cauchy sequences in $\mathbb{Q}$ which do not converge. We now seek to complete $\mathbb{Q}$ - that is to find a complete metric space which contains $\mathbb{Q}$ as a proper subspace. This is achieved very easily by a general procedure called completion which we describe now.

Consider the set of all Cauchy sequences (a_n) in $\mathbb{Q}$. We define an equivalence relation $\sim$ by saying $(a_n) \sim (b_n)$ if the sequence $(a_n - b_n)$ converges to 0 in $\mathbb{Q}$. It is then easy to see that $\sim$ is an equivalence relation and hence partitions the set of all Cauchy sequences into disjoint equivalence classes. Let $\mathbb{R}$ stand for this set of equivalence classes of Cauchy sequences. Thus every element $A \in \mathbb{R}$ is an equivalence class of Cauchy sequences.

For every $r \in \mathbb{Q}$, the constant sequence (r) is obviously Cauchy which clearly converges to r. Let $\widetilde{r}$ denote the equivalence class containing (r) and denote by $\widetilde{\mathbb{Q}}$ the set of all such $\widetilde{r}, r \in \mathbb{Q}$. Then we can identify $\mathbb{Q}$ with $\widetilde{\mathbb{Q}}$ and $\widetilde{\mathbb{Q}} \subset \mathbb{R}$. We can now define the operations addition and multiplication as follows: if $A, B \in \mathbb{R}$ and if $(a_n) \in A$ and $(b_n) \in B$ define their sum $A + B$ to be the equivalence class containing $(a_n + b_n)$ and AB that containing $(a_n b_n)$. One needs to verify that this definition is independent of the sequences $(a_n) \in A$ and $(b_n) \in B$ which can be easily checked. Note that $\widetilde{0}$ corresponding to $0 \in \mathbb{Q}$ is the additive identity and $\widetilde{1}$ containing (1) is the multiplicative identity. The map $r \mapsto \widetilde{r}$ is a homomorphism from $\mathbb{Q}$ into $\widetilde{\mathbb{Q}}$.

Now we want to define a metric in $\mathbb{R}$; given $A, B \in \mathbb{R}$ we want to define $\widetilde{d}(A, B)$. If $(a_n) \in A$ and $(b_n) \in B$ then it is clear that $(|a_n - b_n|)$ is Cauchy and hence defines an equivalence class in $\mathbb{R}$. We simply define $\widetilde{d}(A, B)$ to be the equivalence class containing $(|a_n - b_n|)$. Again one has to verify that $\widetilde{d}$ is well defined. Suppose $(a'_n) \in A$ and $(b'_n) \in B$ so that $(a_n) \sim (a'_n)$ and $(b_n) \sim (b'_n)$. Then $|a_n - b_n| \leq |a_n - a'_n| + |a'_n - b'_n| + |b'_n - b_n|$ so that $|a_n - b_n| - |a'_n - b'_n| \leq |a_n - a'_n| + |b_n - b'_n|$. Changing the roles of $(a_n), (b_n)$ and $(a'_n), (b'_n)$ we also get

$$\big||a_n - b_n| - |a'_n - b'_n|\big| \leq |a_n - a'_n| + |b_n - b'_n|.$$

This shows that $(|a_n - b_n|) \sim (|a'_n - b'_n|)$ and hence $\widetilde{d}$ is well defined.

The order relation $\geq$ can also be extended from $\mathbb{Q}$ into $\mathbb{R}$ in a natural way. We say that $(a_n) > (b_n)$ if for some positive integer p we have $a_n \geq b_n + \frac{1}{p}$ for all but finitely many values of n. With this definition we say $A > B,\ A, B \in \mathbb{R}$ if there exist $(a_n) \in A$ and $(b_n) \in B$ such that $(a_n) > (b_n)$. It then follows that $A > B$ if and only if $(a_n) > (b_n)$ for any $(a_n) \in A$ and $(b_n) \in B$. We say that $A \geq B$ if either $A = B$ or $A > B$. Clearly $\widetilde{d}(A, B) \geq \widetilde{0}$ and $\widetilde{d}(A, B) = \widetilde{0}$ if and only if $A = B$. Using the definition we can check that

$$\widetilde{d}(A, B) \leq \widetilde{d}(A, C) + \widetilde{d}(C, B)$$

for all $A, B, C \in \mathbb{R}$. This makes $(\mathbb{R}, \widetilde{d})$ into a metric space. Note that for $r, s \in \mathbb{Q}$

$$d(r, s) = \widetilde{d}(\widetilde{r}, \widetilde{s}).$$

Thus $\widetilde{d}$ can be thought of as an extension of d to the bigger space $\mathbb{R}$.

Convergence in $\mathbb{R}$ is now defined using the metric $\widetilde{d}$.

Theorem 1.3.1. $\mathbb{R}$ *equipped with* $\widetilde{d}$ *is a complete metric space.*

Let us recall the definition of convergence of sequences in $\mathbb{R}$. We say that $A_n \in \mathbb{R}$ converges to $A \in \mathbb{R}$ if given any $\widetilde{r} \in \mathbb{R}, r \in \mathbb{Q}, r > 0$ there exists

$N \in \mathbb{N}$ such that

$$\widetilde{d}(A_n, A) < \widetilde{r},\ n \geq N.$$

Cauchy sequences are defined similarly.

Let $\widetilde{\mathbb{Q}}$ be the image of $\mathbb{Q}$ in $\mathbb{R}$ under the map $r \mapsto \widetilde{r}$. We observe that $\widetilde{\mathbb{Q}}$ is dense in $\mathbb{R}$. To see this, let $(a_n) \in A$ and choose N large enough so that $|a_n - a_m| < \frac{1}{p}$ for all $m, n \geq N$ where $p \in \mathbb{N}$ is given. Now consider $\widetilde{d}(A, \widetilde{a}_N)$. As $|a_n - a_N| < \frac{1}{p}$ for $n \geq N$ it follows that $\widetilde{d}(A, \widetilde{a}_N) < \frac{1}{p}$. As p is arbitrary, this proves the density of $\widetilde{\mathbb{Q}}$.

Suppose now (A_n) is a Cauchy sequence in $\mathbb{R}$. For each n, let us choose $B_n \in \widetilde{\mathbb{Q}}$ so that $\widetilde{d}(A_n, B_n) < \frac{1}{n}$. Let $b_n \in \mathbb{Q}$ be such that $B_n = \widetilde{b}_n = (b_n, b_n, \cdots)$. Note that

$$|b_n - b_m| = \widetilde{d}(B_n, B_m) \leq \frac{1}{n} + \widetilde{d}(A_n, A_m) + \frac{1}{m}.$$

As (A_n) is Cauchy, this means that (b_n) is a Cauchy sequence of rationals. Let B be the equivalence class which contains (b_n). Then

$$\widetilde{d}(A_m, B) \leq \widetilde{d}(A_m, B_m) + \widetilde{d}(B_m, B) \leq \frac{1}{m} + \widetilde{d}(B_m, B).$$

But then

$$\widetilde{d}(B_m, B) = \lim_{n \to \infty} |b_m - b_n|$$

which can be made arbitrarily small by choosing m large. Hence $\widetilde{d}(A_m, B) \to 0$ as $m \to \infty$ and therefore, $\mathbb{R}$ is complete.

In $\mathbb{R}$ we have defined addition and multiplication which makes it a ring. It is natural to ask whether $\mathbb{R}$ is also a field. That is given $A \in \mathbb{R}, A \neq 0$ does there exist $B \in \mathbb{R}$ such that $AB = \widetilde{1}$. Let $A \in \mathbb{R}, A \neq 0$ and take any $(a_n) \in A$. Then (a_n) cannot converge to 0; otherwise $(a_n) \in \widetilde{0}$ and hence $A = 0$. Therefore, there exists $p \in \mathbb{N}$ such that $|a_n| > \frac{1}{p}$ for all values of n, save for finitely many. As (a_n) is Cauchy, this is possible only if $a_n \geq \frac{1}{p}$(or $a_n \leq -\frac{1}{p}$) except for finitely many values of n. That is either $A > 0$ or $A < 0$. Assume $a_n \geq \frac{1}{p}, n \geq n_0$. Then the sequence $(1, 1, \cdots, 1, a_{n_0+1}^{-1}, \cdots)$ is Cauchy. Take A^{-1} to be the equivalence class containing this. It is easy to check $AA^{-1} = \widetilde{1}$.

Theorem 1.3.2. *$\mathbb{R}$ is a field which contains $\widetilde{\mathbb{Q}}$ as a dense subfield.*

The construction of $\mathbb{R}$ was motivated by two reasons: (i) $(\mathbb{Q}, d)$ is not complete as a metric space (ii) simple equations like $x^2 = 2$ has no solution in $\mathbb{Q}$. We have achieved completeness by defining $\mathbb{R}$ suitably. But now we may ask if equations like $x^2 = a$ has a solution in $\mathbb{R}$. Our definition of multiplication in $\mathbb{R}$ (which extends the definition from $\mathbb{Q}$) shows that $A^2 \geq 0$ for any $A \in \mathbb{R}$, hence the equation $x^2 = a$ has no solution when $a < 0$. But for $a > 0$ we do have.

Theorem 1.3.3. *Let $a \in \mathbb{R}, a \geq 0$. Then the equation $x^2 = a$ has a solution in $\mathbb{R}$.*

Proof. We can assume $a > 0$ since $x^2 = 0$ is satisfied by $A = 0$. When $a = m^2 \in \mathbb{N}$, $x = \pm m$ are solutions. So we can assume, dividing a by m^2, m large if necessary, that $0 < a < 1$. Since $\widetilde{\mathbb{Q}}$ is dense in $\mathbb{R}$ choose a rational a_1, such that $a < a_1 < 1$. Define

$$a_{n+1} = \frac{1}{2}\left(a_n + \frac{a}{a_n}\right).$$

Using induction we can easily verify that $a < a_n < 1$ for all n. If we can show that (a_n) converges to $A \in \mathbb{R}$ then the above will lead to

$$A = \frac{1}{2}\left(A + \frac{a}{A}\right) \text{ or } A^2 = a$$

and the proof will be complete.

So we have a sequence (a_n) with $a < a_n < 1$ but this boundedness is not enough to guarantee convergence as we have seen even in $\mathbb{Q}$. But we can say something more about (a_n) namely, (a_n) is monotone decreasing: $a_{n+1} < a_n$ for all n. To see this

$$4a_{n+1}^2 = a_n^2 + \frac{a^2}{a_n^2} + 2a > 4a$$

since $x^2 + y^2 \geq 2xy$ for any $x, y \in \mathbb{R}$ with equality only when $x = y$. So we have $a_{n+1}^2 > a$ and using this

$$a_{n+1} = \frac{1}{2a_n}(a_n^2 + a) \leq \frac{1}{2a_n}(a_n^2 + a_n^2) = a_n.$$

Thus, (a_n) is a monotone decreasing sequence and $a < a_n < 1$ for all n. This forces (a_n) to be Cauchy (verify!) and hence (a_n) converges to some $A \in \mathbb{R}$. This completes the proof. □

In the above proof we have been fortunate to prove that (a_n) is monotone. But if we only know (a_n) is bounded, we may not be able to prove that (a_n) converges. But something weaker is true, which is quite useful in several situations and is one of the important properties of the real number system.

Theorem 1.3.4. *Every bounded sequence (a_n) in $\mathbb{R}$ has a convergent subsequence.*

First let us recall the definition of a subsequence. Let (a_n) be a sequence and let $\varphi : \mathbb{N} \to \mathbb{R}$ be the function such that $\varphi(n) = a_n$. If $\psi : \mathbb{N} \to \mathbb{N}$ is a strictly increasing function (i.e., $\psi(k) < \psi(j)$ for $k < j$) the sequence $\varphi \circ \psi : \mathbb{N} \to \mathbb{R}$ is called a subsequence of (a_n). We usually denote the subsequence by $(a_{n_k}), n_1 < n_2 < \cdots$.

In order to prove the theorem it is enough to show that (a_n) has a monotonic (either increasing or decreasing) subsequence. The boundedness of (a_n) will then force the subsequence to be Cauchy and the completeness of $\mathbb{R}$ proves the theorem. Therefore, it remains to prove the following:

Proposition 1.3.5. *Every sequence (a_n) in $\mathbb{R}$ has a monotone subsequence.*

Proof. If the given sequence (a_n) has an increasing subsequence then there is nothing to prove, so let us assume (a_n) has no such subsequence. We then prove that (a_n) has a decreasing subsequence.

There is n_1 such that $a_{n_1} > a_n$ for all $n \geq 1$ for otherwise we can extract an increasing subsequence. Then choose n_2 such that $a_{n_2} < a_{n_1},\ n_2 > n_1$ and define a_{n_k} inductively. Clearly, (a_{n_k}) is a decreasing subsequence of (a_n). □

When we undertake the study of function spaces defined on $\mathbb{R}$, the set of all polynomials is going to play an important role - actually, they will play a role similar to that of $\mathbb{Q}$ in the construction of real numbers. So, it is important to study some properties of polynomials. In $\mathbb{R}$, even the simplest quadratic $x^2 + 1 = 0$ has no solution. So we now proceed to extend the real number system into something bigger where we can solve all polynomial equations - equations of the form $p(x) = 0$ where $p(x) = a_0 + a_1 x + \cdots + a_n x^n$ are polynomials. This leads us to the field of complex numbers and the fundamental theorem of algebra!

1.4 The field of complex numbers

Within the field of real numbers we are not able to solve such simple equations as $x^2 + 1 = 0$. Our aim now is to enlarge $\mathbb{R}$ defining the field of complex numbers in which not only the equation $x^2 + 1 = 0$ but any general polynomial equation $p(x) = 0$ can be solved.

To this end let $M_2(\mathbb{R})$ stand for the set of all 2×2 matrices with real entries. By this we mean the following: every element $A \in M_2(\mathbb{R})$ is a symbol of the form $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ where $a, b, c, d \in \mathbb{R}$. We can make $M_2(\mathbb{R})$ into a ring by defining addition and multiplication as follows. Let $A_j = \begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix}, j = 1, 2$. Define

$$A_1 + A_2 = \begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{pmatrix},$$

$$A_1 A_2 = \begin{pmatrix} a_1 a_2 + b_1 c_2 & a_1 b_2 + b_1 d_2 \\ c_1 a_2 + d_1 c_2 & c_1 b_2 + d_1 d_2 \end{pmatrix}.$$

Then it is easy to check that $A + 0 = 0 + A = A$ and $AI = IA = A$ where $0 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$ and $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

However, $M_2(\mathbb{R})$ is not a field. If $A \in M_2(\mathbb{R})$ and $A \neq 0$ then it is not true that there exists $B \in M_2(\mathbb{R})$ such that $AB = BA = I$. To see this let us introduce the concept of determinant. Define $\det A = (a_1 d_1 - b_1 c_1)$ if $A = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix}$. Then by routine calculation one proves that $\det(AB)$ $= \det A \cdot \det B$. Once we have this it is easy to see that there is a necessary condition for the existence of B with $AB = I$. Indeed, $AB = I$ gives $\det A \cdot \det B = \det I = 1$ which forces $\det A \neq 0$. It can be shown that this necessary condition is also sufficient: i.e., A has a multiplicative inverse iff $\det A \neq 0$.

Let $GL(2, \mathbb{R})$ stand for all $A \in M_2(\mathbb{R})$ with $\det A \neq 0$ so that $GL(2, \mathbb{R})$ becomes a group under multiplication. It is interesting to observe that this group is nonabelian/ noncommutative meaning AB=BA need not be true in general. Though $GL(2, \mathbb{R})$ is a group under multiplication, it does not contain 0, the additive inverse; it is not true that $GL(2, \mathbb{R})$ is closed under addition.

We therefore look for a subgroup of $GL(2, \mathbb{R})$ which is closed under addition so that when augmented with 0 it will form a field. To achieve this

we consider $\mathbb{C}$ to be the set of all $A \in GL(2,\mathbb{R})$ of the form $\begin{pmatrix} x & y \\ -y & x \end{pmatrix}$. The condition $\det A \neq 0$ translates into $x^2 + y^2 \neq 0$, i.e., either x or y is nonzero. Now if $A' = \begin{pmatrix} x' & y' \\ -y' & x' \end{pmatrix}$ then

$$A + A' = \begin{pmatrix} x + x' & y + y' \\ -(y + y') & x + x' \end{pmatrix} \in \mathbb{C}$$

$$AA' = \begin{pmatrix} xx' - yy' & xy' + yx' \\ -(xy' + yx') & xx' - yy' \end{pmatrix} \in \mathbb{C}.$$

This shows that $\mathbb{C}$ is a field, called the field of complex numbers.

Let us write $J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ so that any $A \in \mathbb{C}$ can be written in the form

$$A = xI + yJ$$

where the scalar multiplication of a matrix B by a real $\lambda \in \mathbb{R}$ is defined by

$$\lambda B = \begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix} \text{ if } B = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The set $\{xI : x \in \mathbb{R}\}$ sits inside $\mathbb{C}$ as a subfield which is isomorphic to $\mathbb{R}$. We can therefore think of $\mathbb{C}$ as an extension of $\mathbb{R}$, the field of real numbers.

In $\mathbb{C}$ the matrix J has the interesting property $J^2 = -I$ and hence solves $J^2 + I = 0$. Thus we have achieved one of our goals - being able to solve $x^2 + 1 = 0$.

We now proceed to show that $\mathbb{C}$ can be made into a complete metric space. Let us denote the elements of $\mathbb{C}$ by z, w etc. If $z = xI + yJ$ we define $|z|$ by

$$|z| = (x^2 + y^2)^{\frac{1}{2}}.$$

Now that $(x^2 + y^2) > 0$ and hence there exist real numbers A such that $A^2 = (x^2 + y^2)$. There are exactly two solutions of this - one positive and one negative. In the above we take the positive solution.

We claim that $|z|$ satisfies the property

$$|z + w| \leq |z| + |w|, \ z, w \in \mathbb{C}.$$

Once this is proved, the function d defined by $d(z, w) = |z - w|$ gives a metric on $\mathbb{C}$.

Let $z = xI + yJ$ and $w = uI + vJ$ so that $|z|^2 = x^2 + y^2, |w| = u^2 + v^2$ and we have to show that

$$((x+u)^2 + (y+v)^2)^{\frac{1}{2}} \leq (x^2+y^2)^{\frac{1}{2}} + (u^2+v^2)^{\frac{1}{2}}.$$

Expanding out

$$(x+u)^2 + (y+v)^2 = (x^2+y^2) + (u^2+v^2) + 2(xu+yv).$$

If we can show that

$$|(xu+yv)| \leq (x^2+y^2)^{\frac{1}{2}}\ (u^2+v^2)^{\frac{1}{2}}$$

then

$$(x+u)^2 + (y+v)^2) \leq (x^2+y^2) + (u^2+v^2) + 2(x^2+y^2)^{\frac{1}{2}}\ (u^2+v^2)^{\frac{1}{2}}$$

or

$$|z+w|^2 \leq |z|^2 + |w|^2 + 2|z||w| = (|z| + |w|)^2$$

which proves our claim. So it remains to prove

Lemma 1.4.1. *(Cauchy-Schwarz inequality) For $z = xI + yJ, w = uI + vJ$ in $\mathbb{C}$,*

$$|xu + yv| \leq |z||w|.$$

We remark that an easy calculation reveals

$$(xI + yJ)(uI - vJ) = (xu + yv)I + (uy - xv)J.$$

Let us define $\overline{z} = (xI - yJ)$ whenever $z = (xI + yJ)$ so that

$$z\overline{w} = (xu + yv)I + (uy - xv)J.$$

By defining $Re(z) = x$ and $Im(z) = y$ we have

$$Re(z\overline{w}) = (xu + yv),\ Im(z\overline{w}) = (uy - vx)$$

and the Cauchy-Schwarz inequality takes the form

$$|Re(z\overline{w})| \leq |z||w|.$$

We now proceed to the proof of the lemma.
For $t \in \mathbb{R}$ consider $|z + tw|$:

$$|z + tw|^2 = (x + tu)^2 + (y + tv)^2$$

which after simplification gives

$$at^2 + 2bt + c \geq 0$$

where $a = |w|^2,\ b = Re(z\overline{w}),\ c = |z|^2$. Now

$$at^2 + 2bt + c = a(t + \frac{b}{a})^2 + c - \frac{b^2}{a} \geq 0$$

As this is true for all t, taking $t = -\frac{b}{a}$ we get $c - \frac{b^2}{a} \geq 0$ or $b^2 \leq ac$ which proves the lemma.

We can now prove

Theorem 1.4.2. $\mathbb{C}$ *equipped with the metric* d *is a complete metric space.*

Proof. Let $(z_n) \in \mathbb{C}$ be a Cauchy sequence. Writing $z_n = x_nI + y_nJ,\ x_n, y_n \in \mathbb{R}$ we see that (x_n) and (y_n) are Cauchy sequences in $\mathbb{R}$. Hence there exist $x, y \in \mathbb{R}$ such that $(x_n) \to x$ and $(y_n) \to y$. Then clearly $(z_n) \to (xI + yJ)$. □

From now on we use a simplified notation. In place of xI let us write x; denote J by i so that $i^2 = -1$. Write $z = x + iy$ in place of $xI + yJ$ always remembering that $(x + iy)$ stands for the matrix $(xI + yJ)$.

1.5 The fundamental theorem of algebra

We consider polynomials $p(z)$ with coefficients from $\mathbb{C}$:

$$p(z) = a_0 + a_1 z + \cdots + a_n z^n,\ a_j \in \mathbb{C}.$$

We say that $p(z)$ is of degree n if $a_n \neq 0$. Our aim is to prove the following result which is known as the fundamental theorem of algebra.

Theorem 1.5.1. *For every polynomial* $p(z)$ *as above, the equation* $p(z) = 0$ *has at least one solution in* $\mathbb{C}$.

We prove this theorem in two steps. First we show that there exists $a \in \mathbb{C}$ such that $|p(z)| \geq |p(a)|$ for all $z \in \mathbb{C}$ i.e., the minimum value of $|p(z)|$ is attained at some $a \in \mathbb{C}$. Then we show that $p(a) = 0$. This will then prove the theorem.

Let $\overline{B_R(0)} = \{z \in \mathbb{C} : |z| \leq R\}$ which is called the closed disc of radius R centred at 0 (here $R \in \mathbb{R},\ R > 0$). As $0 \in \overline{B_R(0)}$ and $p(0) = a_0$

$$\min_{z \in \overline{B_R(0)}} |p(z)| \leq |a_0|.$$

Suppose we can choose R large enough so that $|p(z)| \geq |a_0|$ for all z not in $\overline{B_R(0)}$, then the minimum value of $|p(z)|$ can occur only in $\overline{B_R(0)}$. Our first claim that there exists $a \in \mathbb{C}$ with $|p(a)| \leq |p(z)|$ for all $z \in \mathbb{C}$ will be proved in two sets.

Lemma 1.5.2. *There exists $R > 0$ such that $|p(z)| \geq |a_0|$ for all z with $|z| > R$.*

Lemma 1.5.3. *In $\overline{B_R(0)}$, $|p(z)|$ attains its minimum, that is there exists $a \in \overline{B_R(0)}$ such that $|p(a)| \leq |p(z)|$ for all $z \in \overline{B_R(0)}$.*

The first lemma is easy to prove. We need only to make precise the intuitive idea that $|p(z)|$ behaves like $|z|^n$ for large values of $|z|$. Defining $b_j = a_j a_n^{-1}, j = 0, 1, 2, \cdots, n$,

$$|p(z)| = |a_n||z^n + b_{n-1}z^{n-1} + \cdots + b_0|.$$

Using triangle inequality,

$$|p(z)| \geq |a_n|(|z|^n - |b_{n-1}z^{n-1} + \cdots + b_0|).$$

Again by triangle inequality,

$$|b_{n-1}z^{n-1} + \cdots + b_0| \leq \sum_{j=0}^{n-1} |b_j||z|^j \leq \beta \, \frac{|z|^n - 1}{|z| - 1}$$

where $\beta = \max\limits_{0 \leq j \leq n-1} |b_j|$. Thus

$$|p(z)| \geq |a_n| \left(|z|^n - \frac{\beta |z|^n}{|z| - 1} + \frac{\beta}{|z| - 1} \right) \geq |a_n| \left(1 - \frac{\beta}{|z| - 1} \right) |z|^n$$

provided $|z| > 1$. Thus we have

$$|p(z)| \geq \frac{1}{2} \, |a_n||z|^n$$

whenever $(1 - \frac{\beta}{|z|-1}) > \frac{1}{2}$ or $|z| > 2\beta + 1$. We can achieve $|p(z)| \geq |a_0|$ by choosing z such that $|z| > 2\beta + 1$ and $|z|^n > \frac{2|a_0|}{|a_n|}$. That we can choose R satisfying $R > 2\beta + 1$ and $R^n > \frac{2|a_0|}{|a_n|}$ follows from the fact that for $|z| > 1$ the set $\{|z|^n\}$ is unbounded i.e., given any $M \in \mathbb{N}$ we can choose ζ with $|\zeta| > 1$ such that $|\zeta|^n > M$. This proves Lemma 1.5.2.

Lemma 1.5.3 is not so easy to prove. We need some preparation. So, we take a digression, make a couple of definitions, prove one or two important results concerning subsets of $\mathbb{R}$ and return to Lemma 1.5.3.

Let $A \subseteq \mathbb{R}$ be a nonempty set. We say that A is bounded above (below) if there exists C (c) such that $a \leq C$ ($a \geq c$) for all $a \in A$. Any such C (c) is called an upper bound (lower bound) for A. We can ask if A has a least upper bound (or greatest lower bound). If such a thing exists we call it $\sup A$ (or $\inf A$) as the case may be. An important property of the real number system is that sup (inf) exists for sets bounded above (below). When exist they are obviously unique!

Theorem 1.5.4. *Every nonempty set $A \subseteq \mathbb{R}$ which is bounded above (below) has a supremum (infimum).*

Proof. Assume that A is bounded above, the bounded below case can be handled similarly. Let a_1 be a non-upperbound and b_1 be an upper bound. Consider the interval $I_1 = [a_1, b_1]$ and divide it into two equal parts: $I_1 = [a_1, \frac{1}{2}(a_1 + b_1)] \cup [\frac{1}{2}(a_1 + b_1), b_1]$. Of these choose the interval for which the left end point is not an upper bound but the right one is an upper bound. Call it $I_2 = [a_2, b_2]$. Repeat the process choosing $I_j = [a_j, b_j] \subset I_{j-1}$ such that a_j is not an upper bound of A and b_j is an upper bound. Note that (a_j) is increasing and (b_j) is decreasing. As both sequences are bounded they converge. Let $(a_j) \to a$ and $(b_j) \to b$. But $(b_j - a_j) \to 0$ which implies $a = b$.

It is easy to see that this common limit is the supremum of A. Why? Since $x \leq b_j$ for any j and $x \in A$, we get $x \leq a$ for all $x \in A$. If $a' < a$ then there exists j large so that $a' < a_j < a$ but then as a_j is not an upper bound, a' also is not an upper bound. Hence $a = \sup A$. □

Corollary 1.5.5. *If A is bounded above then there exists a sequence $(a_j) \in A$ such that $(a_j) \to \sup A$. (Similar statement holds for sets bounded below).*

Let us try to prove Lemma 1.5.3 now. Let $R > 1$ and consider

$$A = \{|p(z)| : z \in \overline{B_R(0)}\}.$$

It is clear that A is bounded from below since $y \geq 0$ for all $y \in A$. We can also show that A is bounded above: there exists $C > 0$ such that $|p(z)| \leq CR^n$. By Theorem 1.5.4, A has an infimum, i.e., there exists $m \geq 0$ such that $|p(z)| \geq m$ for all $z \in \overline{B_R(0)}$ and m is the largest with this property. Lemma 1.5.3 claims that this inf is attained, i.e., there exists $a \in \overline{B_R(0)}$ such that $|p(a)| = m$.

In view of the above Corollary we know that there exists $(a_n) \in A$ such that $(a_n) \to m$. But then there exists $z_n \in \overline{B_R(0)}$ with $a_n = |p(z_n)|$ by the definition of A. If we can show that (z_n) has a subsequence (z_{n_k}) converging to some $a \in \overline{B_R(0)}$ and that $|p(z_{n_k})| \to |p(a)|$ then we are done.

Thus we are left with proving the following two lemmas.

Lemma 1.5.6. *Given a sequence $(z_n) \in \overline{B_R(0)}$ there is a subsequence (z_{n_k}) which converges to some $a \in \overline{B_R(0)}$.*

Proof. Let $z_n = x_n + iy_n$ so that $|x_n| \leq R$ and $|y_n| \leq R$. As $\{x_n\}$ is bounded there exists a subsequence (x_{n_k}) which converges to some $x \in \mathbb{R}$. The sequence (y_{n_k}) will then have a subsequence which converges to some $y \in \mathbb{R}$. Put together there exists a subsequence of (z_n) converging to $a = x + iy$. As $|z_n| \leq R$ for all n, we get $|a| \leq R$ as well. □

Lemma 1.5.7. *Let $p(z)$ be a polynomial. If (z_n) converges to a then $(p(z_n))$ converges to $p(a)$.*

Proof.

$$p(z) - p(a) = \sum_{j=0}^{n} a_j(z^j - a^j)$$

and

$$z^j - a^j = (z-a)(z^{j-1} + z^{j-2}a + \cdots + a^{j-1})$$

as can be verified directly. Hence

$$p(z) - p(a) = (z - a)q(z)$$

where q is a polynomial of degree $n-1$. On the set $|z-a| \leq 1$, q is bounded and so

$$|p(z) - p(a)| \leq C|z - a|, \ |z - a| \leq 1.$$

From this follows our claim. □

As $||p(z)| - |p(a)|| \leq |p(z) - p(a)|$ Lemma 1.5.3 is proved completely.

The proof of Theorem 1.5.1 will be complete if we can prove

Lemma 1.5.8. *Let $p(z)$ be a polynomial and $a \in \mathbb{C}$ is such that $|p(z)| \geq |p(a)|$ for all $z \in \mathbb{C}$. Then $p(a) = 0$.*

To prove this we proceed as follows. Writing $z = z - a + a$ we get

$$p(z) = p(z - a + a) = A_0 + A_1(z - a) + \cdots + A_n(z - a)^n$$

where $A_0 = p(a)$ and $A_n \neq 0$. If A_k is the first nonzero coefficient (after A_0) we have

$$p(z) = A_0 + A_k(z-a)^k + A_{k+1}(z-a)^{k+1} + \cdots + A_n(z-a)^n.$$

We can write this as

$$p(z) = p(a) + A_k(z-a)^k + A_k(z-a)^{k+1}Q(z)$$

where $$Q(z) = \frac{A_{k+1}}{A_k} + \frac{A_{k+2}}{A_k}(z-a) + \cdots + \frac{A_n}{A_k}(z-a)^{n-k-1}.$$

When z is close enough to a, $A_k(z-a)^{k+1}Q(z)$ is very small. If we can choose z in such a way that $p(a) + A_k(z-a)^k$ is equal to $(1-c)p(a)$ where $0 < c < 1$, then for such an z, $|p(z)|$ will be smaller than $|p(a)|$ which will force $|p(a)| = 0$.

We can arrange such a thing to happen if we can solve the equation

$$A_k(z-a)^k = -cp(a),\ 0 < c < 1.$$

Note that when $|z-a| < 1$,

$$|Q(z)| \leq \frac{|A_{k+1}|}{|A_k|} + \cdots + \frac{|A_n|}{|A_k|} = C \quad \text{(say)}.$$

Let δ be chosen so that $0 < \delta < 1$ and $\delta C < 1$. Then for $|z-a| < \delta$ we have $|z-a||Q(z)| < 1$. Let us choose $0 < c < 1$ such that $c\frac{|p(a)|}{|A_k|} < \delta^k$. Then any solution of $A_k(z-a)^k = -cp(a)$ will satisfy $|z-a| < \delta$. If z is such a solution then

$$p(z) = p(a) - cp(a) - cp(a)(z-a)Q(z)$$

which gives

$$|p(z)| \leq (1-c)|p(a)| + c|p(a)||z-a||Q(z)| < |p(a)|$$

leading to a contradiction.

The equation $A_k(z-a)^k = -cp(a)$ can be solved if we can solve $z^k = \alpha$ for any $\alpha \in \mathbb{C}$. Given $z \in \mathbb{C}$ we can factorise z as $z = r\omega$, where $r > 0$ and $\omega \in \mathbb{C}$ satisfies $|\omega| = 1$. We only need to define $r = (x^2+y^2)^{\frac{1}{2}}$ and $\omega = r^{-1}z$ where $z = xI + yJ$. The equation $z^k = \alpha$ reduces to two equations: $z^k = |\alpha|$ and $z^k = |\alpha|^{-1}\alpha$.

The equation $z^k = |\alpha|$ can be easily solved: Indeed, it has a real solution as the following proposition shows.

Proposition 1.5.9. *Let k be a positive integer. For any $r > 0$, the equation $x^k = r$ has a real solution.*

Proof. Consider the polynomial $p(x) = x^k - r$ which satisfies $p(0) < 0$. Choose $b_1 > 0$ such that $p(b_1) = b_1^k - r > 0$. As in the proof of Theorem 1.5.4 we define nested intervals $I_{j+1} \subset I_j,\ I_j = [a_j, b_j]$ with the property that $p(a_j) < 0$ and $p(b_j) > 0$. (We let $a_1 = 0$ to start with). Let a be the common limit point of the Cauchy sequences (a_j) and (b_j). Clearly, $p(a) = 0$ or equivalently, $a^k = r$. □

Thus we are left with solving the equation $z^k = \alpha$, where $\alpha \in \mathbb{C}$, $|\alpha| = 1$. In order to solve this, we need some more preparation. Hence, we postpone this to a later section.

1.6 $\mathbb{R}$ and $\mathbb{C}$ as topological spaces

We describe another important property for subsets of $\mathbb{R}$ or $\mathbb{C}$. If we take $A \subset \mathbb{R}$ or $\mathbb{C}$ then A becomes a metric space in its own right and so we can ask if A is complete or not. If so then every Cauchy sequence (x_n) in A will converge to some point in A. Under these circumstances we say that A is closed. (Imagine walking along a Cauchy sequence to get out of A).

When A is closed we say that its complement A' is open. In otherwords, a subset $G \subset \mathbb{R}$ or $\mathbb{C}$ is called open if G' is closed. It is easy to see that G is open if and only if the following condition is verified:

For every $x \in G$ there exists $\delta > 0$ (depending on x) such that $B_\delta(x) \subset G$ where $B_\delta(x) = \{y : |x - y| < \delta\}$.

When considering subsets of $\mathbb{R}$, $B_\delta(x)$ will be denoted by $(x - \delta, x + \delta)$. With the above definition it is easy to prove that any union of open sets is open; any finite intersection of open sets is open.

We have seen that any $A \subset \mathbb{R}$ or $\mathbb{C}$ which is bounded has the Bolzano-Weierstrass property, viz., any sequence $(x_n) \subset A$ has a Cauchy subsequence (which may or may not converge in A). If A is closed and bounded then every sequence (x_n) has a convergent subsequence (which converges in A). This property can be characterised in another way which is very general so that it can be used whenever we have the notion of open sets.

A family $G_\alpha, \alpha \in \Lambda$ of open sets is said to be an open cover for A if $A \subset \bigcup_{\alpha \in \Lambda} G_\alpha$. If Λ is finite we say that $G_\alpha, \alpha \in \Lambda$ is a finite cover.

Theorem 1.6.1. *(Heine-Borel) A set $A \subset \mathbb{R}$ is closed and bounded if and only if every open cover of A has a finite subcover.*

Proof. Assume first that A is closed and bounded. Assume also that $A \subset \mathbb{R}$. (The case $A \subset \mathbb{C}$ can be done similarly). If there is an open cover $\{G_\alpha : \alpha \in \Lambda\}$ which does not have a finite subcover, we will arrive at a contradiction. As A is bounded, $A \subset I$ where I is an interval, say $I = [a, b]$. Divide I into two equal subintervals I_{11} and I_{12}. Then at least one of $A \cap I_{11}$ or $A \cap I_{12}$ will not have a finite subcover (from G_α). Call that I_2. Subdivide I_2 and proceed as before. So we get $I_j \subset I_{j-1}$ such that $I_j \cap A$ cannot be covered by finitely many G_α. Let a be the unique point which belongs to all $I_j \cap A$ (completeness of $\mathbb{R}$ is used here). Since $a \in A \subset \bigcup G_\alpha$ there exists α_0 such that $a \in G_{\alpha_0}$. But G_{α_0} is open and so there exists $\delta > 0$ such that $(a - \delta, a + \delta) \subset G_{\alpha_0}$. As I_j's decrease to a, for large enough j we should have $I_j \subset (a - \delta, a + \delta) \subset G_{\alpha_0}$. But this means $I_j \cap A$ is covered by the single G_{α_0}, a contradiction to our construction of I_j's.

Conversely, assume that every open cover of A has a finite subcover. First we show that A is bounded. Let $G_n = B_n(0), n = 1, 2, \cdots$. Clearly $\{G_n : n = 1, 2, \cdots\}$ is an open cover of A and there exists N such that $A \subset \bigcup_{n=1}^{N} B_n(0)$ which simply means $A \subset B_N(0)$ or $|a| \leq N$ for all $a \in A$, i.e., A is bounded.

To show that A is closed, assume that (x_n) is a Cauchy sequence in A which converges to $a \in \mathbb{C}$ which does not lie in A. Then for any $x \in A$, $|x - a| > 0$ so that with $\delta(x) = \frac{1}{2}|x - a|$ we get $A \subset \bigcup_{x \in A} B_{\delta(x)}(a)$. This open cover has a finite subcover - there exist $y_1, y_2, \cdots, y_n \in A$ such that

$$A \subset \bigcup_{j=1}^{n} B_{\delta(y_j)}(a).$$

But this is not possible as $(x_n) \in A$ converges to a. Hence A has to be closed. □

This result is true only for the metric spaces $\mathbb{R}, \mathbb{C}$ (or $\mathbb{R}^n, \mathbb{C}^n$ etc.,) but not true in general. We will come across some metric spaces where 'closed and bounded' is not equivalent to 'every open cover has a finite subcover'. The latter property is called compactness and the above theorem says that in $\mathbb{R}$ or $\mathbb{C}$ a set A is compact if and only if it is closed and bounded.

Chapter 2

The space of continuous functions

We let $\mathcal{P}$ stand for the set of all polynomials $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n,\ n \in \mathbb{N}, a_j \in \mathbb{C}$ of a real variable x. For any $a, b \in \mathbb{R}, a < b$, we can define a metric d on $\mathcal{P}$ by

$$d(f,g) = \sup_{a \le x \le b} |f(x) - g(x)|,\ f, g \in \mathcal{P}.$$

It is easy to see that d is a metric. The only nontrivial thing to verify is that $d(f,g) = 0$ implies $f = g$. Recall that if $f, g \in \mathcal{P}$ and if $f(x) = \sum_{k=0}^{n} a_kx^k,\ g(x) = \sum_{k=0}^{m} b_kx^k$ we say $f = g$ whenever $n = m$ and $a_k = b_k$ for $0 \le k \le n$. $d(f,g) = 0$ gives $f(x) - g(x) = 0$ for all $a \le x \le b$ from which we need to conclude that $f = g$. To see this, suppose $p \in \mathcal{P}$ is such that $p(x) = 0$ for all $a \le x \le b$. We prove $p = 0$ by using induction on the degree of p. If $p(x) = a_1x + a_0 = 0$ for all $a \le x \le b$, clearly $p = 0$. Note that

$$p(x) - p(y) = (x - y)q(x, y)$$

where q is a polynomial in x of degree $(n-1)$ with coefficients depending on y. For $a \le x < y, (x - y) \neq 0$ and hence $q(x, y) = 0$ which implies all the coefficients of q are zero. By explicitly calculating the coefficients of q we can show that all the coefficients of p are zero.

Thus d defined above is indeed a metric. $\mathcal{P}$ equipped with this metric is denoted by $\mathcal{P}[a, b]$. We claim:

Theorem 2.0.2. *The metric space $(\mathcal{P}[a, b], d)$ is not complete.*

To prove this theorem we need to produce at least one Cauchy sequence (p_n) of polynomials which does not converge to any element of $\mathcal{P}$. One such sequence is provided by

$$p_n(x) = \sum_{k=0}^{n} \frac{1}{k!} x^k.$$

Without loss of generality let us take $a = 0$ and $b = 1$ and show that (p_n) is Cauchy in $\mathcal{P}[0,1]$ but (p_n) cannot converge to any polynomial p in the metric d.

It is easy to see that (p_n) is Cauchy. Indeed, if $m > n$

$$p_m(x) - p_n(x) = \sum_{k=n+1}^{m} \frac{1}{k!}\ x^k$$

so that $d(p_m, p_n) \leq \sum_{k=n+1}^{m} \frac{1}{k!}$. Since the numerical sequence $a_n = \sum_{k=0}^{n} \frac{1}{k!}$ converges to e the above shows $d(p_m, p_n) \to 0$ as $n, m \to \infty$. Hence the claim that (p_n) is Cauchy.

Since $|p_n(x) - p_m(x)| \leq d(p_n, p_m)$ it is clear that for each $x \in [0,1]$ (for any $x \in \mathbb{R}$ in fact)$(p_n(x))$ converges. Let us set $E(x)$ to be the limit function. Our theorem will be proved if we can show that $E(x)$ is not a polynomial. In order to prove this we require

Proposition 2.0.3. *For $x, y \in [0,1]$ with $x+y \in [0,1]$ we have $E(x)E(y) = E(x+y)$.*

Assume the proposition for a moment. If possible let $E(x) = a_0 + a_1 x + \cdots + a_m x^m$. As $p_n(0) = 1$ for all n, $E(0) = 1$ so that $a_0 = 1$. In view of the proposition we get

$$E(x)^2 = E(2x),\ 0 \leq x \leq \frac{1}{2}$$

$$\text{or}\ \ (1 + a_1 x + \cdots a_m x^m)^2 = (1 + 2a_1 x + 4a_2 x^2 + \cdots 2^m a_m x^m).$$

From this it is easy to show that $a_k = 0,\ 1 \leq k \leq m$ but then $E(x) = 1$ which is not true.

Coming to the proof of the proposition let us write

$$p_n(x) = \sum_{k=0}^{n} a_k,\ p_n(y) = \sum_{k=0}^{n} b_k$$

where $a_k = \frac{1}{k!}\ x^k,\ b_k = \frac{1}{k!}\ y^k$. Define

$$\begin{aligned} c_k &= \sum_{i=0}^{k} a_i b_{k-i} = \sum_{i=0}^{k} \frac{1}{i!}\ x^i\ \frac{1}{(k-i)!}\ y^{k-i} \\ &= \frac{1}{k!} \sum_{i=0}^{k} \left(\begin{array}{c} k \\ i \end{array} \right) x^i\ y^{k-i} = \frac{1}{k!}\ (x+y)^k \end{aligned}$$

so that $p_n(x+y) = \sum_{k=0}^{n} c_k$. Our result will be proved if we could show that

$$\lim_{n\to\infty} p_n(x+y) = \lim_{n\to\infty} (p_n(x), p_n(y)).$$

Now, from the definition

$$c_m = a_0 b_m + a_1 b_{m-1} + \cdots + a_m b_0$$

so that

$$p_n(x+y) = a_0 b_0 + (a_0 b_1 + a_1 b_0) + \cdots + (a_0 b_n + \cdots + a_n b_0).$$

This gives

$$p_n(x+y) = a_0 p_n(y) + a_1 p_{n-1}(y) + \cdots + a_n p_0(y)$$

Defining $\beta_n = p_n(y) - E(y)$ we have

$$\begin{aligned} p_n(x+y) &= a_0(E(y)+\beta_n) + a_1(E(y)+\beta_{n-1}) + \cdots + a_n(E(y)+\beta_0) \\ &= p_n(x)E(y) + \gamma_n \end{aligned}$$

where $\gamma_n = a_0\beta_n + a_1\beta_{n-1} + \cdots + a_n\beta_0.$

As $p_n(x+y) \to E(x+y)$ and $p_n(x) \to E(x)$ we only need to prove $\gamma_n \to 0$ as $n \to \infty$.

Given $\epsilon > 0$, choose N large enough so that $\beta_n < \epsilon$ for all $n \geq N$. Now

$$\begin{aligned} \gamma_n &= a_0\beta_n + \cdots + a_{n-N-1}\beta_{N+1} + a_{n-N}\beta_N + \cdots + a_n\beta_0 \\ &\leq \epsilon(a_0 + \cdots + a_{n-N-1}) + a_{n-N}\beta_N + \cdots a_n\beta_0. \end{aligned}$$

Since $a_0 + \cdots + a_m$ converges to $E(x) \leq e$

$$\gamma_n \leq e\epsilon + a_{n-N}\beta_N + \cdots a_n\beta_0.$$

Fixing N, let n go to infinity. As $a_n \to 0$ as $n \to \infty$ we obtain that $\gamma_n \leq 2e\epsilon$ for large n. Hence $\gamma_n \to 0$.

Remark 2.0.1. The above function $E(x)$ is defined for every $x \in \mathbb{R}$. In fact $p_n(x)$ is Cauchy for every x and hence converges to a function which is denoted by $E(x)$. The above property in the proposition is true for any $x, y \in \mathbb{R}$:

$$E(x)E(y) = E(x+y).$$

We will make use of this function later.

Given any metric space (M, d) which is not complete, we can embed M into a complete metric space. The construction is similar to that of $\mathbb{R}$ from $\mathbb{Q}$. Let $\overline{M}$ denote the set of equivalence classes of Cauchy sequences (x_n) in M, under the relation "$(x_n) \sim (y_n)$ iff $(d(x_n, y_n)) \to 0$." Then it can be shown that $\overline{M}$ can be made into a metric space by defining $\rho(A, B) = \lim_{n\to\infty} d(x_n, y_n)$ where $(x_n) \in A, (y_n) \in B$. Then $(\overline{M}, \rho)$ is a complete metric space.

The incomplete metric space $\mathcal{P}[a, b]$ can be completed as above getting a complete metric space $\overline{\mathcal{P}}[a, b]$. In this particular case we can get a concrete realisation of the completion so we dot not use this abstract approach.

Let $(p_n) \in \mathcal{P}[a, b]$ be a Cauchy sequence. Then for any $x \in [a, b]$, $(p_n(x))$ is a Cauchy sequence in $\mathbb{C}$ which converges. Let us define a function $f : [a, b] \to \mathbb{C}$ by $f(x) = \lim_{n\to\infty} p_n(x)$. We claim that (i) f is bounded and (ii) $\lim_{n\to\infty} \|f - p_n\| = 0$ where $\|g\| = \sup_{a\le x\le b} |g(x)|$ for any bounded function g on $[a, b]$. As (p_n) is Cauchy, there exists N such that $\|p_n - p_m\| \le 1$ for all $n, m \ge N$. Given $x \in [a, b]$, as $p_n(x) \to f(x)$, there exists $N(x) > 0$ such that $|f(x) - p_n(x)| \le 1$ for all $n \ge N(x)$. Taking $N_1 = \max(N, N(x))$ we have

$$|f(x)| \le |f(x) - p_{N_1}(x)| + |p_{N_1}(x)| \le 1 + \|p_{N_1}\|$$

As (p_n) is Cauchy $\|p_n\| \le C$ for a C independent of n. Hence $\|f\| \le 1 + C$.

Also, given $\epsilon > 0$ choose $N > 0$ such that $\|p_n - p_m\| < \frac{1}{2}\epsilon$, for all $n, m \ge N$ and $N(x)$, for $x \in [a, b]$ such that $|f(x) - p_n(x)| < \frac{1}{2}\epsilon,\ n \ge N(x)$. Then for $N_1 = \max(N, N(x))$,

$$|f(x) - p_n(x)| \le |f(x) - p_{N_1}(x)| + |p_{N_1}(x) - p_n(x)| < \epsilon$$

provided $n > N$. As this N is independent of x

$$\sup_{a\le x\le b} |f(x) - p_n(x)| < \epsilon \ \text{ for all } \ n \ge N.$$

The above considerations lead us to define the space $\mathcal{C}[a, b]$ as the set of all functions $f : [a, b] \to \mathbb{C}$ for which there is a sequence (p_n) of polynomials

such that $\|f - p_n\| \to 0$ as $n \to \infty$. In other words, elements of $\mathcal{C}[a,b]$ are functions that can be approximated by polynomials uniformly on $[a,b]$. Obviously $\mathcal{P}[a,b] \subset \mathcal{C}[a,b]$ and the inclusion is proper since $E \in \mathcal{C}[a,b]$ is not a polynomial. We make $\mathcal{C}[a,b]$ into a metric space by defining $d(f,g) = \|f-g\| = \sup_{a\le x\le b} |f(x)-g(x)|$. (Note that $\mathcal{C}[a,b]$ is a vector space over $\mathbb{C}$.)

Theorem 2.0.4. *$\mathcal{C}[a,b]$ is a complete metric space (with respect to the metric d).*

Proof. Let (f_n) be a Cauchy sequence. Then certainly, we can define a function $f(x)$ by $f(x) = \lim_{n\to\infty} f_n(x)$. The theorem will be proved once we show that $f \in \mathcal{C}[a,b]$. First we note that f is bounded. Given $x \in [a,b]$, choose $N(x)$ so that $|f(x) - f_{N(x)}(x)| \le 1$. Then

$$|f(x)| \le |f(x) - f_{N(x)}(x)| + |f_{N(x)}(x)| \le 1 + \sup_n \|f_n\|.$$

As (f_n) is Cauchy, $\sup_n \|f_n\| < \infty$ and hence for any $x \in [a,b]$, $|f(x)| \le 1 + \sup_n \|f_n\|$ or f is bounded.

Our next claim is that $\|f - f_n\| \to 0$ as $n \to \infty$. Given $\epsilon > 0$, choose N so that $\|f_n - f_m\| < \frac{1}{2}\epsilon$ for all $n, m \ge N$. Let $x \in [a,b]$ and choose $N(x)$ so that $|f(x) - f_{N(x)}(x)| < \frac{1}{2}\epsilon$. Then

$$|f(x) - f_n(x)| \le |f(x) - f_{N_1(x)}(x)| + |f_{N_1(x)}(x) - f_n(x)| < \epsilon$$

if $N_1(x) > \max(N, N(x))$ and $n > N$. Thus $\|f - f_n\| < \epsilon$ for $n > N$ which proves the claim.

Finally, for any $k \in \mathbb{N}$ choose f_{n_k} so that $\|f - f_{n_k}\| < 2^{-k-1}$. As $f_{n_k} \in \mathcal{C}[a,b]$, choose a polynomial p_k such that $\|f_{n_k} - p_k\| < 2^{-k-1}$. Then $\|f - p_k\| \le \|f - f_{n_k}\| + \|f_{n_k} - p_k\| < 2^{-k}$ and hence (p_k) converges to f in $\mathcal{C}[a,b]$. This completes the proof. □

Given a function $f : [a,b] \to \mathbb{C}$ it is not easy to determine whether $f \in \mathcal{C}[a,b]$ or not. (How to find a sequence of polynomials (p_n) so that $\|f - p_n\| \to 0$?) It is therefore, desirable to give a different characterisation of elements of $\mathcal{C}[a,b]$ which is relatively easy to verify. One such property is the so called uniform continuity of members of $\mathcal{C}[a,b]$.

If p is a polynomial and if $x, y \in [a,b]$ we know that

$$p(x) - p(y) = (x - y)\ q(x,y),$$

where q is a polynomial in two variables. So we have

$$|p(x) - p(y)| \leq C|x - y|$$

where $C = \sup_{a \leq x \leq b, a \leq y \leq b} |q(x, y)| < \infty$. This implies:

Given $\epsilon > 0$, there exists $\delta > 0$ such that $|p(x) - p(y)| < \epsilon$ whenever $|x - y| < \delta$.

(The above estimate shows that $\delta = C^{-1}\epsilon$ works.) The above property continues to hold for every element of $\mathcal{C}[a, b]$. Indeed, if $f \in \mathcal{C}[a, b]$

$$|f(x) - f(y)| \leq |f(x) - p_n(x)| + |p_n(x) - p_n(y)| + |p_n(y) - f(y)|$$

we only need to choose n first so that $\|f - p_n\| < \frac{\epsilon}{3}$ and then $\delta > 0$ so that $|p_n(x) - p_n(y)| < \frac{\epsilon}{3}$ for $|x - y| < \delta$.

These considerations lead us to a general definition.

Definition 2.0.5. *Let $f : [a, b] \to \mathbb{C}$ be a function. We say that f is uniformly continuous on $[a, b]$ if the above property holds.*

We have seen that every $f \in \mathcal{C}[a, b]$ is uniformly continuous. The converse is also true.

Theorem 2.0.6. *$f \in \mathcal{C}[a, b]$ iff f is uniformly continuous on $[a, b]$.*

This theorem is known as Weierstrass approximation theorem. If f is uniformly continuous on $[a, b]$ then $g(x) = f(a + (b - a)x)$ is uniformly continuous on $[0, 1]$ and hence enough to prove that g can be approximated by polynomials uniformly on $[0, 1]$. Therefore, we can assume $[a, b] = [0, 1]$ to start with.

We give a constructive proof of this theorem. We explicitly construct a sequence p_n of polynomials, depending on f, of course, so that $\|f - p_n\| \to 0$. Let

$$p_n(x) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \; f\left(\frac{k}{n}\right) x^k (1 - x)^{n-k}.$$

These are called Bernstein polynomials. Note that

$$\sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1 - x)^{n-k} = (1 - x + x)^n = 1$$

and hence

$$f(x) - p_n(x) = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \left(f(x) - f\left(\frac{k}{n}\right) \right) x^k (1-x)^{n-k}.$$

Our aim is to show that given $\epsilon > 0$, there exists N so that $|f(x) - p_n(x) < \epsilon$ for all $x \in [0,1]$ and $n \geq N$.

If $|f(x) - f(\frac{k}{n})| < \epsilon$ for all k then the sum will be bounded by $\epsilon \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k} = \epsilon$. But $|f(x) - f(\frac{k}{n})| < \epsilon$ need not be true for all $\frac{k}{n}$ but certainly for those k for which $|x - \frac{k}{n}| < \delta$ where ϵ and δ are related via the definition of uniform continuity, i.e., given $\epsilon > 0$ choose δ such that $|f(x) - f(y)| < \epsilon$ whenever $|x - y| < \delta$. This suggests that we split the sum into two parts: Let $I = \{k : 0 \leq k \leq n, |x - \frac{k}{n}| < \delta\},\ J = \{0, 1, \cdots, n\} \setminus I$.

$$\begin{aligned} |f(x) - p_n(x)| \quad \leq \quad & |\sum_{k \in I} \left(\begin{array}{c} n \\ k \end{array} \right) \left(f(x) - f\left(\frac{k}{n}\right) \right) x^k (1-x)^{n-k}| \\ & + |\sum_{k \in J} \left(\begin{array}{c} n \\ k \end{array} \right) \left(f(x) - f\left(\frac{k}{n}\right) \right) x^k (1-x)^{n-k}|. \end{aligned}$$

Note that

$$|\sum_{k \in I} \left(\begin{array}{c} n \\ k \end{array} \right) \left(f(x) - f\left(\frac{k}{n}\right) \right) x^k (1-x)^{n-k}| \leq \epsilon \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k} = \epsilon.$$

On the other hand

$$\begin{aligned} & |\sum_{k \in J} \left(\begin{array}{c} n \\ k \end{array} \right) \left(f(x) - f\left(\frac{k}{n}\right) \right) x^k (1-x)^{n-k}| \\ & \leq 2\|f\| \sum_{|x - \frac{k}{n}| \geq \delta} \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k} \\ & \leq \frac{2}{\delta^2} \|f\| \sum_{k=0}^{n} \left(x - \frac{k}{n} \right)^2 \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k}. \end{aligned}$$

We now claim

Lemma 2.0.7.

$$\sum_{k=0}^{n} (k - nx)^2 \left(\begin{array}{c} n \\ k \end{array} \right) x^k (1-x)^{n-k} = nx(1-x).$$

Assuming the Lemma for a moment we see that

$$\sum_{k=0}^{n} \left(x - \frac{k}{n}\right)^2 \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k} \leq \frac{x(1-x)}{n} \leq \frac{1}{4n}.$$

Thus

$$|f(x) - p_n(x)| \leq \epsilon + \frac{2}{\delta^2} \|f\| \frac{1}{4n} < 2\epsilon$$

if n is large (independent of x). Hence $\|f - p_n\| < 2\epsilon$ for n large proving the theorem.

We prove the Lemma by brute force calculation:

$$\begin{aligned}
&\sum_{k=0}^{n} (k - nx)^2 \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k} \\
&= \sum_{k=0}^{n} k^2 \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k} - 2nx \sum_{k=0}^{n} k \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k} \\
&\quad + \sum_{k=0}^{n} n^2 x^2 \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k}.
\end{aligned}$$

The last sum is of course n^2x^2. The second one equals, as
$k \begin{pmatrix} n \\ k \end{pmatrix} = \frac{k\ n!}{k!(n-k)!} = \frac{n(n-1)!}{(k-1)!(n-1-(k-1))!}$,

$$\begin{aligned}
&n \sum_{k=1}^{n} \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} x^k (1-x)^{n-1-(k-1)} \\
&= nx \sum_{k=0}^{n-1} \begin{pmatrix} n-1 \\ k \end{pmatrix} x^k (1-x)^{n-1-k} = nx.
\end{aligned}$$

The first sum is, writing $k^2 = k(k-1) + k$ and proceeding as above, given by $n(n-1)x^2 + nx$.

Hence

$$\begin{aligned}
&\sum_{k=0}^{n} (k - nx)^2 \begin{pmatrix} n \\ k \end{pmatrix} x^k (1-x)^{n-k} \\
&= n(n-1)x^2 + nx - 2n^2x^2 + n^2x^2 = nx(1-x).
\end{aligned}$$

We now introduce the space $\mathcal{C}(\mathbb{R})$. We say that $f \in \mathcal{C}(\mathbb{R})$ if $f \in \mathcal{C}[a,b]$ for any $a < b$. That is, f can be approximated by a sequence of polynomials (in the uniform norm) over any interval $[a,b]$. Of course, on different intervals, the sequence may be different.

In view of Weierstrass approximation theorem $f \in \mathcal{C}(\mathbb{R})$ if and only if f is uniformly continuous over every $[a,b]$. If $x \in \mathbb{R}$ and if we fix an interval $[a,b]$ containing x then we know: given $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for all $|y - x| < \delta$. As x varies this δ will also vary; it can be chosen independent of x, for all $x \in [a,b]$, for fixed $a < b$. This leads to the definition of continuity at a point:

We say that f is continuous at a point x if given $\epsilon > 0$ there exists δ depending on x (and ϵ) such that $|f(x) - f(y)| < \epsilon$ for all y with $|y - x| < \delta$.

Thus every uniformly continuous function is continuous at every point but the converse is not true. Examples: any polynomial $p(x)$ whose degree is greater than one, and $E(x)$ are continuous at every $x \in \mathbb{R}$ but they are not uniformly continuous.

Elements of $\mathcal{C}(\mathbb{R})$ are called continuous functions on $\mathbb{R}$. Since we have defined continuity at every point, we can define $\mathcal{C}(A)$ whenever $A \subset \mathbb{R}$: $\mathcal{C}(A)$ is just the set of all $f : A \to \mathbb{C}$ such that f is continuous at every point of A. It is easy to prove the following

Proposition 2.0.8. *If $K \subset \mathbb{R}$ is compact then every $f \in \mathcal{C}(K)$ is uniformly continuous.*

The proof is simple. By definition given $\epsilon > 0$ and $x \in K$ there exists $\delta = \delta(x) > 0$ such that $|f(x) - f(y)| < \frac{\epsilon}{2}$ for $|y - x| < \delta(x)$. But then K is covered by the open sets $(x - \delta(x), x + \delta(x))$ as x varies over K. The compactness of K shows that there are points $x_1, x_2, \cdots x_n \in K$ such that

$$K \subset \bigcup_{j=1}^{n} (x_j - \frac{1}{2}\delta_j, x_j + \frac{1}{2}\delta_j).$$

Choose $\delta = \min\{\delta_j : 1 \leq j \leq n\}$. If $x, y \in K$ and $|x - y| < \delta$ then both $x, y \in (x_j - \delta_j, x_j + \delta_j)$ for some j. Hence

$$|f(x) - f(y)| \leq |f(x) - f(x_j)| + |f(x_j) - f(y)| < \epsilon.$$

If $K \subset \mathbb{R}$ is compact, every $f \in \mathcal{C}(K)$ is bounded (why?) and so we can define $\|f\| = \sup_{x \in K} |f(x)|$ and make $\mathcal{C}(K)$ into a metric space with $d(f,g) = \|f - g\|$. It is easy to prove that $\mathcal{C}(K)$ is a complete metric space.

Can we make $\mathcal{C}(\mathbb{R})$ into a metric space? If $f \in \mathcal{C}(\mathbb{R})$ there is not guarantee that f is bounded. But $f \in \mathcal{C}[-n, n]$ for any n, by definition. Let us define $d_n(f, g) = \sup_{-n \leq x \leq n} |f(x) - g(x)|$ which is just the metric on $\mathcal{C}[-n, n]$. There is another metric on $\mathcal{C}[-n, n]$ equivalent to d_n : that is they define the same family of open sets; the convergence is the same in both metric. The new metric is better in the sense that it is bounded. Simply define

$$\rho_n(f, g) = \frac{d_n(f, g)}{1 + d_n(f, g)}.$$

It is easy to verify all our claims. It is obvious that $\rho_n(f, g) \leq 1$ for any $f, g \in \mathcal{C}[-n, n]$.

Let us use this sequence of metrics to define a metric on $\mathcal{C}(\mathbb{R})$. Let

$$\rho(f, g) = \sum_{n=1}^{\infty} 2^{-n} \rho_n(f, g) = \sum_{n=1}^{\infty} 2^{-n} \frac{d_n(f, g)}{1 + d_n(f, g)}.$$

Then (1) ρ is a metric on $\mathcal{C}(\mathbb{R})$. (2) $(\mathcal{C}(\mathbb{R}), \rho)$ is a complete metric space. (3) $f_n \to f$ in ρ iff $f_n \to f$ in $\mathcal{C}[-m, m]$ for every m. Thus convergence in $(\mathcal{C}(\mathbb{R}), \rho)$ is uniform convergence over compact subsets of $\mathbb{R}$ or simply compact convergence.

It is natural to ask if an analogue of Weierstrass approximation theorem is valid for the spaces $\mathcal{C}(K)$ or $\mathcal{C}(\mathbb{R})$. Stone obtained a far reaching generalisation of Weierstrass approximation theorem proving such a result for $\mathcal{C}(K)$ for any 'compact Hausdorff' space. We restrict ourselves to the case of compact subsets K of $\mathbb{R}$ or $\mathbb{C}$, though the same proof works in the general case.

Stone made the following observations: $\mathcal{C}[a, b]$ is an algebra, i.e., $\mathcal{C}[a, b]$ is a vector space over $\mathbb{C}$ (or $\mathbb{R}$) and $fg \in \mathcal{C}[a, b]$ whenever $f, g \in \mathcal{C}[a, b]$. Here fg is the function defined by $(fg)(x) = f(x)g(x)$. The polynomials $\mathcal{P}[a, b]$ then forms a subalgebra of $\mathcal{C}[a, b]$ and Weierstrass theorem says that the subalgebra $\mathcal{P}[a, b]$ is dense in $\mathcal{C}[a, b]$. He realised that two important properties of $\mathcal{P}[a, b]$ are needed for the proof: (i) $\mathcal{P}[a, b]$ separates points in the sense that given $x, y \in [a, b], x \neq y$, there is a $p \in \mathcal{P}[a, b]$ such that $p(x) \neq p(y)$ (in this case we can simply take $p(x) = x$) (ii) the constant function $f(x) \equiv 1$ belongs to $\mathcal{P}[a, b]$.

He was able to prove the following generalisation. Let $\mathcal{C}(K, \mathbb{R})$ stand for all real-valued continuous functions on K.

Theorem 2.0.9. *(Stone-Weierstrass) Let K be a compact metric space. Let B be any subalgebra of $\mathcal{C}(K, \mathbb{R})$ which satisfies*

(i) *B separates points, i.e., given $x, y \in K, x \neq y$, there is $f \in B$ such that $f(x) \neq f(y)$.*

(ii) *B contains a nonzero constant function.*

Then B is dense in $\mathcal{C}(K, \mathbb{R})$, i.e., every $f \in \mathcal{C}(K, \mathbb{R})$ can be approximated by members of B in the uniform norm (over K).

The proof proceeds in several steps. As $B \subset \mathcal{C}(K, \mathbb{R})$, $f \in B$ implies $|f| \in \mathcal{C}(K, \mathbb{R})$, but $|f|$ need not be in B itself. First we show that $|f|$ can be approximated by elements of B. To this end we apply Weierstrass approximation theorem to the function $g(t) = |t|$ on the interval $[-\|f\|, \|f\|]$ where $\|f\| = \sup_{x \in K} |f(x)|$. We know that there is a sequence $q_n(t)$ of polynomials such that $q_n \to g$ uniformly over $[-\|f\|, \|f\|]$. As $g(0) = 0$, the sequence p_n defined by $p_n(t) = q_n(t) - q_n(0)$ also converges to g. Note that $p_n(0) = 0$ for all n. Given $\epsilon > 0$ we can choose p_n such that $|g(t) - p_n(t)| < \epsilon$ for all $t \in [-\|f\|, \|f\|]$ so that for all $x \in K$, $||f(x)| - p_n(f(x))| < \epsilon$. If $p_n(t) = \sum_{j=1}^{m} a_j \, t^j$, $p_n(f(x)) = \sum_{j=1}^{m} a_j \, f(x)^j$ is an element of the algebra B. Thus $|f|$ is approximated by elements of B.

Next we make the following observation. Define $f \vee g$ and $f \wedge g$ by

$$\begin{aligned} (f \vee g)(x) &= \max\{f(x), g(x)\} \\ (f \wedge g)(x) &= \min\{f(x), g(x)\}. \end{aligned}$$

Then we can easily verify that

$$\begin{aligned} (f \vee g)(x) &= \frac{1}{2}((f + g) + |f - g|)(x) \\ (f \wedge g)(x) &= \frac{1}{2}((f + g) - |f - g|)(x). \end{aligned}$$

This shows that if f and g are in B then both $f \vee g$ and $f \wedge g$ can be approximated by elements of B. The same is true if we consider $f_1 \vee f_2 \vee \cdots \vee f_n$ and $f_1 \wedge f_2 \wedge \cdots \wedge f_n$ for any finite number of functions $f_j \in B$.

We now prove the theorem in the following way. Given $f \in \mathcal{C}(K, \mathbb{R})$ and $\epsilon > 0$ we show that we can find a function g such that $\|f - g\| < \epsilon$. We then show that this g can be approximated by elements of B, proving theorem.

So far we haven't used any hypothesis on B or on K. Let $x, y \in K, x \neq y$. We claim that there exists $g \in B$ such that $g(x) = f(x), g(y) = f(y)$, i.e.,

g agrees with f at the points x and y. Since $x \neq y$ and B separates points there exists $h \in B$ such that $h(x) \neq h(y)$. Consider

$$g(t) = f(x)\frac{h(t) - h(y)}{h(x) - h(y)} + f(y)\frac{h(t) - h(x)}{h(y) - h(x)}$$

Then clearly, $g(x) = f(x)$, $g(y) = f(y)$ and $g \in B$ (as B is an algebra and B contains constant functions).

Let us fix $x \in K$ and let $y \in K$ be different from x. Let f_y be the function constructed above: i.e., $f_y \in B$, $f_y(x) = f(x)$ and $f_y(y) = f(y)$. Consider

$$G_y = \{t \in K : f_y(t) < f(t) + \epsilon\}.$$

Since f_y and f are continuous G_y is open. Further, $x, y \in G_y$ and so $\{G_y : y \in K, y \neq x\}$ is an open cover of K which has a finite subcover say $G_{y_1}, G_{y_2}, \cdots, G_{y_n}$. Let $f_1, f_2, \cdots f_n$ be the corresponding functions. Define

$$g_x = f_1 \wedge f_2 \wedge \cdots \wedge f_n.$$

Then $g_x(x) = f(x)$ and $g_x(t) < f(t) + \epsilon$ for all $t \in K$.

Next consider the open sets

$$H_x = \{t \in K : g_x(t) > f(t) - \epsilon\}.$$

As $x \in H_x$, $\{H_x : x \in K\}$ is an open cover of K which has a finite subcover $H_{x_1}, H_{x_2}, \cdots H_{x_m}$. Let $g_1, g_2, \cdots g_m$ be the corresponding functions and define $g = g_1 \vee g_2 \vee \cdots \vee g_m$. Then clearly $f(t) - \epsilon < g(t) < f(t) + \epsilon$ for all $t \in K$, i.e., $\|g - f\| < \epsilon$.

Finally, as $f_1, f_2, \cdots f_n \in B$, the functions g_x can be approximated by elements of B. Consequently, $g_1 \vee g_2 \vee \cdots \vee g_m = g$ can be approximated by members of B. As g approximates f, we are done.

In the above theorem we have considered only real-valued functions on K. If we consider $\mathcal{C}(K, \mathbb{C})$ then the result is not true without further assumptions on the subalgebra B. Given $f : K \to \mathbb{C}$ define $\overline{f}$ by $\overline{f}(x) = \overline{f(x)}$ where $\overline{a}$ is the complex conjugate of $a \in \mathbb{C}$.

Theorem 2.0.10. *(Stone-Weierstrass) Let K be a compact metric space. Let B be a subalgebra of $\mathcal{C}(K, \mathbb{C})$ which satisfies the following (i) B separates points (ii) B contains a nonzero constant function (iii) $f \in B$ implies $\overline{f} \in B$. Then B is dense in $\mathcal{C}(K, \mathbb{C})$.*

Proof. If $f \in \mathcal{C}(K, \mathbb{C})$ define $Ref(x) = \frac{1}{2}(f(x)+\overline{f}(x))$ and $Imf(x) = \frac{1}{2i}(f(x) - \overline{f}(x))$. Define $A = B \cap \mathcal{C}(K, \mathbb{R})$. Enough to show that A is dense in $\mathcal{C}(K, \mathbb{R})$ because then Ref and Imf can be approximated by elements of A which will then prove the theorem.

We only need to verify that A satisfies the conditions of the real Stone-Weierstrass theorem. Given $x, y \in K$, $x \neq y$, there exists $f \in B$ such that $f(x) \neq f(y)$. Then either $Ref(x) \neq Ref(y)$ or $Imf(x) \neq Imf(y)$. Since $f \in B$ implies $\overline{f} \in B$, both $Ref = \frac{1}{2}(f + \overline{f})$ and $Imf = \frac{1}{2i}(f - \overline{f})$ are in A. Hence A separates points. Also A contains a non-zero constant function. To see this, there exists a nonzero constant function say $g \in B$ by hypothesis. But then $|g|^2 = g\overline{g} \in A$. Hence A is dense in $\mathcal{C}(K, \mathbb{R})$, proving the theorem. □

So far we have considered Stone-Weierstrass theorem for $\mathcal{C}(K)$ where K is compact. It is natural to ask what happens when K is not necessarily compact. We take up only the case $K = \mathbb{R}$ and prove a version of Stone-Weierstrass theorem not for all of $\mathcal{C}(\mathbb{R})$ but for a subspace $\mathcal{C}_0(\mathbb{R})$.

We set up a one to one correspondence between functions in $\mathcal{C}[-1, 1]$ and $\mathcal{C}_0(\mathbb{R})$ and then appeal to the theorem in $\mathcal{C}_0[-1, 1]$ to prove a similar theorem for $\mathcal{C}_0(\mathbb{R})$. We first note that the map $g : (-1, 1) \to \mathbb{R}$ defined by $g(x) = \frac{x}{1-|x|}$ sets up a one to one correspondence between $(-1, 1)$ and $\mathbb{R}$. Using this we can set up a one to one correspondence between functions in $\mathcal{C}(\mathbb{R})$ and functions continuous on $(-1, 1)$, viz., if $f \in \mathcal{C}(\mathbb{R})$ then $\widetilde{f}(x) = f \circ g(x) = f(g(x))$ is a continuous function on $(-1, 1)$. If h denotes the inverse of g, $g \circ h(x) = x = h \circ g(x)$ (which exists as g is one to one and onto) then $\widetilde{f} \circ h = f$ giving the inverse to the map $f \mapsto \widetilde{f}$.

Consider the collection $\{\widetilde{f} : f \in \mathcal{C}(\mathbb{R})\}$. It is not true that every $\widetilde{f}$ defined on $(-1, 1)$ can be extended to a function on $[-1, 1]$. This can be easily seen by taking, for example, $f(x) = e^x$ and noting that $\widetilde{f}(x) = e^{g(x)} = e^{\frac{x}{1-|x|}}$ does not have a limit as $x \to 1$. If $\widetilde{f}$ has such an extension then $\widetilde{f}$ will be an element of $\mathcal{C}[-1, 1]$. It is easy to see what functions f on $\mathbb{R}$ have this property: $\widetilde{f}$ has an extension to $[-1, 1]$ if and only if $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$ both exist. If that is the case we simply define $\widetilde{f}(1) = \lim_{x\to\infty} f(x)$ and $\widetilde{f}(-1) = \lim_{x\to-\infty} f(x)$.

Let $\mathcal{C}_*(\mathbb{R})$ stand for all continuous functions on $\mathbb{R}$ for which $\lim_{x\to\infty} f(x)$ and $\lim_{x\to-\infty} f(x)$ exist. Then $\mathcal{C}_*(\mathbb{R})$ is in one-to-one correspondence with $\mathcal{C}[-1, 1]$. We denote by $\mathcal{C}_0(\mathbb{R})$ the subspace of $\mathcal{C}_*(\mathbb{R})$ consisting of f with

$\lim_{x\to\pm\infty} f(x) = 0$. Functions in $\mathcal{C}_0(\mathbb{R})$ are called continuous functions vanishing at ∞ (for the obvious reason). Some examples are $f(x) = e^{-x^2}$, $f(x) = (1+x^2)^{-1}$ etc. $\mathcal{C}_0(\mathbb{R})$ then corresponds to those $f \in \mathcal{C}[-1,1]$ with $f(\pm 1) = 0$.

It is for this $\mathcal{C}_0(\mathbb{R})$ we can prove a version of Stone-Weierstrass theorem. Note that $\mathcal{C}_0(\mathbb{R})$ is an algebra of continuous functions. Let A be a subalgebra of $\mathcal{C}_0(\mathbb{R})$. Define $\widetilde{A} = \{\widetilde{f} : f \in A\}$ so that $\widetilde{A}$ is a subalgebra of $\mathcal{C}[-1,1]$. If we can find conditions on A so that $\widetilde{A}$ satisfies conditions of the Stone-Weierstrass theorem for $\mathcal{C}[-1,1]$ then A will be dense in $\mathcal{C}_0(\mathbb{R})$.

Assume that A separates points. Then if $x, y \in (-1,1)$, $x \neq y$, we can find a function $\widetilde{f} \in \widetilde{A}$ so that $\widetilde{f}(x) \neq \widetilde{f}(y)$. To see this let $g(x) = x'$, $g(y) = y'$ with $x', y' \in \mathbb{R}$, $x' \neq y'$. Then as A separates points there exists $f \in \mathcal{C}_0(\mathbb{R})$ such that $f(x') \neq f(y')$. This gives $\widetilde{f}(x) = f(x') \neq f(y') = \widetilde{f}(y)$. This argument does not work if either $x = 1$ or -1. If we want $\widetilde{A}$ to separate $x = 1$ and $y \in (-1,1)$ we need to assume further assumption on A viz., given any $y \in \mathbb{R}$ there exists $f \in A$ such that $f(y) \neq 0$. Under this extra condition $\widetilde{f}(y) \neq 0$ and $\widetilde{f}(x) = 0$ so that $x = 1$ and $y \in (-1,1)$ can be separated. Similarly $x = -1$ and $y \in (-1,1)$ can be separated. However $x = 1$ and $y = -1$ can never be separated as $\widetilde{f}(1) = \widetilde{f}(-1) = 0$ for any $f \in A$.

Instead of identifying $\mathcal{C}_0(\mathbb{R})$ with functions in $\mathcal{C}[-1,1]$ vanishing at the end points, let us identify it with $\mathcal{C}(-1,1]$, i.e., continuous functions on $(-1,1]$ which vanish at $x = 1$. But then the compactness of $[-1,1]$ is lost and we cannot appeal to the Stone-Weierstrass theorem known for $\mathcal{C}(K)$ where K is compact. There is a way out.

We define a new metric ρ on $(-1,1]$ which will make $(-1,1]$ into a compact metric space. This ρ has to be defined in such a way that functions in $\mathcal{C}(-1,1]$ are precisely those functions that are continuous with respect to ρ. Let us define $\rho(x,y) = d(1, d(x,y))$ where $d(x,y) = |x-y|$, $x, y \in (-1,1]$. Then we leave it to the reader to verify that this ρ is a metric which makes $(-1,1]$ into compact and the above remark about $\mathcal{C}(-1,1]$ holds. To help the reader we remark that

$$\rho(x,y) = \min\{2 - |x-y|, |x-y|\}.$$

Once this is done we can prove the following version of Stone-Weierstrass theorem for $\mathcal{C}_0(\mathbb{R})$.

Theorem 2.0.11. *(Stone-Weierstrass) Let A be a subalgebra of $\mathcal{C}_0(\mathbb{R})$ which satisfies the following three conditions: (i) A separates points on $\mathbb{R}$ (ii) given*

any $x \in \mathbb{R}$ there exists $f \in A$ such that $f(x) \neq 0$ (iii) $f \in A$ implies $\overline{f} \in A$. Then A is dense in $\mathcal{C}_0(\mathbb{R})$.

Proof. Let $\widetilde{A} = \{\widetilde{f} : f \in A\}$ so that $\widetilde{A}$ is a subalgebra of $\mathcal{C}(-1,1]$ (here $(-1,1]$ is the metric space equipped with the metric ρ so that it is compact). Conditions (i) and (ii) imply that $\widetilde{A}$ separates points in $(-1,1]$. However, $\widetilde{A}$ does not contain any nonzero constant function (why?). To remedy this consider $B = \{\widetilde{f} + \lambda : \widetilde{f} \in A, \lambda \in \mathbb{C}\}$. Then B is a subalgebra which satisfies all the conditions in the Stone-Weierstrass theorem for $\mathcal{C}(-1,1]$.

Given $f \in \mathcal{C}_0(\mathbb{R})$, $\widetilde{f} \in \mathcal{C}(-1,1]$. Let $\epsilon > 0$ be given. Then there exists $\widetilde{h} + \lambda \in B$ ($h \in A, \lambda \in \mathbb{C}$) such that $\sup_{(-1,1]} |\widetilde{f}(x) - \widetilde{h}(x) - \lambda| < \frac{\epsilon}{2}$. Taking $x = 1$ we see that $|\lambda| < \frac{\epsilon}{2}$. Consider for $x \in (-1,1]$, $|\widetilde{f}(x) - \widetilde{h}(x)| \leq |\widetilde{f}(x) - \widetilde{h}(x) - \lambda| + |\lambda| < \epsilon$. This means, $\sup_{x \in \mathbb{R}} |f(x) - h(x)| < \epsilon$. As $h \in A$ our theorem is proved. □

2.1 Compact subsets of $\mathcal{C}[a,b]$

We know that a subset $K \subset \mathbb{R}$ or $\mathbb{C}$ is compact iff it is closed and bounded. We are interested in finding such a characterisation for compact subsets of the metric space $\mathcal{C}[a,b]$. First let us make the definition:

We say that $K \subset \mathcal{C}[a,b]$ is compact if every sequence (f_n) in K has a subsequence (f_{n_k}) which converges in K.

The above definition is usually referred to as sequential compactness of K. The traditional definition of compactness being 'every open cover of K has a finite subcover'. However for a metric space these two notions coincide and so we don't have to bother too much about open covers.

With our experience with subsets of $\mathbb{R}$ or $\mathbb{C}$ we may be tempted to think that $K \subset \mathcal{C}[a,b]$ is compact iff it is closed and bounded. But unfortunately, only half of this statement is true: if K is compact then it is closed and bounded. The converse need not be true.

It is easy to see that every compact K is closed and bounded. If (f_n) is a sequence in K which is Cauchy, then the definition of compactness of K implies the existence of a subsequence (f_{n_k}) of (f_n) and an $f \in K$ such that $f_{n_k} \to f$. But then (f_n) itself should converge (check!) to f which means K is closed. If K is compact and if we assume, if possible, that K is not bounded, then for each $n \in \mathbb{N}$ there exists $f_n \in K$ with $\|f_n\| > n$. But

then this sequence (f_n) cannot have any convergent subsequence (as f_n is unbounded).

We have the following counterexample to show that the converse of the above is not true. Let $K = \{f \in \mathcal{C}[0,1] : \|f\| = 1\}$ and consider $f_n(x) = x^n$. Then $f_n \in K$ for all n. K is certainly closed and bounded (why?) but (f_n) cannot have any convergent subsequence. For, if (f_{n_k}) is a subsequence which converges to f then $f(x) = \lim_{k\to\infty} f_{n_k}(x) = 0$ for $0 \leq x < 1$ and $f(1) = \lim_{k\to\infty} f_{n_k}(1) = 1$ which means that f is not an element of $\mathcal{C}[0,1]$, i.e., (f_{n_k}) cannot converge in $\mathcal{C}[0,1]$.

Compact subsets of $\mathcal{C}[a,b]$ have a property stronger than boundedness. Indeed let K be compact and let $\epsilon > 0$. Let $f_1 \in K$ and consider $B_\epsilon(f_1)$. There may be points of K outside $B_\epsilon(f_1)$. Let f_2 be such a point and consider $B_\epsilon(f_1) \cup B_\epsilon(f_2)$. There may (or may not) be points outside this union. Choose f_{j+1} so that f_{j+1} does not belong to $B_\epsilon(f_1) \cup B_\epsilon(f_2) \cup \cdots \cup B_\epsilon(f_j)$ for each j. We claim that this cannot go on indefinitely because otherwise we would have a sequence (f_j) in K such that $\|f_j - f_k\| > \epsilon$ for any k and j. This will then contradict the compactness of K. Thus for every $\epsilon > 0$, K can be covered by finitely many $B_\epsilon(f_j)$ where $f_j \in K$.

We call this property 'total boundedness'. Thus compactness of K implies K is totally bounded.

An example of K which is not totally bounded is provided by $\{f \in \mathcal{C}[0,1] : \|f\| = 1\}$. In this take f_n defined by

$$f_n(x) = nx,\ 0 \leq x \leq \frac{1}{n};\ f_n(x) = 1,\ \frac{1}{n} \leq x \leq 1.$$

Then for $m > n$, $\|f_n - f_m\| = 1 - \frac{n}{m}$ which shows that K is not totally bounded. We now have the following

Theorem 2.1.1. *$K \subset \mathcal{C}[a,b]$ is compact if and only if K is closed and totally bounded.*

Proof. We only need to show that if K is closed and totally bounded then it is compact. Let (f_n) be any sequence in K. As K is closed enough to show that there exists a Cauchy subsequence of (f_n). As K is totally bounded K is contained in a finite union of balls $B_1(g_j), g_j \in K$. At least one of them should contain infinitely many f_n's from our sequence. Call it (f_{1j}) - which is a subsequence of (f_n) and all of them lie within $B_1(g)$ for some $g \in K$. Take $\epsilon = \frac{1}{2}$ and cover K by finitely many $\epsilon-$balls. Then (f_{1j}) will have a

subsequence, call it (f_{2j}) all of which will be in some $\frac{1}{2}$ ball. Proceeding like this we get subsequences (f_{kj}) all of whose elements lie in a $\frac{1}{k}-$ball.

Now, take the diagonal sequence (f_{kk}) which is a subsequence of (f_n) and by construction $\|f_{kk} - f_{jj}\| < \frac{2}{k}$ if $k < j$. This means that (f_{kk}) is Cauchy proving the theorem. □

Total boundedness of a set $K \subset \mathcal{C}[a,b]$ has the following interesting consequence. We know that every $f \in K$ is uniformly continuous - i.e., given $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for all x, y with $|x - y| < \delta$. But we cannot expect to choose δ independent of $f \in K$ so that $|f(x) - f(y)| < \epsilon$ holds for all $f \in K$ and $|x - y| < \delta$. When K is assumed to be totally bounded such a thing is possible.

In fact, let K be totally bounded so that $K \subset \bigcup_{j=1}^{m} B_{\frac{\epsilon}{3}}(f_j)$ for some $f_j \in K$ and $m \in \mathbb{N}$. Each f_j is uniformly continuous so that we can choose δ_j such that $|x - y| < \delta_j$ implies $|f_j(x) - f_j(y)| < \frac{\epsilon}{3}$. Let us take $\delta = \min\{\delta_j : 1 \leq j \leq m\}$ and consider $|x - y| < \delta$. If $f \in K$ then $f \in B_{\frac{\epsilon}{3}}(f_j)$ for some j and so

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_j(x)| + |f_j(x) - f_j(y)| + |f_j(y) - f(y)| \\ &\leq 2\|f - f_j\| + |f_j(x) - f_j(y)| < \epsilon. \end{aligned}$$

Thus we have shown that if K is totally bounded then δ can be chosen independent of $f \in K$. This warrants the following definition.

A set $K \subset \mathcal{C}[a,b]$ is said to be equicontinuous if given $\epsilon > 0$ there exists $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for all $f \in K$ and for all $|x - y| < \delta$.

A celebrated theorem of Ascoli shows that boundedness together with equicontinuity implies total boundedness for subsets of $\mathcal{C}[a,b]$.

Theorem 2.1.2. *(Ascoli) A subset $K \subset \mathcal{C}[a,b]$ is compact iff it is closed, bounded and equicontinuous.*

Proof. We only need to prove the sufficiency of the three conditions. Let (f_n) be a sequence in K. We have to show that (f_n) has a subsequence which is Cauchy. We make use of the fact that $\mathbb{Q}$, the set of all rationals is countable. Let $(x_n)_1^\infty$ be an enumeration of $\mathbb{Q} \cap [a,b]$. We look at the sequence $(f_n(x_1))$ of complex numbers which is bounded since $\|f_n\| \leq M$ for all n, owing to the fact that K is bounded.

By Bolzano-Weierstrass theorem $(f_n(x_1))$ has a subsequence say $(f_{1j}(x_1))$ which converges. Now look at $(f_{1j}(x_2))$ which is again bounded. Hence there

exists a subsequence $(f_{2j}(x_2))$ which converges. Proceeding like this we get subsequences (f_{kj}) with the following properties: (i) (f_{kj}) is a subsequence of $(f_{k-1,j})$ for any $k \geq 2$ (ii) $(f_{kj}(x))$ converges for $x = x_1, x_2, \cdots, x_k$.

We claim that the diagonal sequence (f_{kk}) is Cauchy in $\mathcal{C}[a,b]$ which will then prove the theorem.

So far we haven't used the equicontinuity of K. Let $\epsilon > 0$ be given. Then there exists $\delta > 0$ such that $|f(x) - f(y)| < \epsilon$ for all $|x - y| < \delta$ and all $f \in K$. With this δ we form an open cover $B_\delta(x_j)$ of $[a,b]$ - this is an open cover since $\{x_j\} = \mathbb{Q} \cap [a,b]$ is dense in $[a,b]$ - and extract a finite subcover. Let

$$[a,b] \subset \bigcup_{j=1}^{n} B_\delta(x_j)$$

If n is large enough we can ensure

$$|f_{kk}(x_j) - f_{mm}(x_j)| < \frac{\epsilon}{3},\ k,m \geq N,\ 1 \leq j \leq n.$$

Finally, given $x \in [a,b]$, there exists j such that $x \in B_\delta(x_j)$ and hence

$$\begin{aligned} |f_{kk}(x) - f_{mm}(x)| \quad \leq \quad & |f_{kk}(x) - f_{kk}(x_j)| + |f_{kk}(x_j) - f_{mm}(x_j)| \\ & + |f_{mm}(x_j) - f_{mm}(x)|. \end{aligned}$$

The middle term is less than $\frac{\epsilon}{3}$ if $k,m > N$ whereas the first and last are less than $\frac{\epsilon}{3}$ each since $|x - x_j| < \delta$. Hence $\|f_{kk} - f_{mm}\| < \epsilon$ for all $k,m \geq N$ and this proves the claim. □

2.2 The space of real analytic functions

Suppose $f \in \mathcal{C}(\mathbb{R})$. Then by definition given any $[a,b] \subset \mathbb{R}$, there exists a sequence of polynomials p_n converging to f uniformly over $[a,b]$. The sequence (p_n) will certainly depend on $[a,b]$ and it will be nice if we can choose a single sequence independent of $[a,b]$. We are interested in functions $f \in \mathcal{C}(\mathbb{R})$ for which such a sequence can be found. Even if we can find a single sequence converging to f uniformly over every compact subset of $\mathbb{R}$ there may not be any simple relation between p_{n+1} and p_n. We therefore impose one more condition on p_n requiring that $p_{n+1}(x) = p_n(x) + a_{n+1}x^{n+1}$. Then the sequence (p_n) is completely determined by a sequence (a_n) of complex numbers: $p_0(x) = a_0,\ p_1(x) = a_0 + a_1 x, \cdots, p_n(x) = \sum_{k=0}^{n} a_k x^k$.

Let us say that $f \in \mathcal{C}^\omega(\mathbb{R};0)$ if there exists a sequence $p_n(x)$ of polynomials of the form $p_n(x) = \sum_{k=0}^{n} a_k\ x^k$ such that $p_n \to f$ in $\mathcal{C}[a,b]$ for every $[a,b] \subset \mathbb{R}$ (or equivalently p_n converges to f uniformly over every compact subset of $\mathbb{R}$). Such functions are called real analytic functions.

Obviously, every polynomial p belongs to $\mathcal{C}^\omega(\mathbb{R};0)$. We also know that $E \in \mathcal{C}^\omega(\mathbb{R};0)$ for which the required sequence p_n is given by $p_n(x) = \sum_{k=0}^{n} \frac{1}{k!}\ x^k$.

It is natural to ask: how to produce real analytic functions. As the above considerations show we have to start with a sequence (a_k) of complex numbers and form the sequence $p_n(x) = \sum_{k=0}^{n} a_k\ x^k$ of polynomials. If we have luck, this sequence may converge uniformly over every compact subset K of $\mathbb{R}$. Then the limit function, will be an element of $\mathcal{C}^\omega(\mathbb{R};0)$.

A simple necessary condition for the sequence (p_n) to converge is that $a_k \to 0$ as $k \to \infty$ but this is not sufficient. Some more conditions need to be imposed if we want (p_n) to converge uniformly over compact subsets.

A sufficient condition can be obtained using the so-called root test for infinite series.

Consider an infinite series $\sum_{k=0}^{\infty} a_k$ of complex numbers. By this we mean the limit of a sequence of the form $s_n = a_0 + a_1 + \cdots + a_n$, called the partial sums of the series. We say that the series $\sum_{k=0}^{\infty} a_k$ converges if (s_n) converges. The series is said to converge absolutely if $\sum_{k=0}^{\infty} |a_k|$ converges.

In order to state the root test we require the notion of lim sup and lim inf. Given a sequence (a_n) which is bounded then by Bolzano-Weierstrass we know that there is at least one subsequence (a_{n_k}) of (a_n) which converges. Let E be the set of all such limits, i.e., $a \in E$ if and only if there is a subsequence (a_{n_k}) of (a_n) which converges to a. If (a_n) is unbounded then depending on whether it fails to be bounded above or below we can also get subsequences tending to $+\infty$ or $-\infty$. If so we include ∞ and $-\infty$ also in E. Let $s^* = \sup E$ and $s_* = \inf E$. We define

$$\limsup a_n = s^*,\ \liminf a_n = s_*.$$

We can now prove the following.

Proposition 2.2.1. *Given a series* $\sum_{k=0}^{\infty} a_k$ *let* $\alpha = \limsup |a_n|^{\frac{1}{n}}$. *Then the series converges absolutely if* $\alpha < 1$; *it diverges if* $\alpha > 1$ *(if* $\alpha = 1$ *we cannot say anything).*

Proof. Let $\alpha < 1$ and choose $\beta, \alpha < \beta < 1$. Then we claim that $|a_n|^{\frac{1}{n}} < \beta$ for all but finitely many values of n. For, otherwise there will be a subsequence (a_{n_k}) with $\beta \leq a_{n_k}$ which will lead to the conclusion that $\limsup a_n \geq \beta > \alpha$ which is not true. Hence the claim is proved and starting from some N we have $|a_n| < \beta^n$. Since the geometric series $\sum_{k=0}^{\infty} \beta^n$ converges so does $\sum_{k=0}^{\infty} |a_k|$.

Let us now assume $\alpha > 1$. We claim that $\alpha \in E$, the set of all limit points of $|a_n|^{\frac{1}{n}}$. To see this take any interval $(\alpha - \delta, \alpha + \delta)$ containing α. Since α is the supremum of E, we can find $s \in E \cap (\alpha - \delta, \alpha + \delta)$. But now s is a limit point and so we can find at least one $|a_n|^{\frac{1}{n}}$ in $(\alpha - \delta, \alpha + \delta)$. As this is true for every δ, we can get a sequence $(|a_{n_k}|^{\frac{1}{n_k}})$ converging to α. This simply means that $\alpha \in E$.

Along such a sequence $|a_{n_k}| > 1$ for infinitely many values of k and hence $\sum_{k=0}^{\infty} a_k$ cannot hope to converge. □

Applying the above proposition we can prove the following theorem.

Theorem 2.2.2. *Given a sequence* (a_k) *consider the power series* $\sum_{k=0}^{\infty} a_k\, x^k$. *Let* α *be defined by* $\alpha = \limsup |a_n|^{\frac{1}{n}}$. *Let* $R = \frac{1}{\alpha}$. *Then the above series converges for every* x, *with* $|x| < R$ *and diverges for* $|x| > R$. *The convergence is uniform over compact subsets of the interval* $(-R, R)$.

Proof. We only need to apply the root test to the series: it converges provided

$$\limsup |a_k\, x^k|^{\frac{1}{k}} < 1$$

(which translates into $|x| < R$) and diverges when $\limsup |a_k\, x^k|^{\frac{1}{k}} > 1$ (which is the same as $|x| > R$). □

If the power series $\sum_{k=0}^{\infty} a_k\, x^k$ converges and if R is as above, then R is called the radius of convergence. We can define a function $f : (-R, R) \to \mathbb{C}$ by $f(x) = \sum_{k=0}^{\infty} a_k\, x^k$. As the series converges uniformly over compact subsets

of $(-R, R)$, $f \in \mathcal{C}(-R, R)$. We say that f has the power series expansion $\sum_{k=0}^{\infty} a_k\ x^k$.

Consider a polynomial $p(x) = \sum_{k=0}^{n} a_k\ x^k$. Given $y \in \mathbb{R}$ we can rewrite the polynomial in the form $p(x) = \sum_{k=0}^{n} b_k\ (x-y)^k$. To do this we write

$$x^k = (x - y + y)^k = \sum_{j=0}^{k} \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j}\ (x-y)^j$$

so that

$$\begin{aligned} p(x) &= \sum_{k=0}^{n} a_k \sum_{j=0}^{k} \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j}\ (x-y)^j \\ &= \sum_{j=0}^{n} \left(\sum_{k=j}^{n} a_k \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j} \right) (x-y)^j. \end{aligned}$$

Thus with $b_j = b_j(y) = \sum_{k=j}^{n} a_k \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j}$, $p(x)$ has the expansion $\sum_{j=0}^{n} b_j(y)(x-y)^j$.

If $f(x)$ is represented by a power series $\sum_{k=0}^{\infty} a_k\ x^k$ over the interval $(-R, R)$ and if $y \in (-R, R)$ we would like to represent $f(x)$ in the form $\sum_{k=0}^{\infty} b_k(y)(x-\ y)^k$. Proceeding as in the case of a polynomial we get

$$\sum_{k=0}^{\infty} a_k\ x^k = \sum_{k=0}^{\infty} a_k \left(\sum_{j=0}^{k} \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j} (x-y)^j \right).$$

If we can change the order of summation we get

$$\sum_{k=0}^{\infty} a_k\ x^k = \sum_{j=0}^{\infty} \left(\sum_{k=j}^{\infty} a_k \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j} \right) (x-y)^j.$$

In order to justify this 'change of order' we require a result on double series.

Theorem 2.2.3. *Consider the series* $\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{ij}$. *Suppose we know that* $\sum_{j=0}^{\infty}|a_{ij}| = b_i < \infty$ *and* $\sum_{i=0}^{\infty} b_i < \infty$. *Then*

$$\sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{ij} = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} a_{ij}.$$

Proof. In order to prove this, let $x_0 = 0, x_k = \frac{1}{k}, k = 1, 2, \cdots$ and consider the metric space $E = \{x_k : k = 0, 1, 2, \cdots\} \subset \mathbb{R}$. Define functions $f_i : E \to \mathbb{C}$ by

$$f_i(x_n) = \sum_{j=0}^{n} a_{ij},\ f_i(x_0) = \sum_{j=0}^{\infty} a_{ij}.$$

Note that f_i are continuous functions on E. We also define

$$g(x) = \sum_{i=0}^{\infty} f_i(x).$$

For each $x \in E$, $|f_i(x)| \leq b_i$ and hence the series defining g converges uniformly over E and hence g is continuous.

On the one hand

$$g(x_0) = \sum_{i=0}^{\infty} f_i(x_0) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} a_{ij}$$

and on the other hand

$$g(x_0) = \lim g(x_n) = \lim \sum_{i=0}^{\infty} f_i(x_n) = \lim \sum_{i=0}^{\infty}\sum_{j=0}^{n} a_{ij}.$$

In the last sum we can change the order of summation:

$$\lim g(x_n) = \lim \sum_{j=0}^{n}\sum_{i=0}^{\infty} a_{ij} = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} a_{ij}.$$

This proves the theorem. □

Returning to our discussion on power series consider the coefficients $b_j(y)$ defined by

$$\begin{aligned} b_j(y) &= \sum_{k=j}^{\infty} a_k \begin{pmatrix} k \\ j \end{pmatrix} y^{k-j} \\ &= \sum_{k=0}^{\infty} a_{k+j} \begin{pmatrix} k+j \\ j \end{pmatrix} y^k. \end{aligned}$$

It is easy to see that the radius of convergence of this series defining $b_j(y)$ is the same as that of $\sum a_k\ x^k$. Hence for any $y \in (-R, R)$ the coefficients are well defined.

We can obtain

$$\sum_{k=0}^{\infty} a_k\ x^k = \sum_{j=0}^{\infty} b_j(y)\ (x-y)^j$$

provided, in view of the above result on double series,

$$\sum_{k=0}^{\infty}\sum_{j=0}^{k} |a_k| \begin{pmatrix} k \\ j \end{pmatrix} |y|^{k-j}\ |x-y|^j \ < \ \infty$$

i.e.,

$$\sum_{k=0}^{\infty} |a_k|\ (|y| + |x-y|)^k \ < \ \infty.$$

This happens precisely when $|x-y| + |y| < R$. Thus we have the following

Theorem 2.2.4. *Suppose $f(x) = \sum_{k=0}^{\infty} a_k\ x^k$ has R as its radius of convergence. Given any $y \in (-R, R)$ we can write*

$$f(x) = \sum_{j=0}^{\infty} b_j(y)\ (x-y)^j$$

for all x satisfying $|x-y| < R - |y|$. Moreover, the coefficients $b_j(y)$ are given by the power series

$$b_j(y) = \sum_{k=j}^{\infty} a_k \begin{pmatrix} k \\ j \end{pmatrix} y^{k-j}.$$

The above theorem says that if $f(x)$ can be represented as $\sum_{k=0}^{\infty} a_k\ (x-a)^k$ converging over an interval $(a-R, a+R)$ then for any y in this interval we

can also represent $f(x)$ as $\sum_{k=0}^{\infty} b_k\ (x-y)^k$ valid for all values of x in a small interval centred at y. This motivates us to make the following definition.

We say that a continuous function f on (a,b) belongs to $\mathcal{C}^\omega(a,b)$ if for any $y \in (a,b)$, f can be represented as $\sum_{k=0}^{\infty} a_k(x-y)^k$ valid for all x in some interval containing y. Members of $\mathcal{C}^\omega(a,b)$ are called real analytic functions on (a,b).

It is convenient to denote the power series expansion of f in the form

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_k\ x^k.$$

If we take y from the interval of convergence of the above series and expand $f(x)$ in the form

$$f(x) = \sum_{j=0}^{\infty} \frac{1}{j!}\ b_j(y)\ (x-y)^j$$

then we get

$$b_j(y) = j! \sum_{k=j}^{\infty} \frac{1}{k!}\ a_k \left(\begin{array}{c} k \\ j \end{array} \right) y^{k-j}$$

which after simplification gives

$$b_j(y) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_{k+j}\ y^k.$$

We have shown earlier that if two polynomials p and q agree over an interval then $p = q$ everywhere. This property is shared by power series: if two power series represent the same function over an interval, they coincide over the whole common interval of convergence. Actually some thing much stronger is true.

Theorem 2.2.5. *Suppose we have two power series* $f(x) = \sum_{k=0}^{\infty} a_k\ (x-a)^k$ *and* $g(x) = \sum_{k=0}^{\infty} b_k\ (x-a)^k$ *both converging in* $|x-a| < r$. *If* $f(x_n) = g(x_n)$ *along a sequence* $x_n \to a$ *then* $a_k = b_k$ *for all* k *and consequently* $f(y) = g(y)$ *for all* y *in* $|x-a| < r$.

Proof. Both f and g are continuous. Hence $f(x_n) = g(x_n)$ for all n gives $f(a) = g(a)$ which means $a_0 = b_0$. Now consider $f_1(x) = (x-a)^{-1}(f(x)- a_0)$, $g_1(x) = (x-a)^{-1}(g(x) - b_0)$. Both are continuous, given by power series and $f_1(x_n) = g_1(x_n)$ for all n. As before this gives, $a_1 = b_1$. Proceeding by induction we get $a_k = b_k$ for all k. □

Consider now a real analytic function $f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_k\ x^k$ with radius of convergence R. Then for any $y \in (-R, R)$ we can also represent f in the form $f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ b_k(y)\ (x-y)^k$ where

$$b_j(y) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_{k+j}\ y^k.$$

This suggests that we introduce an operator $D : \mathcal{C}^\omega(-R, R) \to \mathcal{C}^\omega(-R, R)$ as follows. If $f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_k\ x^k$ we simply define

$$Df(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ a_{k+1}\ x^k.$$

It is clear that $Df \in \mathcal{C}^\omega(-R, R)$ whenever f is in $\mathcal{C}^\omega(-R, R)$. Let $D^2(f) = D(Df)$ and so on. Then

$$b_j(y) = D^j f(y).$$

Let us write $D^0 f(x) = f(x) = b_0(x)$. Then the function $f(x)$ has the expansion

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!}\ D^k f(y)\ (x-y)^k.$$

This series is known as the Taylor series of f centred at y. Note that as $f \in \mathcal{C}^\omega(-R, R)$, the series converges in a neighbourhood of the point y.

The function $D^k f$ is given by the power series

$$D^k f(x) = \sum_{j=0}^{\infty} \frac{1}{j!}\ a_{j+k}\ x^j$$

and is called the k^{th} derivative of f.

Given $f \in \mathcal{C}^\omega(a,b)$ and $y \in (a,b)$ by definition there is a power series expansion

$$f(x) = \sum_{k=0}^{\infty} b_k(y)\ (x-y)^k.$$

It is worthwhile to find a formula for the various coefficients $b_k(y)$ appearing in the above expansion. First of all we note that $b_0(y) = f(y)$ by simply evaluating f at y. We then have

$$f(x) - f(y) = \sum_{k=1}^{\infty} b_k(y)\ (x-y)^k$$

which we can write as $f(x) - f(y) = (x-y)\ g(x,y)$, where $g(x,y) = \sum_{k=0}^{\infty} b_{k+1}(y)\ (x-y)^k$. This g is a continuous function of x in a neighbourhood of y and as $x \to y$ we get

$$\lim_{x\to y} g(x,y) = g(y,y) = b_1(y).$$

In otherwords

$$b_1(y) = \lim_{x\to y} \frac{f(x)-f(y)}{x-y}$$

and as $b_1(y) = Df(y)$ we get the formula $Df(y) = \lim_{x\to y} \frac{f(x)-f(y)}{x-y}$. As we have noted earlier $Df(x)$ is again given by a power series and $D^2f(y) = D(Df)(y)$ etc. We also have

$$D^j f(y) = \lim_{x\to y} \frac{\left(f(x) - \sum_{k=0}^{j-1} \frac{D^k f(y)}{k!}(x-y)^k\right)}{(x-y)^j}.$$

We can write the above as

$$f(x) - \sum_{k=0}^{j-1} \frac{1}{k!}\ D^k f(y)\ (x-y)^k = (x-y)^j\ g_j(x,y)$$

where $g_j(x,y)$ is a continuous function in a neighbourhood of y. If $|g_j(x,y)| \leq C_\delta$ for $|x-y| \leq \delta$ we get

$$|f(x) - \sum_{k=0}^{j-1} \frac{1}{k!}\ D^k f(y)\ (x-y)^k| \leq C_\delta\ |x-y|^j$$

which gives a rate at which the polynomials

$$p_n(x) = \sum_{k=0}^{n} \frac{1}{k!} D^k f(y) \ (x-y)^k$$

converge to the function $f(x)$.

The above discussions lead us to the following definition. We set $\mathcal{C}^0(a,b) = \mathcal{C}(a,b)$, the space of all continuous functions on (a,b). For $m = 1, 2, 3, \cdots$ we define $\mathcal{C}^m(a,b)$ to be the space of all f for which $Df \in \mathcal{C}^{m-1}(a,b)$. Here Df is the function defined by

$$Df(y) = \lim_{x \to y} \frac{f(x) - f(y)}{x - y}, \ y \in (a,b).$$

Thus, $f \in \mathcal{C}^m(a,b)$ if and only if f is 'm times differentiable' and $D^m f$ is continuous. We let $\mathcal{C}^\infty(a,b) = \bigcap_{m=1}^{\infty} \mathcal{C}^m(a,b)$. We note that

$$\mathcal{C}^\omega(a,b) \subset \mathcal{C}^\infty(a,b) \subset \mathcal{C}^m(a,b) \subset \mathcal{C}^0(a,b)$$

and all the inclusions are proper.

The function $f(x) = |x|$ is not differentiable at $x = 0$ though it is obviously continuous. The function f defined by

$$f(x) = e^{-\frac{1}{x^2}}, \ x \neq 0, \ f(0) = 0$$

is a $\mathcal{C}^\infty$ function on $\mathbb{R}$. Nevertheless, the function is not real analytic in any interval containing the origin. This can be directly verified by assuming that there is a power series expansion of the form $\sum a_k \ x^k$ converging to $f(x)$ in a neighbourhood of $x = 0$ and arriving at a contradiction.

Another striking result is the fact that there are continuous functions which are nowhere differentiable. First example of such a function was given by Weierstrass around 1875. He showed that when a is an odd positive integer and $0 < b < 1$ is such that $ab > 1 + \frac{3}{2}\pi$ the function

$$f(x) = \sum_{k=0}^{\infty} b^k \ \cos(a^k \pi x)$$

is nowhere differentiable. Later in 1916 Hardy showed that the above f is nowhere differentiable if $ab \geq 1$. He had also proved that

$$f(x) = \sum_{n=1}^{\infty} \frac{\sin n^2 \pi x}{n^2}$$

is nowhere differentiable. Existence of such functions can be proved by appealing to Baire's category theorem to the complete metric space $\mathcal{C}(\mathbb{R})$.

We can easily construct one nowhere differentiable function. Let $\varphi(x) = |x|$ for $|x| \leq 1$. Extend φ to the whole real line as a $2-$periodic function: $\varphi(x+2) = \varphi(x)$. Let $f(x) = \sum_{n=1}^{\infty} \left(\frac{3}{4}\right)^n \varphi(4^n x)$. Then this f is not differentiable at any point.

To prove this first note that $|\varphi(x)| \leq 1$ so that the above series converges uniformly and hence f is continuous. We also note that

$$|\varphi(s) - \varphi(t)| \leq |s - t|.$$

Fix a real number x and define for each positive integer m, $\delta_m = \pm\frac{1}{2} \cdot 4^{-m}$ where the sign is chosen so that no integer lies between $4^m x$ and $4^m(x+\delta_m)$ (which is possible since $4^m \delta_m = \pm\frac{1}{2}$). Consider

$$\gamma_n = \frac{\varphi(4^n(x+\delta_m)) - \varphi(4^n x)}{\delta_m}.$$

When $n > m$, $4^n \delta_m = 2p$ for some integer p and as φ is $2-$periodic, $\varphi(4^n(x+\delta_m)) = \varphi(4^n x)$ or $\gamma_n = 0$. Thus

$$\frac{f(x+\delta_m) - f(x)}{\delta_m} = \sum_{n=0}^{m} \frac{\varphi(4^n(x+\delta_m)) - \varphi(4^n x)}{\delta_m} \cdot \left(\frac{3}{4}\right)^n.$$

As $|\gamma_n| \leq 4^n$, and $|\gamma_m| = 4^m$ by the choice of δ_m we get

$$\begin{aligned} \frac{|f(x+\delta_m) - f(x)|}{\delta_m} &\geq |\gamma_m| - \sum_{n=0}^{m-1} |\gamma_n| \left(\frac{3}{4}\right)^n \\ &\geq 3^m - \sum_{n=0}^{m-1} 3^n = \frac{1}{2}(3^m + 1). \end{aligned}$$

Letting $m \to \infty$ we see that $\delta_m \to 0$ but the derivative of f doesn't exist at x.

When $f \in \mathcal{C}^\omega(a,b)$, f can be expanded as $f(x) = \sum_{j=0}^{\infty} \frac{1}{j!} D^j f(y)(x-y)^j$ for any $y \in (a,b)$ and the series converges uniformly over an interval containing y. Defining

$$P_{m-1}(x) = \sum_{j=0}^{m-1} \frac{1}{j!} D^j f(y) \, (x-y)^j$$

we can write

$$f(x) = P_{m-1}(x) + g_m(x,y)\ (x-y)^m$$

where $g_m(x,y) = \sum_{j=m}^{\infty} \frac{1}{j!}\ D^j f(y)\ (x-y)^{j-m}$. As g_m is a continuous function, we get $|g_m(x,y)| \leq C_\eta$ in a neighbourhood $|x-y| < \eta$ of y. Thus

$$|f(x) - P_{m-1}(x)| \leq C_\eta\ |x-y|^m,\ |x-y| < \eta.$$

Given $\epsilon > 0$ we can choose $\delta < \eta$ such that $C_\eta\ \delta^m < \epsilon$. Then we have

$$|f(x) - P_{m-1}(x)| < \epsilon,\ |x-y| < \delta.$$

Thus the polynomial $P_{m-1}(x)$ approximates $f(x)$ very well near the point y.

For the above approximation to be valid we only require f to be $\mathcal{C}^m$ near the point y. Indeed we have

Theorem 2.2.6. *(Taylor). Let $f \in \mathcal{C}^m(a,b)$ be a real-valued function and let $y \in (a,b)$. Given $\epsilon > 0$ we can choose $\delta > 0$ such that*

$$|f(x) - \sum_{j=0}^{m-1} \frac{1}{j!}\ D^j f(y)\ (x-y)^j| < \epsilon$$

for all x with $|x-y| < \delta$.

The polynomial $P_{m-1}(x)$ is called the Taylor polynomial centred at y associated to the function $f(x)$.

In order to prove the theorem we claim that given any $x \in (a,b)$ there exists a point ξ lying between x and y such that

$$f(x) - P_{m-1} = \frac{1}{m!}\ D^m f(\xi)\ (x-\xi)^m.$$

Here ξ depends on x (and also on y which is fixed now). Once this claim is proved it is easy to prove the theorem. As $D^m f \in \mathcal{C}(a,b)$ we know that in a neighbourhood $|x-y| < \eta,\ |D^m f(x)| \leq C_\eta$. Hence

$$|f(x) - P_{m-1}(x)| \leq C_\eta\ \frac{1}{m!}\ |x-y|^m$$

which immediately gives the theorem.

To prove the claim we first take up the case $m = 1$ which is called mean value theorem. We actually prove a generalised mean value theorem.

Theorem 2.2.7. *(Mean value theorem). Let $f, g \in \mathcal{C}[a,b]$ be real-valued functions. Assume that f and g are differentiable at every point of (a,b). Then for any $x < y$ in (a,b) there exists $\xi \in (x,y)$ such that*

$$(f(x) - f(y))\ g'(\xi) = (g(x) - g(y))\ f'(\xi).$$

Proof. Consider the function

$$h(t) = (f(x) - f(y))\ g(t) - (g(x) - g(y))\ f(t).$$

This h is continuous on $[a,b]$, differentiable on (a,b) and satisfies $h(x) = h(y)$. If h is a constant throughout $[x,y]$ there is nothing to prove. Otherwise, it either reaches a maximum or minimum at an interior point ξ of $[x,y]$. Let us say $h(\xi) - h(t) > 0$ for all $t \in (\xi - \delta, \xi + \delta)$ for some $\delta > 0$ (i.e., h has a local maximum at $t = \xi$). Then on the one hand

$$\frac{h(t) - h(\xi)}{(t - \xi)} \geq 0 \text{ for } \ t \in (\xi - \delta, \xi)$$

and on the other hand

$$\frac{h(t) - h(\xi)}{(t - \xi)} \leq 0 \text{ for } \ t \in (\xi, \xi + \delta).$$

Therefore, in the limit $0 \leq h'(\xi) \leq 0$, i.e., $h'(\xi) = 0$. This means

$$(f(x) - f(y))\ g'(\xi) = (g(x) - g(y))\ f'(\xi)$$

which proves the theorem. □

If we take $g(t) = t$ we get the usual mean value theorem: $f(x) - f(y) = (x - y)f'(\xi)$. We are now in a position to prove our claim: Define

$$g(t) = f(t) - p_{m-1}(t) - C(t - y)^m$$

where we choose C in such a way that

$$g(x) = f(x) - p_{m-1}(x) - C(x - y)^m = 0.$$

Note that $g(y) = g'(y) = \cdots = g^{m-1}(y) = 0$. Since $g(x) = 0 = g(y)$, by mean value theorem we get a point x_1 in between x and y such that $g'(x_1) = 0$. Now $g'(x_1) = 0 = g'(y)$ and again there exists x_2 such that $g''(x_2) = 0$. Proceeding like this we get a point ξ at which $g^{(m)}(\xi) = 0$. But

$g^{(m)}(t) = f^{(m)}(t) - m!C$ or $C = \frac{1}{m!}\ f^{(m)}(\xi)$ and this is what we wanted to prove.

When $f \in \mathcal{C}^{\omega}(a,b)$ and $y \in (a,b)$ we have $f(x) = \sum_{j=0}^{\infty} f_j(x)$ where $f_j(x) = \frac{1}{j!}\ D^j f(y)\ (x-y)^j$. The series being convergent uniformly in a neighbourhood containing y we can calculate derivatives of f at y by simply taking derivatives of f_j and summing them up. That is for any k,

$$D^k f(y) = \sum_{j=0}^{\infty} D^k f_j(y).$$

(In fact $D^k f_j(y) = 0$ unless $j = k$ and so the above series reduces to just one term). In general if f_j are differentiable and $f = \sum f_j$ it may not be true that $Df = \sum Df_j$. This is equivalent to saying that $f_n \to f$ need not imply $Df_n \to Df$ even if $f_n \to f$ uniformly. A simple example is given by $f_n(x) = \frac{1}{\sqrt{n}} e^{inx}$ which converges to 0 uniformly but f_n' does not converge.

We can consider D as an operator taking $\mathcal{C}^1(a,b)$ into $\mathcal{C}(a,b)$. Suppose we equip $\mathcal{C}^1(a,b)$ with the metric inherited from $\mathcal{C}(a,b)$ so that it is a subspace of $\mathcal{C}(a,b)$. Then we can ask if $D : \mathcal{C}^1(a,b) \to \mathcal{C}(a,b)$ is continuous or not. Unfortunately it is not continuous. Indeed, consider $f_n(x) = x^n$ in $\mathcal{C}^1(-1,1)$ which converges to 0. However, $Df_n(x) = nx^{n-1}$ does not converge in $\mathcal{C}(-1,1)$. To remedy the situation we define a new metric on $\mathcal{C}^1(a,b)$.

Recall that the metric ρ on $\mathcal{C}(a,b)$ is defined by $\rho(f,g) = \sum_{n=1}^{\infty} 2^{-n}\ \frac{d_n(f,g)}{1+d_n(f,g)}$ where $d_n(f,g) = \sup_{x \in [a+\frac{1}{n}, b-\frac{1}{n}]} |f(x) - g(x)|$. Define ρ_1 on $\mathcal{C}^1(a,b)$ by

$$\rho_1(f,g) = \rho(f,g) + \rho(Df, Dg).$$

More generally, we define

$$\rho_m(f,g) = \sum_{j=0}^{m} \rho(D^j f, D^j g)$$

on $\mathcal{C}^m(a,b)$. Then it is easy to see that $\mathcal{C}^{m+1}(a,b) \subset \mathcal{C}^m(a,b)$ and the inclusion is continuous. When we consider $\mathcal{C}^{m+1}(a,b)$ with the metric ρ_{m+1} the operator $D : \mathcal{C}^{m+1}(a,b) \to \mathcal{C}^m(a,b)$ becomes continuous. Actually, we have a slightly stronger result. In the following theorem we give a sufficient condition on a sequence f_n so that $\lim_{n\to\infty} Df_n = D(\lim_{n\to\infty} f_n)$.

Theorem 2.2.8. *Let f_n be a sequence of differentiable functions on (a,b) so that f_n' converges to g uniformly over (a,b). If there is a point $x_0 \in (a,b)$ at which $(f_n(x_0))$ converges, then (f_n) converges uniformly to a function f, which is differentiable and $g = f'$. That is $D(\lim_{n\to\infty} f_n) = \lim_{n\to\infty} Df_n$.*

Proof. First we show that (f_n) is uniformly Cauchy so that there exists f to which it converges. Consider

$$f_n(x) - f_m(x) = (f_n(x) - f_m(x)) - (f_n(x_0) - f_m(x_0)) + (f_n(x_0) - f_m(x_0)).$$

As $(f_n(x_0))$ converges we can choose n and m large so that $|f_n(x_0) - f_m(x_0)| < \frac{\epsilon}{2}$. On the other hand by mean value theorem

$$|(f_n(x_0) - f_m(x)) - (f_n(x_0) - f_m(x_0))| \leq |x - x_0||f_n'(\xi) - f_m'(\xi))|$$

which converges to 0 uniformly as $n, m \to \infty$ since f_n' converges to g uniformly.

Consider now

$$\varphi_n(t) = \frac{f_n(t) - f_n(x)}{t - x},\ \varphi(t) = \frac{f(t) - f(x)}{t - x}.$$

Then we have $\lim_{t\to x} \varphi_n(t) = f_n'(x)$ which converges to $g(x)$ i.e., $\lim_{n\to\infty} \lim_{t\to x} \varphi_n(t) = g(x)$. We need to show that

$$\lim_{n\to\infty} \lim_{t\to x} \varphi_n(t) = \lim_{t\to x} \lim_{n\to\infty} \varphi_n(t)$$

which will prove the theorem as $\varphi_n(t) \to \varphi(t)$ for $t \neq x$. We claim that $\varphi_n \to \varphi$ uniformly on $(a,b) \setminus \{x\}$. Indeed,

$$\varphi_n(t) - \varphi_m(t) = \frac{f_n(t) - f_m(t) + f_m(x) - f_n(x)}{t - x}$$

gives by mean value theorem

$$|\varphi_n(t) - \varphi_m(t)| \leq |f_n'(\xi) - f_m'(\xi)| < \epsilon$$

if n, m are large. Thus the convergence of (φ_n) to φ is uniform in $(a,b)\setminus\{x\}$. We complete the proof by appealing to the following proposition. □

Proposition 2.2.9. *Suppose (f_n) is a sequence of functions converging to a function f uniformly on a set E. Let x be a limit point of E. Let $\lim_{t\to x} f_n(t) = A_n$. Then (A_n) converges and $\lim_{n\to\infty} A_n = \lim_{t\to x} f(t)$, i.e., $\lim_{t\to x} \lim_{n\to\infty} f_n(t) = \lim_{n\to\infty} \lim_{t\to x} f_n(t)$.*

Proof. It is easy to see that (A_n) is Cauchy. Indeed, given $\epsilon > 0$ there exists N such that $n, m \geq N$ implies

$$|f_n(t) - f_m(t)| \leq \epsilon \text{ for all } t \in E.$$

Letting $t \to x$ we get $|A_n - A_m| \leq \epsilon$ for $n, m \geq N$. Hence there exists A such that $A_n \to A$. To show that $A = \lim_{t \to x} f(t)$ we write

$$|A - f(t)| \leq |f(t) - f_n(t)| + |f_n(t) - A_n| + |A_n - A|.$$

The first and last terms are less than $\frac{\epsilon}{3}$ each if n is large enough. Then $|f_n(t) - A_n| < \frac{\epsilon}{3}$ if t is close to x. This proves the proposition. □

We complete our discussion on differentiable functions by proving inverse function theorem for functions of one variable.

Theorem 2.2.10. *Let $f : (a, b) \to (c, d)$ be a differentiable function which is onto. If f' never vanishes on (a, b), then f has an inverse g which is also differentiable.*

If $f : [a, b] \to \mathbb{R}$ is a continuous function which is one to one, then f has an inverse $g : E \to [a, b]$ where E is the range of f, i.e., $E = f([a, b])$. In this case it is easy to see that g is continuous.

If g is not continuous at some point $y \in E$, then there will be a sequence $y_n \in E$ which converges to y but $g(y_n)$ will not converge to $g(y)$. Let $x_n = g(y_n)$, $x = g(y)$ so that $f(x_n) = y_n$, $f(x) = y$. Since x_n does not converge to x, there exists $\delta > 0$ and a subsequence (x_{n_k}) such that $|x_{n_k} - x| > \delta$ for all k. As $(x_{n_k}) \in [a, b]$ it will have a convergent subsequence converging to some point x_0. But then by continuity of f, $f(x_0)$ which is the limit of a subsequence of (y_n) should be equal to $y = f(x)$. By one to one-ness, $x = x_0$ contradicting that $|x_{n_k} - x| > \delta$.

The result is true even if we assume that $f : (a, b) \to (c, d)$ is continuous and bijective. To show that g is continuous we proceed as above. There are two cases to consider. If f is bounded on (a, b) then f is defined at a and b (since f is monotone on (a, b) we can extend the definition). In this case the above argument goes through. If f is unbounded, say $\lim_{t \to b} f(t) = \infty$, then as above we will get a subsequence $\widetilde{x_{n_k}}$ along which f will be unbounded. This will contradict the fact that $f(\widetilde{x_{n_k}})$ which converges to y is bounded.

Coming to the proof of the inverse function theorem consider $f : (a, b) \to (c, d)$ for which $f'(x) \neq 0$ for all $x \in (a, b)$. Then by intermediate value

theorem for derivative we know that either $f'(x) > 0$ or $f'(x) < 0$ for all x. In either case f is one to one and hence it has an inverse. Let $f : (c, d) \to (a, b)$ be the inverse such that $g(f(x)) = x,\ x \in (a, b)$ and $f(g(y)) = y,\ y \in (c, d)$. We already know that g is continuous.

To show that g is differentiable at $y = f(x)$ consider $g(y + k) - g(y)$. Define h by $y + k = f(x + h)$. We then have

$$g(y + k) - g(y) = x + h - x = h$$

and $k = f(x + h) - f(x)$. Therefore,

$$\frac{g(y + k) - g(y)}{k} = \frac{h}{f(x + h) - f(x)} = \left(\frac{f(x + h) - f(x)}{h}\right)^{-1}.$$

As $k \to 0$, h also tends to 0 (by continuity of g) and hence $g'(y)$ exists and equals $f'(x)^{-1}$. This proves the theorem.

2.3 The exponential and trigonometric functions

Recall that the function $e(x)$ defined by the power series $\sum_{k=0}^{\infty} \frac{1}{k!}\, x^k$ satisfies the equation $e(x)e(y) = e(x+y)$. It follows either from this equation or from the above equation that $De(x) = e(x)$. It can be shown that it is the unique function, upto a constant multiple, satisfying $Du(x) = u(x)$ for all $x \in \mathbb{R}$.

The above power series converges uniformly over compact subsets of $\mathbb{R}$. It can be easily checked that $e(x)$ can be extended to $\mathbb{C}$ simply by defining

$$e(z) = \sum_{k=0}^{\infty} \frac{1}{k!}\, z^k,\ z \in \mathbb{C}.$$

The power series with x replaced by z also converges uniformly on every compact subset of $\mathbb{C}$. Further $e(z)e(w) = e(z + w)$ continues to hold even for complex z, w.

Of particular interest is the function $e(ix), x \in \mathbb{R}$. We can write this as $e(ix) = c(x) + is(x)$ where

$$c(x) = \sum_{k=0}^{\infty} (-1)^k\, \frac{x^{2k}}{(2k)!},\ s(x) = \sum_{k=0}^{\infty} (-1)^k\, \frac{x^{2k+1}}{(2k+1)!}.$$

Since $\overline{e(z)} = e(\overline{z})$ as can be easily verified it follows that $|e(ix)|^2 = c(x)^2 + s(x)^2 = 1$. We now show that the functions $c(x)$ and $s(x)$ are periodic and consequently $e(z)$ is also periodic.

Theorem 2.3.1. *There exists a positive real number π such that $e(z+2\pi i) = e(z)$ for all $z \in \mathbb{C}$.*

Proof. We show that $c(x)$ has a zero on $(0,\infty)$ and by taking x_0 to be the smallest such zero and defining $\pi = 2x_0$ we prove our theorem. To show that $c(x)$ has zeros on $(0,\infty)$ assume the contrary. As can be easily checked $Dc(x) = -s(x)$ and $Ds(x) = c(x)$. Since $c(0) = 1$, it follows that $c(x) > 0$ for all $x > 0$ and hence $s(x)$ is strictly increasing on $(0,\infty)$. By mean value theorem, for $0 < x < y$

$$c(x) - c(y) = -s(\xi)(x-y),\ \xi \in (x,y)$$

and by the monotonicity of $s(x)$ and the boundedness of $c(x)$ we get

$$s(x)(y-x) < c(x) - c(y) \leq 2$$

for all $y > x$ which is clearly not possible. Hence $c(x)$ has positive zeros and as $c(0) = 1$, there is a smallest zero, say x_0, which is positive.

It then follows that $c(\frac{\pi}{2}) = 0$ and hence $s(\frac{\pi}{2})^2 = 1$. As $s(x) > 0$ on $(0, \frac{\pi}{2})$ we have $s(\frac{\pi}{2}) = 1$ and so $e(i\frac{\pi}{2}) = i$. Consequently, $e(i\pi) = 1$ and $e(2\pi i) = 1$. This proves that $e(z + 2\pi i) = e(z)$. By the choice of $\frac{\pi}{2}$ we see that 2π is the smallest positive real number with this property. □

Corollary 2.3.2. *The map $t \mapsto e(it)$ is a bijection from $[0, 2\pi)$ onto $S^1 = \{z \in \mathbb{C} : |z| = 1\}$.*

Proof. When $z \in S^1$ with $z = x + iy, x \geq 0, y \geq 0$ we can find $t \in [0, \frac{\pi}{2}]$ such that $c(t) = x$ by the continuity of c. As $x^2 + y^2 = 1$ it follows that $s(t) = y$ and $e(it) = x + iy$. When $x < 0$ and $y \geq 0$, we can find $t \in [0, \frac{\pi}{2}]$ so that $-iz = y - ix = e(it)$. As $i = e(\frac{\pi}{2}i)$ it follows that $e(i(t + \frac{\pi}{2})) = ie(it) = z$. By similar arguments we can handle the remaining cases. □

The surjectivity of the map $t \mapsto e(it)$ shows that the equation $z^n = \alpha$ has a solution whenever $\alpha \in S^1$ and $n = 1, 2, \cdots$. Indeed, simply take $z = e(\frac{it}{n})$ where t is chosen in $[0, 2\pi)$ so that $e(it) = \alpha$. With this observation the fundamental theorem of algebra stands completely proved.

2.4 Functions of several variables

So far we have considered real (or complex) valued functions defined on $\mathbb{R}$. Now we will consider functions defined on subsets of $\mathbb{R}^n$ taking values in $\mathbb{R}^m$. Let us quickly set up notation.

By $\mathbb{R}^n$ we mean the cartesian product $\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ (n times) the elements of which are n-tuples of real numbers. We define $x + y = (x_1 + y_1, \cdots, x_n + y_n)$ if $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ are from $\mathbb{R}^n$. With this addition $\mathbb{R}^n$ is an abelian group. The scalar multiplication $\lambda x,\ \lambda \in \mathbb{R}$ is defined by $\lambda x = (\lambda x_1, \cdots, \lambda x_n)$. Thus $\mathbb{R}^n$ becomes a vector space. The standard basis for $\mathbb{R}^n$ is given by the vectors $e_j = (0, \cdots, 1, \cdots, 0)$ (where 1 is in the j^{th} place), $j = 1, 2, \cdots, n$.

For $x \in \mathbb{R}^n$ we define its Euclidean norm $|x|$ to be the positive square root of $\sum_{j=1}^{n} x_j^2$. Then $d(x, y) = |x - y|$ defines a metric which makes $\mathbb{R}^n$ into a complete metric space. The space $\mathcal{C}(\mathbb{R}^n; \mathbb{C})$ of complex valued functions defined on $\mathbb{R}^n$ can be made into a complete metric space by defining

$$\rho(f, g) = \sum_{m=1}^{\infty} 2^{-m} \frac{d_m(f, g)}{1 + d_m(f, g)}$$

where $d_m(f, g) = \sup_{|x| \leq m} |f(x) - g(x)|$. We can also consider the space $\mathcal{C}(\mathbb{R}^n, \mathbb{R}^m)$ of continuous functions $f : \mathbb{R}^n \to \mathbb{R}^m$.

We are interested in extending the notion of derivative for functions $f : \mathbb{R}^n \to \mathbb{R}^m$. When $n = 1$, it is very easy to do this. Suppose $f : \mathbb{R} \to \mathbb{R}^m$ is a function taking values in $\mathbb{R}^m$. Then defining $f_j(t) = (P_j \circ f)(t)$ where $P_j : \mathbb{R}^m \to \mathbb{R}$ is the projection of $\mathbb{R}^m$ onto the j^{th} coordinate space we have $f(t) = (f_1(t), \cdots, f_m(t))$. It is therefore natural to define

$$Df(t) = (Df_1(t), \cdots, Df_m(t))$$

so that Df is again a function defined on $\mathbb{R}$ taking values in $\mathbb{R}^m$. It is clear that if $f : \mathbb{R} \to \mathbb{R}^m$ is differentiable at a point t, then it is continuous there.

The situation changes drastically if we consider functions of several variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real valued function. For each $j = 1, 2, \cdots, n$ we can consider f as a function of x_j, keeping the other variables fixed. The derivative with respect to the j^{th} variable is denoted by $\frac{\partial}{\partial x_j} f$ or $\partial_j f$. Thus

$$\partial_j f(x) = \lim_{t \to 0} \frac{f(x + te_j) - f(x)}{t}$$

where $e_j = (0, \cdots, 1, 0, \cdots, 0)$ are the co-ordinate vectors. Contrary to the expectations, the existence of partial derivatives $\partial_j f(x), j = 1, 2, \cdots, n$ does not guarantee the continuity of the function at x. For example, if $f(x, y) =$

$xy(x^2+y^2)^{-1},\ (x,y) \in \mathbb{R}^2, (x,y) \neq (0,0)$ and $f(0,0) = 0$, then $\partial_1 f(0)$ and $\partial_2 f(0)$ both exist. However, it is clear that f is not continuous at 0.

As the following proposition shows, if $\partial_j f(x)$ exist for all $x \in E$ and if they are all bounded then f becomes continuous.

Proposition 2.4.1. *Let $E \subset \mathbb{R}^n$ be open, $f : E \to \mathbb{R}$ be such that $\partial_j f(x)$ exist and bounded on E for $j = 1, 2, \cdots, n$. Then f is continuous on E.*

Proof. We just make use of the mean value theorem in one variable. Let us set $v^0 = 0$ and define $v^k = \sum_{j=1}^{k} v_j\ e_j$ where $v = (v_1, \cdots, v_n)$. Note that $\|v^k\| \leq \|v\|$ for $k = 1, 2, \cdots, n$. Consider $f(x+v) - f(x)$ which can be written as

$$f(x+v) - f(x) = \sum_{j=1}^{n} f(x+v^j) - f(x+v^{j-1}).$$

Since $v^j = v^{j-1} + v_j e_j$, in view of mean value theorem applied to f as a function of the j^{th} variable we have

$$f(x+v^j) - f(x+v^{j-1}) = v_j\ \partial_j f(x + v^{j-1} + \theta_j\ v_j\ e_j)$$

for some $0 \leq \theta_j \leq 1$. As we are assuming that $|\partial_j f(y)| \leq C$ for all $y \in E$, we get

$$|f(x+v) - f(x)| \leq C \sum_{j=1}^{n} v_j,$$

which clearly shows that f is continuous. □

In the above proof, we observe that if $\partial_j f$ are further assumed to be continuous then we can write

$$\partial_j f(x + v^{j-1} + v_j\ \theta_j\ e_j) = \partial_j f(x) + r_j(x, v).$$

Therefore,

$$f(x+v) - f(x) - \sum_{j=1}^{n} \partial_j f(x) v_j = \sum_{j=1}^{n} v_j\ r_j(x, v),$$

where $r_j(x, v) \to 0$ as $v \to 0$. Note that we can consider $\sum_{j=1}^{n} \partial_j f(x) v_j$ as the image of v under a linear transformation. Thus defining $Df(x) : \mathbb{R}^n \to \mathbb{R}$

by $Df(x)y = \sum_{j=1}^{n} \partial_j f(x)\ y_j$ we have

$$f(x+v) - f(x) - Df(x)v = \sum_{j=1}^{n} v_j\ r_j(x,v)$$

where $\|v\|^{-1} \sum_{j=1}^{n} v_j\ r_j(x,v) \to 0$ as $v \to 0$. This motivates us to make the following definition.

Let $E \subset \mathbb{R}^n$ be open and $f : E \to \mathbb{R}^m$. We say that f is differentiable at $x \in E$ if there is a linear transformation $Df(x) : \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(x+v) - f(x) - Df(x)v = r(x,v)$$

where $\|v\|^{-1}\ r(x,v) \to 0$ as $v \to 0$. First of all we observe that if $Df(x)$ is defined then it is unique. Infact if $A : \mathbb{R}^n \to \mathbb{R}^m$ is another linear transformation such that

$$f(x+v) - f(x) - Av = r_{_A}(x,v)$$

with $\|v\|^{-1} r_{_A}(x,v) \to 0$ as $v \to 0$, then

$$r(x,v) - r_{_A}(x,v) = (A - Df(x))v.$$

Therefore, $\|v\|^{-1}(A - Df(x))v \to 0$ as $v \to 0$. Since $A - Df(x)$ is a linear transformation, the above is possible only when $(A - Df(x))v = 0$ for all $v \in \mathbb{R}^n$, that is, $A = Df(x)$.

Let $\widetilde{e_j}$ be the coordinate vectors in $\mathbb{R}^m$. Then we can write f as $f(x) = \sum_{i=1}^{m} f_i(x)\ \widetilde{e_i}$ where $f_i : E \to \mathbb{R}$ are the components of f. With this notation we have, under the assumption that $\partial_j f_i$ are all continuous on E,

$$f(x+v) - f(x) = \sum_{i=1}^{m} (f_i(x+v) - f_i(x))\ \widetilde{e_i}$$

which is equal to

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{n} \partial_j f_i(x) v_j \right) \widetilde{e_i} + \sum_{i=1}^{m} \left(\sum_{j=1}^{n} v_j\ r_{_{ij}}(x,v) \right) \widetilde{e_i}.$$

If we let $Df(x)$ to stand for the linear transform which sends y into

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{n} \partial_j f_i(x) y_j \right) \widetilde{e_i}$$

then, the above gives

$$f(x+v) - f(x) - Df(x)v = r(x,v)$$

where the remainder, given by

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{n} v_j \; r_{ij}(x,v) \right) \widetilde{e_i},$$

satisfies $\|v\|^{-1} \; r(x,v) \to 0$ as $v \to 0$.

Let us say that $f \in \mathcal{C}^1(E)$ if $Df(x)$ exists at every $x \in E$ and is a continuous function from E into $L(\mathbb{R}^n, \mathbb{R}^m)$, the space of all linear transformations from $\mathbb{R}^n$ into $\mathbb{R}^m$. This space can be given the metric $d(T,S) = \|T - S\|$ where $\|T\|$ is the norm defined by

$$\|T\| = \sup_{\|x\| \leq 1} \|Tx\|.$$

We have almost proved the following result.

Theorem 2.4.2. *A function $f : E \subset \mathbb{R}^n \to \mathbb{R}^m$, belongs to $\mathcal{C}^1(E)$ if and only if $\partial_j f_i, \; 1 \leq i \leq m, 1 \leq j \leq n$ are all continuous on E.*

Proof. Note that the $m \times n$ matrix $(\partial_j f_i(x))$ defines a linear transformation from $\mathbb{R}^n$ into $\mathbb{R}^m$. If all the partial derivatives $\partial_j f_i$ are continuous, the above calculations preceding the theorem show that $Df(x)$ is given by the matrix $(\partial_j f_i(x))$. And hence f is differentiable and the continuity of $x \mapsto Df(x)$ follows from that of $\partial_j f_i$.

To prove the converse we observe that

$$f(x + te_j) - f(x) - Df(x)te_j = r(x,t)$$

which gives

$$\sum_{i=1}^{m} \left(f_i(x + te_j) - f_i(x) - t(Df(x)e_j, \widetilde{e_i}) \right) \widetilde{e_i} = \sum_{i=1}^{m} (r(x,t), \widetilde{e_i}) \widetilde{e_i}.$$

The existence of $Df(x)$ shows that

$$\lim_{t \to 0} \frac{f_i(x + te_j) - f_i(x)}{t} = (Df(x)e_j, \widetilde{e_i})$$

which means $\partial_j f_i(x)$ exists and is given by $(Df(x)e_j, \widetilde{e_i})$. This proves the theorem completely. □

Chapter 3

The space of integrable functions

3.1 Riemann integral of continuous functions

Consider the differential operator $D : \mathcal{C}^\omega(a,b) \to \mathcal{C}^\omega(a,b)$. This operator is not one to one as it kills all constant functions. However, it is onto: this means given $f \in \mathcal{C}^\omega(a,b)$ the differential equation $Du = f$ has a solution. Indeed, it has infinitely many solutions since $u + C, C \in \mathbb{C}$ is a solution whenever u is. If $f(x) = \sum_{j=0}^{\infty} a_j \ (x-y)^j$ is the expansion of f around $y \in (a,b)$ then a solution of the above equation is given by

$$u(x) = Sf(x) = \sum_{j=0}^{\infty} \frac{a_j}{j+1} \ (x-y)^{j+1}.$$

Note that this particular solution has the additional property that $u(y) = 0$. This means that the initial value problem

$$Du = f, \ u(y) = 0$$

has the unique solution

$$u(x) = \sum_{j=0}^{\infty} \frac{a_j}{j+1} \ (x-y)^{j+1},$$

whenever f is given by the expansion

$$f(x) = \sum_{j=0}^{\infty} a_j \ (x-y)^j.$$

We are interested in solving the above initial value problem when f is merely a continuous function. That is we are interested in finding a $\mathcal{C}'$ function such that $Du = f$. This can be done on every closed interval $I \subset (a, b)$ which contains y.

When $f = p$ is a polynomial then the function Sp is again a polynomial: if $f(x) = \sum_{k=0}^{n} a_k \ x^k$, then $Sp(x) = \sum_{k=0}^{n} \frac{a_k}{k+1} \ x^{k+1}$. Let us call this $P(x)$. Then the unique solution of $Du = p, \ u(y) = 0$ is given by

$$u(x) = P(x) - P(y).$$

To proceed further let us prove an elementary

Lemma 3.1.1. *Let I be a closed interval containing y. Then there exists $C = C_I > 0$ such that*

$$\|S_I \ p\| \leq C_I \|p\|$$

where $S_I \ p$ is the unique solution of $Du = p, \ u(y) = 0$ and $\|\cdot\|$ is the norm in $\mathcal{C}(I)$.

Proof. As observed above,

$$S_I \ p(x) = P(x) - P(y) = P'(\xi)(x-y)$$

by mean value theorem. (Here ξ lies between x and y). Since $P'(\xi) = p(\xi)$ we get

$$|S_I \ p(x)| \leq |x-y||p(\xi)| \leq (\beta - \alpha)\|p\|$$

for all x where $I = [\alpha, \beta]$. This proves the lemma. □

If $f \in \mathcal{C}(I)$ then by definition there exists a sequence of polynomials (p_n) such that $p_n \to f$ in $\mathcal{C}(I)$. In view of the above lemma, if $u_n = S_I \ p_n$, we have

$$\|u_n - u_m\| \leq C_I \|p_n - p_m\|$$

and hence (u_n) converges to a continuous function u on I. As $Du_n = p_n$ converges uniformly on I, by a theorem we have proved in the previous

section we get u is differentiable and $Du = \lim_{n\to\infty} Du_n = f$. It is obvious that $u(y) = 0$. Thus $u = \lim_{n\to\infty} u_n$ is the unique solution of the initial value problem $Du = f,\ u(y) = 0$.

The result of the previous lemma remains true for all continuous functions.

Proposition 3.1.2. *Let $f \in \mathcal{C}(I)$ and $y \in I$. Then the solution u of the initial value problem satisfies $\|u\| \leq C\|f\|$.*

Proof. Take a sequence (p_n) of polynomials converging to u uniformly. Since $\|u_n\| \leq C\|p_n\|$, we can choose N large enough so that

$$|u(x) - u_n(x)| \leq \frac{\epsilon}{2},\ \|p_n - f\| \leq \frac{\epsilon}{2}$$

for $n \geq N$. We have

$$\begin{aligned} |u(z)| &\leq |u(x) - u_n(x)| + |u_n(x)| \\ &\leq |u(x) - u_n(x)| + C\|p_n - f\| + C\|f\| \\ &\leq (C+1)\epsilon + C\|f\|. \end{aligned}$$

As $\epsilon > 0$ is arbitrary we get the result. □

We now make the following definition: If f is continuous on $I = [\alpha, \beta]$ we define

$$\int_\alpha^x f = u(x),\ x \in I$$

to be the unique solution of the problem

$$Du = f,\ u(\alpha) = 0.$$

$u(\beta) = \int_\alpha^\beta f$ is called the Riemann integral of the continuous function f over the interval I. We use the alternative notation $\int_\alpha^\beta f(t)\ dt$ more often.

Here are some important properties of $\int_\alpha^\beta f(t)\ dt$.

(1) The integral is linear as a function of f :

$$\begin{aligned}\int_\alpha^\beta (f+g)(t)\ dt &= \int_\alpha^\beta f(t)\ dt + \int_\alpha^\beta g(t)\ dt \\ \int_\alpha^\beta \lambda f(t)\ dt &= \lambda \int_\alpha^\beta f(t)\ dt\end{aligned}$$

for $f, g \in \mathcal{C}(I),\ \lambda \in \mathbb{C}$.

(2) The integral preserves nonnegativity of f : that is

Proposition 3.1.3. $\int_\alpha^\beta f(t)\ dt \geq 0$ *whenever* $f \geq 0$.

Proof. Look at $u(x) = \int_\alpha^x f(t)\ dt$ which satisfies $Du(x) = f(x),\ u(\alpha) = 0$. As $f(x) \geq 0$, u is increasing and as $u(\alpha) = 0,\ u(x) \geq 0$ for all $x \in I$. This proves the proposition. □

(3) Fundamental theorem of Calculus

Theorem 3.1.4. *If $f \in \mathcal{C}(I)$ is the derivative of another function, say F then*

$$\int_\alpha^\beta f(t)\ dt = F(\beta) - F(\alpha).$$

Proof. By definition $u(x) = \int_\alpha^x f(t)\ dt$ is the unique solution of $Du = f$, $u(y) = 0$. As $v(x) = F(x) - F(\alpha)$ also solves the same equation we get $u = v$. Hence

$$\int_\alpha^\beta f(t)\ dt = u(\beta) = v(\beta) = F(\beta) - F(\alpha)$$

as claimed. □

(4)

Theorem 3.1.5. *For every $f \in \mathcal{C}(I)$,*

$$|\int_I f(t)\ dt| \leq \int_I |f(t)|\ dt.$$

Proof. First assume f is real valued. Choose $c = \pm 1$ in such a way that

$$c\int_I f(t)\ dt = |\int_I f(t)\ dt|.$$

Since $\int_I cf(t)\ dt = c\int_I f(t)\ dt$ we have

$$|\int_I f(t)\ dt| = \int_I cf(t)\ dt.$$

and $cf \leq |f|$ or $|f| - cf \geq 0$ so that

$$\int_I (|f| - cf)(t)\ dt \geq 0$$

i.e., $c\int_I f(t)\ dt \leq \int_I |f(t)|\ dt$ as desired.

If f is complex, $\int_I f\ dt$ may be complex and choose c complex so that

$$c\int_I f(t)\ dt = |\int_I f(t)\ dt|.$$

To do this we need the polar representation of complex numbers which we haven't done but we will do it soon. □

(5) Change of variables formula

Theorem 3.1.6. *Let $I = [\alpha, \beta]$, $J = [\gamma, \delta]$. Let $\varphi : [\alpha, \beta] \to [\gamma, \delta]$ be a homeomorphism which is $\mathcal{C}'$. Then for any $f \in \mathcal{C}(J)$,*

$$\int_\gamma^\delta f(t)\ dt = \int_\alpha^\beta f(\varphi(t))\ \varphi'(t)\ dt.$$

Proof. As $f \circ \varphi$, φ' are continuous all the integrals involved are defined. Let $v(s) = \int_{\gamma}^{s} f(t)\ dt$, $s \in J$ and define $u(x) = v(\varphi(x))$. Then by chain rule u satisfies

$$Du(x) = f(\varphi(x))\ \varphi'(x),\ u(\alpha) = 0.$$

By uniqueness,

$$u(x) = \int_{\alpha}^{x} f(\varphi(t))\ \varphi'(t)\ dt.$$

Taking $x = \beta$ and noting that $\varphi(\beta) = \delta$ we get

$$\int_{\alpha}^{\beta} f(\varphi(t))\ \varphi'(t)\ dt = \int_{\gamma}^{\delta} f(t)\ dt.$$

This proves the theorem. □

(6) Consider $\varphi : [0,1] \to [\alpha, \beta]$ given by $\varphi(t) = \alpha + (\beta - \alpha)t$ which satisfies the conditions of the above theorem. Hence for any $f \in \mathcal{C}[\alpha, \beta]$,

$$\begin{aligned} \int_{\alpha}^{\beta} f(t)\ dt &= \int_{0}^{1} f(\alpha + (\beta - \alpha)t)\ (\beta - \alpha)\ dt \\ \text{i.e.,}\ \int_{\alpha}^{\beta} f(t)\ dt &= (\beta - \alpha) \int_{0}^{1} f(\alpha + (\beta - \alpha)t)\ dt. \end{aligned}$$

Therefore, in order to calculate $\int_{I} f(t)\ dt$ we can always assume that $I = [0,1]$.

(7) Integration by parts

Proposition 3.1.7. *If $f, g \in \mathcal{C}'(I)$ then (if $I = [\alpha, \beta]$)*

$$\int_{\alpha}^{\beta} f'g\ dt + \int_{\alpha}^{\beta} fg'\ dt = f(\beta)g(\beta) - f(\alpha)g(\alpha).$$

Proof. $D(fg) = f'g + g'f$ which means

$$\int_{\alpha}^{x} (f'g + g'f)\ dt = \int_{\alpha}^{x} (fg)'\ dt.$$

By fundamental theorem of calculus

$$\int_\alpha^x (fg)'\ dt = (fg)(x) - (fg)(\alpha).$$

□

This proves the proposition.

(8) Let us do one calculation

Lemma 3.1.8. *For any* $n \in \mathbb{N},\ 0 \le k \le n$

$$\int_0^1 x^k\ (1-x)^{n-k}\ dx = \frac{k!(n-k)!}{(n+1)!}.$$

Proof. Let $a_{n,k} = \int_0^1 x^k\ (1-x)^{n-k}\ dx$. Take $f(x) = \frac{1}{k+1}\ x^{k+1}$ and $g(x) = (1-x)^{n-k}$ in the formula for integration by parts. Then

$$\begin{aligned} a_{n,k} &= \frac{n-k}{k+1} \int_0^1 x^{k+1}\ (1-x)^{n-k-1}\ dx = \frac{n-k}{k+1}\ a_{n,k+1} \\ \text{or}\quad a_{n,k} &= \frac{k}{n-k+1}\ a_{n,k-1}. \\ \text{This gives}\quad a_{n,k} &= \frac{k}{n-k+1} \cdot \frac{k-1}{n-k+2} \cdots \frac{1}{n} \cdot a_{n,0}. \end{aligned}$$

As $a_{n,0} = \int_0^1 (1-x)^n\ dx = \int_0^1 x^n\ dx = \frac{1}{n+1}$ we get the lemma. □

(9) Riemann sums

If $f \in \mathcal{C}[a,b]$, then $\int_a^b f(t)\ dt$ is approximated by $\int_a^b p_n(t)\ dt$ where (p_n) is a sequence of polynomials converging to f uniformly. The approximating sequence $\int_a^b p_n(t)\ dt$ takes a particularly simple form if we use Bernstein polynomials in place of p_n. So, let us assume $a = 0,\ b = 1$ and consider

$$B_n f(x) = \sum_{k=0}^n\ f\left(\frac{k}{n}\right) \binom{n}{k} x^k\ (1-x)^{n-k}$$

which converges to f uniformly on $[0,1]$.

$$\int_0^1 f(t)\ dt = \lim_{n\to\infty} \int_0^1 B_n f(t)\ dt.$$

But in view of the Lemma above

$$\begin{aligned}\int_0^1 B_n f(x)\ dx &= \sum_{k=0}^{n}\ f\left(\frac{k}{n}\right)\ \binom{n}{k} \int_0^1 x^k\ (1-x)^{n-k}\ dx \\ &= \sum_{k=0}^{n}\ f\left(\frac{k}{n}\right)\ \frac{n!}{k!(n-k)!}\ \frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1}\ \sum_{k=0}^{n}\ f\left(\frac{k}{n}\right).\end{aligned}$$

Note that the points $\{\frac{k}{n} :\ k = 0,1,2,\cdots,n\}$ form a partition of $[0,1]$: $[0,1] = \bigcup_{k=0}^{n-1} \left[\frac{k}{n}, \frac{k+1}{n}\right]$ and the sums

$$\sum_{k=0}^{n-1}\ f\left(\frac{k}{n}\right)\cdot\frac{1}{n} = \sum_{k=0}^{n-1}\ f\left(\frac{k}{n}\right)\cdot\left(\frac{k+1}{n} - \frac{k}{n}\right)$$

and called the lower Riemann sums and

$$\sum_{k=1}^{n}\ f\left(\frac{k}{n}\right)\cdot\frac{1}{n} = \sum_{k=1}^{n}\ f\left(\frac{k}{n}\right)\cdot\left(\frac{k}{n} - \frac{k-1}{n}\right)$$

are called upper Riemann sums. And

$$\begin{aligned}\int_0^1 f(t)\ dt &= \lim_{n\to\infty} \int_0^1 B_n f(t)\ dt = \lim_{n\to\infty} \sum_{k=0}^{n-1}\ f\left(\frac{k}{n}\right)\ \frac{1}{n} \\ &= \lim_{n\to\infty} \sum_{k=1}^{n}\ f\left(\frac{k}{n}\right)\ \frac{1}{n}.\end{aligned}$$

Thus we have shown that when f is a continuous function on $[0,1]$ then

$$\int_0^1 f(t)\ dt = \lim_{n\to\infty} \sum_{k=0}^{n-1}\ f\left(\frac{k}{n}\right)\ \frac{1}{n}.$$

Note that $P = \{0, \frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}, 1\}$ forms a partition of $[0,1]$ and $\frac{1}{n} = \frac{k+1}{n} - \frac{k}{n}$ is the length of the kth subinterval $[\frac{k}{n}, \frac{k+1}{n}]$. We can generalise

the above sum by considering arbitrary partitions of $[0,1]$. Let $P = \{x_0 = 0, x_1, \cdots, x_n = b\}, x_k < x_{k+1}$ be any partition of $[a,b]$. Given a continuous function f on $[a,b]$ we can consider the sum

$$R(P,f) = \sum_{k=0}^{n-1} f(x_k)\ (x_{k+1} - x_k)$$

which we call the Riemann sum of f associated to the partition P. We defined the norm of the partition $N(P)$ to be the maximum length of the subintervals: $N(P) = \max\limits_{0\le k\le n-1}(x_{k+1} - x_k)$. We would like to know if $R(P,f)$ converges to $\int\limits_a^b f\ dt$ when $N(P) \to 0$. If that is the case, we can ask a similar question when f is not necessarily a continuous function. If the answer is affirmative then we have a definition of $\int\limits_a^b f\ dt$ for not necessarily continuous functions.

Let $f \in \mathcal{C}[a,b]$ be real valued and consider $M_k = \sup\limits_{x\in[x_k,x_{k+1}]} f(x)$ and $m_k = \inf\limits_{x\in[x_k,x_{k+1}]} f(x)$. Then it is clear that

$$L(P,f) \le R(P,f) \le U(P,f)$$

where $L(P,f)$ and $U(P,f)$ are the special Riemann sums called lower and upper sums:

$$L(P,f) = \sum_{k=0}^{n-1} m_k\ (x_{k+1} - x_k),\ U(P,f) = \sum_{k=0}^{n-1} M_k\ (x_{k+1} - x_k).$$

As f is a continuous function, we can find $\xi_k, \eta_k \in [x_k, x_{k+1}]$ such that $M_k = f(\eta_k), m_k = f(\xi_k)$. Thus,

$$\begin{aligned} U(P,f) - R(P,f) &= \sum_{k=0}^{n-1} (f(\eta_k) - f(x_k))\ (x_{k+1} - x_k) \\ R(P,f) - L(P,f) &= \sum_{k=0}^{n-1} (f(x_k) - f(\xi_k))\ (x_{k+1} - x_k). \end{aligned}$$

Given $\epsilon > 0$, by the uniform continuity of f, we can choose $\delta > 0$ such that $|f(x) - f(y)| < \frac{\epsilon}{b-a}$ whenever $|x - y| < \delta$. Therefore, for any partition P with $N(P) < \delta$ we have the inequalities

$$U(P,f) - R(P,f) < \epsilon;\ R(P,f) - L(P,f) < \epsilon.$$

Thus we see that if $R(P,f)$ converges to a limit then both $U(P,f)$ and $L(P,f)$ converge to the same limit and consequently $U(P,f)-L(P,f)$ converges to 0. Conversely, it turns out that if $U(P,f)-L(P,f)$ converges to 0 then they have a common limit and consequently, $R(P,f)$ also converges to the same limit.

Thus it is reasonable to make the following definition: We say that a bounded function f on $[a,b]$ is *Riemann integrable* (we write $f\in\mathcal{R}[a,b]$) if $U(P,f)-L(P,f)\to 0$ as $N(P)\to 0$, i.e., given $\epsilon>0$, there exists a $\delta>0$ such that $U(P,f)-L(P,f)<\epsilon$ whenever $N(P)<\delta$.

Theorem 3.1.9. *Suppose $f\in\mathcal{R}[a,b]$. Then*

$$\sup_P L(P,f)=\inf_P U(P,f)=\lambda \text{ (say)}.$$

Consequently, $\lim R(P,f)=\lambda$.

Proof. Note that for any partition, $L(P,f)\le U(P,f)$. If P and Q are two partitions, then it follows from the definition that $L(P,f)\le L(P\cup Q,\ f)$, $U(P\cup Q,f)\le U(Q,f)$. Consequently, $L(P,f)\le U(Q,f)$ for any two partitions. This leads to the inequality

$$\sup_P L(P,f)\le\inf_P U(P,f).$$

Since $f\in\mathcal{R}[a,b]$, given $\epsilon>0$ we can find a partition Q such that $U(Q,f)-L(Q,f)<\epsilon$. Hence we have

$$\inf_P U(P,f)-\sup_P L(P,f)\le U(Q,f)-L(Q,f)<\epsilon$$

which proves $\inf_P U(P,f)=\sup_P L(P,f)$. It remains to be shown that

$$\lim_{N(P)\to 0}U(P,f)=\lambda=\lim_{N(P)\to 0}L(P,f).$$

In order to do this we need the following lemma.

Note that when $P\subset Q$ are two partitions then $U(Q,f)\le U(P,f)$ $(L(P,f)\le L(Q,f))$. The lemma says that $U(Q,f)$ (resp. $L(Q,f)$) cannot be very much smaller (resp. larger) than $U(P,f)$ (resp. $L(P,f)$).

Lemma 3.1.10. *Let $f(x)\le M$ on $[a,b]$. If $P\subset Q$ are two partitions and if $Q\cap P^c$ has m points then*

$$L(Q,f)\le L(P,f)+2mMN(P),\ U(P,f)\le U(Q,f)+2mMN(P).$$

Proof. We prove the lemma for upper sums. The proof is similar for lower sums.

Consider the case when $Q = P \cup \{c\}$ where $x_k < c < x_{k+1}$. Then it follows that

$$U(P,f) - U(Q,f) = M_k(x_{k+1} - x_k) - A_k(c - x_k) - B_k(x_{k+1} - c)$$

where $A_k = \sup_{x \in [x_k, c]} f(x)$ and $B_k = \sup_{x \in [c, x_{k+1}]} f(x)$. Thus

$$U(P,f) - U(Q,f) = (M_k - B_k)(x_{k+1} - c) + (M_k - A_k)(c - x_k).$$

Since $|M_k - B_k| \leq 2M,\ |M_k - A_k| \leq 2M$ we obtain

$$U(P,f) - U(Q,f) \leq 2M(x_{k+1} - x_k) \leq 2MN(P).$$

This proves the lemma when $m = 1$.

If $m > 1$, let $P = P_0 \subset P_1 \subset \cdots \subset P_m = Q$ be the partitions where P_{j+1} is obtained from P_j by adding an extra point. By the above, we know that

$$U(P_j,f) - U(P_{j+1},f) \leq 2MN(P),$$

and hence

$$U(P,f) - U(Q,f) = \sum_{j=0}^{m-1} U(P_j,f) - U(P_{j+1},f) \leq 2mMN(P).$$

□

Let us return to the proof of the theorem. Since $\lambda = \inf_P U(P,f)$, given $\epsilon > 0$ we can choose a partition P_0 such that

$$U(P_0,f) < \lambda + \frac{\epsilon}{2}.$$

If P is any partition with $N(P) < \delta$, δ to be chosen in a moment, we have by the lemma

$$U(P,f) \leq U(P \cup P_0, f) + 2mM\delta$$

where $m = \sharp(P \cup P_0) \cap P^c \leq \sharp P_0$. Thus we have

$$\lambda \leq U(P,f) \leq U(P \cup P_0, f) + 2mM\delta \leq U(P_0,f) + 2mM\delta.$$

If we choose δ such that $2mM\delta < \frac{\epsilon}{2}$ then it follows that

$$\lambda \leq U(P,f) < \lambda + \epsilon.$$

□