Modern Applied Analysis.

(Draft version of December 18, 1998)

John Hunter and Bruno Nachtergaele

Department of Mathematics
University of California, Davis

Contents

List of Figures

1 The real numbers as a metric space

In this section we recall some basic properties of the real numbers and real-valued functions. The familiar notions of limit, convergence, and continuity are expressed in terms of the metric or distance function. This allows us to define the above mentioned notions in the more general context of metric spaces, which we will need later on.

Good references for this and several of the following sections are
G.F. Simmons, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.
J.E. Marsden, *Elementary Classical Analysis*, 2nd ed., W.H. Freeman and Company, New York, 1993.

The majority of the material of Sections 1 and 2 is also covered in
W. Rudin, *Principles of Mathematical Analysis*, 2nd ed. McGraw-Hill, New York, 1964

1.1 Definition of Metric

On a first reading, substitute $\mathbb{R}$ for the arbitrary set X in the following definition and subsequent discussion. This week we will discuss almost exclusively $\mathbb{R}$ and $\mathbb{R}^n$ (Euclidean space), but more general (and usually infinite-dimensional) metric spaces will follow soon.

The following is the definition of *metric*, also called *distance function*, on an arbitrary set X.

Definition 1.1 (metric) *A metric on X is a function $d : X \times X \to \mathbb{R}$, with the following properties:*
(M1) $d(x, y) \geq 0$ for all $x, y \in X$, and $d(x, x) = 0$.
(M2) If $d(x, y) = 0$, then $x = y$.
(M3) $d(x, y) = d(y, x)$, for all $x, y \in X$.
(M4) $d(x, y) \leq d(x, z) + d(z, y)$, for all $x, y, z \in X$.

In English the four defining properties read:
M1: Distances are non-negative, and the distance between a point and itself vanishes
M2: The only point at distance 0 from x is x itself.
M3: The distance is a *symmetric* function.
M4: This is the so-called *Triangle Inequality* . When applied to the Euclidean plane, i.e., $X = \mathbb{R}^2$ with the usual distance function, this inequality expresses the fact that the length of any side of a triangle is less than the sum of the lengths of the two other sides. M2 is also called *strict positivity* or *non-degeneracy.*

X, equipped with a metric d, is called a metric space, often denoted by (X, d), especially when it is important to specify which metric is being considered.

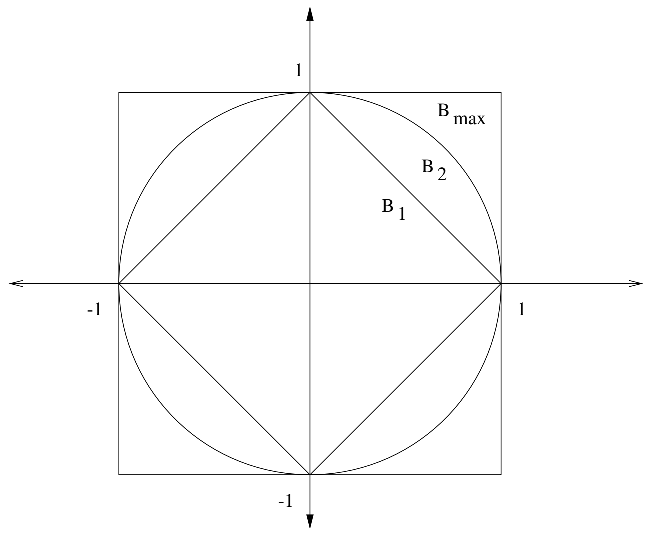

Figure 1.1: The unit ball in $\mathbb{R}^2$ for the Euclidean norm (B_2), the sum norm (B_1), and the maximum norm ($B_{\max}$).

Examples:

1. $X = \mathbb{R}$, and $d(x, y) = |x - y|$.

2. $X = \mathbb{R}^n$, $n = 1, 2, 3, \ldots$, and the Euclidean metric given by
$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$
Note: the distance of $x \in \mathbb{R}^n$ to $0 \in \mathbb{R}^n$ (the origin), is called the length or *norm* of x, denoted by $\|x\|$. The obvious relation between the norm and the metric is then
$$d(x, y) = \|x - y\|$$

3. $X = \mathbb{R}^n$, $n = 1, 2, 3, \ldots$, and
$$d(x, y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$
Note: the corresponding norm $\|x\| = d(0, x)$ is called the *sum norm* or the L^1 norm .

4. $X = \mathbb{R}^n$, $n = 1, 2, 3, \ldots$, and
$$d(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_n - y_n|\}$$
Note: the corresponding norm $\|x\| = d(0, x)$ is usually called the *maximum norm* or the L^∞ norm. The unit balls in $\mathbb{R}^2$ for each of the metrics defined in Examples 2-4 are shown in Figure 1.1.

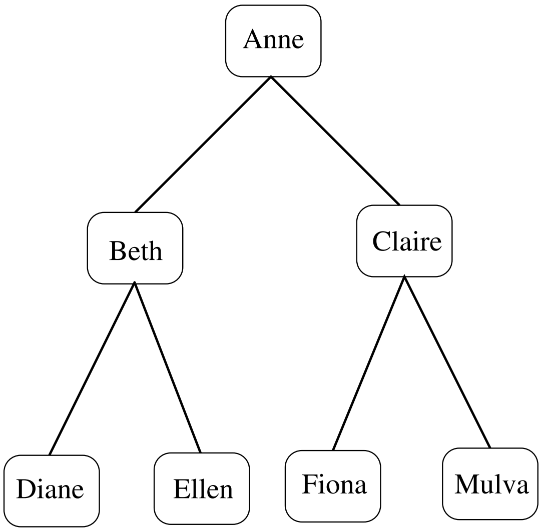

Figure 1.2: An example of a family tree as used in the definition of the ultrametric defined in (1.1).

5. Let X be a set people of one and the same generation and with a common ancestor, e.g. all greatgrandchildren of greatgrandmother $\mathcal{M}$. A metric d can then be defined on this set by putting $d(x, y) =$ the number of generations one has to go back to find the first common ancestor, along female lines, of any two individuals x and y. E.g., the distance between two sisters is 1. Then

$$d(x, y) \leq \max\{d(x, z), d(z, y)\} \quad \text{for all } x, y, z \in X \tag{1.1}$$

   and a fortiori M4 holds. See Figure 1.2 for an illustration. (1.1) is stronger than the triangle inequality. When it holds the metric d is called an *ultrametric*. In taxonomy ultrametrics have been used to characterize genetic proximity of species.

6. Let X be the set of all N-letter words in an alphabet A with k characters, i.e.,

$$X = \{(x_1, x_2, \ldots, x_N) \mid x_i \in A\}$$

   and $A = \{a_1, a_2, \ldots, a_k\}$. A natural metric on X is then defined by

$$d(x, y) = \#\{i \mid x_i \neq y_i\}$$

7. The metric spaces most important for this course are infinite-dimensional ones, of which there is a large variety. Before we can start introducing those we need to review just a little more real analysis.

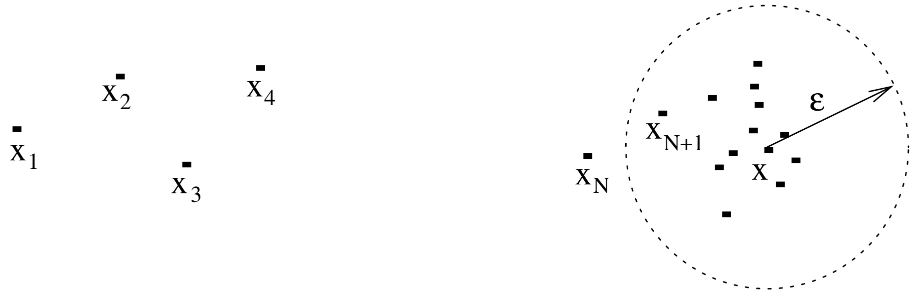

Figure 1.3: A sequence x_n converging to x.

1.2 When distances tend to zero: convergence

1.2.1 Sequences

For concreteness let us consider again the real numbers with their usual metric. All definitions, however, do carry over to arbitrary metric spaces. Just replace $\mathbb{R}$ by X, $|x-y|$ by $d(x,y)$, etc.

Let (x_n) be a *sequence* of real numbers, i.e., for each $n \in \mathbb{N}, x_n \in \mathbb{R}$ (other common notations are $\{x_n\}, (x_n)_{n\geq 1}, \{x_n\}_{n=0}^{\infty}$ etc.

Definition 1.2 (convergence) *The sequence of real numbers (x_n) converges to $x \in \mathbb{R}$ iff for all $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n \geq N$, $|x_n - x| < \epsilon$.*

Note that in this definition N depends on ϵ, and you could write $N(\epsilon)$ to make this dependence explicit. The point x is called the *limit* of (x_n). Common notations expressing that (x_n) converges to x are:

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

Sequences that become eventually larger than any number do not converge according to the definition given above. They are usually called *divergent*. But it is sometimes more useful to consider them as converging to $+\infty$, i.e., $x_n \to +\infty$ iff for all $M \in \mathbb{R}$, there is $N \in \mathbb{N}$, such that for all $n \geq N$, $x_n > M$. And similarly one defines $x_n \to -\infty$ iff for all $M \in \mathbb{R}$, there is $N \in \mathbb{N}$, such that for all $n \geq N$, $x_n < M$.

For an arbitrary metric space (X,d), convergence of a sequence (x_n), $x_n \in X$, to $x \in X$ means that for all $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n \geq N$, $d(x_n, x) < \epsilon$. Figure 1.3 shows a convergent sequence in $\mathbb{R}^2$, considered with the Euclidean metric (or any one of the other metrics defined in Examples 3 and 4).

Next, we define the concept of a *Cauchy* sequence.

Definition 1.3 (Cauchy sequence) *(x_n) is a Cauchy sequence iff for all $\epsilon > 0$ there is $N \in \mathbb{N}$ such that for all $n, m \geq N$, $|x_n - x_m| < \epsilon$.*

Cauchy sequences are sequences that *ought* to converge. The following propositon says that, in general, a necessary condition for any sequence to converge is that it is Cauchy.

Proposition 1.4 *If (x_n) converges to $x \in \mathbb{R}$, then (x_n) is Cauchy.*

This proposition holds for sequences in any metric space (X, d). A natural question now is whether the converse is also true: does every Cauchy sequence have a limit it converges to? For real sequences the answer is *yes*, but this is not so in general. The property that every Cauchy sequence in a metric space (X, d) converges to a limit *in* X, singles out a particularly interesting class of metric spaces. Such (X, d) are called *complete* . So, $\mathbb{R}$ with the usual metric is complete, but the rational numbers, $\mathbb{Q}$, considered as a subset of $\mathbb{R}$ and with the same distance function, is *not* complete. In fact, one way of constructing the set of real numbers is by considering all Cauchy sequences in $\mathbb{Q}$ and grouping those together in equivalence classes of sequences that ought to have the same limit. We will discuss the notion of completeness and its applications in Chapter 4.

Exercise 1.1 *Prove that $\mathbb{Q}$ with the usual distance function is not complete as a metric space. (Hint: find a sequence of rational numbers that is Cauchy, and hence converges in $\mathbb{R}$, but with an irrational limit.)*

Exercise 1.2 *Starting from the fact that $\mathbb{R}$ equipped with its usual distance function is complete, prove that*
a) $\mathbb{R}^n$ *with the Euclidean norm is a complete metric space,*
b) $\mathbb{R}^n$ *with the sum norm or the maximum norm is also complete.*

Another natural question is whether a sequence could converge to two different limits. In a metric space, the answer is *no*, in particular due to property M2 of the metric.

Proposition 1.5 *If $x_n \to x$ and $x_n \to y$, then $x = y$*

A *subsequence* of a sequence (x_n) is a sequence of the form $(x_{n_k})_{k=1}^{\infty}$, where for each $k \in \mathbb{N}$, $n_k \in \mathbb{N}$ and $n_k < n_{k+1}$, for all k. I.e., $k \mapsto n_k$ is a strictly increasing function from the set of natural numbers to itself. Clearly, if (x_n) is a convergent sequence with limit x, then all its subsequences also converge to x.

1.2.2 Upper and lower bound, inf and sup, lim inf and lim sup

There is a natural order relation on $\mathbb{R}$ that is nicely related to the metric structure. First, we define the notion of *upper* and *lower bound* for a subset A of $\mathbb{R}$.

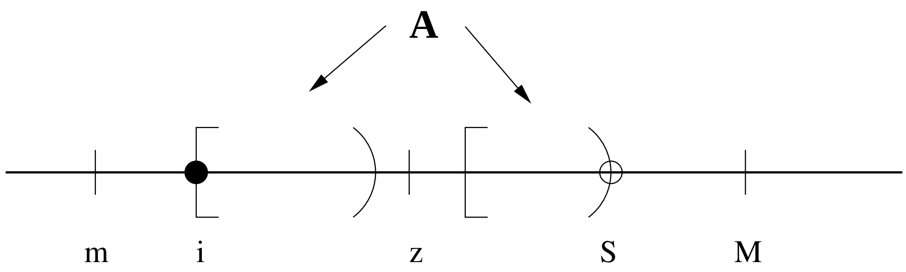

Figure 1.4: M is an upper bound for A and m is lower bound for A. z is neither an upper bound nor a lower bound. S is the supremum of A (but does not belong to A). i is the infimum (and, as $i \in A$, also the minimum) of A.

Definition 1.6 (upper and lower bound) *Let $A \subset \mathbb{R}$. $M \in \mathbb{R}$ is an upper bound for A iff $x \leq M$, for all $x \in A$, and $m \in \mathbb{R}$ is a lower bound for A iff $m \leq x$, for all $x \in A$. The set A is called bounded iff both an upper and a lower bound for it exist.*

If A has an upper bound M, then A has many upper bounds. E.g, any $M' \geq M$ is also an upper bound. The *least upper bound* for A is called the *supremum* of A, denoted by $\sup A$. I.e., $\sup A$ is an upper bound with the property that if M is any upper bound for A, then $M \geq \sup A$.

Similarly one defines $\inf A$, the *infimum* of A, as the *greatest lower bound* for A. I.e., $\inf A$ is a lower bound with the property that if m is any lower bound for A, then $m \leq \inf A$.

By convention we set $\sup A = +\infty$ if A does not have an upper bound, i.e., if A contains arbitrarily large real numbers. In the same way, one sets $\inf A = -\infty$ if A does not have a lower bound.

Obviously, for each $A \neq \emptyset$, there can be only one least upper bound (greatest upper bound). This is due to the fact that $\mathbb{R}$ is *totally* ordered, i.e., for all $x, y \in \mathbb{R}$, $x \leq y$ or $y \leq x$. One also has $\inf A \leq \sup A$, for all $A \neq \emptyset$. For $A = \emptyset$ the convention $\inf A = +\infty$ and $\sup A = -\infty$ is sometimes useful.

That $\sup A$ ($\inf A$) exists for any set with an upper (lower) bound, is a consequence of (and actually equivalent to) the completeness of $\mathbb{R}$.

If $\sup A \in A$, $\sup A$ is also the *maximum* of A, defined as the largest element of A, and often denoted by $\max A$. If $\inf A \in A$, $\inf A$ is also the *minimum* of A, defined as the smallest element of A, and often denoted by $\min A$. Note that every bounded set has a supremum and an infimum, but that it does not necessarily have a maximum or a minimum.

The concepts of maxiumum, minimum, supremum, infimum, and upper and lower bound, are illustrated in Figure 1.4.

The following two quantities can again be defined in an arbritrary metric space (X, d).

For any $A \subset X$, the *diameter* of A is defined by

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

For a subset $A \subset X$ and a point $x \in X$ the distance from x to A is naturally defined by

$$\text{dist}(x, A) \equiv d(x, A) = \inf\{d(x,y) \mid y \in A\}$$

Note that the distance function dist is *not* a metric on the power set of X.

Examples:

1. $A = \{1, 1/2, 1/3, 1/4, \ldots\}$. Then $\sup A = 1 = \max A$ and $\inf A = 0$. Note that, in this example, A does not have a minimum.

2. With A as given in 1), $\text{diam}A = 1$. Note that $d(0, A) = 0$, although $0 \notin A$.

In order to define lim inf and lim sup we first observe that monotone sequences always converge (allowing for convergence to $\pm\infty$). Indeed, if (x_n) is *monotone increasing*, i.e., $x_n \leq x_{n+1}$, for all n, and (y_n) is *monotone decreasing*, i.e., $x_n \geq x_{n+1}$, for all n, then

$$\lim_{n\to\infty} x_n = \sup\{x_n\} \tag{1.2}$$

$$\lim_{n\to\infty} y_n = \inf\{y_n\} \tag{1.3}$$

lim inf (*limes inferior*) and lim sup (*limes superior*) can now be defined for an arbitrary sequence (x_n), as the following two limits:

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf\{x_k \mid k \geq n\} \tag{1.4}$$

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup\{x_k \mid k \geq n\} \tag{1.5}$$

Another common notation is: $\liminf = \underline{\lim}, \limsup = \overline{\lim}$.

Exercise 1.3 *Let (x_n) be a sequence of real numbers. A point $c \in \mathbb{R} \cup \{\pm\infty\}$ is called a* cluster point *of (x_n) if there is a convergent subsequence of (x_n) with limit c. Let C denote the set of all cluster points of (x_n). Prove that*

$$\limsup x_n = \max C, \quad \textit{and} \quad \liminf x_n = \min C \quad .$$

Exercise 1.4 *Let (x_n) be a sequence of real numbers. Prove that, for every $\epsilon > 0$, and every N, there exists $n_1, n_2 > N$, such that*

$$x_{n_1} + \epsilon \geq \limsup_m x_m, \quad x_{n_2} - \epsilon \leq \liminf_m x_m$$

In contrast to the limit, the lim inf and lim sup always exist. Clearly, one always has $\liminf_n x_n \le \limsup_n x_n$. Note that a sequence (x_n) converges to a limit x iff $\liminf_n x_n = \limsup_n x_n = x$.

Let A be an arbitrary set, and $x_\alpha \in \mathbb{R}$, for all $\alpha \in A$. Then $\sup\{x_\alpha \mid \alpha \in A\}$ is usually abbreviated to $\sup_{\alpha\in A} x_\alpha$, or, when this does not lead to confusion, by $\sup_\alpha x_\alpha$. Now consider a family of real numbers indexed by $n \in \mathbb{N}$ and $\alpha \in A$: $x_{n,\alpha}$. The following two relations are very useful and easy to prove:

$$\limsup_{n\to\infty} \inf_\alpha x_{n,\alpha} \le \inf_\alpha \limsup_{n\to\infty} x_{n,\alpha} \tag{1.6}$$

$$\sup_\alpha \liminf_{n\to\infty} x_{n,\alpha} \le \liminf_{n\to\infty} \sup_\alpha x_{n,\alpha} \tag{1.7}$$

Exercise 1.5 *Prove the inequalities (1.6) and (1.7)*

Finally, we mention the notion of *subsequence*. A sequence (y_k) is called a subsequence of the sequence (x_n), if there is a strictly increasing sequence of positive integers n_k such that $y_k = x_{n_k}$, for all k.

1.3 Series

Let (x_n) be a sequence in $\mathbb{R}$. The sequence of *partial sums* of (x_n) is the new sequence (s_n) defined by

$$s_n = \sum_{k\le n} x_k$$

If (s_n) converges to a limit $s \in \mathbb{R}$, we say that the *series* $\sum_{n=1}^\infty x_n$ converges to s. If (s_n) does not converge the series is called *divergent*.

A series $\sum_{n=1}^\infty x_n$ is said to be *absolutely convergent* iff the series of absolute values $\sum_{n=1}^\infty |x_n|$ converges. Absolute convergence implies convergence, but not vice versa.

Exercise 1.6 *Show that the series*

$$\sum_{n=1}^\infty (-1)^n \frac{1}{n}$$

is not absolutely convergent. Show that by permuting the terms of this series one can obtain series with different limits.

A useful feature of absolutely convergent series is that any series obtained from it by a permutation of its terms, is also convergent and has the same limit. For series of real numbers the property that any permutation converges to the same limit (this property is called *unconditional convergence*) is in fact equivalent to absolute convergence. This is no longer true for series in infinite-dimensional spaces.

1.4 Open and closed sets

Consider a metric space (X, d), e.g., $\mathbb{R}^n$. The *open ball*, $B_r(x_0)$, with radius $r > 0$ and center $x_0 \in X$ is defined by

$$B_r(x_0) = \{x \in X \mid d(x_0, x) < r\}$$

The *closed ball*, $\overline{B}_r(x_0)$, is defined by

$$\overline{B}_r(x_0) = \{x \in X \mid d(x_0, x) \leq r\}$$

The basic concept of (set) topology is the notion of *open set.*

Definition 1.7 (open and closed sets) *A subset $G \subset X$ is open iff for all $x \in G$ there exists $r > 0$ such that $B_r(x) \subset G$.*
A subset $F \subset X$ is closed iff $F^c \equiv X \setminus F$ is open.

Proposition 1.8 *A subset F of a metric space is closed iff all convergent sequences of elements in F converge to a limit in F, i.e., if $x_n \to x$, and $x_n \in F$, for all n, then $x \in F$.*

A set $A \subset X$ is called bounded if it is contained in a ball $B_r(x_0)$, with $r < +\infty$. The following properties are easy to prove:

1. The empty set $\emptyset$ and X itself are open *and* closed.

2. The intersection $G_1 \cap G_2$ of two open sets $G_1, G_2 \subset X$ is open. Consequently, the intersection of any finite family of open sets is also open.

3. The union of an arbitrary family (finite or infinite) of open sets is open.

4. The union $G_1 \cup G_2$ of two closed sets $G_1, G_2 \subset X$ is closed, and hence any finite family of closed sets has a closed union.

5. The intersection of an arbitrary family (finite or infinite) of closed sets is closed.

Exercise 1.7 *Prove the properties 1-5.*

The closure $\overline{A}$ of a set $A \subset \mathbb{R}^n$, is, by definition, the smallest closed set containing A. The last property above implies that such a set indeed exists. In any metric space the closure of a subset A can be obtained by adding to A all limits of converging sequences of elements of A. I.e,

$$\overline{A} = \{x \in X \mid \text{there exist } x_n \in A, x_n \to x\}$$

Therefore, $\overline{A}$ is sometimes *defined* as the set of *points of adherence* of A. By definition, $x \in X$ is a *point of adherence* of A if every open set containing x also contains points of A.

For any $x \in X$ a *neighborhood* of x is a subset $V \subset X$ such that there is an open set $G \subset V$ containing x. Equivalently, V is a neighborhood of x if there is a ball $B_r(x) \subset V$, for some $r > 0$.

A set $A \subset X$ is called *dense* if $\overline{A} = X$.

1.5 Countably and uncountably infinite sets

This is a brief aside on infinite sets. Our only aim here is to clarify the distinction between countably infinite and uncountably infinite sets.

A set A is finite, i.e., has a finite number of elements, iff there is a bijection (a map that is one-to-one and onto), from A to one of the sets $\emptyset$ or $\{1, 2, \ldots, n\}$. Of course, n is the number of elements of A. If a set is not finite, it is called *infinite*.

A set A is *countable* if it is either finite or if there is a bijection from A to $\mathbb{N}$. In the latter case it is called *countably infinite*, i.e., in order to enumerate all the elements of A, you need an infinite number of natural numbers.

Sets that are not countable are *uncountable*. To label all the elements of an uncountable set, the natural numbers do not suffice, i.e., they cannot be enumerated.

Examples:

1. Any infinite subset of $\mathbb{N}$ (or any other countable set) is countably infinite. So, e.g., the even natural numbers form a countably infinite set.

2. $\mathbb{Z}$ is countably infinite.

3. $\mathbb{Q}$ is countably infinite. Do you know an enumeration of the rational numbers?

4. $\mathbb{R}$ is uncountably infinite. Do you know how to prove this?

1.6 Compactness

The full importance of the *compact sets* will become clear only in an infinite-dimensional setting, and we will have ample opportunity to see how useful they are later. It is nevertheless interesting to start with a finite-dimensional warmup.

Definition 1.9 *A subset K of a metric space X is called compact iff every sequence (x_n) of elements of K has a convergent subsequence with a limit $x \in K$.*

Sets satisfying the definition above are sometimes called *sequentially compact* in order to make a distinction with the more general definition of compactness in the framework of general topological spaces (of which the metric spaces are a special case). Using the general definition here would force us to also prove that for metric spaces compactness is equivalent to sequential compactness, which is the more useful notion for the applications we have in mind. Therefore, it is more economic not to go beyond the scope of metric spaces now. We will discuss the more general situation later.

Exercise 1.8 *Prove that every compact set is closed.*

The compact subsets of $\mathbb{R}^n$ have a simple and practical characterization given by the following theorem.

Theorem 1.10 (Heine-Borel) *A subset $K \subset \mathbb{R}^n$ is compact iff it is closed and bounded.*

The non-trivial part in the proof of the Heine-Borel Theorem is the following result attributed to Weierstrass. The property that, in a given metric space, every bounded sequence has a convergent subsequence is sometimes called the *Bolzano-Weierstrass property*.

Theorem 1.11 (Bolzano-Weierstrass) *Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Proof: This proof is constructive in the sense that we will give a procedure to construct a convergent subsequence starting from an arbitrary bounded sequence. In fact, this subsequence will be a Cauchy sequence by construction, and hence, as $\mathbb{R}^n$ is complete, it converges.

Let (x_i) denote the given sequence. As (x_i) is bounded, there is $M > 0$ such that $x_i \in [-M, M]^n$, for all i. The set $[-M, M]^n$ is a cube of side $2M$. The first element of the subsequence is chosen as follows: partition C_0 into 2^n cubes of side M, and pick C_1 to be one of these smaller cubes that contains an infinite number of elements of the sequence (x_i), i.e., $x_i \in C_1$, for infinitely many i. Such a cube exists because there are only a finite number of cubes and an infinite number of elements in the sequence. Let i_1 be the smallest index such that $x_{i_1} \in C_1$. This is the first element of the subsequence.

To find the second element, first form a new sequence (y_i) by deleting from (x_i), the element x_{i_1} and all $x_j \notin C_1$. Now, repeat the procedure described in the previous paragraph but with (x_i) replaced by (y_i), and $[-M, M]^n$ replaced by C_1. This will yield a subcube C_2 of C_1 of size $M/2$, and an element y_{i_1}. Put $x_{i_2} = y_{i_1}$. This is the second element of the subsequence.

Repeat this ad infinitum to obtain a subsequence (x_{i_j}) of (x_i). By construction we will have that for every j, and every $k \geq j$, $x_{i_k} \in C_j$, where C_j is a cube of size $2M/2^j$. Therefore, (x_{i_j}) is a Cauchy sequence and hence it converges. ∎

Exercise 1.9 *Complete the proof of Theorem 1.10.*

Note that for any metric space (X, d), and any $X_0 \subset X$ we can use the same distance function d on X_0, i.e., the restriction d_0 (often simply denoted by d) of d to X_0, to make X_0 into a metric space. If X_0 is a compact subset of X, (X_0, d) is a compact metric space.

The following is a useful property.

Exercise 1.10 *Let (K, d) be a compact metric space, and G a closed subset of K. Prove that G is compact.*

2 Continuous functions

For further reading and more exercises see
J.E. Marsden, *Elementary Classical Analysis*, 2nd ed., W.H. Freeman and Company, New York, 1993.
or
W. Rudin, *Principles of Mathematical Analysis*, 2nd ed. McGraw-Hill, New York, 1964

2.1 Continuity and semi-continuity

Let D be a subset of $\mathbb{R}$. A function $f : D \to \mathbb{R}$ is *continuous* at a point $x_0 \in D$ iff for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $x \in D$, $|x - x_0| < \delta$ implies $|f(x) - f(x_0)| < \epsilon$. Continuity at x_0 is thus the property that the value of a function f in a point x close to x_0, is close to the value of f in x_0. If f is not continuous at x one says that it is *discontinuous* at x.

It is now obvious how to define continuity of a function from one metric space into another. Let (X, d_X) and (Y, d_Y) be two metric spaces.

Definition 2.1 (continuity) *A function $f : X \to Y$ is continuous at $x_0 \in X$ iff for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $x \in X$, $d_X(x, x_0) < \delta$ implies $d_Y(f(x), f(x_0)) < \epsilon$.*

Remark: In the sequel we will usually denote all metrics by d when it is clear from the context (e.g., the arguments of d) which metric is meant.

We say that f is continuous on $D \subset X$ if f is continuous at all points $x \in D$. Note that this is not the same as saying that f restricted to D is continuous, e.g., if D contains only one point any function f restricted to D is continuous.

There are always continuous functions on X. E.g., any constant function is continuous. A less trivial example is the following. For any $x_0 \in X$, the function f defined by $f(x) = d(x_0, x)$ is continuous on X.

Definition 2.2 (uniform continuity) *A function $f : X \to Y$ is uniformly continuous on X iff for all $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in X$, $d_X(x, y) < \delta$ implies $d_Y(f(x), f(y)) < \epsilon$.*

Note the subtle difference between this and the previous definition.

Examples:

1. A function $f : \mathbb{R}^n \to \mathbb{R}^m$, satisfying

$$f(tx + (1-t)y) = tf(x) + (1-t)f(y), \qquad \text{for all } x, y \in \mathbb{R}^n, t \in [0,1]$$

   is called *affine*. Affine functions are uniformly continuous.

2. The function $x \mapsto x^2, x \in \mathbb{R}$, is continuous but not uniformly continuous.

A function $f : D \to \mathbb{R}$, where $D \subset \mathbb{R}$ is an open set, is called *differentiable* at $x_0 \in D$, iff the following limit exists (and is finite):

$$f'(x_0) := \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}$$

Exercise 2.1 *Prove that if f is differentiable at x_0, then f is continuous at x_0.*

A useful characterization of continuous functions, which could also serve as an equivalent definition, is the following.

Proposition 2.3 (say it with sequences) *$f : X \to Y$ is continuous at $x \in X$ iff for every convergent sequence $x_n \to x$, the sequence $f(x_n)$ converges to $f(x)$.*

Proof: We will denote the metrics of both X and Y by d. The arguments of d unambiguously indicate which metric is meant.
The 'if' part: we show by construction that, if f is discontinuous at x, there is sequence (x_n), converging to x, but such that the sequence $(f(x_n))$ does not converge to $f(x)$. If f is discontinuous at x, there exists $\epsilon > 0$ such that for all $n \geq 1$, we can find $x_n \in X$ satisfying

$$d(x, x_n) < \frac{1}{n}, \text{ but } d(f(x), f(x_n)) > \epsilon$$

Clearly, the sequence x_n converges to x, but $f(x_n)$ does not converge to $f(x)$, as it stays at least a distance ϵ away from $f(x)$.
The 'only if' part: Suppose f is continuous at x, and $x_n \to x$. For any $\epsilon > 0$, there is $\delta > 0$ such that

$$\text{if } d(x, x_n) < \delta \text{ then } d(f(x), f(x_n)) < \epsilon$$

As $x_n \to x$, we can find N such that for all $n \geq N$, $d(x, x_n) < \delta$. Therefore, $f(x_n) \to f(x)$. ■

From the fact that continuous functions preserve convergence of sequences it immediately follows that they also preserve compactness. This is expressed in the following theorem.

Theorem 2.4 *Let $f : K \to Y$ be continuous on K, where K is a compact metric space and Y is any metric space. Then, $f(K) \subset Y$ is compact.*

In the rest of this section the target space Y will always be taken to be the real numbers. A function $f : D \to \mathbb{R}$ is called *bounded* on D iff there exist $M \in \mathbb{R}$ such that $|f(x)| \leq M$ for all $x \in D$, which is equivalent to saying that $f(D)$ is a bounded set.

Here is our first practical use of compactness:

Theorem 2.5 *Let K be a compact metric space (e.g., a compact subset of a metric space). If $f : K \to \mathbb{R}$ is continuous, then f is bounded and attains its supremum (i.e, maximum) and infimum (i.e., minimum). I.e.,*

$$\sup_{x \in K} f(x) = f(y) \quad \text{for some } y \in K \tag{2.1}$$

$$\inf_{x \in K} f(x) = f(z) \quad \text{for some } z \in K \tag{2.2}$$

Proof: First, we show that f is bounded. By Theorem 2.4, $f(K)$ is a compact subset of $\mathbb{R}$. Therefore, $f(K)$ is bounded by the Heine-Borel Theorem 1.10.

As (2.1) follows from (2.2) applied to the function $-f$, we only need to prove (2.2). We already know that f is bounded. In particular, f is bounded from below, i.e., there exists a constant $M \in \mathbb{R}$ such that $f(x) \geq M$, for all $x \in K$. Hence, $m := \inf_{x \in K} f(x) \geq M > -\infty$. By the definition of infimum, for all n there is $x_n \in K$ such that

$$|m - f(x_n)| < \frac{1}{n}, \text{ and hence } f(x_n) \to m \tag{2.3}$$

As K is compact it follows that (x_n) has a convergent subsequence. Let (y_n) be such a subsequence and denote its limit by y. Clearly, (2.3) implies $f(y_n) \to m$. As f is continuous $f(y_n) \to f(y)$. By the uniqueness of the limit, Proposition 1.5, we conclude $f(y) = m$, i.e., f attains its infimum in the point y. ■

Examples:

1. The function $f(x) = 1 - \sqrt{|x|}$ is continuous on $[-1, 1]$ and bounded. f attains its maximum in $x = 0$. (The theorem applies)
2. $f(x) = e^x$ is continuous on the whole of $\mathbb{R}$ but not bounded. ($\mathbb{R}$ is not compact)
3. $f(x) = 1/x$ on $(0, 1]$ is continuous but not bounded. ($(0, 1]$ is not compact)

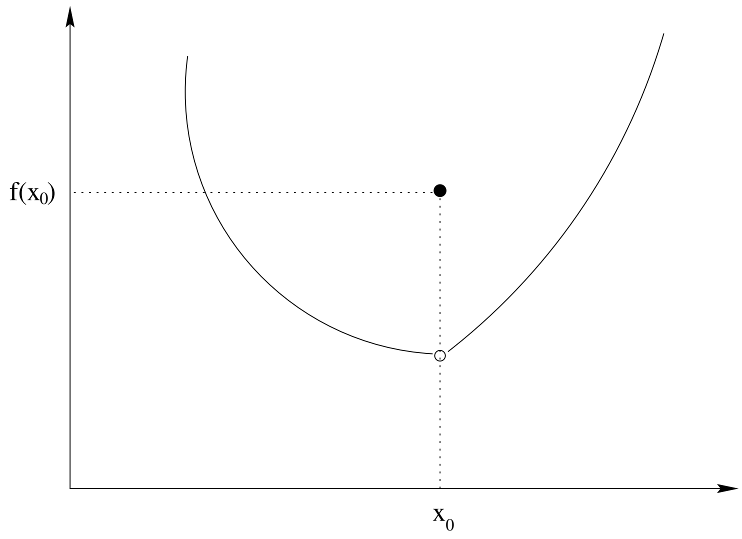

Figure 2.1: The function f is upper semicontinuous, but not continuous, in the point x_0.

4. $f(x) = \text{sign}x$ on $[-1, 1]$ is bounded, not continuous, but attains its maximum on any $x \in (0, 1]$ (by convention we put $\text{sign}0 = 0$). (continuity is not a necessary condition for attaining the maximum)

For functions that have only a certain kind of discontinuities, part of the conclusions of Theorem 2.5 still holds. There are two kinds of "half-continuous" functions, defined as follows.

Definition 2.6 (upper and lower semi-continuity) *$f : X \to \mathbb{R}$ is* upper *semi-continuous on X if for all $x \in X$ and any sequence $x_n \to x$, $x_n \neq x$, one has*

$$\limsup_{n\to\infty} f(x_n) \leq f(x)$$

Similarly, f is lower *semi-continuous on X if for all $x \in X$ and any sequence $x_n \to x$, $x_n \neq x$, one has*

$$\liminf_{n\to\infty} f(x_n) \geq f(x)$$

Note that f is continuous iff f is both upper and lower semicontinuous. The analogue of Theorem 2.5 is then:

Theorem 2.7 *Let K be a compact metric space. If $f : K \to \mathbb{R}$ is upper semi-continuous, then f is bounded from above and attains its supremum. If $f : K \to \mathbb{R}$ is lower semicontinuous, then f is bounded from below and attains its infimum.*

Bounded (from) above here means that there is $M \in \mathbb{R}$ such that $f(x) \leq M$ for all $x \in K$, and f is *bounded (from) below* iff $-f$ is bounded from above.

Theorem 2.5 can be generalized in yet another way: by adding an additional assumption, called *coersiveness* , expressed by (2.4), we can also treat continuous functions on non-compact domains. See the following exercise.

Exercise 2.2 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function satisfying*

$$\lim_{\|x\| \to \infty} f(x) = +\infty \tag{2.4}$$

a) *Prove that f is bounded below.*

b) *Prove that f attains its infimum.*

Exercise 2.3 *Let $p : \mathbb{R}^2 \to \mathbb{R}$ be a polynomial function of two real variables. Suppose $p(x, y) \geq 0$ for all $x, y \in \mathbb{R}$. Do all such functions p attain their infimum? Prove or disprove.*

Exercise 2.4 *Let $(X, d_X), (Y, d_Y)$, and (Z, d_Z) be metric spaces and let $f : X \to Y$, and $g : Y \to Z$ be two continuous functions. Show that $h : X \to Z$, defined by $h(x) = g(f(x))$ is also continuous.*

2.2 Convergence of sequences of functions

In this section we consider sequences of real-valued functions on a common domain $D \subset X$ in a metric space (X, d), i.e., $f_n : D \to \mathbb{R}$, for $n \geq 1$. What would we mean by $f_n \to f$? Your "first thought" on this question is likely to be one of the following two:

1) The functions f_n are defined by the values they take on D. So, let us say that the functions converge if the values converge, i.e., if $f_n(x) \to f(x)$ for all $x \in D$. Thus, convergence of real-valued functions is reduced to convergence of real numbers, which we already know. This is called *pointwise* convergence of functions.

2) Find a set of functions with a reasonable metric on it and define $f_n \to f$ in the sense of that metric.

1) and 2) are both natural and useful ideas. But they are not compatible! It turns out that for most relevant domains D (e.g., whenever D is not a countable set), pointwise convergence cannot be expressed as convergence in a metric space (see later). This could be good motivation to study more general topological spaces (beyond metric spaces). Let us postpone this study for another while, and first consider 1) and 2), separately, in some more detail.

Besides being non-metric, pointwise convergence has another limitation: it does not preserve continuity. Consider, e.g., the functions $f_n : [0, 1] \to \mathbb{R}, f_n(x) = x^n$.

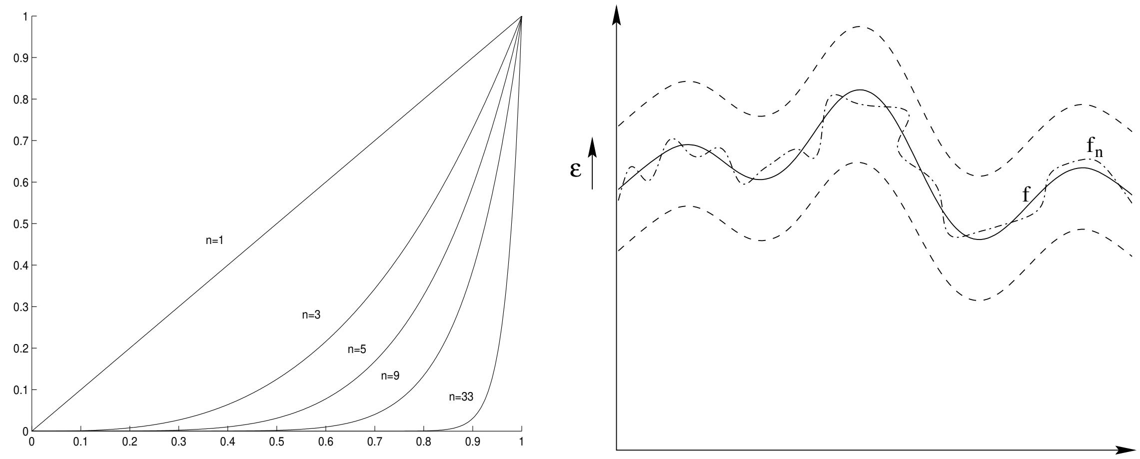

Figure 2.2: Left: the sequence of functions $f_n(x) = x^n$ converges pointwise but not uniformly on $[0, 1]$. Right: graphically, uniform convergence means that for an arbitrarily narrow tubular neighborhood of the limiting function, the functions f_n will be contained in it for all sufficiently large n.

Then, (f_n) converges pointwise to the function δ_1 defined by

$$\delta_1(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x = 1 \end{cases} \tag{2.5}$$

which is clearly discontinuous at $x = 1$.

A natural metric structure on the space of all continuous functions is given in terms of the *uniform* or *supremum* norm defined by:

$$\|f\| = \sup_{x \in D} |f(x)| \tag{2.6}$$

This norm is often denoted by $\|\cdot\|_{\text{sup}}$ or $\|\cdot\|_\infty$. The origin of the latter notation will become clear when we introduce L^p spaces in Chapter 22. From the definition (2.6) it is clear that a function f is bounded iff $\|f\|_{\text{sup}} < +\infty$.

Definition 2.8 (uniform convergence) *A sequence of bounded functions (f_n) converges uniformly to a function f iff*

$$\lim_{n\to\infty} \|f_n - f\|_{\text{sup}} = 0$$

Note that uniform convergence implies pointwise convergence. The example defined in (2.5) shows that the opposite does not hold. Here is another example.

Example:
$f_n(x) = \tanh nx \to \text{sign} x$, pointwise, but not uniformly.

A crucial property of uniform convergence is that it preserves continuity. Below, we state the theorem, and its proof. Note the strategy of the proof.

Theorem 2.9 *Let (f_n) be a sequence of continuous real-valued functions on a metric space (X, d), and suppose $f_n \to f$ in the uniform norm. Then, f is continuous.*

Proof: In order to show that f is continuous at $x \in X$, we need to prove that for every $\epsilon > 0$, there is a $\delta > 0$ such that for all $y \in X$, with $d(x, y) < \delta$, $|f(x) - f(y)| < \epsilon$.

We start from

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

First, use the *uniform* convergence to choose an n such that the first and the third term are less than $\epsilon/3$ *for all* $x, y \in X$. Then, use the continuity of f_n at x to find δ such that the middle term is less than $\epsilon/3$ for $d(x, y) < \delta$. The desired result follows. ■

2.3 The spaces $C(K)$

Let K be a compact metric space, and let $C(K)$ denote the set of all continuous functions $f : K \to \mathbb{R}$. Equipped with the metric derived from the uniform norm, $C(K)$ becomes a metric space. Explicitly:

$$d(f, g) = \|f - g\| = \sup_{x \in K} |f(x) - g(x)|$$

for all $f, g \in C(K)$. Note that due to Theorem 2.5, $\|f\| < +\infty$, for all $f \in C(K)$. It is straightforward to verify the properties M1-4 of Definition 1.1.

Theorem 2.10 *$C(K)$, with the uniform norm, is a complete metric space.*

Proof: Let (f_n) be a Cauchy sequence in $C(K)$ for the uniform norm. We have to show that (f_n) converges in the uniform norm. By Theorem 2.9 the limit function f will then automatically be continuous and hence $f \in C(K)$. We do this in two steps a) find a candidate function f for the limit of the sequence, and b) show that the sequence indeed converges uniformly to f.

a) Note that (f_n) converges pointwise due to the completeness of $\mathbb{R}$ and the observations that for all $x \in K$ the sequence of real numbers $(f_n(x))$ is Cauchy. Call the limit of $(f_n(x))$, $f(x)$. This defines f.

b) We show explicitly that $\lim_n \|f_n - f\|_{\mathrm{sup}} = 0$.

$$\|f_n - f\|_{\mathrm{sup}} = \sup_{x \in K} |f_n(x) - f(x)|$$

$$
\begin{aligned}
&= \sup_{x\in K} \lim_{m\to\infty} |f_n(x) - f_m(x)| \\
&= \sup_{x\in K} \liminf_{m\to\infty} |f_n(x) - f_m(x)| \\
&\leq \liminf_{m\to\infty} \sup_{x\in K} |f_n(x) - f_m(x)| \qquad (2.7)
\end{aligned}
$$

For the last step we have used (1.7) The fact that (f_n) is Cauchy in the uniform norm means that, for all $\epsilon > 0$,

$$\sup_{x\in K} |f_n(x) - f_m(x)| < \epsilon, \text{ for all } n, m \text{ large enough.}$$

This proves that $\|f_n - f\|_{\sup} \to 0$. ∎

$C(K)$ has additional structure: it is also a linear space for the natural operations of addition and multiplication by a scalar of functions, and its metric is derived from a norm. Normed linear spaces that are complete are called *Banach spaces*, and so, $C(K)$ is a Banach space. In fact, one can also multiply two functions (pointwise) and thus give $C(K)$ an *algebra* structure. One says that $C(K)$ is a *Banach algebra*.

Instead of real-valued functions on K, one can equally well consider complex-valued functions and the same result holds: the complex-valued continuous functions on K form a Banach space.

Exercise 2.5 *Let $f_n \in C([a,b])$ be a sequence of functions converging uniformly to a function f. Show that*

$$\lim_{n\to\infty} \int_a^b f_n(x)\, dx = \int_a^b f(x)\, dx \quad ,$$

i.e., show that $f \mapsto \int_a^b f(x)\, dx$, is continuous on $C([a,b])$.

If X is not compact, continuous functions defined on it may be unbounded. But it is still true that the space of all *bounded continuous functions* on X, denoted by $C_b(X)$, is a Banach space for the supremum norm. In the case of $X = \mathbb{R}$, there is another commonly used Banach space of continuous functions: the space $C_0(\mathbb{R})$ consisting of all continuous functions vanishing at infinity, i.e.,

$$C_0(\mathbb{R}) = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous and } \lim_{x\to\pm\infty} f(x) = 0\}$$

Exercise 2.6 *Prove that $C_0(\mathbb{R})$ is a closed subspace of $C_b(\mathbb{R})$ with the sup norm.*

Exercise 2.7 *Show that $(C_0(\mathbb{R}), \|\cdot\|_{\sup})$ and $(C_b(\mathbb{R}), \|\cdot\|_{\sup})$, are Banach spaces. (Hint: adapt the proof of Theorem 2.10.)*

Theorem 2.11 *Let K be a compact metric space and let $f : K \to \mathbb{R}$ be a continuous real-valued function on K. Then, f is uniformly continuous.*

Proof: We prove the theorem by contraposition. Suppose f is not uniformly continuous. Then, there exists $\epsilon > 0$ such that for all $\delta > 0$, there are x, y, $d(x, y) < \delta$, but $d(f(x), f(y)) > \epsilon$. We use this for the sequence of δ's, $\delta_n = 1/n$ to obtain two sequences (x_n) and y_n in K with the properties:

$$d(x_n, y_n) < 1/n \quad \text{and} \quad d(f(x_n), f(y_n)) > \epsilon \tag{2.8}$$

Since K is compact there are convergent subsequences of (x_n) and (y_n), which for simplicity we again denote by (x_n) and (y_n), with limits x and y in K, and satisfying (2.8). Clearly, $x = y$, because

$$d(x, y) \leq d(x, x_n) + d(x_n, y_n) + d(y_n, y)$$

and the three terms on the RHS of this inequality converge to 0. We conclude the proof by showing that assuming that all this holds for a continuous function f implies that f is not continuous at x (*reductio ad absurdum*)! Indeed, starting from (2.8), we have

$$\epsilon < d(f(x_n), f(y_n)) \leq d(f(x_n), f(x)) + d(f(x), f(y_n))$$

As the two terms on the RHS converge to) as $m \to \infty$, this contradicts the continuity of f. ■

Let $[a, b] \subset \mathbb{R}$ be a closed interval. Consider the polynomial functions on $[a, b]$, i.e., functions p of the form

$$p(x) = \sum_{k=0}^{n} a_k x^k$$

where the coefficients a_k are arbitrary real numbers and $n \in \mathbb{N}$ is called the degree of p if $a_n \neq 0$.

Our next goal is the classical Weierstrass Approximation Theorem, which states that any continuous function $f : [a, b] \to \mathbb{R}$ can be approximated in the sense of the uniform norm by a polynomial (function), to arbitrary precision. Such an approximation property can be concisely formulated using the notion of dense set.

Theorem 2.12 (Weierstrass Approximation Theorem) *The set of all polynomial functions is dense in $C([a, b])$.*

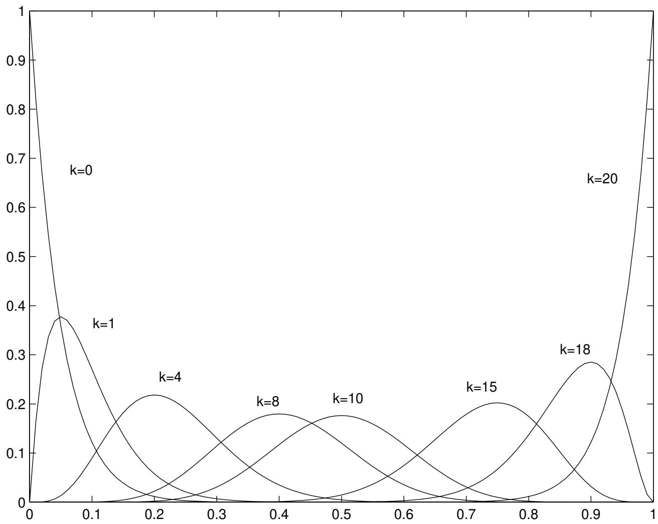

Figure 2.3: The polynominals $x^k(1-x)^{n-k}$, for the case $n = 20$, appearing in the definition of the Bernstein polynomials (2.9).

Proof: We need to show that for any $f \in C([a, b])$ there is a sequence of polynomials p_n such that $\lim_n \|f - p_n\| = 0$.

We first show that it is sufficient to prove the theorem in the case $a = 0, b = 1$. This is so because rescaling and shifting the argument x are norm-preserving transformations from $C([a, b] \to C([0, 1])$ that map the polynomial functions into themselves. More explicitly, define $T : C([a, b] \to C([0, 1])$ by

$$Tf(x) = f(a + (b - a)x)$$

Then, $\|Tf\|_{\text{sup}} = \|f\|_{\text{sup}}$ and T is linear, in particular $T(f - g) = Tf - Tg$, and T is invertible with inverse

$$T^{-1}f = f(\frac{x-a}{b-a})$$

It is also easy to see that t and T^{-1} preserve continuity, and that for any polynomial p, Tp and $T^{-1}p$ are again polynomials. Then, for any $f \in C([a, b])$, $Tf \in C([0, 1])$, and assuming we have shown that the polynomials are dense in $C([0, 1])$, we then have a sequence of polynomials p_n such that $p_n \to Tf$ in $C([0, 1])$ with the uniform norm. From the properties discussed above it follows that the sequence $T^{-1}p_n$ is a sequence of polynomials converging to f.

So, we have to prove that the polynomial functions are dense in $C([0, 1])$. We follow the proof by Bernstein, who gave, for every f, an explicit formula for a sequence of polynomials converging to f. These are called the Bernstein polynomials and are defined as follows:

$$B_n(x;f) = \sum_{k=0}^{n} f(\frac{k}{n})\binom{n}{k}x^k(1-x)^{n-k} \tag{2.9}$$

$B_n(x;1) = 1$, by the binomial theorem. Therefore,

$$|f(x) - B_n(x;f)| \le \sum_{k=0}^{n} |f(x) - f(\frac{k}{n})|\binom{n}{k}x^k(1-x)^{n-k} \tag{2.10}$$

Hence,

$$\begin{aligned} \|B_b(\cdot\,;f) - f\|_{\sup} &= \sup_{0\le x\le 1}\sum_{k=0}^{n}\left(f(\frac{k}{n}) - f(x)\right)\binom{n}{k}x^k(1-x)^{n-k} \\ &\le \sup_{0\le x\le 1}\sum_{k=0}^{n}\left|f(\frac{k}{n}) - f(x)\right|\binom{n}{k}x^k(1-x)^{n-k} \end{aligned} \tag{2.11}$$

Next, we use the fact that f is *uniformly* continuous (Theorem 2.11), and pick a $\delta > 0$ such that $|f(k/n) - f(x)| < \epsilon$, for all k such that $|x - (k/n)| < \delta$. And define $I(\delta, x) = \{k||x - n/k| \ge \delta\}$. Then we can estimate the RHS of (2.11) as follows:

$$\begin{aligned} \|B_b(\cdot\,;f) - f\|_{\sup} &\le \sup_{0\le x\le 1}\epsilon\sum_{0\le k\le n, k\notin I(\delta,x)}\binom{n}{k}x^k(1-x)^{n-k} \\ &\quad + \sup_{0\le x\le 1}\sum_{k\in I(\delta,x)}\left|f(\frac{k}{n}) - f(x)\right|\binom{n}{k}x^k(1-x)^{n-k} \\ &\le \epsilon\sum_{k=0}^{n}\binom{n}{k}x^k(1-x)^{n-k} \\ &\quad + 2\|f\|_{\sup}\sup_{0\le x\le 1}\sum_{k\in I(\delta,x)}\binom{n}{k}x^k(1-x)^{n-k} \end{aligned} \tag{2.12}$$

The first term is equal to ϵ, and the second term can be estimated as follows:

$$\begin{aligned} &2\|f\|_{\sup}\sup_{0\le x\le 1}\sum_{k\in I(\delta,x)}\binom{n}{k}x^k(1-x)^{n-k} \\ &\le 2\|f\|_{\sup}\sup_{0\le x\le 1}\frac{1}{\delta^2}\sum_{k=0}^{n}(x - \frac{n}{k})^2\binom{n}{k}x^k(1-x)^{n-k} \\ &\le 2\|f\|_{\sup}\sup_{0\le x\le 1}\frac{1}{\delta^2}\sum_{k=0}^{n}(x^2 - 2\frac{k}{n}x + \frac{k^2}{n^2})\binom{n}{k}x^k(1-x)^{n-k} \end{aligned}$$

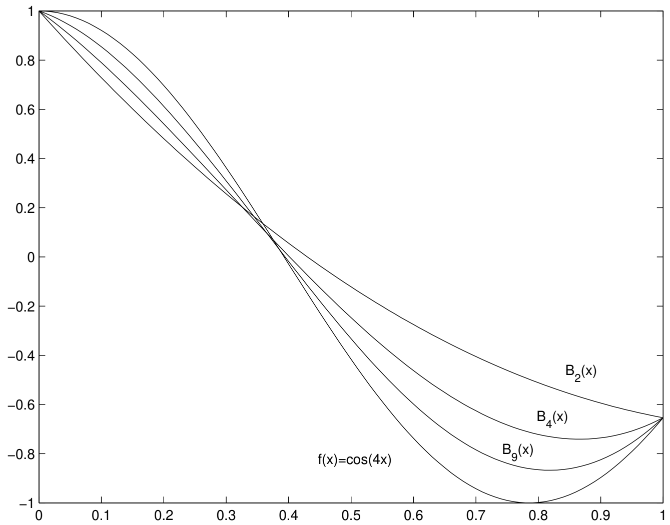

Figure 2.4: Some approximations of the function $x \mapsto \cos(4x)$ by its Berstein polynomials.

$$\leq \frac{2\|f\|_{\sup}}{\delta^2} \sup_{0\leq x\leq 1} \big(x^2 B_n(x;1) - 2xB_n(x;x) + B_n(x;x^2)\big) \tag{2.13}$$

Here, $1, x$, and x^2 denote the corresponding polynomial functions, when they serve as the second argument of B_n. We can now compute $B_n(x;1), B_n(x;x)$, and $B_n(x;x^2)$ explicitly to finish the estimate.

$$\begin{aligned} B_n(x;1) &= 1 && \text{for all } n \geq 0 \\ B_n(x;x) &= x && \text{for all } n \geq 1 \\ B_n(x;x^2) &= \tfrac{n-1}{n}x^2 + \tfrac{1}{n}x && \text{for all } n \geq 1 \end{aligned} \tag{2.14}$$

All combined we obtain the estimate

$$\|B_n(\cdot\,;f) - f\|_{\sup} \leq \epsilon + \frac{\|f\|_{\sup}}{2n\delta^2}$$

which concludes the proof.

[Hint: prove (2.14) using the binomial theorem to expand $(x+y)^n$ and take derivatives with respect to x to obtain the identities:

$$\begin{aligned} x(x+y)^{n-1} &= \sum_{k=0}^{n} \binom{n}{k} \frac{k}{n} x^k y^{n-k} \\ \frac{1}{n}x(x+y)^{n-1} + \frac{n-1}{n}x^2(x+y)^{n-2} &= \sum_{k=0}^{n} \binom{n}{k} \frac{k^2}{n^2} x^k y^{n-k} \quad] \end{aligned} \tag{2.15}$$

■

Note that the proof works for both real-valued and complex-valued functions.

The space of all continuous functions (with values in $\mathbb{R}$ or $\mathbb{C}$) defined on the unit circle, T, is closely related to $C([0,1])$. A very usefull density result for $C(T)$ is the following.

If we parametrize the unit circle by an angle θ, i.e.,

$$T = \{e^{i\theta} \mid \theta \in [0, 2\pi)\}$$

any continuous function $f : T \to \mathbb{C}$ can be considered as a continuous periodic function of the real variable θ. E.g., the trigonometric polynomials are continuous. A *trigonometric polynomial* of degree n is a function of the form

$$P_n(\theta) = \sum_{k=-n}^{n} a_k e^{ik\theta}$$

the a_k are arbitrary complex coefficients.

The following density result, which has many applications, will be proved later in Chapter 8, when we discuss Fourier series.

Theorem 2.13 *The set of all trigonometric polynomials is dense in $C(T)$ with the uniform norm.*

2.4 Compact subsets of $C(K)$

Recall that the proof of the Heine-Borel Theorem used the finite-dimensionality of $\mathbb{R}^n$ in a crucial way. In infinite-dimensional spaces compact sets are also necessarily bounded and closed, but these properties are no longer sufficient. The aim of this subsection is to prove a characterization of the compact subsets of $C(K)$, the space of all real-valued continuous functions on a compact space K with the uniform norm (the Arzelá-Ascoli Theorem). For this we will need the fact that every compact set has the *finite-open-subcover* property. First, we explain what the finite-open-subcover property is.

A family of open sets $G_\alpha \subset X$ is called an *open cover* of $A \subset X$ if

$$A \subset \bigcup_\alpha G_\alpha$$

The index α runs through an arbitrary index set that need not be countable. For any subset A of a metric space (X, d), it is obvious that for any $\epsilon > 0$, A can be covered with a family of open balls in X of radius ϵ, meaning that there exists a family of points $x_\alpha \in X$ such that

$$A \subset \bigcup_\alpha B_\epsilon(x_\alpha) \tag{2.16}$$

Indeed, it is always possible to take $\{x_\alpha\} = A$. Let us call such a cover an *ϵ-cover* of A. A subset $\{x_i\} \subset A$ with the property (2.16) is called an *ϵ-net* for A, and a finite ϵ-net if the set of points is finite. $A \subset X$ is called *totally bounded* if A has a finite ϵ-net for every $\epsilon > 0$.

We say that A has the finite-open-subcover property if for all open covers $\{G_\alpha\}$ of A there is a finite subcover. By this we mean that we can find a finite subset of the G_α, say $\{G_{\alpha_1}, \ldots, G_{\alpha_k}\}$ such that

$$A \subset \bigcup_{i=1}^k G_{\alpha_i}$$

The finite-open-subcover property is equivalent to compactness. We will prove one direction of this equivalence in Theorem 2.16. First, we prove some preliminary results.

Theorem 2.14 *Every compact metric space K is totally bounded, i.e., has a finite ϵ-net for every $\epsilon > 0$.*

Proof: Let $\epsilon > 0$, and suppose that K has no finite ϵ-net. Then, for every finite subset $\{x_1, \ldots, x_n\} \subset K$, there exists

$$x_{n+1} \in K \setminus \bigcup_{i=1}^n B_\epsilon(x_i)$$

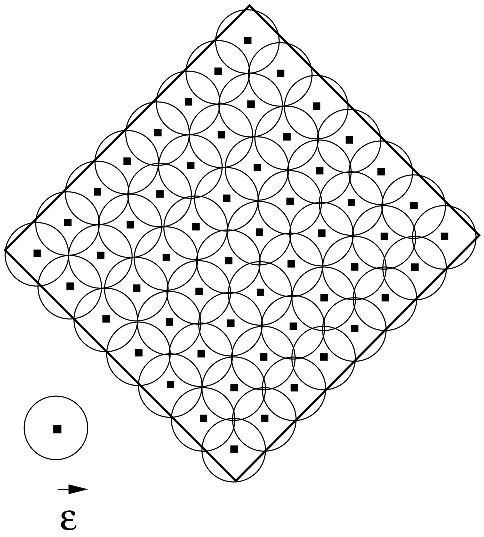

Figure 2.5: A set with an ϵ-net for it.

So, we can find an infinite sequence (x_n) in K with the property that $d(x_n, x_m) > \epsilon$ for all $n \neq m$. Clearly, this sequence does not contain any Cauchy subsequence, and hence no convergent subsequence. We conclude that K must have a finite ϵ-net for every $\epsilon > 0$. ∎

A metric space is called *separable* if it contains a countable dense subset.

Corollary 2.15 *Every compact metric space K is separable, i.e., contains a countable dense subset.*

Proof: For every $n \geq 1$, let A_n be a finite $\frac{1}{n}$-net of K. Such A_n exist due to Theorem 2.14. Then, $A := \bigcup_{n \in \mathbb{N}} A_n$ is dense in K, and, as a countable union of finite sets, A is countable. ∎

The notation LHS := RHS is often used when the RHS *defines* the LHS. Other commonly used equivalent notations are $\equiv$ and $\stackrel{\text{def}}{=}$.

Theorem 2.16 *Let (X, d) be a metric space and suppose $K \subset X$ is compact. Then K has the finite-open-subcover property.*

Proof: We will first show that any open cover $\{G_\alpha\}$ of K contains a countable subcover. From Corollary 2.15, we know that K contains a countable dense subset A. Consider the family $\mathcal{B}$ of open balls with center in A and radius in $\mathbb{Q}$. Then, $\mathcal{B}$ is a countable family. Consider the subset $\mathcal{C}$ of $\mathcal{B}$ consisting of all $B \in \mathcal{B}$ that are contained in one of the G_α. As a subset of a countable set, $\mathcal{C}$ is still countable. For every $B \in C$, pick α_B such that $B \subset G_\alpha$. Then, $\{G_{\alpha_B}\}_{B\in\mathcal{C}}$ is a countable subset of $\{G_\alpha\}$, and that it is a cover of K can be seen as follows. For every $x \in K$, there is a member of the open cover G_α, such that $x \in G_\alpha$. As G_α is open, there is an open ball $B_\epsilon(x) \subset G_\alpha$, and we can assume that ϵ is rational. As A is dense, there exists $y \in A$ such that $d(x, y) < \epsilon/2$. Then $x \in B_{\epsilon/2}(y) \subset G_\alpha$, and, hence, $B_{\epsilon/2}(y) \in \mathcal{C}$, and it follows that $\mathcal{C}$ is a cover. By construction we have

$$\bigcup_{B\in\mathcal{C}} B \subset \bigcup_{B\in\mathcal{C}} G_{\alpha_B}$$

Therefore, $\{G_{\alpha_B}\}_{B\in\mathcal{C}}$ is a countable subcover of $\{G_\alpha\}$.

Next, we show that any countable open cover $\{G_i \mid i \in \mathbb{N}\}$ contains a finite subcover. To prove this we will perform a construction that either selects a finite subcover or yields a sequence (x_n) of elements $x_n \in K$ without a convergent subsequence with limit in K. If K is compact the last possibility must be excluded and hence there must be a finite subcover.

Define

$$R_n = K \setminus \bigcup_{i=1}^{n} G_i$$

If $R_n = \emptyset$ for some n, we have found a finite subcover and we are done. If $R_n \neq \emptyset$, for all n, we can construct a sequence $y_k \in K$ as follows. Pick $y_1 \in R_1$. Let n_1 be such that $y_1 \in G_{n_1}$. Then pick $y_2 \in R_{n_1}$, and so on. The result is a sequence y_k such that

$$y_{k+1} \notin \bigcup_{i=1}^{n_k} G_i \quad \text{and} \quad y_{k+1} \in \bigcup_{i=1}^{n_{k+1}} G_i \tag{2.17}$$

Suppose (y_{k_n}) is a convergent subsequence of (y_k), with limit y. Then, $y \notin G_i$, for any i, for if $y \in G_i$ for some i, (y_{k_n}) cannot converge to y. This follows from (2.17) and the fact that the G_i are open. We conclude that $y \notin K \subset \bigcup_i G_i$.

As K is assumed to be compact the possibility that all $R_n \neq \emptyset$ must be excluded and hence there is a finite subcover of K. ■

Exercise 2.8 (Dini's Monotone Convergence Theorem)
Suppose $f_n \in C([0,1])$ converges pointwise and monotone decreasingly *to $f \in C([0,1])$. I.e., for all $x \in [0,1]$, the sequence $(f_n(x))$ is monotone decreasing and converges to $f(x)$. Prove that f_n converges uniformly to f.*

For the characterization of the compact subsets of $C(K)$ we will need the notion of *equicontinuity.*

Definition 2.17 (Equicontinuity) *Let F be a family of continous functions from (X, d) to (Y, d). F is equicontinuous iff for all $x \in X$, and for all $\epsilon > 0$, there is a $\delta > 0$ such that for all y with $d(x, y) < \delta$, one has $d(f(x), f(y)) < \epsilon$, for all $f \in F$.*

The crucial point in this definition is that δ can be chosen independent of f, while it may depend on x. If δ can be chosen independent of x as well, the family is *uniformly equicontinuous*. If X is compact equicontinuity implies uniform equicontinuity, as is shown in the following theorem.

Theorem 2.18 *Let F be a family of continuous functions from (K, d) to (Y, d), where K is compact. If F is equicontinuous, then F is uniformly equicontinuous.*

Proof: Suppose the equicontinuous family F is not uniformly equicontinuous. Then, there must be $\epsilon > 0$, such that for all $\delta > 0$ there are x_δ, and f_δ, such that there exist $y, d(x_\delta, y) < \delta$, but $d(f_\delta(y), f_\delta(x_\delta)) > 2\epsilon$.

In particular there are sequences x_n, y_n, and f_n, such that for all n, $d(x_n, y_n) < 1/n$, and

$$d(f_n(y_n), f_n(x_n)) > 2\epsilon \tag{2.18}$$

As K is compact, (x_n) has a convergent subsequence. Let such a subsequence again be denoted by (x_n). We then have $x_n \to x_0$, and hence $y_n \to 0$. Choose n large enough such that $d(x_n, x_0) < \delta$ and $d(y_n, x_0) < \delta$. Because of (2.18) we must have at least one of the following two: $d(f_n(y_n), f_n(x_0)) > \epsilon$ or $d(f_n(x_n), f_n(x_0)) > \epsilon$. As $\delta > 0$ is arbitrary, this is in contradiction with F being equicontinuous in x_0. ∎

Note that this theorem implies Theorem 2.11. It suffices to take $F = \{f\}$.

A set A is called *relatively compact*, or *precompact*, if its closure $\overline{A}$ is compact.

Theorem 2.19 (Arzelà-Ascoli) *Let K be compact. A set $A \subset C(K)$ with the uniform norm is compact iff A has the following two properties:*
i) A is closed and bounded, and
ii)A is equicontinuous.
Equivalently, A is precompact iff it satisfies:
i′) A is bounded, and
ii′) A is equicontinuous.

Proof: We divide the proof into four parts:
1) precompact $\Rightarrow$ bounded
2) compact $\Rightarrow$ closed
3) precompact $\Rightarrow$ equicontinuous
4) bounded + equicontinuous $\Rightarrow$ precompact.

1) If A is not bounded there is a sequence of functions $f_n \in A$, satisfying $\|f_{n+1}\| \geq \|f_n\| + 1$. Then, $\|f_n - f_m\| \geq 1$, for all $n \neq m$. It follows that (f_n) that does not have a Cauchy subsequence, and hence does not have a convergent subsequence.

2) If A is not closed, there is a sequence $f_n \in A$ converging to a limit $f \in \overline{A} \setminus A$. All subsequences of this sequence also converge to f, and hence do not have a limit *in* A.

3) If A is precompact, then $\overline{A}$ is compact by definition. Fix $\epsilon > 0$. A is dense in $\overline{A}$. Therefore

$$\overline{A} \subset \bigcup_{f \in A} B_\epsilon(f)$$

By Theorems 2.14 and 2.16 there is then a finite subset $\{f_1, \ldots, f_k\} \subset A$ such that

$$\overline{A} \subset \bigcup_{i=1}^{k} B_\epsilon(f_i)$$

As each f_i is uniformly continuous (Theorem 2.11), there exist $\delta_i > 0$ such that for all $x, y \in K$, we have

$$d(x,y) < \delta_i \Rightarrow |f_i(x) - f_i(y)| < \epsilon. \tag{2.19}$$

Then, for any $f \in A$, there is an i such that $\|f - f_i\| < \epsilon$ (because $\{f_i\}$ is an ϵ-net). Take $\delta = \min_i \delta_i$. We conclude that for $d(x,y) < \delta$

$$\begin{aligned} |f(x) - f(y)| &\leq |f(x) - f_i(x)| + |f_i(x) - f_i(y)| + |f_i(y) - f(y)| \\ &\leq 3\epsilon \end{aligned} \tag{2.20}$$

As ϵ is arbitrary, and δ is independent of f, this shows that A is equicontinuous.

4) We will show, using boundedness and equicontinuity, that every sequence in A has a convergent subsequence (then, necessarily, with a limit in $\overline{A}$). By Corollary 2.15 there is a countable dense set $\{x_1, x_2, x_3, \ldots\} \subset K$. We now choose a first subsequence $(f_{1,n})$ of (f_n) such that the sequence $(f_{1,n}(x_1))$ converges in $\mathbb{R}$. Such a subsequence exists because A, and hence $(f_n(x_1))$, is bounded. Next, we choose a subsequence $(f_{2,n})$ of $(f_{1,n})$ such that $(f_{2,n}(x_2))$ converges, which exists for the same reason. And so on. $(f_{(k+1),n})$ is a subsequence of $(f_{k,n})$ such that $(f_{(k+1),n}(x_{k+1}))$ converges as $n \to \infty$. Finally, the subsequence we are after is the "diagonal" of the "table" of functions $f_{i,j}$, i.e., define the subsequence (g_k) of (f_n) by $g_k = f_{k,k}$.

By construction (g_k) has the property that $g_k(x_i)$ converges in $\mathbb{R}$ as $k \to \infty$ for all x_i in a dense subset of K. So far, we only used the boundedness of A. The equicontinuity is needed to ensure that (g_k) converges uniformly to a limit function g.

Let $\epsilon > 0$. Since A is equicontinuous, A is uniformly equicontinuous by Theorem 2.18. Therefore, and since all $g_k \in A$, there exists $\delta > 0$ such that for all x, y such that $d(x,y) < \delta$,

$$|g_k(x) - g_k(y)| < \epsilon$$

Since $\{x_i\}$ is dense in K we have

$$K \subset \bigcup_{i=1}^{\infty} B_\delta(x_i)$$

By Theorem 2.16 there is a finite subset $\{x_{i_1}, \ldots, x_{i_l}\}$ such that

$$K \subset \bigcup_{j=1}^{l} B_\delta(x_{i_j})$$

Since for all $j = 1, \ldots, l$, $(g_k(x_{i_j}))$ is convergent, and hence Cauchy, we can now find N such that

$$\text{for all } p, q \geq N, \quad |g_p(x_{i_j}) - g_q(x_{i_j})| < \epsilon \text{ for all } j = 1, \ldots, l$$

Now, for any $x \in K$,

$$\begin{aligned} |g_p(x) - g_q(x)| &\leq |g_p(x) - g_p(x_{i_j})| + |g_p(x_{i_j}) - g_q(x_{i_j})| + |g_q(x_{i_j}) - g_q(x)| \\ &\leq 3\epsilon \end{aligned} \tag{2.21}$$

where we have made use of the δ-cover to choose x_{i_j} appropriately. It follows that (g_k) is a Cauchy sequence for the uniform norm. As $C(K)$ is complete (Theorem 2.10), it converges. ■

Remarks:

1. In the proof of 4) we used Cantor's "*diagonal argument*". Keep it in mind! It is used in other proofs. Also note the repeated use of the $\epsilon/3$-trick (in the disguise of a 3ϵ-trick).

2. The same result can be proved (in exactly the same way) for a more general situation: consider the space $C(K, Y)$ of all continuous functions $f : K \to Y$, with K a compact and Y an arbitrary metric space. Define a metric on $C(K, Y)$ by
$$d(f, g) = \sup_{x \in K} d(f(x), g(x))$$
Then $(C(K, Y), d)$ is a complete metric space and the compact subsets of $C(K, Y)$ are exactly those that have the properties i) and ii) stated in the Arzelà-Ascoli Theorem.

3. Parts 1) and 2) of the proof work as stated for arbitrary normed spaces.

Examples and counterexamples

We start with two examples of bounded sets in $C([0, 1])$ that are not precompact (and hence their closures are bounded and closed but not compact). Both examples make use of the fact that, as a linear space, $C([0, 1])$ is infinite-dimensional.

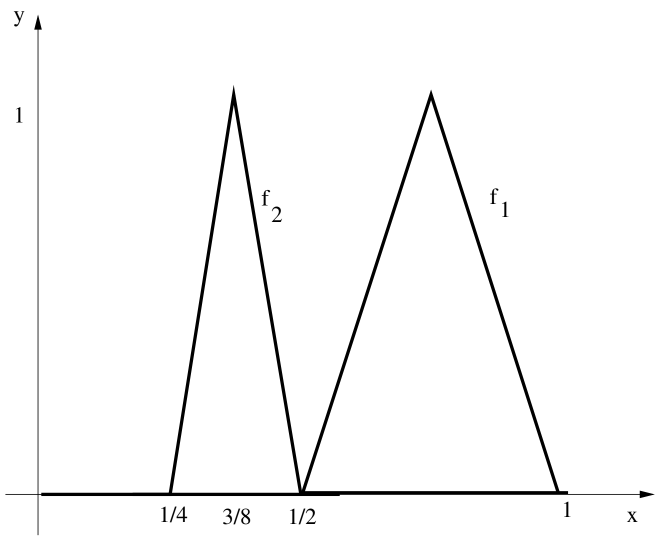

Figure 2.6: The sequence of bounded continuous functions defined in Example 1 is not equicontinuous.

1. For each $n \geq 0$ define a function $f_n : [0,1] \to \mathbb{R}$ by

$$f_n(x) = \begin{cases} 0 & \text{if } \ 0 \leq x \leq 2^{-(n+1)} \\ 2^{n+2}(x - 2^{-(n+1)}) & \text{if } \ 2^{-(n+1)} \leq x \leq 3 \cdot 2^{-(n+2)} \\ 2^{n+2}(2^{-n} - x) & \text{if } \ 3 \cdot 2^{-(n+2)} \leq x \leq 2^{-n} \end{cases} \tag{2.22}$$

Then $\|f_n\|_\infty = 1$, for all $n \geq 1$, and so $\{f_n\}$ is bounded, but $\|f_n - f_m\|_\infty = 1$ for all $n \neq m$. Hence, the sequence (f_n) does not have any Cauchy subsequences, and therefore, as a set it can not be precompact. Look at the graphs of the functions f_n (Figure 2.6) and show explicitly that $\{f_n\}$ is not equicontinuous.
2. The same phenomenon occurs for the set of functions f_n defined on $[0,1]$ by $f_n(x) = \sin(n\pi x)$.

For a subset of an infinite-dimensional linear space to be (pre)compact "most of it" must be contained in a finite-dimensional subspace. Without making this more precise, we illustrate this with the following example.

3. Define a subset A of $C([0,1])$ as follows:

$$A = \{f \in C([0,1]) \mid f(x) = \textstyle\sum_{n=1}^\infty a_n \sin(n\pi x) \text{ with } a_n \in \mathbb{R}, \text{ and } \sum_{n=0}^\infty n^2 a_n^2 \leq 1\}$$

It can be shown that is precompact in $(C([0,1]), \|\cdot\|_\infty)$, by verifying the assumptions of the Arzelà-Ascoli Theorem.

Next, we give two examples of useful sufficient conditions for equicontinuity.

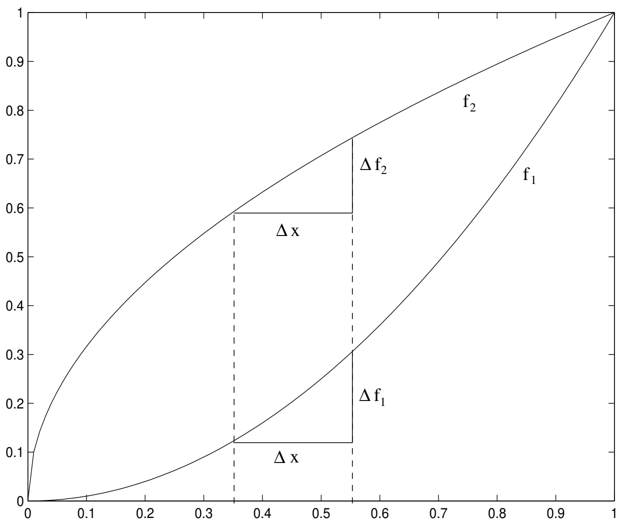

Figure 2.7: The graph of two continuous functions on $[0,1]$: $f_1(x) = x^2$, and $f_2(x) = \sqrt{x}$. f_1 is Lipschitz on $[0,1]$, but f_2 is not Lipschitz in the point 0. This follows from the observation that the ratio $\Delta f_1/\Delta x$, is bounded (for arbitrarily small Δx, and everywhere in $[0,1]$), while $\Delta f_2/\Delta x$ is unbounded for small Δx near 0.

4. Let (X, d) be a metric space. A function $f : X \to \mathbb{R}$ is called (globally) *Lipschitz continuous* if there exists a constant $C \geq 0$ such that

$$|f(x) - f(y)| \leq Cd(x, y)$$

It is obvious that Lipschitz continuous $\Rightarrow$ continuous. For a Lipschitz continuous f we can define the *Lipschitz constant* $\mathrm{Lip}(f)$ of f by

$$\mathrm{Lip}(f) = \sup_{x \neq y \in X} \frac{|f(x) - f(y)|}{d(x, y)}$$

or, equivalently

$$\mathrm{Lip}(f) = \inf\{C \geq 0 \mid |f(x) - f(y)| \leq Cd(x, y), \text{ for all } x, y \in X\}$$

For any compact metric space K and any constant M define

$$A = \{f \in C(K) \mid \mathrm{Lip}(f) \leq M\} \tag{2.23}$$

Then, A is equicontinuous. Indeed, for all $\delta > 0$ we have

$$d(x, y) < \delta \Rightarrow |f(x) - f(y)| \leq M\delta \quad \text{for all } f \in A; x, y \in K$$

The set A defined by (2.23) is not bounded. E.g., it is easy to see that $f \in A \Rightarrow f+a \in A$, for any $a \in \mathbb{R}$, and hence, A is unbounded. Here is a bounded set $B \subset A$, in the case $K = [0,1]$:

$$B = \{f \in A \mid f(0) = 0\}$$

Now, B is bounded because

$$|f(x)| = |f(x) - f(0)| \leq M|x - 0| \leq M$$

In fact, B is also closed (and hence compact). In fact, A is closed for arbitrary K and M. To check this we use once more our favorite relation involving lim inf. Let f_n be a sequence in B converging uniformly to f. Then

$$\begin{aligned}
\mathrm{Lip}(f) &= \sup_{x \neq y \in K} \frac{|f(x) - f(y)|}{d(x,y)} \\
&= \sup_{x \neq y \in K} \liminf_{n \to \infty} \frac{|f_n(x) - f_n(y)|}{d(x,y)} \\
&\leq \liminf_{n \to \infty} \sup_{x \neq y \in K} \frac{|f_n(x) - f_n(y)|}{d(x,y)} \\
&\leq \liminf_{n \to \infty} M = M
\end{aligned}$$

5. Let $C^1([0,1])$ denote the space of all continuously differentiable real-valued functions on $[0,1]$, and suppose $f \in C^1([0,1])$. Then, by the mean-value theorem, for every $x, y \in [0,1]$, $x < y$, there is z, $x < z < y$, such that

$$f(x) - f(y) = f'(z)(x - y)$$

It follows that $\mathrm{Lip}(f) \leq \|f'\|_\infty$, which is finite because $f' \in C([0,1])$ by assumption. For any constants M and N define a set $D \subset C^1([0,1])$ by

$$D = \{f \in C^1([0,1]) \mid \|f\|_\infty \leq M, \|f'\|_\infty \leq N\}$$

It follows that D is bounded and equicontinuous, and hence precompact in $C([0,1])$ with the uniform norm.

2.5 Existence of local solutions of ODE's

This section is devoted to a simple application of the Arzelà-Ascoli Theorem. Consider a scalar ordinary differential equation (ODE) of the type

$$y' = F(x,y)$$

on an open domain $D \subset \mathbb{R}$, and with "initial" condition $y(x_0) = y_0$. Solving this ODE on $D_0 \subset D$ means finding a function (or functions) $f : D_0 \to \mathbb{R}$ such that f and its derivative f' satisfy

$$\left.\begin{array}{rcl} f'(x) & = & F(x, f(x)) \quad \text{for all } x \in D_0 \\ f(x_0) & = & y_0 \end{array}\right\} \tag{2.24}$$

We made a distinction between D and D_0, because it is a priori not clear one can find a solution f defined on the whole of D on which the problem is posed. But under mild conditions on F there is always an open interval D_0, containing x_0, on which at least one solution exists.

Theorem 2.20 (Local existence for scalar ODE) *If F is continuous on an open set $R \in \mathbb{R}^2$, containing (x_0, y_0), then there exists an open interval $D_0 \subset \mathbb{R}$, containing x_0, and a function $f : D_0 \to \mathbb{R}$ solving the system (2.24).*

Proof: The proof is mainly constructive, in the following sense. We will first construct a sequence of functions that "approximately" solve the ODE. This sequence will consist of elements in a precompact subset (Arzelà-Ascoli) of some $C([a, b])$, and hence it contains a uniformly convergent subsequence. We then show that any limit of such a subsequence must solve (2.24).

For any $\epsilon > 0$ we will define an approximate solution f_ϵ in the sense that

$$|f'_\epsilon(x) - F(x, f_\epsilon(x))| \leq \epsilon\,, \quad f_\epsilon(x_0) = y_0$$

for all x in some interval $[x_0 - d, x_0 + d]$, except for a finite number of points (where f'_ϵ does not exist). The f_ϵ are piecewise linear continuous functions obtained by the following, rather naive, construction:

$$f_\epsilon(x) = a_k + b_k(x - (x_0 + k\delta))\,, \qquad \text{for } x_0 + k\delta \leq x \leq x_0 + (k+1)\delta$$

for some appropriate $\delta > 0$, and a similar expression for $x \leq x_0$. The parameters a_k and b_k are recursively defined by

$$\begin{array}{rcl} a_0 & = & y_0 \\ b_0 & = & F(x_0, y_0) \\ a_k & = & a_{k-1} + b_{k-1}\delta \\ b_k & = & F(x_0 + k\delta, a_k) \end{array}$$

for $k \geq 1$. Then $f'_\epsilon(x) = b_k$, for $x_0 + k\delta < x < x_0 + (k+1)\delta$, and therefore we have

$$\begin{array}{rcl} |f'_\epsilon(x) - F(x, f_\epsilon(x))| & = & |b_k - F(x, a_k + b_k(x - (x_0 + k\delta))| \\ & = & |F(x_0 + k\delta, a_k) - F(x, a_k + b_k(x - (x_0 + k\delta))| \end{array} \tag{2.25}$$

for $x_0 + k\delta \leq x \leq x_0 + (k+1)\delta$, and

$$\begin{aligned} |x_0 + k\delta - x| &\leq \delta && (2.26) \\ |a_k - a_k + b_k(x - (x_0 + k\delta))| &\leq b_k\delta && (2.27) \end{aligned}$$

Let $R \subset D \subset \mathbb{R}^2$ be a closed rectangle with center (x_0, y_0). We will choose R such that the graph of each f_ϵ, for $x \in [x_0 - d, x_0 + d]$ is contained in R (See Figure 2.8). Then

$$|b_k| \leq \sup_{(x,y)\in R} |F(x,y)| \leq M$$

for some $M \in \mathbb{R}$, because F is continuous. In fact, R is compact, and hence, F restricted to R is uniformly continuous. It then follows form (2.26) and (2.27) that for every ϵ, there exist $\delta > 0$ such that

$$|F(x_0 + k\delta, a_k) - F(x, a_k + b_k(x - (x_0 + k\delta))| \leq \epsilon$$

Combined with (2.25) this shows that f_ϵ is an approximate solution.

A suitable rectangle R is found as follows. First, consider a rectangle $R_1 \subset D$, with center (x_0, y_0). Define

$$M = \sup_{(x,y)\in R_1} |F(x,y)|$$

Then choose $R \subset R_1$ of the form $[x_0 - d, x_0 + d] \times [y_0 - dM, y_0 + dM]$. Figure 2.8 makes it crystal clear why this choice of R works.

Let us summarize the properties of the approximate solution f_ϵ: for each ϵ, δ is found as described above. Then f_ϵ is piecewise linear on intervals of length δ. The slope of each linear piece is $\leq M$, and, therefore, the graph of f_ϵ lies in the cone bounded by $x_0 + Mx$ and $x_0 - Mx$ (see Figure 2.8). This also means that each f_ϵ is Lipschitz continuous with $\mathrm{Lip}(f_\epsilon) \leq M$. By construction we also have $f_\epsilon(x_0) = y_0$, for all ϵ. From example 4 in Section 2.4 we then know that the set $\{f_\epsilon \mid 0 < \epsilon \leq Md\}$ is bounded and equicontinuous, and hence, by the theorem of Arzelà-Ascoli (Theorem 2.19), precompact. It follows that there is a sequence $\epsilon_n \downarrow 0$ such that $f_{\epsilon_n} \to f$ uniformly on $[x_0 - d, x_0 + d]$.

It remains to be shown that this limiting function f solves (2.24). This goes as follows. As f_ϵ is piecewise linear we have

$$\begin{aligned} f_\epsilon(x) &= f_\epsilon(x_0) + \int_{x_0}^{x} f'_\epsilon(s)\, ds \\ &= y_0 + \int_{x_0}^{x} F(s, f_\epsilon(s))\, ds + \int_{x_0}^{x} [f'_\epsilon - F(s, f_\epsilon(s))]\, ds \end{aligned}$$

where f'_ϵ is not necessarily defined in points of the form $x_0 \pm k\delta$, but this does not affect the value of the integral (e.g., one could set the derivative equal to 0 in those

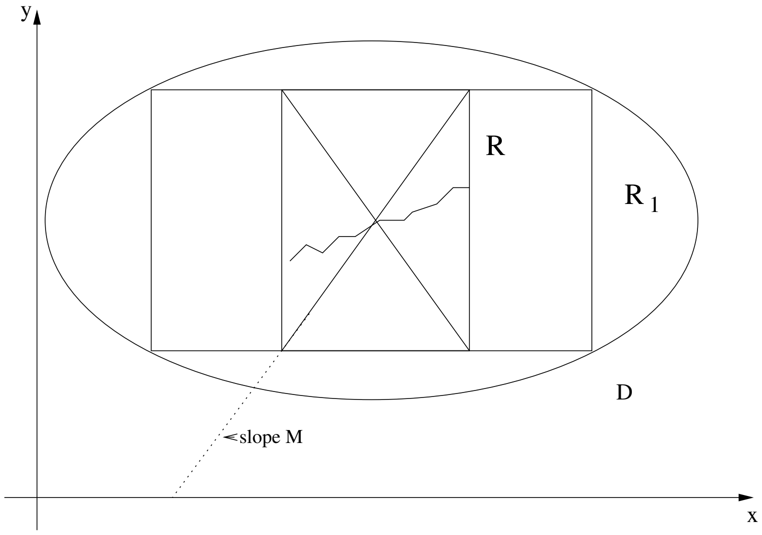

Figure 2.8: The construction of the rectangle R used in the proof of Theorem 2.20.

points). Now, Exercises 2.4 and 2.5 come in handy to show that

$$\begin{aligned} f(x) &= \lim_{n\to\infty}\left\{y_0 + \int_{x_0}^{x} F(s, f_{\epsilon_n}(s))\,ds + \int_{x_0}^{x}\left[f'_{\epsilon_n} - F(s, f_{\epsilon_n}(s))\right]\,ds\right\} \\ &= y_0 + \int_{x_0}^{x} F(s, f(s))\,ds \end{aligned} \tag{2.28}$$

Using the Fundamental Theorem of Calculus we conclude that the RHS of (2.28) is differentiable and therefore also the LHS is, and

$$f'(x) = F(x, f(x))$$

■

We observe that we also proved that there exists a differentiable solution.

Note that uniqueness of the solution is not part of Theorem 2.20. In fact, there are continuous F for which (2.24) has more than one solution on any open interval around x_0, as is shown in the following examples.

Examples of non-uniqueness:

1. Consider (2.24) with $F(x, y) = \sqrt{y}, y \geq 0$, and $x_0 = y_0 = 0$. I.e., we look for solutions of

$$f'(x) = \sqrt{f(x)} \tag{2.29}$$
$$f(0) = 0 \tag{2.30}$$

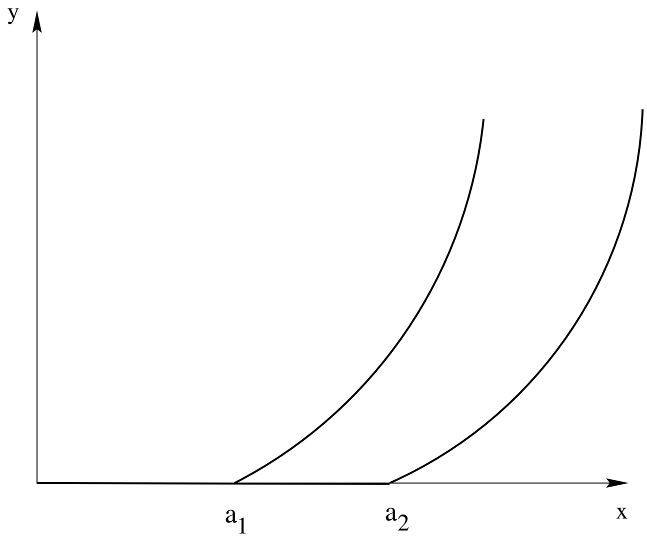

Figure 2.9: Two solutions of the ODE of Example 1.

Clearly, $f \equiv 0$ is a solution, but there is also the following one-parameter family of non-zero solutions (See Figure 2.9):

$$f(x) = \begin{cases} 0 & \text{if } x \leq a \\ \frac{1}{4}(x-a)^2 & \text{if } x > a \end{cases}$$

for any $a \geq 0$. The construction used in the proof of Theorem 2.20 would only yield the zero solution. One could slightly modify it to produce other solutions as well. In any case this example illustrates clearly that under the conditions stated in Theorem 2.20, one does not necesarily specify a unique function by the system (2.24).

To get a clue as to what extra condition would be needed to guarantee uniqueness of the solution, consider the following example, which is a generalization of Example 1.

2. For $\alpha \geq 0$, we consider

$$\left.\begin{array}{rcl} f'(x) & = & |f(x)|^\alpha \\ f(0) & = & 0 \end{array}\right\} \tag{2.31}$$

Again there is always the solution $f \equiv 0$. When one tries to find other solutions of the form

$$f(x) = \begin{cases} 0 & \text{if } x \leq a \\ C(x-a)^\nu & \text{if } x > a \end{cases}$$

one arrives at the following condition on the parameters C and ν:

$$\alpha = 1 - \frac{1}{\nu} \quad , c^{1-\frac{1}{\nu}} = C\nu \tag{2.32}$$

For $f \not\equiv 0$, we need $C \neq 0$, and the initial condition $f(0) = 0$ is only satisfied if $\nu > 0$. From (2.32) we conclude then conclude that $\alpha < 1$. If $\alpha \geq 1$ there are no non-zero solutions of the proposed from. In fact, as a direct application of the following proposition, one can show that (2.31) has indeed only one solution when $\alpha \geq 1$.

Proposition 2.21 *If F is continuous on an open set $R \subset \mathbb{R}^2$, containing (x_0, y_0), and F is Lipschitz continuous in its second argument, with a Lipschitz constant independent of x (on the domain R). Then, on any interval containing x_0 there exists at most one solution of the system (2.24).*

Proof: The strategy of this uniqueness proof is a very common one: suppose there are two solutions, say f and g, show that the difference $\Delta := f - g$ satisfies an equation that implies $\Delta = 0$.

If f and g both satisfy (2.24) on a interval I containing x_0, it is easy to see that Δ must satisfy

$$\left.\begin{array}{rcll} \Delta'(x) & = & F(x, f(x)) - F(x, g(x)) & , \text{ for all } x \in I \\ \Delta(x_0) & = & 0 & \end{array}\right\} \tag{2.33}$$

As F is Lipschitz continuous in its second argument, there is a constant C such that

$$|F(x, f(x)) - F(x, g(x))| \leq C|f(x) - g(x)| = C|\Delta(x)|$$

for all $x \in I$. This implies that the function G defined by

$$G(x) = \frac{F(x, f(x)) - F(x, g(x))}{\Delta(x)}$$

is bounded on I. On the other hand we have

$$\frac{\Delta'(x)}{\Delta(x)} = G(x)$$

for all $x \in I, x \neq x_0$. This equation can be integrated to obtain

$$\frac{\Delta(y)}{\Delta(x)} = \exp\{\int_x^y G(s)\, ds\}$$

for all $x \neq x_0$. Using the boundedness of G we conclude that Δ, which is by assumption a differentiable and hence continuous function, satisfies

$$e^{-M}|\Delta(x)| \leq |\Delta(y)| \leq e^M|\Delta(x)|$$

for all $x, y \in I$, and where M can be taken to be $C|I|$, with $|I|$ denoting the length of the interval I. Hence, Δ either never vanishes or is identically 0. As $\Delta(x_0) = 0$, we must have $\Delta = 0$. ■

We will see another, slightly more general proof of Proposition 2.21 in Section 5 as an application of the Contraction Mapping Theorem.

3 Topological Spaces

In the previous weeks we have generalized concepts from the finite-dimensional spaces $\mathbb{R}^n$ with the usual Euclidean metric, to general metric spaces. We are fundamentally interested in convergence of sequences and continuity of functions. These properties have been expressed in terms of the metric. And as an important auxiliary tool we explored the notion of compactness.

As a particularly useful example of infinite-dimensional metric spaces we studied the function spaces $C(K)$ in detail. We also mentioned that pointwise convergence of functions (e.g., real-valued functions defined on an interval) can generally not be expressed in terms of a metric on some function space, suggesting that there is an even more general framework for expressing convergence, continuity, and compactness. Indeed there is: topological spaces.

Here, we limit ourselves to the basic definitions. For more information on point set topology and its use, see, e.g.,
W. Rudin, *Functional Analysis*, 2nd ed, McGraw-Hill, New York 1991
or
J.L. Kelley and I. Namioka, *Linear Topological Spaces*, D. Van Nostrand Company, Princeton, NJ 1963

3.1 General definition of a topological space

Let X be an arbitrary set.

Definition 3.1 *A set $\mathcal{T}$ of subsets of X is a topology on X if*
i) $\emptyset \in \mathcal{T}$ and $X \in \mathcal{T}$,
ii) for every family $\{G_\alpha\} \subset \mathcal{T}$, $\bigcup_\alpha G_\alpha \in \mathcal{T}$, and
iii) for every finite family $\{G_1, \ldots, G_n\} \subset \mathcal{T}$, $\bigcap_{1\le i\le n} G_i \in \mathcal{T}$.

The elements of $\mathcal{T}$ are called *open sets*. A set $A \subset X$ is called *closed* iff its complement in X, $A^c := X \setminus A$, is open. The pair $(X, \mathcal{T})$ is called a *topological space.*

Examples:

1. For any X, $\mathcal{T} = \{\emptyset, X\}$ is a topology and it is obviously the smallest $\mathcal{T}$ that satisfies Definition 3.1. $\mathcal{T}$ is called the *trivial topology* or *indiscrete topology.*

2. For any X, $\mathcal{T} = \mathcal{P}(X)$, the set of *all* subsets of X, obviously also satisfies Definition 3.1. It is called the *discrete topology.*

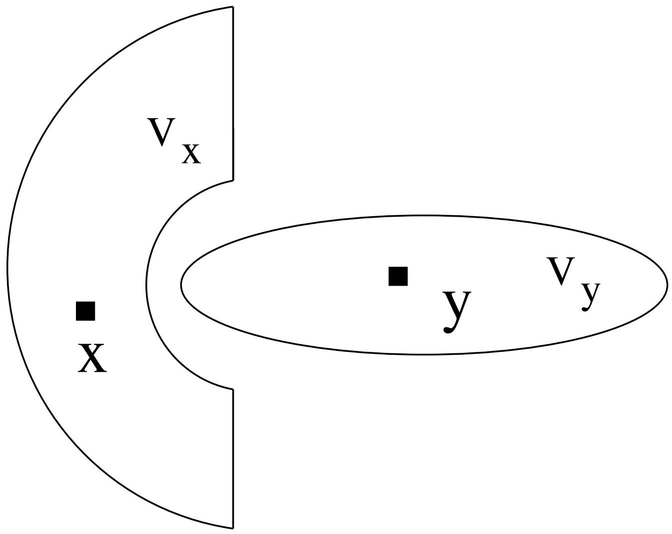

Figure 3.1: The Hausdorff property.

3. Let $(X, \mathcal{T})$ be a topological space and $X_0 \subset X$. Then $\mathcal{T}_0 = \{G_0 \subset X_0 \mid G_0 = G \cap X_0,$ for some $G \in \mathcal{T}\}$ is a topology on X_0. So, the open subsets of X_0 are all sets obtained by intersecting the open sets of X with X_0. This topology is called the *induced topology* or *relative topology* on X_0 and X_0 is said to be a topological subspace of X.

4. Let (X, d) be a metric space. Then the set of all open sets, as defined in Section 1, is a topology on X.

We will encounter many concrete examples of topologies later on.

Some more terminology:

- A set $V \subset X$ is called a *neighborhood* of a point $x \in X$ if there exixts an open set contained in V and containing x, i.e., if there exists $G \in \mathcal{T}$ such that $x \in G \subset V$.

- A topology $\mathcal{T}$ is called *Hausdorff* (or $(X, \mathcal{T})$ is called a Hausdorff space) if every pair of distinct points $x, y \in X$, has a pair of non-intersecting neighborhoods, i.e., there exists a neighborhood V_x of x and a neighborhood V_y of y such that $V_x \cap V_y = \emptyset$. Almost all topological spaces used in analysis are Hausdorff. E.g., all metric topologies are Hausdorff (why ?). On the other hand, if X has at least two elements, the trivial topology on X is not Hausdorff.

Next, we express the basic notions of convergence, continuity, and compactness in terms of open sets. Let X and Y be a topological spaces.

Definition 3.2 (convergence) *A sequence (x_n) in X converges to a limit $x \in X$ iff for every neighborhood V of x, there exists N such that for all $n \geq N$, $x_n \in V$.*

This definition says that eventually the sequence lies entirely in any neighborhood of x.

Exercise 3.1 *What are the convergent sequences in a topological space X with the trivial topology? And with the discrete topology?*

Definition 3.3 (continuity) *A function $f : X \to Y$ is continuous at $x \in X$ if for each neighborhood W of $f(x)$ there exists a neighborhood V of x such that $f(V) \subset W$. f is continuous on X if it is continuous at every $x \in X$.*

Theorem 3.4 *Let $(X, \mathcal{T})$ and $(Y, \mathcal{S})$ be two topological spaces and $f : X \to Y$. Then f is continuous iff for every $G \subset \mathcal{S}$, $f^{-1}(G) \in \mathcal{T}$.*

Exercise 3.2 *Prove Theorem 3.4*

If a function $f : X \to Y$ is one-to-one and onto (a bijection), and both f and f^{-1} are continuous, f is called a *homeomorphism*, and any two topological spaces X and Y such that there exists a homeomorphism $f : X \to Y$, are called *homeomorphic*. Homeomorphic spaces are indistinguishable as topological spaces. E.g., if X and Y are homeomorphic, one has that for any sequence (x_n) in X, (x_n) converges to x iff the sequence $(f(x_n))$ in Y converges to $f(x)$ (prove this).

Remark:
A bijection $f : X \to Y$ always has an inverse f^{-1}, but f^{-1} need not be continuous even if f is. E.g., consider $f : [0, 2\pi) \to S^1, f(\theta) = e^{i\theta}$, where $[0, 2\pi) \subset \mathbb{R}$ is considered with the topology induced by the standard topology on $\mathbb{R}$, and $S^1 \subset \mathbb{C}$ is the unit circle with the topology induced by the topology of the complex plane (or, equivalently, $\mathbb{R}^2$). f is continuous, but f^{-1} is not. For other examples see Section 3.3.

Definition 3.5 (compactness) *$K \subset X$ is compact if every open cover of K contains a finite subcover.*

Note that $K \subset X$ is compact in the topology $\mathcal{T}$ on X iff K is compact as a subset of itself for the induced topology of X on K.

Exercise 3.3 *Let (X, d_1) and (Y, d_2) be metric spaces. Prove that the definitions of convergence and continuity are equivalent to the ones previously given (Definition 1.2 and 2.1).*

The following proposition states that, in the case of metric spaces, compactness as defined here, is equivalent to (sequential) compactness as defined in Definition 1.9. Hence, we can keep using the term compactness (instead of sequential compactness) when we are considering metric spaces.

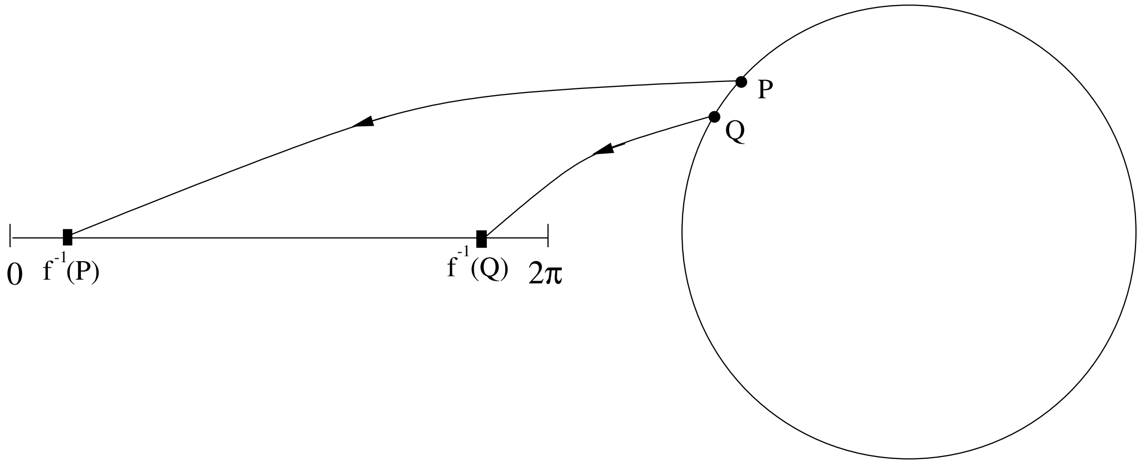

Figure 3.2: The interval and the circle are not homeomorphic. There are arbitrarily close points on the circle, P and Q in the figure, which have inverse images near the left and right end points of the interval respectively. Hence, the inverse images are not close and the inverse map cannot be continuous.

Proposition 3.6 *Let (K, d) be a metric space. Then K is compact in the sense of Definition 3.5 iff every sequence (x_n) in K has a convergent subsequence.*

Proof: We already proved in Theorem 2.16 that sequential compactness implies the finite-open-subcover property, i.e., compactness. It remains to prove the other direction, which we do by contraposition.

Suppose there is a sequence (x_n) in K without any convergent subsequence. Then the sequence necessarily contains an infinite number of distinct x_n. Therefore, without loss of generality, we may as well assume that all elements of (x_n) are distinct. We will show that the existence of such a sequence implies that K does not have the finite-open-subcover property, i.e., that K is not compact.

We start by constructing an open cover of K as follows. For each $x \in K$, either $x \in \{x_n \mid n \geq 1\}$ or $x \notin \{x_n \mid n \geq 1\}$. In both cases, as (x_n) has no convergent subsequences, only a finite number of the x_n belong to $B_{\epsilon_x}(x)$. Therefore, there is $\epsilon_x > 0$ such that $x_n \notin B_{\epsilon_x}(x)$ for all $x_n \neq x$. Clearly $\{B_{\epsilon_x}(x) \mid x \in K\}$ is an open cover of K. This cover has the property that every finite subcover $\{B_{\epsilon_x}(x) \mid x \in \{y_1, \ldots, y_k\} \subset K\}$ contains at most k elements of the sequence (x_n). As we assumed that all the elements of (x_n) are distinct, this implies that any finite subcover does not contain all elements of (x_n). Therefore, this open cover of K does not contain a finite subcover, which is in contradiction with the assumption that K is compact. ∎

We repeat that in general, for non-metric topological spaces, sequential compactness does not imply compactness. In Proposition 1.8 we gave a characterization of closed sets, and the closure of a set, in terms of convergent sequences. Again, this breaks down for general topological spaces. Similarly, continuous functions $f : X \to Y$, in the case X and Y are metric spaces, were characterized in Proposition 2.3 as exactly those functions that preserve convergent sequences. Again, not so in arbitrary topological spaces. There are however objects that replace sequences, called nets, that can be used to express all the above-mentioned properties in an alternative way. We will not make use of nets in this course.

3.2 Bases of open sets

The set of all open sets can be huge and unwieldy. Recall that for metric spaces everything could be expressed in terms of open balls alone, a rather small subset of the open sets. In this section we introduce subsets of $\mathcal{T}$ (called *bases*) that can play the same role for an arbitrary topology as the open balls for a metric topology. We will also see that such special sets of open sets arise naturally when one defines new topologies.

Definition 3.7 *A family of open sets $\mathcal{B} \subset \mathcal{T}$ is a base for the topology $\mathcal{T}$, if for every $G \in \mathcal{T}$ there is a family of $B_\alpha \in \mathcal{B}$, such that $G = \bigcup_\alpha B_\alpha$.*

Examples:
1. The family of all open intervals $(a, b) \subset \mathbb{R}$, $a, b \in \mathbb{R}$, is a base for the standard topology on $\mathbb{R}$.
2. The family of all open intervals $(a, b) \subset \mathbb{R}$, with $a, b \in \mathbb{Q}$, is a base for the standard topology on $\mathbb{R}$.
3. Let (x, d) be a metric space. The set of all open balls $B_{1/n}(x)$, with $n \geq 1$ and $x \in A$, where A is dense in X, is a base for the metric topology on X.

Definition 3.8 *A family $\mathcal{N}$ of neighborhoods of a point $x \in X$ is called a neighborhood base (or a base of neighborhoods) of x if for any neighborhood V of x there exists a $W \in \mathcal{N}$ such that $W \subset V$.*

Example:
Let (X, d) be a metric space. The open (or closed) balls with center x and radius $1/n$, $n \geq 1$, form a neighborhood base for x.

Definition 3.9 (first countable) *A topological space X is first countable if every $x \in X$ has a countable neighborhood base.*

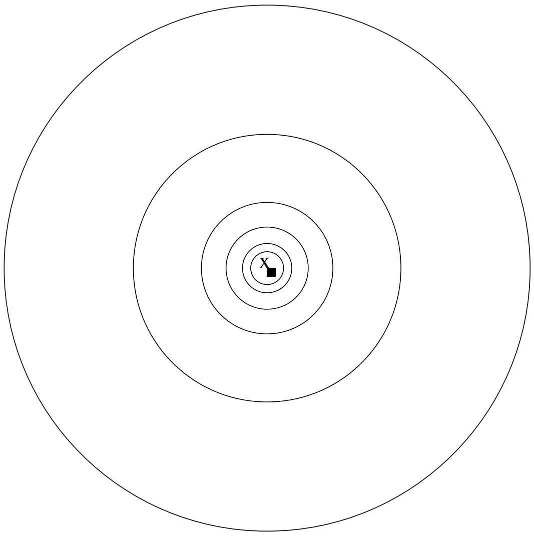

Figure 3.3: A base of neighborhoods of x.

Examples:
1. As the examples above show, any metric space is first countable.
2. Any X with the discrete topology is first countable.
3. Any X with the trivial topology is first countable.

Often, it is useful to *define* a topology in terms of a base or a neighborhood base for every point.

Examples:
1. Let X be the space of all real-valued functions on $[a, b] \subset \mathbb{R}$. Let $f \in X$, $\epsilon > 0$, $n \geq 1$, and $x := \{x_1, \ldots, x_n\} \subset [a, b]$. Define a subset $B_{f,x,\epsilon} \subset X$ by

$$B_{f,x,\epsilon} = \{g \in X \mid |g(x_i) - f(x_i)| < \epsilon, \text{ for all } i = 1, \ldots, n\}$$

Then, define $\mathcal{T}$ as the smallest set of subsets of X satisfying the conditions of Definition 3.1 and containing $B_{f,x,\epsilon}$, for all $f \in X$, $\epsilon > 0$, and finite sets $x \subset X$. This defines the topology of pointwise convergence of functions in the sense that $f_n \to f$ in the topological space $(X, \mathcal{T})$ iff $f_n(x) \to f(x)$ for all $x \in [a, b]$. The sets $B_{f,x,\epsilon}$ are a neighborhood base for $f \in X$. A base for the topology can be obtained by considering arbitrary finite intersections of sets of the form $B_{f,x,\epsilon}$. In fact, it is easy to see that it is sufficient to consider sets of the form

$$\{g \in X \mid |g(x_i) - f(x_i)| < \epsilon_i, \text{ for all } i = 1, \ldots, n\}$$

for all $f \in X$, $\epsilon_i > 0$, and finite sets $x \subset X$.
2. Let V be a linear space. A function $p : V \to \mathbb{R}$ is called a *seminorm* on V if it

has the following properties:
i) $p(x) \geq 0$ for all $x \in V$.
ii) $p(ax) = |a|p(x)$, for all $x \in V$, $a \in \mathbb{R}$.
iii) $p(x+y) \leq p(x) + p(y)$, for all $x, y \in V$.
If, in addition, p satisfies $p(x) > 0$ for all $x \neq 0$, p is called a *norm*. We already know how to make a normed linear space into a metric space (and hence a topological space). More generally, suppose P is a set of seminorms on a linear space V. One can then define a neighborhood base for any $x \in V$ parametrized by $\epsilon > 0$ and finite subsets $\{p_1, \ldots, p_n\} \subset P$:

$$B_{\epsilon,p_1,\ldots,p_n}(x) = \{y \in V \mid p_1(x-y) < \epsilon, \ldots, p_n(x-y) < \epsilon\}$$

We will define various interesting topologies on linear spaces exactly in this way, i.e., by defining appropriate sets of seminorms. Note that the topology of pointwise convergence on the space of all real-valued functions on some common domain is of this type. If the set of seminorms used is uncountable, chances are that the topology is not first countable and non-metric.

Theorem 3.10 *A family of open sets $\mathcal{B} \subset \mathcal{T}$ is a base for the topology $\mathcal{T}$ iff for every $x \in X$, $\mathcal{B}$ contains a base of neighborhoods for x.*

Proof: Suppose $\mathcal{B}$ is a base for $\mathcal{T}$. Then, if V is a neighborhood of x then there exists $G \in \mathcal{T}$, such that $x \in G \subset V$. As $\mathcal{B}$ is a base, there are B_α such that $\bigcup_\alpha B_\alpha = G$. Therefore, there exists $B \in \mathcal{B}$, such that $x \in B$ and $B \subset V$. We conclude that $\mathcal{B}$ contains a base of neighborhoods of x.

Conversely, if a family of open sets $\mathcal{B}$ contains a base of neighborhoods for every $x \in X$, then, for every open set $G \in \mathcal{T}$, and any $x \in G$, there exists $B_x \in \mathcal{B}$ such that $x \in B_x \subset G$. Therefore, $\bigcup_x B_x = G$, and we have proved that $\mathcal{B}$ is a base for the topology. ■

3.3 Comparing topologies

It is clear that different metrics on a space X can lead to one and the same notion of convergence of sequences and continuity of functions. E.g., if d_1 an d_2 are two metrics on X such that for all $x, y \in X$, $d_1(x,y) = 2d_2(x,y)$, this will certainly be the case. And in general, if two metrics lead to exactly the same set of open sets all topological properties should be the same. Ee will see that indeed the notion of topology (as the set of open sets) is more fundamental than, e.g., a norm or a metric.

As noted in the previous section, a topology if often given indirectly by a base of neighborhoods for every point. A natural question is then: how can one compare topologies by investigating a base of neighborhoods for them? When are the topologies the same?

Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies on the same space X. Then, $\mathcal{T}_2$ is called *finer* than $\mathcal{T}_1$ if $\mathcal{T}_1 \subset \mathcal{T}_2$, i.e., if it has more open sets (one then also says that $\mathcal{T}_1$ is *coarser* than $\mathcal{T}_2$). Of course, not every pair of topologies are comparable, i.e., it is quite possible that $\mathcal{T}_1 \not\subset \mathcal{T}_2$ and $\mathcal{T}_2 \not\subset \mathcal{T}_1$.

Proposition 3.11 *Let X and Y be two spaces, each with two topologies, $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{S}_1, \mathcal{S}_2$ respectively. If $f : (X, \mathcal{T}_1) \to (Y, \mathcal{S}_1)$ is continuous and*
i) if $\mathcal{T}_2$ is finer than $\mathcal{T}_1$, then $f : (X, \mathcal{T}_2) \to (Y, \mathcal{S}_1)$ is continuous
ii) if $\mathcal{S}_2$ is coarser than $\mathcal{S}_1$, then $f : (X, \mathcal{T}_1) \to (Y, \mathcal{S}_2)$ is continuous

Proof: These statements are a direct consequence of the general definition of continuity (Definition 3.3), or, more directly, of Theorem 3.4. ∎

The most regular bijection from a space X to itself one can imagine is the identity map $\mathrm{id} : X \to X, x \mapsto \mathrm{id}(x) = x$. As a map $\mathrm{id}^{-1} = \mathrm{id}$, and for every topology $\mathcal{T}$ on X, $\mathrm{id}^{-1} = \mathrm{id} : (X, \mathcal{T}) \to (X, \mathcal{T})$ are continuous, and hence *id* is a homeomorphism. The situation changes drastically when we consider two different topologies on X, say $\mathcal{T}_1$ and $\mathcal{T}_2$. E.g., consider

$$\mathrm{id} : (X, \{\emptyset, X\}) \to (X, \mathcal{P}(X))$$

Then, if X contains at least two points, id is not continuous at any point $x \in X$. More generally, as the following theorem shows, the continuity of the map id is a perfect tool for comparing topologies.

Theorem 3.12 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies on X. Then the map* $\mathrm{id} : (X, \mathcal{T}_1) \to (X, \mathcal{T}_2)$ *is continuous iff $\mathcal{T}_1$ is finer than $\mathcal{T}_2$.*

Proof: This is a direct consequence of Theorem 3.4. ∎

Corollary 3.13 $\mathrm{id} : (X, \mathcal{T}_1) \to (X, \mathcal{T}_2)$ *is a homeomorphism iff $\mathcal{T}_1 = \mathcal{T}_2$.*

Theorem 3.14 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies on X. Then $\mathcal{T}_1 = \mathcal{T}_2$ iff the following is true: for all topological spaces $(Y, \mathcal{S})$, a function $f : (X, \mathcal{T}_1) \to (Y, \mathcal{S})$ is continuous iff $f : (X, \mathcal{T}_2) \to (Y, \mathcal{S})$ is continuous.*

Proof: If $\mathcal{T}_1 = \mathcal{T}_2$, then the statement about continuous functions $f : X \to Y$ is trivial. For the other direction we use Corollary 3.13. So, we have to show that $\mathrm{id} : (X, \mathcal{T}_1) \to (X, \mathcal{T}_2)$ is a homeomorphism. This follows by once taking $(Y, \mathcal{S}) = (X, \mathcal{T}_1)$, and once $(Y, \mathcal{S}) = (X, \mathcal{T}_2)$. ∎

The next result is of practical importance, because often topologies are defined indirectly through bases of neighborhoods for every point. So, it is particularly useful to be able to distinguish (or identify) topologies on the basis of bases of neighborhoods alone.

Theorem 3.15 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies on X. And suppose that for every $x \in X$ there exist bases of neighborhoods of x, $\mathcal{N}_1$ and $\mathcal{N}_2$, for $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively, such that for every $V \in \mathcal{N}_1$ there is a $W \in \mathcal{N}_2$ such that $W \subset V$. Then $\mathcal{T}_2$ is finer than $\mathcal{T}_1$.*

Proof: We simply use Theorem 3.12. It is obvious that under the condition stated $\mathrm{id} : (X, \mathcal{T}_2) \to (X, \mathcal{T}_1)$ is continuous and the result follows. ■

Corollary 3.16 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two topologies on X. Then, $\mathcal{T}_1 = \mathcal{T}_2$ iff for every $x \in X$ there are bases of neighborhoods of x, $\mathcal{N}_1$ and $\mathcal{N}_2$, for $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively, such that for every $V \in \mathcal{N}_1$ there is a $W \in \mathcal{N}_2$, such that $W \subset V$, and, vice versa, there are bases of neighborhoods $\mathcal{M}_1$ and $\mathcal{M}_2$, for $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively, such that for every $W \in \mathcal{M}_2$ there is a $V \in \mathcal{M}_2$, such that $V \subset W$.*

For metric spaces one can charatcerize the topology also by the set of all convergent sequences. Note that in general, the same topology can be obtained from many different metrics. Metrics leading to the same topology, i.e., the same set of open sets, are called *equivalent* .

Theorem 3.17 *Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be two metric topologies on X, defined by metrics d_1 and d_2 respectively. Then, $\mathcal{T}_1 = \mathcal{T}_2$, i.e., d_1 and d_2 are equivalent metrics, iff $x_n \to x_0$ in $(X, d_1) \Leftrightarrow x_n \to x_0$ in (X, d_2).*

Proof: The proof is a direct application of Corollary 3.13. ■

For general topological spaces there is the analogous theorem with sequences replaced by nets.

3.4 Comments on non-metric topologies

1. The notion of Cauchy sequence is defined in terms of a metric. Therefore, also completeness is defined only for metric spaces.
2. Recall that the closure $\overline{A}$ of a set $A \in X$ is defined as the smallest closed set containing A. For metric spaces $\overline{A}$ can be obtained from A through addition of all the limits of converging sequences with elements in A (See Section 1.4). For a non-metric topological space this procedure might fail. In fact, the set thus obtained, sometimes called the *sequential closure* of A and denoted by $\overline{A}^S$, need not be closed. In general one has $A \subset \overline{A}^S \subset \overline{A}$. An example of $\overline{A}^S \neq \overline{A}$ is the following. Define

$$f_{m,n}(x) = [\cos(m!\pi x)]^{2n}$$

on $[0,1]$, for all $m, n \geq 1$. The following pointwise limits are easy to compute:

$$\lim_{n\to\infty} f_{m,n}(x) := f_m(x) = \begin{cases} 1 & \text{if} \quad x = k/m!, k = 0, \ldots, m! \\ 0 & \text{else} \end{cases}$$

and

$$\lim_{m\to\infty} f_n(x) := f(x) = \begin{cases} 1 & \text{if} \quad x \in \mathbb{Q} \cap [0,1] \\ 0 & \text{else} \end{cases} \tag{3.1}$$

For $A = \{f_{m,n} \mid m, n \geq 1\}$, we then have $f_n \in \overline{A}^S$, and certainly also

$$f \in \overline{\overline{A}^S}^S$$

The claim is that $f \notin \overline{A}^S$, and hence $\overline{A}^S \neq \overline{\overline{A}^S}^S$ and therefore, $\overline{A}^S \neq \overline{A}$. To show that $f \notin \overline{A}^S$ one can resort to the following proposition:

Proposition 3.18 *Any function $f : [0,1] \to \mathbb{R}$, that is the pointwise limit of continuous functions, is continuous on a dense subset of* $[0,1]$.

As f defined in (3.1) is nowhere continous in $[0,1]$, it is not the pointwise limit of some sequence of functions $f_{m,n}$, which are all continuous. This example also proves that the topology of pointwise convergence on the real-valued functions on $[0,1]$ is not metric.

4 Completeness and the Contraction Mapping Theorem

Recall that a metric space (X, d) is called *complete* iff every Cauchy sequence in X converges. Our next objective is the *Contraction Mapping Theorem* that we will first state and discuss and prove later in this section.

References for this section are

G.F. Simmons, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963,

and

J.E. Marsden and M. J. Hoffman, *Elementary Classical Analysis*, 2nd ed., W.H. Freeman and Company, New York, 1993.

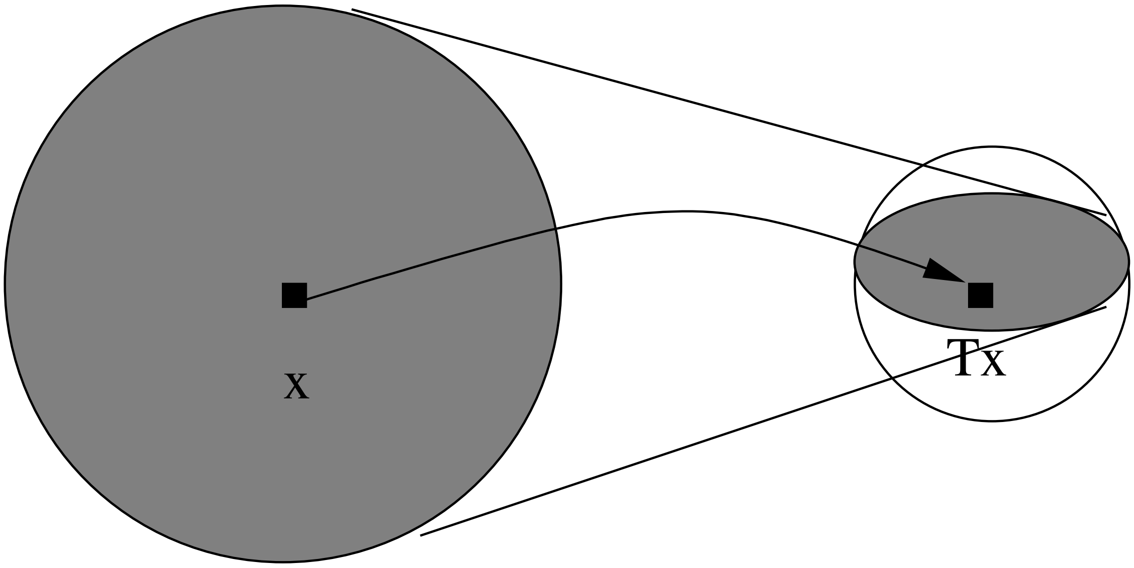

Figure 4.1: T is a contraction.

Definition 4.1 (contraction) *Let (X, d) be a metric space. A map $T : X \to X$ is a contraction mapping (or contraction), if there exists a constant c, $0 \leq c < 1$, such that*

$$d(T(x), T(y)) \leq cd(x, y) \tag{4.1}$$

for all $x, y \in X$.

Exercise 4.1 *Prove that if T is a contraction on a metric space X, then T is uniformly continuous.*

Keep in mind that in some texts a map satisfying (4.1) with $c = 1$ is also called a contraction. Sometimes a map satisfying (4.1) for some $c < 1$ is then called *strictly contracting* or a a *strict contraction.*

Theorem 4.2 (Contraction Mapping Theorem) *If $T : X \to X$ is a contraction mapping and (X, d) is complete, then the equation*

$$T(x) = x \tag{4.2}$$

has exactly one solution $x \in X$.

A solution of (4.2) is called a *fixed point* of T. So, strict contractions have exactly one fixed point. The Contraction Mapping Theorem is only one example of what are more generally called fixed-point theorems. There are fixed-point theorems for maps satisfying (4.1) with $c = 1$, or even for arbitrary continuous transformations on certain metric spaces. In general, however, $c < 1$ is needed for uniqueness or even existence of a fixed point. E.g., consider the metric space $X = \{0, 1\}$, with the metric determined by $d(0, 1) = 1$. The map T defined by $T(0) = 1, T(1) = 0$, satisfies (4.1) with $c = 1$. But T does not have any fixed points. On the other extreme, the identity map on any metric space trivially satisfies (4.1) with $c = 1$, and every point is a fixed point.

It is worth noting that (4.2), and hence its solutions, do not refer to the metric d. If one finds a metric on X such that T is a contraction according to Definition 4.1, one can apply the theorem and obtain existence and uniqueness of the fixed point *if* X is complete for that same metric d. And there is the catch: the frustrating situation may occur that for none of the metrics for which one manages to prove that T is a contraction, X is complete. (X, d) might be complete for some metric, but for that metric T is then not a contraction. This can be an indication that the natural setting for the fixed point problem at hand is not X, but another, larger, space $(\tilde{X}, \tilde{d})$, called the *completion* of (X, d). Therefore, before we prove Theorem 4.2, we will discuss the notion of *completion* of a metric space.

4.1 Completion of metric spaces

A metric space $(\tilde{X}, \tilde{d})$ is called the *completion* of (X, d) (*the* completion, rather than *a* completion in anticipation of Theorem 4.3), if the following conditions are satisfied:

1. $(\tilde{X}, \tilde{d})$ is a complete metric space.
2. There is an embedding $i : X \to \tilde{X}$, satisfying

$$\tilde{d}(i(x), i(y)) = d(x, y) \tag{4.3}$$

for all $x, y \in X$.

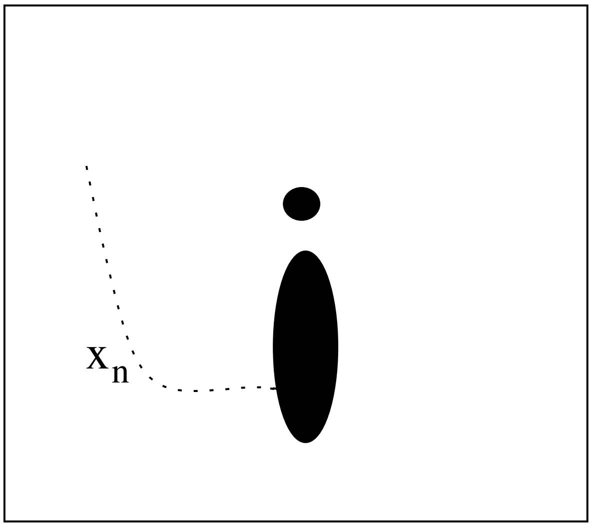

Figure 4.2: The sequence x_n seems to converge, but the candidate limit point is not an element of the space, i.e., x_n converges to a "hole" in the space.

3. $i(X)$ is dense in $\tilde{X}$, i.e., $\overline{i(X)} = \tilde{X}$.

A map i satisfying (4.3) is called *isometric*, or an *isometry*. Two metric spaces (X, d_X) and (Y, d_Y) are *isomorphic* if there exists an isometric bijection ($\equiv$ *isomorphism*) between them. They are then indistinguishable as metric spaces. Note that an isomorphism of metric spaces is also a homeomorphism of the corresponding topological spaces.

Note that isometries are automatically continuous and one-to-one. Some authors include the condition that the map be onto in the definition of isometry. With the latter convention isometries are isomorphisms of metric spaces. Do not let this confuse you. It is usually clear from the context whether an isometry is assumed to be onto or not.

We now have the terminology to state the theorem about the completion of metric spaces.

Theorem 4.3 *Every metric space (X, d) has a completion $(\tilde{X}, \tilde{d})$. The completion is unique up to isomorphism, i.e., any two completions $(\tilde{X}, \tilde{d})$ and $(\hat{X}, \hat{d})$ of (X, d) are isomorphic as metric spaces.*

Proof: We first construct a completion $(\tilde{X}, \tilde{d})$ of (X, d). Consider the set C of all Cauchy sequences in X, and define an equivalence relation on C by

$$x \sim y \quad \text{iff } \lim_{n\to\infty} d(x_n, y_n) = 0$$

for any two Cauchy sequences $x = (x_n), y = (y_n) \in C$. Note that if two Cauchy sequences x and y are convergent, then their limits are the same iff $x \sim y$. Verify that $\sim$ is indeed an equivalence relation. Let $\tilde{X}$ be the set of equivalence classes in C, i.e., $\tilde{X} = C/\sim$. Next, we define $\tilde{d} : \tilde{X} \times \tilde{X} \to \mathbb{R}$ by

$$\tilde{d}(\tilde{x}, \tilde{y}) = \lim_{n\to\infty} d(x_n, y_n) \tag{4.4}$$

where (x_n) and (y_n) are two representatives of $\tilde{x}$ and $\tilde{y}$ respectively. It is elementary to check that $(d(x_n, y_n))_{n\geq 1}$ is a Cauchy sequence of real numbers, and hence the limit exists. For this definition to make sense it is also crucial that the limit is independent of the chosen representatives. This can be checked as follows: suppose $(x_n), (x'_n) \in \tilde{x}$, and $(y_n), (y'_n) \in \tilde{y}$. Then, by the triangle inequality,

$$\begin{aligned} d(x_n, y_n) &\leq d(x_n, x'_n) + d(x'_n, y'_n) + d(y'_n, y_n) \\ d(x_n, y_n) &\geq d(x'_n, y'_n) - d(x_n, x'_n) - d(y'_n, y_n) \end{aligned}$$

Taking the limit $n \to \infty$ of these inequalities, and using that $(x_n) \sim (x'_n)$, and $(y_n) \sim y'_n)$, we see that the limit in (4.4) is indeed independent of the representatives. So, $\tilde{d}$ is well-defined, and it is straightforward to check that it defines a metric on $\tilde{X}$.

Next, we check that the metric space $(\tilde{X}, \tilde{d})$ is indeed a completion of (X, d).
1) To check that $(\tilde{X}, \tilde{d})$ is complete, consider an arbitrary Cauchy sequence $(\tilde{x}_n)$ in $\tilde{X}$. If necessary, we can pick a subsequence, denoted again by $(\tilde{x}_n)$, such that

$$\tilde{d}(\tilde{x}_n, \tilde{x}_m) \leq \frac{1}{n}, \qquad \text{for all } m > n. \tag{4.5}$$

For each element of this sequence, choose a representative $(x_{n,k}) \in \tilde{x}_n$ such that,

$$d(x_{n,k}, x_{n,l}) < \frac{1}{n} \qquad \text{for all } k, l \geq n \tag{4.6}$$

This is possible because any representative is a Cauchy sequences and by selecting an appropriate subsequence, which is automatically also Cauchy and in the same equivalence class, we can make sure that (4.6) holds. We claim that the sequence $(x_{k,k})$ is a Cauchy sequence and that it is a representative of the limit of $(\tilde{x}_n)$. Indeed, for all $k, l \geq n$, note that

$$\begin{aligned} d(x_{n,k}, x_{k,k}) &\leq d(x_{n,k}, x_{n,l}) + d(x_{n,l}, x_{k,l}) + d(x_{k,l}, x_{k,k}) \\ &\leq \frac{2}{n} + d(x_{n,l}, x_{k,l}) \end{aligned}$$

where we have used (4.6). By the definition of $\tilde{d}$, and (4.5), we have

$$\lim_{l\to\infty} d(x_{n,l}, x_{k,l}) = \tilde{d}(\tilde{x}_n, \tilde{x}_k) \leq \frac{1}{n}$$

By combining these estimates we obtain that $(\tilde{x}_n)$ converges to the equivalence class of $(x_{k,k})_{k=1}^{\infty}$:

$$\begin{aligned} \lim_{n\to\infty} \tilde{d}((\tilde{x}_n), \widetilde{(x_{k,k})}) &= \lim_{n\to\infty} \lim_{k\to\infty} d(x_{n,k}, x_{k,k}) \\ &\leq \limsup_{n\to\infty} \frac{3}{n} = 0 \end{aligned}$$

This shows that $(\tilde{X}, \tilde{d})$ is complete.
2) An isometric embedding $i : X \to \tilde{X}$ can be defined by mapping $x \in X$ to the equivalence class of Cauchy sequences containing the constant sequence $x_n = x$. That this map is isometric follows from

$$\tilde{d}(i(x), i(y)) = \lim_{n\to\infty} d(x_n, y_n) = \lim_{n\to\infty} d(x, y) = d(x, y)$$

3) $i(X)$ is dense in $\tilde{X}$ is seen as follows. Let $\tilde{x} \in \tilde{X}$ with representative (x_n). Because (x_n) is a Cauchy sequence the sequence of constant sequences $(y_{n,k} = x_n)_{k=1}^{\infty}$ satisfies

$$\lim_{n\to\infty} \tilde{d}((\tilde{y_n}), (\tilde{x_n})) = 0$$

To conclude the proof of the theorem we still have to show that the completion is unique up to isomorphism, which we do in the following way. Let $(\tilde{X}, \tilde{d})$ be the completion of (X, d) constructed above and let $(\hat{X}, \hat{d})$ be any other completion, with its own isometric embedding $j : X \to \hat{X}$. We will extend j to a map $\tilde{j} : \tilde{X} \to \hat{X}$, in the sense that $\tilde{j} \circ i = j$. To define $\tilde{j}$ let us pick a representative (x_n) for any $\tilde{x} \in \tilde{X}$ and define

$$\tilde{j}(\tilde{x}) = \lim_{n\to\infty} j(x_n)$$

Because j is isometric $(j(x_n))$ is a Cauchy sequence in $\hat{X}$, and therefore converges. For the same reason the limit will be independent of the representative. It is also easy to see that j is isometric:

$$\hat{d}(\tilde{j}(\tilde{x}, \tilde{j}(\tilde{y})) = \lim_n \hat{d}(j(\tilde{x}_n), j(\tilde{y}_n)) = \lim_n d(x_n, y_n) = \tilde{d}(\tilde{x}, \tilde{y})$$

By chosing the constant sequences as representatives of $\tilde{x}, x \in X$ we see that, for $x \in X$, $\tilde{j}(i(x)) = j(x)$. To check that $\tilde{j}$ is onto, note that $\tilde{X}$ contains all Cauchy sequences in X. So, we have $\hat{X} = \overline{j(X)} \subset \tilde{j}(\tilde{X}) \subset \hat{X}$, which shows that $\tilde{j}$ is onto. ∎

It is slightly annoying that $\tilde{X}$ is a space of sequences, rather than some more direct extension of X. E.g., there is no guarantee that the completion of a space of functions will still be a space of functions, just containing more functions than the original one. We will come back to this issue in concrete examples.

Examples:

1. $\mathbb{R}$ is the completion of $\mathbb{Q}$ with the usual metric. This is one of the standard ways of introducing (constructing, and mathematically defining) the real numbers. Note that when you write a real number in decimal notation, you in fact are giving a Cauchy sequence of rational numbers converging to it.

2. Consider the following metric on $C([a,b])$:

$$d(f,g) = \left[\int_a^b (f(x) - g(x))^2\, dx\right]^{1/2}$$

This is a very useful metric but $C([a,b])$, with this metric, is not complete. Its completion is denoted by $L^2([a,b])$, the space of all real-valued square-integrable functions on $[a,b]$. The dirty secret is that it is really not a space of functions, but a space of equivalence classes of (measurable) functions. Many of the usual operations with functions can still be defined, independent of the representative, but the evaluation of an element $f \in L^2([a,b])$ in a point, $f(x_0)$ *cannot* be unambiguously defined! We will meet L^2 again, and eventually come to think of it as a very nice space, consisting of objects as good as functions.

4.2 The Contraction Mapping Theorem

Recall the statement of Theorem 4.2: *If $T : X \to X$ is a contraction mapping and (X,d) is complete, then the equation $T(x) = x$ has exactly one solution $x \in X$.*
Proof: The proof is contructive in the sense that we will explicitly construct a sequence converging to the fixed point.

Let x_0 be any point in X and defined a sequence (x_n) by $x_1 = Tx_0$, $x_{n+1} = Tx_n$ (to alleviate the notations we often ommit the parentheses around the argument of a map). Denote by T^n the n^{th} power (n^{th} iterate) of T. So, $x_n = T^n x_0$.

First, we show that (x_n) is a Cauchy sequence. Let $n \geq m \geq 1$.

$$\begin{aligned}
d(x_n, x_m) &= d(T^n x_0, T^m x_0) \\
&\leq c^m d(T^{n-m} x_0, x_0) \\
&\leq c^m \left[d(T^{n-m}x_0, T^{n-m-1}x_0) + d(T^{n-m-1}x_0, T^{n-m-2}x_0) + \cdots + d(Tx_0, x_0)\right] \\
&\leq c^m \left[\sum_{k=0}^{n-m-1} c^k\right] d(x_1, x_0) \\
&\leq c^m \left[\sum_{k=0}^{\infty} c^k\right] d(x_1, x_0) \\
&\leq d(x_1, x_0) \frac{c^m}{1-c}
\end{aligned}$$

from which it follows that (x_n) is Cauchy. As X is complete x_n converges to a limit $x \in X$. That x is a fixed point of T follows using the continuity of T:

$$Tx = T \lim_n x_n = \lim_n Tx_n = \lim_n x_{n+1} = x$$

The uniqueness of the fixed point is also a consequence of the contraction property. Suppose x and y are two fixed points. Then

$$0 \leq d(x, y) = d(Tx, Ty) \leq cd(x, y)$$

As $c < 1$, this implies $d(x, y) = 0$, and hence $x = y$. So, the fixed point is unique. ■

5 Applications of the Contraction Mapping Theorem

5.1 An algorithm to compute $\sqrt{x}$

The Contraction Mapping Theorem (Theorem 4.2) not only proves the existence of a unique solution of certain equations of the form:

$$Tx = x$$

but also provides us with an algorithm to obtain arbitrarily good approximations of this solution (see the proof of Theorem 4.2). Here we apply the Contraction Mapping Theorem to a method of computing the square root of a real number.

Define a map $T : (0, \infty) \to (0, \infty)$ by

$$Ty = \frac{1}{2}(y + \frac{x}{y}) \tag{5.1}$$

where $x \in (0, \infty)$ is the number we want to compute the square root of. Clearly, $y = \sqrt{x}$ is a fixed point of T. And T has the following structure: let $0 < y_0$ be a first guess for $\sqrt{x}$, and suppose that $y_0^2 \leq x$, i.e., we guessed too small. Then,

$$\frac{x}{y_0} \geq \sqrt{x}$$

and, therefore, Ty, defined by (5.1), is the average of two approximations of $\sqrt{x}$, one that is too big and one that is too small. That should bring us closer to the correct value, and, hopefully, by iterating the process we obtain better and better approximations of $\sqrt{x}$. Before we start using this algorithm we would like to be sure that $T^n y_0 \to \sqrt{x}$. A way to to this is to set up T as a contraction on some metric space. Of course we are considering only real numbers here, and we are interested in approximations in the usual distance on the real numbers. The only remaining freedom is then the choice of a subset $I \subset \mathbb{R}$.

First, let us see whether T contracts at all: for $y_1, y_2 > 0$

$$\begin{aligned} |Ty_1 - Ty_2| &= \left| \frac{1}{2}\left(y_1 + \frac{x}{y_1}\right) - \frac{1}{2}\left(y_2 + \frac{x}{y_2}\right) \right| \\ &= \frac{1}{2}|y_1 - y_2| \left| 1 - \frac{x}{y_1 y_2} \right| \end{aligned}$$

So, T is a contraction if one necessarily has $x < 3y_1y_2$, which means we have to exclude arguments y that are too small. Therefore, we try a metric space of the form $([a,\infty), d)$, with $a > 0$. This is a complete metric space because $[a,\infty) \subset \mathbb{R}$ is closed, and $(\mathbb{R}, d)$ is complete (see Assignment 6, Problem 1).

In order to make a good choice for a we first observe that

$$Ty = \sqrt{x} + \frac{(y - \sqrt{x})^2}{2y} \geq \sqrt{x} \tag{5.2}$$

for all $y > 0$. And for $y_1, y_2 \geq \sqrt{x}$ we have $|1 - x/(y_1, y_2)| < 1$ and hence

$$|Ty_1 - Ty_2| \leq \frac{1}{2}|y_1 - y_2|$$

Therefore, the restriction of T to $[\sqrt{x}, \infty)$ is well-defined (meaning that $([\sqrt{x}, \infty)) \subset [\sqrt{x}, \infty))$, and is a nice contraction ($c = 1/2$). We conclude that for any $y_0 > 0$, $y_i = Ty_0 \in [\sqrt{x}, \infty)$ and $y_n := T^n y_0 \to \sqrt{x}$. Moreover, the convergence is exponentially fast (see the proof of Theorem 4.2):

$$\begin{aligned}
|T^n y_0 - \sqrt{x}| &\leq |T^n y_0 - T^m y_0| \\
&\leq \frac{c^n}{1-c}|Ty_0 - y_0| \\
&\leq \frac{1}{2^n}|\frac{x}{y_0} - y_0|
\end{aligned}$$

Exercise 5.1 *a) Show that for any $y > 0$, the convergence of $T^n y$ to $\sqrt{x}$ is in fact faster than exponential. Start from (5.2) to obtain a good estimate for $|Ty - \sqrt{x}|$.*
b) Take $y = 1$ and $x = 2$. Find a good estimate of how large n needs to be for $T^n 1$ and $\sqrt{2}$ to have identical first d digits. E.g., how many iterations of the map T are sufficient to compute the first 63 digits of $\sqrt{2}$?

5.2 Discrete Dynamical Systems

The term dynamical system refers to models for the evolution in time of the state of some "real life" system, as they arise in many disciplines: physics, chemistry, engineering. biology, physiology, population dynamics, economics and finance, … . In fact, they can also be an auxiliary tool for solving other problems in mathematics, then sometimes referred to as "symbolic" dynamics. A dynamical system is then defined by a state space X (often a metric space), whose elements describe the different states the system can be in, and a prescription relating the state $x_t \in X$ at time t to the state at previous (and/or subsequent times). E.g, this could be given as a differential equation, leading to a function $t \mapsto x_t$ for a suitable initial condition x_0.

The term discrete refers to the fact that time is taken to discrete, i.e., integer multiples of some time unit, i.e., we can take $t = n \in \mathbb{N}$. The dynamics is then usually defined by a map T relating the state x_{n+1} at time $t = n+1$ to the state x_n at time $t = n$:

$$x_{n+1} = Tx_n$$

If T is invertible, the dynamics can be defined for all $t = n \in Ir$.

One of the simplest and at the same time most famous dynamical systems is the *logistic map* of population dynamics. The state space is $[0, \infty) \subset \mathbb{R}$, representing the size of the population, and the dynamics is given by

$$Tx = 4\mu x(1-x)$$

where $\mu \geq 0$ is a parameter. The non-linear dependence of x_{n+1} on x_n expresses the fact that large populations lead to large off-spring, but too large populations have an adverse effect on the reproduction process.

The importance of this model is primarily that it convincingly illustrates the fact that very simple maps can have amazingly complex behavior. Analyzing the full variety of behavior of the logistic map is beyond the scope of this modest application of the Contraction Mapping Theorem. You can read more about it in : M. Schroeder, *Fractals, Chaos, Power Laws*, or R. Devaney, *An Introduction to Chaotic Dynamical Systems*.

Fixed points of the map T describe steady states (equilibrium states) of the system. If T is a contraction on a complete metric space the Contraction Mapping Theorem implies that there is a unique steady state, and that starting from any initial state, the system will evolve toward this steady state as time goes on. In the case of the logistic map we have the following.

Obviously, 0 is a fixed point of T. It is easy to show that when $\mu < 1/4$, T is a contraction on $[0, 1]$. The Contraction Mapping Theorem then tells us that for such values of μ, 0 is the unique fixed point, and that for an arbitrary initial population $x_0 \in [0, 1]$, $T^n x_0 \to 0$, i.e., the population dies out. For $\mu > 1/4$, but not too large, there is a second, non-zero, fixed point. As μ increases further, even more interesting things happen.

5.3 Integral Equations

Consider the (linear) Volterra equation for an unknown function $f : [a, b] \to \mathbb{R}$:

$$f(x) = g(x) + \int_a^b k(x,y) f(y)\, dy \tag{5.3}$$

where $g \in C([a, b])$ and $k : [0, 1] \times [0, 1] \to \mathbb{R}$, are given functions. The integral equation (5.3) is a fixed point equation, i.e., of the form $Tf = f$, for the map T defined by

$$Tf(x) = g(x) + \int_a^b k(x,y) f(y)\, dy \tag{5.4}$$

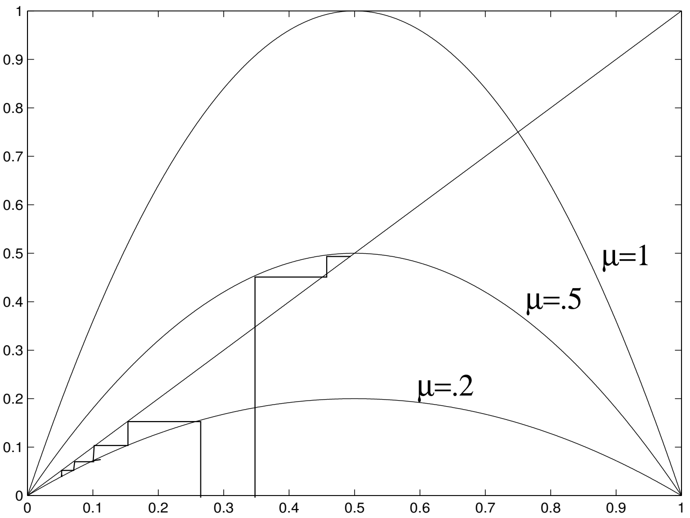

Figure 5.1: The logistic map for three different values of the parameter μ. For $\mu = .2$, 0 is the only fixed point and it is stable. For $\mu = .5$, 0 is an unstable fixed point and there is a non-zero, stable, fixed point. For $\mu = 1$, the two fixed points are unstable and more complex (and more interesting) asymptotic behavior of the dynamical system occurs.

Theorem 5.1 *If there exists a constant c satisfying*

$$\sup_{a \le x \le b} \int_a^b |k(x,y)|\, dy \le c < 1 \tag{5.5}$$

then there is a unique continuous function f satisfying (5.3).

Proof: We prove this result by showing that, when (5.5) is satisfied, the map T is a contraction on the complete metric space $(C([a,b]), |\cdot\|_\infty)$ (see Theorem 2.10). For any $f_1, f_2 \in C([a,b])$ we have

$$\begin{aligned}
\|Tf_1 - Tf_2\|_\infty &= \sup_{a \le x \le b} \left| \int_a^b k(x,y)(f_1(y) - f_2(y))\, dy \right| \\
&\le \sup_{a \le x \le b} \int_a^b |k(x,y)|\, |f_1(y) - f_2(y)|\, dy \\
&\le \|f_1 - f_2\|_\infty \sup_{a \le x \le b} \int_a^b |k(x,y)|\, dy \\
&\le c\|f_1 - f_2\|_\infty
\end{aligned}$$

■

From the proof of the Contraction Mapping Theorem we also know that, when $c < 1$, we can obtain the unique fixed point f as a limit:

$$f = \lim_{n\to\infty} T^n f_0 \tag{5.6}$$

for any $f_0 \in C([a, b])$. It is interesting to reinterprete this limit as a series. Define a map $K : C([a, b]) \to C([a, b])$ by

$$Kf = \int_a^b k(x, y) f(y)\, dy$$

One has that $\|K^n f_0\|_\infty \leq c^n |f_0\|_\infty \to 0$. Then, (5.3) becomes $f = g + Kf$, or

$$(\mathbb{1} - K)f = g \tag{5.7}$$

where $\mathbb{1}$ is the identity map, i.e., $\mathbb{1}f = f$. Solving (5.7) for f, given an arbitrary g, is equivalent to inverting the map $\mathbb{1} - K$, and (5.6) can be reinterpreted as the solution of this problem:

$$f = (\mathbb{1} - K)^{-1} g = \sum_{n=0}^{\infty} K^n g \tag{5.8}$$

This series is called the *Neumann series*, and using its partial sums to approximate f is often called the *Born approximation*. Note the similarity of (5.8) with the geometric series $(1 - x)^{-1} = \sum_{n=0}^{\infty} x^n$, for $|x| < 1$. In fact, we will see later that it *is* a geometric series that is absolutey convergent for a natural norm on the linear operators on a Banach space. This also explains why we only needed a condition on k, and none on g: (5.5) is the condition needed for $\mathbb{1} - K$ to be invertible, which does not involve g.

5.4 Differential Equations (Boundary Value Problems)

Consider a copper rod or pipe, say of length 1 in some units, and partly insulated. The ends of the rod are kept at constant temperatures T_0 and T_1. Denote by x the coordinate along the length of the pipe ($x \in [0, 1]$), and let t be time. We are interested in the temperature $u(x, t)$ as it varies along the rod, and with time. To a reasonable degree of accuracy $u(x, t)$ satisfies

$$u_t = u_{xx} - q(x)u \tag{5.9}$$

where $q(x)$ is a given function describing the cooling rate at x, and

$$u_t(x, t) = \frac{\partial}{\partial t} u(x, t) \tag{5.10}$$

$$u_{xx}(x, t) = \frac{\partial^2}{\partial x^2} u(x, t) \tag{5.11}$$

is a standard notation for partial derivatives. Of course, this equation does not uniquely determine u by itself. It has to be supplemented with an inital condition $u_0(x)$, and boundary conditions, e.g., $u(0,t) = T_0, u(1,t) = T_1$, for all t. If one is interested in time independent solutions $u(x,t) = u(x)$, e.g., describing the situation as $t \to \infty$, one has to solve the following ODE (put $u_t = 0$ in the PDE (5.9)) + boundary conditions:

$$-u''(x) + q(x)u(x) = 0 \tag{5.12}$$

$$u(0) = T_0, \quad u(1) = T_1 \tag{5.13}$$

If $q(x)$ is a constant function, (5.12) is easy to solve explicitly. But if q is not constant explicit integration of this equation is in general impossible. We will show that, when $\|q\|_\infty$ is not too large, the Contraction Mapping Theorem implies the existence of a unique solution, and provides a method for calculating approximations of that solution.

First, it is convenient to get rid of the inhomogeneous boundary conditions (5.13), by deriving an equation for a function v related to u by

$$u(x) = T_0 + (T_1 - T_0)x + v(x)$$

v then satisfies

$$v''(x) = q(x)v(x) + (T_0 + (T_1 - T_0)x)q(x) \tag{5.14}$$

$$v(0) = 0, \quad v(1) = 0 \tag{5.15}$$

Next, we will turn this problem into a fixed point problem of the Volterra type with the aid of the following proposition.

Proposition 5.2 *Let $f : [0,1] \to \mathbb{R}$ be a continuous function, then the unique solution v of*

$$-v'' = f \tag{5.16}$$

$$v(0) = v(1) = 0 \tag{5.17}$$

can be given as

$$v(x) = \int_0^1 g(x,y)f(y)\,dy \tag{5.18}$$

where

$$g(x,y) = \begin{cases} x(1-y) & \text{if } 0 \le x \le y \le 1 \\ y(1-x) & \text{if } 0 \le y \le x \le 1 \end{cases} \tag{5.19}$$

We postpone the proof of this proposition to the end of this section and first see where it takes us. Using this proposition, (5.14) can be written in the form

$$v(x) = -\int_0^1 g(x,y)q(y)v(y)\,dy - \int_0^1 g(x,y)(T_0 + (T_1 - T_0)y)q(y)\,dy$$

which is of the form

$$(\mathbb{1} - K)v = h$$

with

$$Kv(x) = -\int_0^1 \underbrace{g(x,y)q(y)}_{:=k(x,y)} v(y)\,dy \tag{5.20}$$

$$h(x) = -\int_0^1 g(x,y)(T_0 + (T_1 - T_0)y)q(y)\,dy \tag{5.21}$$

We can now apply Theorem 5.1 to obtain the following result.

Theorem 5.3 *If*

$$\sup_{0\le x\le 1}\int_0^1 |g(x,y)q(y)|\,dy < 1 \tag{5.22}$$

the boundary value problem (5.12)-(5.13) has a unique continuous solution.

Note that it is not necessary that q be continuous, as long as h is and the condition (5.22) is satisfied.

We still have to prove Proposition 5.2.

Proof: (Proposition 5.2)
By integrating (5.16) twice we obtain:

$$v(x) = -\int_0^x \underbrace{\int_1^y f(s)\,ds}_{:=a(y)}\,dy + C_1 x + C_2$$

where C_1 and C_2 are two real constants to be determined later. Integration by parts of the integral over y in the RHS gives

$$\begin{aligned} v(x) &= -[ya(y)]_0^x + \int_0^x yf(y)\,dy + C_1 x + C_2 \\ &= x\int_x^1 f(y)\,dy + \int_0^x yf(y)\,dy + C_1 x + C_2 \end{aligned}$$

We can determine C_1 and C_2 by imposing (5.17). One finds

$$\begin{aligned} C_1 &= -\int_0^1 y f(y)\, dy \\ C_2 &= 0 \end{aligned}$$

It follows that

$$v(x) = \int_0^x y(1-x)f(y)\, dy + \int_x^1 x(1-y)f(y)\, dy$$

which is what we had to prove. ■

5.5 Differential Equations (Existence and Uniqueness Theorem)

Another important application of Contraction Mapping Theorem to differential equations is an Existence and Uniqueness Theorem for initial value problems of the type

$$\begin{cases} \dot{x}\ (t) = F(t, x(t)) \\ x(t_0) = x_0 \end{cases} \tag{5.23}$$

with F a Lipschitz continuous function on a suitable domain. Here, x takes values in $\mathbb{R}^n$, and so do its derivative and F. (5.23) is often presented as a set of coupled ODE's, e.g., describing a dynamical system. (2.24) can be reformulated as an integral equation:

$$x(t) = x_0 + \int_{t_0}^t F(t, x(t))\, dt \tag{5.24}$$

By the fundamental theorem of calculus the continuous solutions of (5.23) and (5.24) coincide. (5.24) is easily written as a fixed point equation for the map T defined by

$$Tx(t) = x(t) = x_0 + \int_{t_0}^t F(s, x(s))\, ds \tag{5.25}$$

We are then looking for solutions of

$$Tx = x \tag{5.26}$$

If we can find sufficient conditions that guarantee that T is a contraction on a suitable complete metric space of continuous functions, the Contraction Mapping Theorem immediately implies that (5.26) has a unique solution in that space.

Theorem 5.4 *Let $I \subset \mathbb{R}$ be a (possibly infinite) interval, and t_0 a point in the interior. Suppose $F : I \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous in both arguments and uniformly Lipschitz continuous in its second argument, i.e., suppose there exists a constant $C > 0$ such that for all, $t \in I$, and $x, y \in \mathbb{R}^n$,*

$$\|F(t,x) - F(t,y)\| \leq C\|x - y\| \tag{5.27}$$

Then, there is a unique continuous function $x : I \to \mathbb{R}^n$ solving (5.23).

Proof: As mentioned above, our strategy is to show that T, defined in (5.25), is a contraction on a suitable metric space of continuous functions. Note that T depends on t_0 and t. It will be necessary to first consider solutions for sufficiently short times, i.e., $t \in [t_0, t_0 + \delta] \subset I$, with $\delta > 0$ to be determined later. Let $f, g : [t_0, t_0 + \delta] \to \mathbb{R}^n$, be two continuous functions. Then

$$\begin{aligned}
\|Tf - Tg\|_{\sup} &= \sup_{t_0 \leq t \leq t_0+\delta} |Tf(t) - Tg(t)| \\
&= \sup_{t_0 \leq t \leq t_0+\delta} \left| \int_{t_0}^{t} F(s, f(s)) - F(s, g(s))\, ds \right| \\
&\leq \sup_{t_0 \leq t \leq t_0+\delta} \int_{t_0}^{t} |F(s, f(s)) - F(s, g(s))|\, ds \\
&\leq \sup_{t_0 \leq t \leq t_0+\delta} \int_{t_0}^{t} C|f(s) - g(s)|\, ds \\
&\leq \delta C \|f - g\|_{\sup}
\end{aligned}$$

where we have used the definition of T and the Lipschitz condition (5.27). Now, it is clear that T is a contraction on $C([t_0, t_0 + \delta])$ with the supremum norm as long as $|t - t_0| \leq \delta < 1/C$. Therefore, there is a unique solution $x : [t_0, t_0 + \delta] \to \mathbb{R}^n$. The argument holds for *any* $t_0 \in I$. By covering I with overlapping intervals of length $< 1/C$, we see that (5.23) has a unique continuous solution defined on all of I. ■

Note that, due to (5.24), the solution is in fact C^1 (continously differentiable). In fact, F need not be continuous in t to have a unique solution, but one looses differentiability of the solution at the times where F is not continuous. The condition (5.27) can also be relaxed to hold with a time dependent Lipschitz constant $C(t)$, as long as it is bounded on compact intervals.

Examples:

1. Linear ODE with continuous coefficients have unique global solutions. It is sufficient to consider *first order* ODE's (Why?):

$$\begin{aligned}
\dot{x}\,(t) &= a(t)x(t) + b(t) \\
x(0) &= x_0
\end{aligned}$$

where $a : \mathbb{R} \to \mathbb{R}^{n\times n}$ is a continuous matrix-valued function (with respect to any norm on the $n\times n$ matrices with real entries), and $b : \mathbb{R} \to \mathbb{R}^n$ is continuous as well. Then, for any bounded interval $I \subset \mathbb{R}$, there exists a constant C such that

$$\|a(t)v\| \leq C\|v\|, \quad \text{for all } t \in I, v \in \mathbb{R}^n$$

and it is then obvious that

$$\|a(t)f(t) - a(t)g(t)\|_{\text{sup}} \leq C\|f - g\|_{\text{sup}}$$

Therefore, the hypotheses of Theorem 5.4 are satisfied and we have a unique continuous solution on I. As I was an arbitrary interval, we can conclude that there exists a unique continuous solution for all $t \in \mathbb{R}$.

2. The applications of Theorem 5.4 are by no means limited to linear ODE. E.g., consider the nonlinear scalar ODE given by

$$\begin{aligned} \dot{x}\ (t) &= \sqrt{a(t)^2 + x(t)^2} =: F(t, x(t)) \\ x(0) &= x_0 \end{aligned}$$

where $a : \mathbb{R} \to \mathbb{R}$ is a continuous function. It is not hard to prove that

$$|F(t, x) - F(t, y)| \leq |x - y|, \quad \text{for all } t, x, y, \in \mathbb{R}$$

This bound is obvious for $a^2 = 0$, and the LHS is decreasing as a function of a^2. Again we can conclude by Theorem 5.4 that there is a unique global solution of this ODE.

3. Nonlinear F can easily be continuous without being uniformly Lipschitz continuous. A simple example of this situation shows that global solutions may not exist in this case.

$$\begin{aligned} \dot{x} &= x^2 \\ x(0) &= 1 \end{aligned}$$

It is easy to check that

$$x(t) = \frac{1}{1-t}, \quad t \in [0, 1)$$

is a solution. The singularity at $t = 1$ is possible because the RHS is not uniformly Lipschitz continuous:

$$|x^2 - y^2| = |x + y|\ |x - y|$$

and $|x + y|$ can become arbitrarily large. In this simple example it is easy to understand that the quadratic RHS makes x grow faster the larger it becomes (positive, nonlinear feedback). This leads to blowup of the solution in finite time, at which point the equation is not well-defined.

6 Banach Spaces

In this section we focus on a very important class of complete metric spaces called Banach spaces. A *Banach Space* is linear space V, equipped with a norm, usually denoted by $\|\cdot\|$, that is a complete metric space for the metric derived from this norm ($d(x, y) = \|x - y\|$).

If you do not remember much about linear spaces, see e.g.,
P.D. Lax, *Linear Algebra*, John Wiley & Sons, New York 1997
or
P.R. Halmos, *Finite-dimensional Vector Spaces*, 2nd ed D. Van Nostrand Company, Princeton 1958.

For more information on Banach spaces see
W. Rudin, *Functional Analysis*, 2nd ed, McGraw-Hill Inc., New York 1991
or
G.F. Simmons, *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.

Recall that a *norm* on a linear space V over the field of scalars $\mathbb{R}$ (or $\mathbb{C}$) is a function $\|\cdot\| : V \to \mathbb{R}$ with the following properties:
(N1) $\|x\| \geq 0$, for all $x \in V$. (positivity)
(N2) $\|\lambda x\| = |\lambda|\|x\|$, for all $x \in V$ and $\lambda \in \mathbb{R}$ (or $\mathbb{C}$). (homogeneity)
(N3) $\|x + y\| \leq \|x\| + \|y\|$, for all $x, y \in V$. (triangle inequality)
(N4) $\|x\| = 0 \Rightarrow x = 0$. (strict positivity)
If the (N1–3) hold but not necessarily (N4), $\|\cdot\|$ is called a *seminorm.*

All linear spaces that we will consider will be either over $\mathbb{R}$ or $\mathbb{C}$, respectively called *real* and *complex* linear spaces.

In a normed linear space it makes sense to consider series. The series

$$\sum_{n=1}^{\infty} x_n$$

of elements $x_n \in X$ is said to *converge* to $x \in X$ if the sequence of partial sums (S_n)

$$S_N = \sum_{n=1}^{N} x_n$$

converges to x. The series is said to *converge absolutely* iff the series

$$\sum_{n=1}^{\infty} \|x_n\|$$

converges as a series in $\mathbb{R}$, i.e., $\sum_{n=1}^{\infty} \|x_n\| < +\infty$.

Proposition 6.1 *Let X be a normed linear space. Then, X is complete iff every absolutely convergent series converges.*

Exercise 6.1 *Prove Proposition 6.1.*

6.1 Bounded linear maps

Let X and Y be two normed linear spaces (not necessarily complete). As usual we will denote both norms by $\|\cdot\|$.

Definition 6.2 *A linear map $T : X \to Y$ is bounded if there exists a constant $M \geq 0$ such that*

$$\|Tx\| \leq M\|x\|, \quad \textit{for all } x \in X \tag{6.1}$$

Theorem 6.3 *A linear map $T : X \to Y$ is bounded iff it is continuous.*

Proof:
1) Bounded ⇒ Continuous. For all $x, y \in X$ we have

$$\|Tx - Ty\| = \|T(x-y)\| \leq M\|x-y\|$$

where M is a constant for which (6.1) holds. Therefore, one can take $\delta = \epsilon/M$ in the definition of continuity.
2) Continuous ⇒ Bounded. If T is continuous, then T is continuous at $0 \in X$. So, for all $\epsilon > 0$, there exists $\delta > 0$ such that $\|x\| \leq \delta \Rightarrow \|Tx - T0\| = \|Tx\| \leq \epsilon$. Pick any $\delta > 0$. Then, for any $0 \neq x \in X$, $\tilde{x}$ defined by

$$\tilde{x} = \frac{\delta}{\|x\|} x$$

satisfies $\|\tilde{x}\| \leq \delta$. Therefore, $\|T\tilde{x}\| \leq \epsilon$. It follows that for all $x \in X$

$$\|Tx\| = \|\frac{\|x\|}{\delta} T\tilde{x}\| \leq \frac{\|x\|}{\delta}\epsilon$$

Hence, T satisfies Definition 6.2 with $M = \epsilon/\delta$. ■

Note that we have also proved that if T is continuous at $0 \in X$, then T is continuous on the whole of X. This can also be seen as a direct consequence of linearity.

The set of all bounded linear maps $T : X \to Y$ is denoted by $\mathcal{B}(X, Y)$.

Next, we define the *null space* or *kernel* of a linear map $T \in \mathcal{L}(X, Y)$, denoted by $\ker T$, and the *range* of T, denoted by $\mathrm{ran} T$.

$$\begin{aligned} \ker T &= \{x \in X \mid Tx = 0\} \\ \mathrm{ran} T &= \{y \in Y \mid \text{ there exists } x \in X, \text{ such that } Tx = y\} \end{aligned}$$

Theorem 6.4 *For any $T \in \mathcal{B}(X,Y)$,* $\ker T$ *is a closed linear subspace of X.*

Proof:
• $\ker T$ *is a linear subspace:* For all $\lambda \in \mathbb{R}$ or $\mathbb{C}$, and $x, y \in \ker T$, one has

$$T(x + \lambda y) = Tx + \lambda Ty = 0$$

and therefore, $x + \lambda y \in \ker T$ (here, we only used the fact that T is linear, not that it is bounded).
• $\ker T$ *is closed:* As X is a metric space, we only have to show that for any sequence of elements $x_n \in \ker T$, converging to x, we necesarily have $x \in \ker T$. This follows from the continuity of T:

$$Tx = T \lim_n x_n = \lim_n Tx_n = \lim_n 0 = 0$$

■

When $X = Y$ one usually simplifies the notation: $\mathcal{L}(X,X) = \mathcal{L}(X)$, and $\mathcal{B}(X,X) = \mathcal{B}(X)$. In that case the linear maps T are referred to as *linear transformations* of X, or *linear operators* on X.

Another important special case is when $Y = \mathbb{R}$ or $\mathbb{C}$ (depending on whether X is a real or complex linear space). The elements of $\mathcal{L}(X,\mathbb{C})$ and $\mathcal{B}(X,\mathbb{C})$ are then called linear *functionals* (or linear forms), and bounded ($\equiv$ continuous) linear functionals on X, respectively. $\mathcal{L}(X,\mathbb{C})$ is also called the *algebraic dual*, and $\mathcal{B}(X,\mathbb{C})$ the *topological dual* space of X. Somewhat confusingly, both are commonly denoted by X^*.

Examples:

1. Consider $X = C([a,b])$. For any $\rho \in L_1([a,b])$, the following formula defines a continuous linear functional φ on X:

$$\varphi(f) = \int_a^b f(x)\rho(x)\,dx \tag{6.2}$$

2. Not all continuous linear functionals are of the form (6.2). E.g., for any $x_0 \in [a,b]$, evaluation of f at x_0 is a continuous linear functional. I.e., define

$$\delta_{x_0}(f) = f(x_0),$$

then δ_{x_0} is a continuous linear functional on $C([a,b])$.

3. An example of a discontinuous linear functional on $C([0,1])$ is the following. It is easy to see that the set of monomials $M = \{x^n \mid n \geq 0\}$ is a linearly independent set of vectors in $C([0,1])$. By a general theorem (see, e.g., the reference by Simmons given at the beginning of this section), $C([0,1])$ has a Hamel basis , H, containing M. A *Hamel basis*, is a maximal linearly independent set of vectors. Any f, can be uniquely written as finite linear combination of vectors in H, i.e., for any $f \in C([0,1])$, there is a unique set of elements, $h_i \in H$, $i = 1, \ldots, n$, and unique non-zero coeffcients c_i such that

$$f = c_1 h_1 + \cdots c_n h_n \tag{6.3}$$

For each $n \geq 0$ define $a_n(f)$ by

$$a_n(f) = \begin{cases} c_i & \text{if } h_i = x^n \\ 0 & \text{else} \end{cases}$$

Due to the uniqueness of the decompositon this is well-defined. Define a linear functional φ on $C([0,1])$ by

$$\varphi(f) = \sum_{n=1}^{\infty} n a_n(f)$$

As, for each f, only a finite number of terms in this sum are non-zero, this is well-defined. However, it is quite obvious that φ is unbounded. E.g.,

$$\sup\{|\varphi(f)| \mid \|f\| = 1\} \geq \sup\{|\varphi(f)| \mid f \in M\} = \sup\{n \geq 1\} = +\infty$$

Hence, by Theorem 6.3, this implies that φ is discontinuous. In fact, with this example we have shown that every infinite-dimensional linear space has discontinuous linear functionals defined on it. On the other hand, as will be shown in Theorem 6.14, all linear functionals on a finite-dimensional linear space are continuous.

6.2 Topologies on $\mathcal{B}(X,Y)$

Let X and Y be normed linear spaces. Then, for any $T \in \mathcal{B}(X,Y)$ the following defines a norm on $\mathcal{B}(X,Y)$:

$$\|T\| = \inf\{M \mid \|Tx\| \leq M\|x\|, \text{for all } x \in X\} \tag{6.4}$$

Check that this indeed defines a norm. An alternative expression for $\|T\|$ is

$$\|T\| = \sup\{\|Tx\| \mid \|x\| \leq 1\} \tag{6.5}$$

When $X \neq \{0\}$ one also has

$$\|T\| = \sup\{\|Tx\| \mid \|x\| = 1\} \tag{6.6}$$

$$\|T\| = \sup\{\frac{\|Tx\|}{\|x\|} \mid x \neq 0\} \tag{6.7}$$

This norm on $\mathcal{B}(X, Y)$ is called the *uniform* or *operator* norm, or sometimes just the *usual* norm. And if for a sequence of $T_n \in \mathcal{B}(X, Y)$ one has $\lim_n \|T_n - T\| = 0$, for some $T \in \mathcal{B}(X, Y)$, one says that T_n converges uniformly to T, or that it converges in the uniform (operator) topology.

It is clear that $\mathcal{L}(X)$ is not only a linear space, but that it is also an algebra with the product given by the composition of linear maps. The following theorem implies that $\mathcal{B}(X)$ is a subalgebra of $\mathcal{L}(X)$, i.e., the product of two bounded linear maps is bounded.

Theorem 6.5 *Let X, Y, and Z be normed linear spaces. For any $T \in \mathcal{B}(X, Y)$ and $S \in \mathcal{B}(Y, Z)$, one has that $ST \in \mathcal{B}(X, Z)$, and*

$$\|ST\| \leq \|S\| \, \|T\| \tag{6.8}$$

Proof: For all $x \in X$ we have

$$\|STx\| \leq \|S\| \, \|Tx\| \leq \|S\| \, \|T\| \, \|x\|$$

■

Note that (6.8) also shows that the product of bounded linear maps is a continuous map $\mathcal{B}(Y, Z) \times \mathcal{B}(X, Y) \to \mathcal{B}(X, Z)$.

Theorem 6.6 *If Y is a Banach space, then $\mathcal{B}(X, Y)$ is a Banach space for the uniform norm.*

Proof: We have to prove that every Cauchy sequence in $\mathcal{B}(X, Y)$ converges (to a limit in $\mathcal{B}(X, Y)$). Let (T_n) be a Cauchy sequence. We have that for all $x \in X$

$$\|T_n x - T_m x\| \leq \|T_n - T_m\| \, \|x\|$$

which shows that the sequence $(T_n x)$ is a Cauchy sequence in Y. As Y is complete, there exists $y \in Y$ such that $T_n x \to y$. It is then straightforward to check that $Tx = y$ defines a linear map $X \to Y$. Next, we show that T is bounded. For any $\epsilon > 0$, let N_ϵ be such that $\|T_n - T_m\| < \epsilon/2$ for all $n, m \geq N_\epsilon$. Take $n \geq N_\epsilon$. Then, for all $x \in X$, $\|x\| = 1$, there exist $m(x) \geq N_\epsilon$ such that

$$\|T_n x - Tx\| \leq \|T_n x - T_{m(x)} x\| + \|T_{m(x)} x - Tx\| \leq \epsilon/2\|x\| + \epsilon/2 \leq \epsilon \tag{6.9}$$

Therefore, for all $x \in X$, $\|x\| = 1$, $\|Tx\| \leq \|T_n x\| + \epsilon \leq \|T_n\| + \epsilon$, and T is bounded. Finally, from (6.9) it also follows that $\lim_n \|T_n - T\| = 0$, Hence, $T_n \to T$ in the uniform norm. ■

Exercise 6.2 *Prove that if $T_n \to T$ uniformly, then $\|T_n\| \to \|T\|$.*

The following theorem says that a bounded linear map has a unique bounded extension to the closure (or completion) of its domain. Theorem is very useful when one defines linear maps (one only has to define them on a dense subspace) or when one has to show that two linear maps are equal (one has to show equality only on a dense subspace).

Theorem 6.7 *Let X be a normed linear space and let Y be a Banach space. For any dense linear subspace $X_0 \subset X$ and any $T_0 \in \mathcal{B}(X_0, Y)$, there exists a unique $T \in \mathcal{B}(X, Y)$ such that $Tx_0 = T_0 x_0$, for all $x_0 \in X_0$.*

Proof: We first define a map $T \in \mathcal{B}(X, Y)$ as follows. For all $x \in X$, there is a sequence (x_n) in X_0 converging to x. Define

$$Tx = \lim_n T_0 x_n$$

This limit exists because T_0 is bounded and Y is complete. Also, any two sequences (x_n) and (x'_n) converging to x, satisfy

$$\lim_n \|x_n - x'_n\| \leq \lim_n \|x_n - x\| + \lim_n \|x - x'_n\|$$

and therefore

$$\|\lim_n T_0 x_n - \lim_n T_0 x'_n\| = \lim_n \|T_0 x_n - T_0 x'_n\| \leq \lim_n \|T_0\| \, \|x_n - x'_n\| = 0$$

and Tx is independent of the choice of the sequence (x_n) converging to x.

T is an extension of T_0, i.e., $Tx_0 = T_0 x_0$, for all $x_0 \in X_0$, because we can take the constant sequence $x_n = x_0$, for all n, to define Tx_0. The linearity of T is a consequence of the linearity of the T_n, the fact that X_0 is a linear subspace, and the independence of the chosen sequence that we just proved.

Next, we prove that T is bounded. It follows from

$$\|Tx\| = \lim_n \|T_0 x_n\| \leq \lim_n \|T_0\| \, \|x_n\| = \|T_0\| \, \|x\|$$

that $\|T\| \leq \|T_0\|$. In fact, as $Tx_0 = T_0 x_0$, for $x_0 \in X_0$, we must have $\|T\| = \|T_0\|$.

Finally, we show that T is the unique bounded linear map $X \to Y$, coinciding with T_0 on X_0. Suppose S is another such map. Then, for all $x \in X$, and sequences $x_n \to x$, $x_n \in X_0$,

$$Sx = \lim_n Sx_n = \lim_n T_0 x_n = Tx$$

For the first inequality we used that S is bounded by assumption, for the second that S is an extension of T_0, and the third is the definition of T. ∎

The following exercise asks you to think about the importance of the linearity of T_0 in the previous theorem. Is there a property that can replace it for more general functions?

Exercise 6.3 *What can one say about continuous extensions of functions $f : X_0 \to Y$, where X_0 is a dense subset of a metric space X and Y is a complete metric space?*

Next, we look at an example of a convergent sequence of bounded linear maps on $C([0,1])$, and also an example of a divergent sequence, which, however, convergence in a different topology on $\mathcal{B}(C([0,1]))$.

Example:
Let $X = C([0,1])$, equipped with the supremum norm. For $n = 1, 2, \ldots$, let $k_n(x,y)$ be a sequence of real-valued continuous functions on $[0,1] \times [0,1]$ such that

$$\lim_n \left\{ \sup_{x \in [0,1]} \int_0^1 |k_n(x,y)|\, dy \right\} = 0 \tag{6.10}$$

For $f \in X$, define $K_n \in \mathcal{B}(X)$ by

$$K_n f(x) = \int_0^1 k_n(x,y) f(y)\, dy \tag{6.11}$$

Then $K_n \to 0$ in the uniform norm. Indeed, for all $f \in X$, we have

$$\begin{aligned} |K_n f(x)| &= \left| \int_0^1 k_n(x,y) f(y)\, dy \right| \\ &\leq \int_0^1 |k_n(x,y)|\, |f(y)|\, dy \\ &\leq \|f\| \int_0^1 |k_n(x,y)|\, dy \\ &\leq \|f\| \sup_x \int_0^1 |k_n(x,y)|\, dy \end{aligned}$$

which, by (6.10) implies that

$$\|K_n\| \leq \sup_{x \in [0,1]} \int_0^1 |k_n(x,y)|\, dy \to 0$$

An example of functions k_n satisfying (6.10) is

$$k_n(x,y) = xy^n$$

Next, Consider another sequence of linear transformations $K_n \in \mathcal{B}(X)$, of the form (6.11), but with the functions k_n given by

$$k_n(x,y) = \sin(n\pi y)$$

It is not hard to prove (see the Riemann-Lebesgue Lemma below), that, for any $f \in C([0,1])$

$$\lim_n \int_0^1 \sin(n\pi y)\, f(y)\, dy = 0$$

And, therefore,

$$\lim_n \|K_n f\| = 0 \tag{6.12}$$

for all $f \in X$. On the other hand, for $f_n = \sin(n\pi x)$,

$$\|f_n\| = 1, \quad \text{but} \quad \|K_n f_n\| = 1/2$$

which implies that $\|K_n\| \geq 1/2$, and so K_n does *not* converge to 0 in the uniform norm. This motivates us for the definition of a second notion of convergence of linear maps (and a second topology on $\mathcal{B}(X,Y)$, distinct from the uniform norm topology whenever X is infinite-dimensional).

Definition 6.8 (Strong Convergence) *A sequence (T_n) in $\mathcal{B}(X,Y)$ converges strongly (or in the strong operator topology) to $T \in \mathcal{B}(X,Y)$, if for all $x \in X$*

$$\lim_n T_n x = Tx$$

Note that, by (6.12), K_n of the previous example, converges strongly to zero.

Theorem 6.9 *If $T_n \to T$ uniformly, then $T_n \to T$ strongly.*

Exercise 6.4 *Prove Theorem 6.9.*

Appendix: The Riemann-Lebesgue Lemma

In this appendix we consider complex-valued functions on a finite interval $[a,b]$. Recall the definition of the L_1 norm: for all $f \in C([a,b])$

$$\|f\|_1 = \int_a^b |f(x)|\, dx$$

The completion of $C([a,b])$ with this norm is denoted by $L_1([a,b])$, and, loosely speaking, it is the space of all functions f for which $\|f\|_1$ is finite, and its elements are called L_1 functions. It is easy to see that the product of an L_1 function and a bounded continuous function is again L_1. The functions considered can be either real- or complex-valued, in which case $L_1([a,b])$ is a real or complex Banach space respectively.

The Riemann-Lebesgue Lemma can be considered as a statement about the Fourier coefficients of L_1 functions.

Theorem 6.10 (Riemann-Lebesgue Lemma) *For any $f \in L_1([0, 2\pi])$ one has*

$$\lim_{n\to\pm\infty} \int_0^{2\pi} f(x)e^{inx}\,dx = 0$$

Proof: It is sufficient to prove this for f in a dense subspace of $L_1([0, 2\pi])$ (e.g., use Theorem 6.7).

A convenient choice for the dense subspace is the space of all trigonometric polynomials of the form

$$P_n(x) = \sum_{k=-n}^{n} a_k e^{ikx}$$

the a_k are arbitrary complex coefficients. By Theorem 2.13 the trigonometric polynomials are dense in $C(T)$ with the uniform norm (T is the unit circle). As the L^1-norm is weaker than the uniform norm, the trigonometric polynomials are also dense in $C(T)$ with the L^1 norm. As $C(T)$ is dense in $L_1([0, 2\pi])$ by construction, the trigonometric polynomials are also dense in we conclude $L_1([0, 2\pi])$. For any such P_n it is clear that

$$\int_0^{2\pi} P_n(x)e^{-ikx}\,dx = 2\pi a_k$$

And, therefore,

$$\int_0^{2\pi} P_n(x)e^{-ikx}\,dx = 0$$

for $|k| > n$, and so

$$\lim_{k\to\pm\infty} \int_0^{2\pi} P_n(x)e^{-ikx}\,dx = 0$$

■

6.3 Remarks on Finite-Dimensional Banach Spaces

Let X be a finite-dimensional linear space over $\mathbb{R}$ or $\mathbb{C}$. *Finite-dimensional* means that there is a finite set, $\{e_1, \ldots, e_n\} \subset X$, with the property that for any $x \in X$, there are constants $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ or $\mathbb{C}$, such that

$$x = \alpha_1 e_1 + \cdots + \alpha_n e_n \tag{6.13}$$

In fact, there is n, such that one can choose n *linearly independent* vectors $e_1, \ldots, e_n$ such that every $x \in X$ can be written as a linear combination of them. n is then the *dimension* of X. Linear independence of $e_1, \ldots, e_n$ is equivalent to the property that the coefficients α_i in (6.13) are uniquely determined by x.

Our main objective in these remarks on finite-dimensional Banach spaces is to show that every finite-dimensional normed linear space is a Banach space (i.e., is complete), and that, whenever X is finite-dimensional, every $T \in \mathcal{L}(X,Y)$ is bounded, i.e., $\mathcal{L}(X,Y) = \mathcal{B}(X,Y)$.

For the rest of this chapter, let X be a finite-dimensional normed linear space over $\mathbb{R}$ or $\mathbb{C}$. For simplicity of notation we will use $\mathbb{C}$ as the number field in the statements and proofs below, with the understanding that everywhere one can replace $\mathbb{C}$ by $\mathbb{R}$ and obtain the corresponding result about real linear spaces.

Lemma 6.11 *Let $\{e_1, \ldots, e_n\}$ be a basis for X. For any $i, 1 \leq 1 \leq n$, define a linear functional $\varphi_i : X \to \mathbb{C}$ by*

$$\varphi_i(x) = \alpha_i$$

where α_i is the coefficient of e_i appearing in the decomposition of x (6.13). Then, φ_i is bounded.

Proof: We will show that there exists a constant $m > 0$ such that

$$m \sum_{i=1}^{n} |\alpha_i| \leq \|x\| \tag{6.14}$$

It then follows that, for any i,

$$|\varphi_i(x)| = |\alpha_i| \leq \sum_{i=1}^{n} |\alpha_i| \leq \frac{1}{m}\|x\|$$

and, as $m > 0$, φ_i is bounded.

Clearly, it is sufficient to prove (6.14) for $x \in X$ such that $\sum_{i=1}^{n} |\alpha_i| = 1$ (use linearity and uniqueness of the coefficients α_i). The set

$$C = \{(\alpha_i) \in \mathbb{C}^n \mid \sum_{i=1}^{n} |\alpha_i| = 1\}$$

is a closed and bounded, and hence compact subset of $\mathbb{C}^n$ (Heine-Borel, Theorem 1.10). Define a function $f : C \to X$ by

$$f((\alpha_i)) = \sum_{i=1}^{n} \alpha_i e_i = x$$

As

$$\|f((\alpha_i) - f((\beta_i))\| = \left\| \sum_{i=1}^{n} (\alpha_i - \beta_i) e_i \right\| \leq \sum_{i=1}^{n} |\alpha_i - \beta_i| \|e_i\|$$

f is continuous, hence $(\alpha_i) \mapsto \|f((\alpha_i))\|$ is continuous, and, by Theorem 2.5, $\|f\|$ attains its infimum. Define

$$m = \min_{(\alpha_i)\in C} |f((\alpha_i))|$$

and suppose $|f((\alpha_i^*))| = m$. It follows from the linear independence of the basis vectors e_i that $m \neq 0$, and hence $m > 0$. For $(\alpha_i) \in C$ we have

$$m \sum_{i=1}^{n} |\alpha_i| = m \leq \|x\|$$

which implies (6.14). ∎

Theorem 6.12 *Every finite-dimensional normed linear space is a Banach space.*

Proof: We need to show that every Cauchy sequence $(x_n)_{n\geq 1}$ in a finite-dimensional normed linear space X converges. Let $\{e_1, \ldots, e_k\}$ be a basis for X. Consider the sequences $(\alpha_{i,n})_{n\geq 1}$ in $\mathbb{C}$, where

$$x_n = \sum_{i=1}^{k} \alpha_{i,n} e_i$$

Then, for $i = 1, \ldots k$, $(\alpha_{i,n})_{n\geq 1}$ is Cauchy. Indeed

$$|\alpha_{i,n} - \alpha_{i,m}| = |\varphi_i(x_n - x_m)| \leq M\|x_n - x_m\|$$

where φ_i is the linear function considered in the previous lemma, and M is a common bound for the φ_i. It follows that there are $\beta_i \in \mathbb{C}$, $i = 1 \ldots, k$, such that

$$\lim_n \alpha_{i,n} = \beta_i$$

Define $y = \sum_{i=1}^{k} \beta_i e_i$. Then

$$\|x_n - y\| \leq \sum_{i=1}^{k} |\alpha_{i,n} - \beta_i| \|e_i\|$$

and therefore, $x_n \to y$. ∎

Corollary 6.13 *Every finite-dimensional linear subspace of a Banach space is closed.*

Proof: Use the previous theorem and the following Exercise. ∎

Exercise 6.5 *Let (X, d) be a complete metric space, and $X_0 \subset X$. Prove that (X_0, d) is complete iff X_0 is a closed subset of X.*

A linear map $T : \mathbb{C}^n \to \mathbb{C}^m$, can be defined by specifying a matrix $T = (t_{ij})$, such that

$$T(e_i) = \sum_{j=1}^{m} t_{ji} f_j, \quad \text{for all } i = 1, \ldots, n$$

where $\{e_i\}$ and $\{f_j\}$ are bases for $\mathbb{C}^n$ and $\mathbb{C}^m$ respectively. It is then easy to see that, with the usual norms on $\mathbb{C}^n$ and $\mathbb{C}^m$, T is bounded. E.g., $\|T\| \leq \max_{ij} |t_{ji}|$. The following theorem is a generalization of this observation.

Theorem 6.14 *Let X and Y be normed linear spaces. If X is finite-dimensional, $\mathcal{L}(X, Y) = \mathcal{B}(X, Y)$.*

Proof: Let $\{e_1, \ldots, e_n\}$ be a basis for X, and $x = \sum_{i=1}^n \alpha_i e_i \in X$. Then

$$\|Tx\| \leq \sum_{i=1}^{n} |\alpha_i| \, \|Te_i\| \leq M\|x\| \sum_{i=1}^{n} \|Te_i\|$$

where M is a common bound for the maps φ_i of Lemma 6.11. It follows that

$$\|T\| \leq M \sum_{i=1}^{n} \|Te_i\|$$

■

The next theorem shows that, although one can define many different norms on finite-dimensional linear spaces such as $\mathbb{R}^n$ and $\mathbb{C}^n$, there is only one topology derived from a norm on these spaces.

Theorem 6.15 *Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be two norms on X. If X is finite-dimensional, then there exist constants $C_1 > 0$ and $C_2 > 0$ such that*

$$C_1\|x\|_1 \leq \|x\|_2 \leq C_2\|x\|_1, \quad \textit{for all } x \in X.$$

Proof: Consider the identity map $\mathbb{1} \in \mathcal{L}((X, \|\cdot\|_1), (X, \|\cdot\|_2))$. By Theorem 6.14 $\mathbb{1}$ is bounded. So, there exists a constant M such that

$$\|x\|_2 \leq M\|x\|_1, \quad \text{for all } x \in X.$$

Therefore, one can take $C_2 = M$. By the same argument with the roles of 1 and 2 interchanged one finds C_1. ■

Definition 6.16 *Let X be a linear space. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on X are called equivalent iff there exists constants $C_1, C_2 > 0$ such that*

$$C_1\|x\|_1 \leq \|x\|_2 \leq C_2\|x\|_1, \tag{6.15}$$

for all $x \in X$.

With this definition Theorem 6.15 can be rephrased as follows: any two norms on a finite-dimensional linear space are equivalent. In the following theorem we prove that equivalent norms, and only equivalent norms, generate the same topology on X. In combination with Theorem 6.15 this then proves that there is only one norm topology on a finite-dimensional linear space.

Theorem 6.17 *Let $\|\cdot\|_1$ and $\|\cdot\|_2$ be two norms on a (not necessarily finite-dimensional) linear space X. Then, $\|\cdot\|_1$ and $\|\cdot\|_2$ generate the same topology on X iff $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent norms.*

Proof: Consider the identity map $\mathrm{id} : (X, \|\cdot\|_1) \to (X, \|\cdot\|_2))$. The topologies generated by the two norms are the same iff id and id^{-1} are continuous (Corollary 3.13). As id is linear, id is continuous iff it is bounded. By the same reasoning as in the proof of the previous theorem, the boundedness of the identity map is equivalent to the existence of constants C_1 and C_2 such that (6.15) holds. ∎

Corollary 6.18 *A finite-dimensional linear space has exactly one norm topology.*

Exercise 6.6 *Prove that*
a) *in general, different norms on X lead to different but equivalent norms on $\mathcal{B}(X)$;*
b) *the set $\mathcal{B}(X)$ itself is independent of the chosen norm.*

7 Hilbert Spaces

In MAT203A we have considered different kinds of topological spaces. In increasing order of specialization we studied arbitrary toplogical spaces, metric spaces, normed linear spaces, and Banach spaces. Hilbert spaces are Banach spaces with a norm that is derived from an inner product. Thus, again, they have an extra feature as compared to arbitrary Banach spaces, which makes them more special. On the other hand, the extra structure will lead to new interesting properties.

The material of this sections elementary introduction to Hilbert space is covered in Chapter 4 of
W. Rudin, *Real and Complex Analysis*, 3rd ed, McGraw-Hill, New York 1986.

P.R. Halmos wrote a well-known study book on Hilbert space, which consists of 250 problems with hints and solutions:
P.R. Halmos, *A Hilbert Space Problem Book*, 2nd ed, Springer Verlag, New York 1982.

The new feature of Hilbert spaces is the *inner product*, which, in Section 7.3, we will use to introduce the notion of orthonormal basis of a Hilbert space.

Hilbert spaces are Banach spaces with a norm that is derived from an inner product.

7.1 Inner Products

Definition 7.1 (Inner Product) *An inner product on a linear space X over $\mathbb{C}$ (or $\mathbb{R}$) is a map $(\cdot,\cdot) : X \times X \to \mathbb{C}$ (or $\mathbb{R}$), with the following properties: for all $x, y, z \in X$ and $\lambda \in \mathbb{C}$ (or $\mathbb{R}$) one has*
i) $(x, y + \lambda z) = (x, y) + \lambda(x, z)$ (linear in the second argument)
ii) $(y, x) = \overline{(x, y)}$, where $\overline{\cdot}$ denotes complex conjugation ("hermitian symmetric")
iii) $(x, x) \geq 0$ (non-negative, or positive semi-definite)
iv) $(x, x) = 0 \Rightarrow x = 0$ (positive definite).

A linear space with an inner product is often referred to as an *inner product space.*

From i) and ii) it follows that $(\cdot,\cdot)$ is *antilinear* in the first argument: $(y + \lambda z, x) = (y, x) + \overline{\lambda}(z, x)$, for all $x, y, z \in X$ and $\lambda \in \mathbb{C}$ (or $\mathbb{R}$). In the case of complex X, this means that $(\cdot,\cdot)$ is not bilinear. It is called *sesquilinear*, which literally means $1\frac{1}{2}$ linear.

Note that there are two conventions for the linearity of the inner product in the literature: in most of the mathematics oriented literature $(\cdot,\cdot)$ is linear in the first argument. We will adopt the convention that the inner product is linear in the second argument, which is the most common in applied mathematics and physics oriented works.

The following very useful inequality for inner products on a real or complex linear spaces is a consequence of the non-negative definiteness:

$$|(x,y)|^2 \leq (x,x)(y,y), \quad \text{for all } x,y \in X \tag{7.1}$$

It is usually called the *Cauchy-Schwarz inequality* .

Exercise 7.1 *Prove the Cauchy-Schwarz inequality.*

For simplicity of notation we will henceforth assume that X is a linear space over $\mathbb{C}$. Many statements carry over to the case of real linear spaces.

Let X be a linear space equipped with an inner product $(\cdot\,,\cdot)$. We can define a norm on X by

$$\|x\| = \sqrt{(x,x)}$$

Check that this indeed defines a norm.

Examples:

1. The standard inner product on $\mathbb{C}^n$ is given by

$$(x,y) = \sum_{i=1}^{n} \overline{x_i} y_i, \quad \text{for all } x,y \in \mathbb{C}^n$$

2. Let $C([a,b],\mathbb{C})$ denote the space of all complex-valued continuous functions defined on the interval $[a,b]$. An inner product on this space can be defined by

$$(f,g) = \int_a^b \overline{f(x)} g(x)\,dx, \quad \text{for all } f,g \in C([a,b],\mathbb{C})$$

By continuity this can be extended to $L_2([a,b],\mathbb{C})$, defined as the completion of $C([a,b],\mathbb{C})$ for the norm defined by $\|f\| = (f,f)^{1/2}$. This norm is called the L^2 norm.

3. Define l_2 by

$$l_2 = \{(z_n)_{n=1}^{\infty} \mid \sum_{n=1}^{\infty} |z_n|^2 < \infty\}$$

l_2 is a complex linear space with the obvious operations of addition and multiplication by a scalar, and a natural inner product on it is given by

$$(x,y) = \sum_{n=1}^{\infty} \overline{x_n} y_n$$

l_2 is pronounced *little el two* to distinguish it from L^2 in the previous example.

4. Let M_n denote the space of all $n \times n$ matrices with complex entries. One can define a inner product on M_n by

$$(A, B) = \mathrm{Tr} A^* B$$

where Tr denotes the trace and * is hermitian conjugation of matrices. Check that this is the natural inner product one obtains by identifying M_n with $\mathbb{C}^{n^2}$.

Theorem 7.2 *A normed linear space X is an inner product space with the norm derived from the inner product iff*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2, \quad \text{for all } x, y \in X \quad . \tag{7.2}$$

When this condition is satisfied, the inner product is necessarily given by the formula

$$(x, y) = \frac{1}{4}\left\{\|x+y\|^2 - \|x-y\|^2 - i\|x+iy\|^2 + i\|x-iy\|^2\right\}, \quad \text{for all } x, y \in X \quad . \tag{7.3}$$

Proof: The proof of this theorem is homework. ∎

The relation (7.2) has a geometric content and is called the *Parallelogram Law.* It says that, if one considers the parallelogram spanned by the vectors x and y in the two dimensional subspace ($\equiv$ plane) generated by x and y, the sum of the squares of two adjacent sides of the parallelogram is equal to half the sum of the squares of the diagonals. Equation (7.3) is called the polarization formula. A useful consequence of this formula is that a sequilinear form is uniquely determined by its values on the diagonal, i.e., its values for equal first and second arguments.

Let $(X, (\cdot\,,\,\cdot)_X)$ and $(Y, (\cdot\,,\,\cdot)_Y)$ be two inner product spaces. Then $X \times Y = \{(x, y) \mid x \in X, y \in Y\}$ has a natural inner product defined by

$$((x_1, y_1), (x_2, y_2))_{X\times Y} = (x_1, x_2)_X + (y_1, y_2)_Y$$

which in turn leads to a natural norm on $X \times Y$:

$$\|(x, y)\| = \sqrt{\|x\|^2 + \|y\|^2}$$

Unless otherwise stated, this will be the norm considered on a Cartesian product of two inner product spaces.

Theorem 7.3 *Let X be an inner product space, considered as a normed space with the norm the norm derived from the inner product. Then, the inner product is continuous as a map from $X \times X \to \mathbb{C}$.*

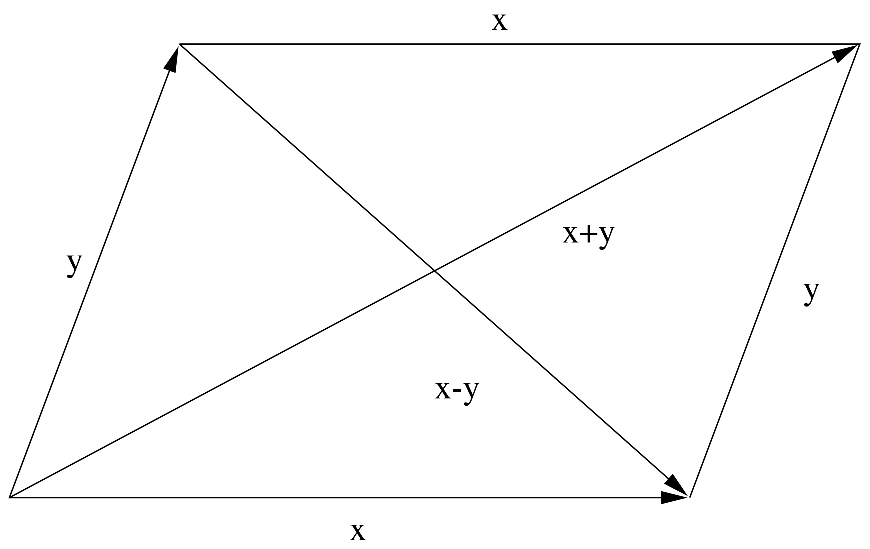

Figure 7.1: The geometric interpretation of the parallelogram law (7.2).

Proof: For all $x, y \in X$ one has the bound

$$|(x, y)| \leq \|x\| \, \|y\|$$

which is nothing else than the Cauchy-Schwartz inequality. Just as for linear maps, continuity follows from this kind of boundedness:

$$|(x_1, y_1) - (x_2, y_2)| = |(x_1 - x_2, y_1) + (x_2, y_1 - y_2)| \leq \|x_1 - x_2\| \, \|y_1\| + \|x_2\| \, \|y_1 - y_2\|$$

■

7.2 Hilbert Spaces

Definition 7.4 *A Hilbert space is an inner product space X that is complete for the norm derived from the inner product.*

In particular, a Hilbert space is a Banach space.

Examples:
Except for $C([a, b], \mathbb{C})$, the examples of inner product spaces given in the previous section are all complete and, therefore, are Hilbert spaces.

Exercise 7.2 *Let $\mathcal{H}_0$ be a linear subspace of a Hilbert space $\mathcal{H}$. Prove that $\mathcal{H}_0$, equipped with the same inner product as $\mathcal{H}$, is a Hilbert space iff $\mathcal{H}_0$ is a closed subspace of $\mathcal{H}$.*

The inner product structure of a Hilbert space allows us to introduce the concept of orthogonality of vectors. Let $\mathcal{H}$ be a Hilbert space. We say that x and $y \in \mathcal{H}$ are *orthogonal* iff $(x, y) = 0$. Shorthand for "x and y are orthogonal" is $x \perp y$. For two linear subspaces $\mathcal{H}_0$ and $\mathcal{H}_1$ of $\mathcal{H}$, we say that $\mathcal{H}_0$ and $\mathcal{H}_1$ are orthogonal if $x \perp y$ for all $x \in \mathcal{H}_0$ and $y \in \mathcal{H}_1$, in which case one writes $\mathcal{H}_0 \perp \mathcal{H}_1$. Similarly, $x \perp \mathcal{H}_0$ means $x \perp y$, for all $y \in \mathcal{H}_0$. The *orthogonal complement* $\mathcal{H}_0^\perp$ of a linear subspace $\mathcal{H}_0$ of $\mathcal{H}$ is defined by

$$\mathcal{H}_0^\perp = \{x \in \mathcal{H} \mid x \perp y, \text{for all } y \in \mathcal{H}_0\}$$

The notion of orthogonality makes it possible to think about vectors and linear subspaces of a Hilbert space in a geometric way. This is sometimes very helpful, much in the same way as our geometric intuition for the Euclidean spaces $\mathbb{R}^n$ helps us to understand linear transformations of $\mathbb{R}^n$. It should be mentioned that the geometry of Banach spaces can be surprisingly complicated and quite different from what naive extrapolations of the finite-dimensional situation would suggest. The following two results, and their applications rely very much on the notion of orthogonality.

Theorem 7.5 *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{H}_0$ a subset of $\mathcal{H}$. Then, $\mathcal{H}_0^\perp$ is a closed linear subspace of $\mathcal{H}$.*

Proof: Let us first prove that $\mathcal{H}_0^\perp$ is indeed a linear subspace. For this we must show that for all $x, y \in \mathcal{H}_0^\perp$, and $\lambda \in \mathbb{C}$, $x + \lambda y \in \mathcal{H}_0^\perp$. But this follows immediately from the linearity of the inner product:

$$(x_0, x + \lambda y) = (x_0, x) + \lambda(x_0, y) = 0 + 0, \quad \text{for all } x_0 \in \mathcal{H}_0$$

Next, we show that $\mathcal{H}_0^\perp$ is closed. A Hilbert space is a metric space, and therefore, it suffices to show that for any convergent sequence of elements $x_n \in \mathcal{H}_0^\perp$, the limit x also belongs to $\mathcal{H}_0^\perp$. Let $x_0 \in \mathcal{H}_0$. From Theorem 7.3 we have that the inner product is continuous and therefore, we can compute (x_0, x) as follows:

$$(x_0, x) = (x_0, \lim_n x_n) = \lim_n (x_0, x_n) = \lim_n 0 = 0$$

Hence, $x \in \mathcal{H}_0^\perp$. ∎

Exercise 7.3 *Prove that for any set $A \subset \mathcal{H}$, one has*

$$(\overline{A})^\perp = A^\perp$$

where $\overline{A}$ denotes the closure of A.

The following theorem expresses a deep property of Hilbert spaces. Beware that there is no analogous theorem for arbitrary Banach spaces.

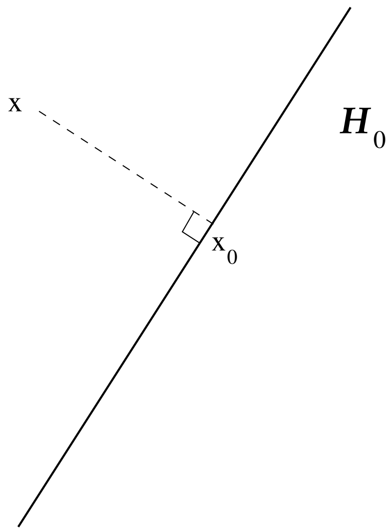

Figure 7.2: x_0 is the point in $\mathcal{H}_0$ closest to x.

Theorem 7.6 *Let $\mathcal{H}_0$ a closed linear subspace of a Hilbert space $\mathcal{H}$.*
i) For each $x \in \mathcal{H}$ there is a unique $x_0 \in \mathcal{H}_0$ such that

$$\|x - x_0\| = \min_{y \in \mathcal{H}_0} \|x - y\| \tag{7.4}$$

ii) The point $x_0 \in \mathcal{H}_0$, closest to $x \in \mathcal{H}$, is the unique element of $\mathcal{H}_0$ with the property that $x - x_0 \perp \mathcal{H}_0$.

Proof:
Step 1. First, we prove that there exists $x_0 \in \mathcal{H}_0$ such that

$$\|x - x_0\| = d := \inf\{\|x - y\| \mid y \in \mathcal{H}_0\} \tag{7.5}$$

From the definition of d it follows that there is a sequence of elements $y_n \in \mathcal{H}_0$, such that $\lim_n \|x - y_n\| = d$. So, for all $\epsilon > 0$, there exists N such that for all $n \geq N$, $\|x - y_n\| \leq d + \epsilon$. Let us show that the sequence (y_n) is Cauchy. By the Parallelogram Law we have

$$\|y_m - y_n\|^2 + \|2x - y_m - y_n\|^2 = 2\|x - y_m\|^2 + 2\|x - y_n\|^2$$

As $(y_m + y_n)/2 \in \mathcal{H}_0$, we have

$$\|x - (y_m + y_n)/2\| \geq d$$

By combining these relations we obtain:

$$\begin{aligned} \|y_m - y_n\|^2 &= 2\|x - y_m\|^2 + 2\|x - y_n\|^2 - \|2x - y_m - y_n\|^2 \\ &\leq 4(d + \epsilon)^2 - 4d^2 \\ &\leq 8\epsilon(d + \epsilon) \end{aligned}$$

for all $m, n \geq N$. This proves that (y_n) is Cauchy. As a Hilbert space is complete by definition, there is y such that $y_n \to y$, and as $\mathcal{H}_0$ is closed by assumption, we have that $y \in \mathcal{H}_0$, and we can take $x_0 = y$ by continuity of the distance.
Step 2. Next, we prove the uniqueness of the vector $x_0 \in \mathcal{H}_0$ minimizing $\|x - x_0\|$. Suppose x_0 and x_0' both minimize the distance to x, i.e., $\|x - x_0\| = \|x - x_0'\| = d$. Again, by the Parallelogram Law, we have

$$2\|x - x_0\|^2 + 2\|x - x_0'\|^2 = \|2x - x_0 - x_0'\|^2 + \|x_0 - x_0'\|^2$$

and hence, as $x_0 + x_0' \in \mathcal{H}_0$,

$$\|x_0 - x_0'\|^2 = 4d^2 - \|2x - x_0 - x_0'\|^2 \leq 0$$

Therefore, $\|x_0 - x_0'\| = 0$, and we must conclude that $x_0 = x_0'$. So there can be only one $x_0 \in \mathcal{H}_0$ such that $\|x - x_0\| = d$.
Step 3. The unique x_0 found in Steps 1 and 2, must satisfy $(z, x - x_0) = 0$, for all $z \in \mathcal{H}_0$. This follows from the inequality

$$\|x - x_0\|^2 \leq \|x - (x_0 - \lambda z)\|$$

for any $z \in H_0, \lambda \in \mathbb{C}$. Indeed, by expanding the RHS and using $\|x - x_0\|^2 = d^2$, we obtain

$$d^2 \leq d^2 + |\lambda|^2\|z\|^2 - 2\mathrm{Re}\,\lambda(x - x_0, z)$$

which simplifies to

$$|\lambda|^2\|z\|^2 \geq 2\mathrm{Re}\,\lambda(x - x_0, z)$$

Take λ of the form $\epsilon e^{i\varphi}$, with $\epsilon > 0$ and $\varphi \in \mathbb{R}$, divide by ϵ, and take the limit $\epsilon \downarrow 0$, and you obtain

$$2\mathrm{Re}\, e^{i\varphi}(x - x_0, z) \leq 0, \quad \text{for all } \varphi \in \mathbb{R}$$

It follows that $(x - x_0, z) = 0$ (consider this inequality for $\varphi = 0, \pi/2, \pi$, and $3\pi/2$).
Step 4. Finally, we show that x_0 is the only element in $\mathcal{H}_0$ such that $x - x_0 \perp \mathcal{H}_0$. Suppose x_0' is another such element in $\mathcal{H}_0$. Then $x_0 - x_0' \in \mathcal{H}_0$, and, for any $y \in \mathcal{H}_0$, we have

$$(y, x_0 - x_0') = (y, x - x_0') - (y, x - x_0) = 0 - 0 = 0$$

In particular, we can take $y = x_0 - x_0'$, and therefore we must have $x_0 = x_0'$. ∎

Note that for the proof of Theorem 7.6, part i), it is sufficient that $\mathcal{H}_0$ is a closed convex subset of $\mathcal{H}$.

Exercise 7.4 *For any linear subspace $\mathcal{H}_0 \subset \mathcal{H}$, prove that*
i) $\mathcal{H}_0^{\perp\perp} := (\mathcal{H}_0^{\perp})^{\perp} = \overline{\mathcal{H}_0}$
ii) $\mathcal{H}_0^{\perp\perp\perp} = \mathcal{H}_0^{\perp}$.
iii) How do the relation i) and ii) have to be changed if $\mathcal{H}_0$ is just a subset and not a linear subspace?

Exercise 7.5 *Let $\mathcal{H}_0$ and $\mathcal{H}_1$ be two linear subspaces of a Hilbert space $\mathcal{H}$, such that $\mathcal{H}_0 \perp \mathcal{H}_1$.*
i) Prove that $\mathcal{H}_0^{\perp\perp} \perp \mathcal{H}_1^{\perp\perp}$.
ii) Is it true, in general, that $\mathcal{H}_0^{\perp\perp\perp} \perp \mathcal{H}_1^{\perp\perp\perp}$?

7.3 Orthonormal Bases

Let $\mathcal{H}$ be a Hilbert space ans denote its inner product by $\langle \cdot, \cdot \rangle$. This is another common notation for inner products, which is generally reserved for Hilbert spaces.

A set $U \subset \mathcal{H}$, $0 \notin U$, is called *orthogonal* if any pair of distinct elements (vectors) $u, v \in U$ are mutually orthogonal, i.e., $\langle u, v \rangle = 0$. A set of vectors U is called *orthonormal* if it is orthogonal and $\|u\| = 1$ for all $u \in U$, in which case the vectors u are said to be *normalized*.

Exercise 7.6 *Prove that if U is an orthogonal set, then U is a set of linearly independent vectors.*

For the finite-dimensional Hilbert spaces $\mathbb{C}^n$ we have the notion of *orthonormal basis*, often abbreviated as o.n. basis. A set of vectors $\{e_1, \dots, e_n\} \subset \mathbb{C}^n$ is an o.n. basis if it has the following properties:
i) $\langle e_i, e_j \rangle = \delta_{ij}$ for $1 \leq i, j \leq n$, and
ii) for all $x \in \mathbb{C}^n$ there are unique coordinates $x_i \in \mathbb{C}$ such that

$$x = \sum_{i=1}^{n} x_i e_i \tag{7.6}$$

Our goal is now to generalize the notion of o.n. basis to arbitrary Hilbert spaces. Of course, we expect that, in general, an o.n. basis will be an infinite set of vectors. In the case of *separable* Hilbert spaces (i.e., containing a countable dense set), we will find that an o.n. basis is at most countably infinite. In this case the sum in (7.6) simply becomes an infinite series in a Banach space, for which we have a natural notion of convergence. But how do we interpret a sum with an uncountable number of terms? We address this question first.

7.3.1 Unordered infinite sums

Let I be a (not necessarily countable) index set, and consider a set of elements of linear topological space X indexed by $I : \{x_\alpha \mid \alpha \in I\} \subset X$. We would like to give sense to

$$\sum_{\alpha \in I} x_\alpha \tag{7.7}$$

For every finite subset $J \subset I$ define a partial sum S_J by

$$S_J = \sum_{\alpha \in J} x_\alpha$$

We say that $\sum_{\alpha \in I} x_\alpha$ converges to $x \in X$ if for any neighborhood V of x there is a finite set J_0 such that for all finite sets J containing J_0 one has $S_J \in V$. This is a special case of convergence of nets. Note that (7.7) is an 'unordered sum', and that it converges iff any permutation of its terms converges. The following exercise shows that, in the case of normed linear spaces, this notion of convergence is not very different from the usual notion of (absolutely) convergent series with a countable number of terms.

Exercise 7.7 *Prove the following:*
i) If X is a normed linear space and $\sum_{\alpha \in I} x_\alpha \to x$, then there exists a countable subset $\{\alpha_1, \alpha_2, \ldots\} \subset I$ such that the series $\sum_{i=1}^{\infty} x_{\alpha_i}$ converges to x.
ii) Under the same conditions as in i) at most a countable number of the x_α are non-zero.

This exercise also shows that in the case of $X = \mathbb{C}$ (or $\mathbb{R}$), $\sum_{\alpha \in I} x_\alpha \to x$ iff $\sum_{\alpha \in I} |x_\alpha| < \infty$, for absolute convergence and unconditional convergence coincide for numerical series.

Exercise 7.8 *Let $\{x_\alpha\}_{\alpha \in I}$ be a family of non-negative real numbers. Prove that*

$$\sum_{\alpha \in I} x_\alpha = \sup\{\sum_{\alpha \in J} x_\alpha \mid J \subset I, J \text{ finite } \}$$

It is quite obvious how to generalize the Cauchy criterion for convergence of unordered sum in a Banach space (such as, e.g., a Hilbert space).

Proposition 7.7 *If X is a Banach space we have that $\sum_{\alpha \in I} x_\alpha$ converges iff for every $\epsilon > 0$ there exists a finite set $J_0 \subset I$ such that $\|S_J\| < \epsilon$ for every $J \subset I \setminus J_0$.*

7.3.2 Bessel's inequality etc.

In this section the sums $\sum_{\alpha \in I}$ have to be interpreted as unordered infinite sums in the sense of above. Although in the future we will mostly consider separable Hilbert spaces (and countable sums) it is worth postponing this restriction for another while. This will take the mystery out of non-separable Hilbert spaces.

Lemma 7.8 *Let $\mathcal{H}$ be a Hilbert space and $U = \{u_\alpha \mid \alpha \in I\} \subset \mathcal{H}$ an orthogonal set. Then, the sum $\sum_{\alpha\in I} u_\alpha$ converges iff $\sum_{\alpha\in I} \|u_\alpha\|^2 < \infty$. In this case one has*

$$\|\sum_{\alpha\in I} u_\alpha\|^2 = \sum_{\alpha\in I} \|u_\alpha\|^2 \tag{7.8}$$

If U is orthonormal one has

$$\|\sum_{\alpha\in I} c_\alpha u_\alpha\|^2 = \sum_{\alpha\in I} |c_\alpha|^2 \tag{7.9}$$

for arbitrary $c_\alpha \in \mathbb{C}$ for which the RHS is finite.

Proof: For any finite set J we have

$$\|\sum_{\alpha\in J} u_\alpha\|^2 = \sum_{\alpha,\beta\in J} \langle u_\alpha, u_\beta\rangle = \sum_{\alpha\in J} \langle u_\alpha, u_\alpha\rangle = \sum_{\alpha\in J} \|u_\alpha\|^2$$

It follows that the Cauchy criterion is fulfilled for $\sum_{\alpha\in I} u_\alpha$ iff it is for $\sum_{\alpha\in I} \|u_\alpha\|^2$, and, therefore, each of the sums converges iff the other does. (7.8) follows because the norm is a continuous function.

(7.9) is a simple application of (7.8): If U is orthonormal, the set $\{c_\alpha u_\alpha\}$ is orthogonal, and we have

$$\sum_{\alpha\in I} \|c_\alpha u_\alpha\|^2 = \sum_{\alpha\in I} |c_\alpha|^2$$

■

Theorem 7.9 *Let $U = \{u_\alpha \mid \alpha \in I\} \subset \mathcal{H}$ be an orthonormal set and $x \in \mathcal{H}$. Then*
i) $\sum_{\alpha\in I} |\langle u_\alpha, x\rangle|^2 \le \|x\|^2$,
ii) $x_U := \sum_{\alpha\in I} \langle u_\alpha, x\rangle u_\alpha$ is a convergent sum, and
iii) $x - x_U \in U^\perp$

Proof: We begin by computing $\|x - \sum_{\alpha\in J} \langle u_\alpha, x\rangle u_\alpha\|$ for any finite subset $J \subset I$:

$$\begin{aligned}
\|x - \sum_{\alpha\in J} \langle u_\alpha, x\rangle u_\alpha\|^2 &= \langle (x - \sum_{\alpha\in J} \langle u_\alpha, x\rangle u_\alpha), (x - \sum_{\beta\in J} \langle u_\beta, x\rangle u_\beta)\rangle \\
&= \langle x, x\rangle - \sum_{\beta\in J} \langle u_\beta, x\rangle \langle x, u_\beta\rangle - \sum_{\alpha\in J} \overline{\langle u_\alpha, x\rangle} \langle u_\alpha, x\rangle \\
&\quad + \sum_{\alpha,\beta\in J} \overline{\langle u_\alpha, x\rangle} \langle u_\beta, x\rangle \langle u_\alpha, u_\beta\rangle \\
&= \|x\|^2 - \sum_{\alpha\in J} |\langle u_\alpha, x\rangle|^2
\end{aligned}$$

Hence

$$\sum_{\alpha\in J}|\langle u_\alpha, x\rangle|^2 = \|x\|^2 - \|x - \sum_{\alpha\in J}\langle u_\alpha, x\rangle u_\alpha\|^2 \leq \|x\|^2$$

and this proves i) by 'taking the limit $J \uparrow I$'.

The convergence claimed in ii) now follows by application of Lemma 7.8.

In order to prove iii) consider any $u_{\alpha_0} \in U$ and relying on the orthonormality of U and the continuity of the inner product we compute:

$$\begin{aligned}\langle x - \sum_{\alpha\in I}\langle u_\alpha, x\rangle u_\alpha, u_{\alpha_0}\rangle &= \langle x, u_{\alpha_0}\rangle - \sum_{\alpha\in I}\overline{\langle u_\alpha, x\rangle}\langle u_\alpha, u_{\alpha_0}\rangle \\ &= \langle x, u_{\alpha_0}\rangle - \langle x, u_{\alpha_0}\rangle = 0\end{aligned}$$

Hence, $x - \sum_{\alpha\in I}\langle u_\alpha, x\rangle u_\alpha \in U^\perp$. ■

The inequality i) is called *Bessel's inequality.*

Lemma 7.10 *For any orthonormal set $U = \{u_\alpha \mid \alpha \in I\}$ in $\mathcal{H}$, the set of vectors of the form*

$$\sum_{\alpha\in I} c_\alpha u_\alpha$$

with $c_\alpha \in \mathbb{C}$, such that the sum converges is a closed linear subspace of $\mathcal{H}$.

Exercise 7.9 *Prove Lemma 7.10.*

Let $[U]$ denote the smallest closed linear subspace of $\mathcal{H}$ that contains U, which is described in Lemma 7.10. By Lemma 7.8, we have

$$[U] = \{\sum_{\alpha\in I} c_\alpha u_\alpha \mid c_\alpha \in \mathbb{C} \text{ such that } \sum_{\alpha\in I}|c_\alpha|^2 < \infty\}$$

By combining Theorem 7.6 and Theorem 7.9 we see that x_U, defined in Theorem 7.9 ii), is the unique element of $[U]$ satisfying

$$\|x - x_U\| = \min_{u\in[U]}\|x - u\|$$

In particular, it follows that $x_U = x$ if $[U] = \mathcal{H}$. In this case U is called a *orthonormal (o.n.) basis* of $\mathcal{H}$.

Here are some equivalent conditions for this property, often referred to as the *completeness* of the set U.

Theorem 7.11 *If $U = \{u_\alpha \mid \alpha \in I\} \subset \mathcal{H}$ is orthonormal, the following properties are equivalent:*
i) $\langle u_\alpha, x\rangle = 0$ *for all* $\alpha \in I$ *implies* $x = 0$, *i.e.,* $U^\perp = \{0\}$.
ii) $x = \sum_{\alpha\in I}\langle u_\alpha, x\rangle u_\alpha$, *for all* $x \in \mathcal{H}$.
iii) $\|x\|^2 = \sum_{\alpha\in I} |\langle u_\alpha, x\rangle|^2$, *for all* $x \in \mathcal{H}$.
iv) $[U] = \mathcal{H}$.
v) U is maximal in the sense that for any $0 \neq v \in \mathcal{H} \setminus U$, the set $U \cup \{v\}$ is no longer orthogonal.

Proof: We prove the string of implications $i) \Rightarrow ii) \Rightarrow iii) \Rightarrow iv) \Rightarrow v) \Rightarrow i)$.
$i) \Rightarrow ii)$: By Theorem 7.9 iii), $U^\perp = \{0\}$ implies ii).
$ii) \Rightarrow iii)$: This is Lemma 7.8 (7.9).
$iii) \Rightarrow iv)$: iii) implies that $U^\perp = 0$ which is equivalent to iv).
$iv) \Rightarrow v)$: iv) implies that any v is of the form $v = \sum_{\alpha\in I} c_\alpha u_\alpha$. Therefore, $\langle u_\alpha, v\rangle = c_\alpha$, and hence $v \perp U$ implies $c_\alpha = 0$ for all α, and hence $v = 0$.
$v) \Rightarrow i)$: i) is a reformulation of v). ∎

i) is often the easiest condition to verify. ii) is the property that is most often used explicitly. iii) is called *Parseval's identity*, iv) simply expresses completeness of the basis, and v) will be used in the proof of existence of an o.n. basis in a arbitrary Hilbert space (Theorem 7.12).

Our next result is that every Hilbert space has an o.n. basis. It also says that any orthonormal set can be extended to an orthonormal basis.

Theorem 7.12 *Every Hilbert space $\mathcal{H}$ has an o.n. basis, and if U is an orthonormal set, then $\mathcal{H}$ has an o.n. basis containing U.*

Proof: The proof is a standard application of Zorn's Lemma (see, e.g., W. Rudin, *Real and Complex Analysis*)). If $\mathcal{H} = \{0\}$, the statement is trivially true with $U = \emptyset$. Suppose $\mathcal{H}_0 \neq \{0\}$. Then, the family of all orthonormal sets of $\mathcal{H}$ is non-empty and has a partial order given by the inclusion of sets. If $\{U_\alpha\}$ is a totally ordered family of orthonormal sets (i.e., a chain for the order relation), then $\bigcup_\alpha U_\alpha$ is an orthonormal set and is an upper bound (in the sense of inclusion) for the family $\{U_\alpha\}$. Zorn's Lemma then implies that that the family of all orthonormal sets in $\mathcal{H}$ has a maximal element, which then satisfies the equivalent conditions of Theorem 7.11.

To prove that any orthonormal set U can be extended to an o.n. basis of $\mathcal{H}$, repeat the above argument starting with the family of all orthonormal sets containing U. ∎

Remark: Do not confuse the notion of o.n. basis with that of Hamel basis of a linear space. U is a Hamel basis if for all $x \in \mathcal{H}$, x can be uniquely written as *finite* linear combination of elements in U.

7.3.3 The Gram-Schmidt orthogonalization procedure

From the previous Theorem we know that every Hilbert space has an o.n. basis (and, hence, many o.n. bases). This would not be very useful in practical applications if we would not have a means of constructing one. The Gram-Schmidt procedure is a recipe for constructing an o.n. basis starting from a countable linearly independent set with the property that its linear span is dense in $\mathcal{H}$.

Let V be a countable set of linearly independent vectors in a Hilbert space $\mathcal{H}$. The Gram-Schmidt orthogonalization procedure is a method to construct an o.n. set U such that $[U] = [V]$, where for any subset $A \subset \mathcal{H}$, $[A]$ denotes the set of all convergent sums of the form

$$\sum_{v \in A} c_v v$$

with $c_v \in \mathbb{C}$.

Denote the elements of V by v_n. The o.n. set $U = \{u_n\}$ is then constructed inductively by putting $u_1 = v_1/\|v_1\|$ and

$$u_{n+1} = C\left(v_{n+1} - \sum_{k=1}^{n} \langle u_k, v_{n+1}\rangle u_k\right)$$

for all $n \geq 1$, and where C is chosen such that $\|u_{n+1}\| = 1$.

It is straightforward to check that $[\{v_1, \ldots, v_n\}] = [\{u_1, \ldots, u_n\}]$ for all $n \geq 1$, and hence

$$[V] = \overline{\bigcup_n [\{v_1, \ldots, v_n\}]} = \overline{\bigcup_n [\{u_1, \ldots, u_n\}]} = [U]$$

Exercise 7.10 *As a metric space a Hilbert space is called separable if it contains a countable dense subset. Prove that a Hilbert space is separable iff it has a finite or countably infinite o.n. basis.*

Exercise 7.11 *Prove that if H_0 is a dense linear subspace of a Hilbert space $\mathcal{H}$, $\mathcal{H}$ has an o.n. basis consisting of elements in $\mathcal{H}_0$. Does the same hold for arbitrary dense subsets of $\mathcal{H}$?*

Consider $\mathcal{H} = L^2([-1, 1])$, and let it be given that the set $V = \{1, x, x^2, \ldots\}$ of all monic monomials is linearly independent and one has $[V] = L^2([-1, 1])$. In the following excercise you are asked to compute the first few elements of a particular o.n. basis of $[V] = L^2([-1, 1])$, called the Legendre polynomials.

Exercise 7.12 *Calculate the first 3 (or more) elements of the o.n. basis obtained from V, with the given ordering, by applying the Gram-Schmidt procedure.*

This example can be generalized as follows. Let $[a, b] \subset \mathbb{R}$ be a finite or infinite interval and $w : [a, b] \to \mathbb{R}$ be a continuous function, $w(x) > 0$ except for at most finitely many x, and such that

$$\int_a^b w(x)\, dx < +\infty \quad .$$

For any such weight function w, one can define an inner product on $C_b([a, b]) = \{f : [a, b] \to \mathbb{C} \mid f \text{ continuous and } |f| \text{ bounded}\}$, by

$$(f, g) = \int_a^b \overline{f(x)} g(x) w(x)\, dx, \qquad \text{for all } f, g \in C_b([a, b]) \quad .$$

Let $L^2([a, b], w)$ be the Hilbert space obtained by completion of $C_b([a, b])$ for the metric derived from this inner product. One can then use the Gram-Schmidt procedure to obtain an o.n. basis of polynomials for this Hilbert space.

7.4 Examples of Hilbert spaces in applications

7.4.1 Quantum Mechanics

In the 20's Quantum Mechanics developed at a feverish pace, and people did hardly have time to pause and think about the mathematical structures they were using and/or developing. Only later was it realized (by John von Neumann) that Hilbert space is the natural setting for Quantum Mechanics. But the harm had been done: generations of physicists have been educated in the "new mechanics" without any knowledge of the proper language, which added even more mystery to the already quite mysterious theory.

One of the simplest mechanical systems where quantum mechanics is needed for an accurate description is a particle, confined to move on a straight line and bouncing back and forth between two parallel walls (the so-called particle in a box). Quantum effects are important when its kinetic energy is not much larger than $m\hbar^2/L^2$, where m is the mass of the particle, $\hbar$ is Planck's constant, and L is the distance between the walls.

In Quantum Mechanics the state of the particle at each instant in time t is described by an element $\psi(t, \cdot) \in L^2([0, L])$, i.e., a vector in the Hilbert space of square-integrable complex-valued funtions on the interval $[0, L]$. This is in constrast to classical mechanics à la Newton, where the state is described by just two numbers: the position of the particle ($0 \leq x \leq L$) and its velocity ($v \in \mathbb{R}$).

The dynamics of the system is determined by a PDE, called the Schrödinger equation,

$$i\hbar \frac{\partial}{\partial t} \psi(x, t) = -\frac{\hbar^2}{2m} \frac{\partial^2}{\partial x^2} \psi(x, t), \quad x \in [0, L], t \in \mathbb{R} \tag{7.10}$$

with the boundary conditions $\psi(0,t)=\psi(L,t)=0$, for all $t\in\mathbb{R}$.

We will discuss this equation and its solutions later on, making use of Hilbert space notions.

7.4.2 Natural L^2-spaces

L^2-spaces often arise naturally in many contexts because

$$\int |f(x)|^2\,dx \tag{7.11}$$

is a fundamental quantity of the model that one wants to be finite. E.g.,

- $f(x)$ is the velocity of a fluid at the point with coordinates x. Then, for an incompressible fluid, $\int_V |f(x)|^2\,dx$ is the total kinetic energy of the fluid, which, for any finite volume V, should be finite.

- $f(x)$ is a random variable with vanishing mean. The space of all such random variables with finite variance is an L^2-space.

- The electromagnetic field in space-time is described by two vector fields $E(x,t)$ and $B(x,t)$. $|E(x,t)|^2+|B(x,t)|^2$ (in suitable units) is the energy density of the electromagnetic field, and hence, again, an integral of the form (7.11) is a fundamental quantity in the theory: the total energy.

7.4.3 Sobolev spaces

Sometimes, the relevant quantity is not just the L^2 norm but rather a combination of the L^2-norm and the L^2-norm of some derivatives. On $C^n([a,b]$, the space of functions with n continuous derivatives, one defines the *Sobolev* norms by

$$\|f\|_{n,2}=\sum_{k=0}^{n}\int_a^b |f^{(k)}(x)|^2\,dx \tag{7.12}$$

where $f^{(k))}$ denotes the k-th derivative of f. The Hilbert space obtained by completion of $C^n([a,b])$ with the norm $\|\cdot\|_{n,2}$, is denoted by $H^n([a,b])$. These Hilbert spaces are called *Sobolev spaces*. They are widely used in the analysis of partial differential operators needed for PDE.

8 Special orthonormal bases 1. Fourier series

What makes Hilbert space so powerful in applications is the possibility to express a problem in a suitable o.n. basis, e.g., one of the bases of orthogonal polynomials introduced in Section 7.3.3. But orthogonal polynomials are by no means the only interesting o.n. bases. One of the most widely used is the *Fourier basis*, consisting of trigonometric functions, which is the subject of this chapter. Here, we will introduce the Fourier basis, which is the mathematical formulation underlying Fourier series, and investigate its general properties. Later, in Chapter 9, we will explore a variety of applications of Fourier series. In Chapter 10 we will give a brief introduction to yet another class of important special o.n. bases, the so-called *wavelets*.

8.1 Completeness of the Fourier basis

Our main objective is to prove that $\{e_n \mid n \in \mathbb{Z}\}$ with

$$e_n(x) = \frac{1}{\sqrt{2\pi}} e^{inx} \tag{8.1}$$

is an o.n. basis of $L^2([0, 2\pi])$. Of course one can then construct an o.n. basis for $L^2([a, b])$, for any $a < b \in \mathbb{R}$, by a simple affine transformation of the functions (i.e., by a shift and a rescaling of the argument).

In concreto, this then means that for any $f \in L^2([0, 2\pi])$, we have

$$f(x) = \sum_{n\in\mathbb{Z}} a_n \frac{1}{\sqrt{2\pi}} e^{inx} = \frac{a_0}{\sqrt{2\pi}} + \sum_{n=1}^{\infty} (a_n \frac{a_0}{\sqrt{2\pi}} e^{inx} + a_{-n} \frac{a_0}{\sqrt{2\pi}} e^{-inx})$$

where the equality means convergence of the partial sums to f in the L^2-norm.

Our proof consists of the following steps:

1. $\{e_n \mid n \in \mathbb{Z}\}$ is an orthonormal set:

$$\begin{aligned} \int_0^{2\pi} \overline{\frac{1}{\sqrt{2\pi}} e^{imx}} \frac{1}{\sqrt{2\pi}} e^{inx}\, dx &= \frac{1}{2\pi} \int_0^{2\pi} e^{i(n-m)x}\, dx \\ &= \begin{cases} 1 & \text{if } n = m \\ \frac{e^{i(n-m)2\pi} - e^{i(n-m)0}}{i(n-m)} = 0 & \text{if } m \neq n \end{cases} \end{aligned}$$

2. Let P denote the set of all finite linear combinations of functions e_n. These are the *trigonometric polynomials*. According to Theorem 7.11 one of the equivalenty conditions for completeness is that P is dense in $L^2([0, 2\pi])$, i.e., $\overline{P} = [\{e_n\}] = L^2([0, 2\pi])$. By construction, the continuous functions $C([0, 2\pi])$ are dense in $L^2([0, 2\pi])$. Therefore, it is sufficient to show that P is dense in the continous functions with the L^2-norm.

3. We will, in fact, prove a stronger result, namely that P is dense in the *periodic* continuous functions (i.e, continuous f such that $f(0) = f(2\pi)$, or, equivalently, the continuous functions defined on the unit circle, which we denote by $C(T)$), with the *uniform* norm, which is a useful fact in its own right. Compare Theorem 8.1 with the Weierstrass approximation theorem (Theorem 1.11). Both theorems are a special case of a much more general density thereom, called the Stone-Weirestrass Theorem (see W. Rudin, Functional Analysis).

Exercise 8.1 *a) Prove that P is dense in the periodic continuous functions on $[0, 2\pi]$ with the uniform norm, implies P is dense in the space of all continous functions on $[0, 2\pi]$ with the L^2-norm.*
b) Is P dense in the space of all continous functions $[0, 2\pi]$ with the uniform norm?

Theorem 8.1 *The trigonometric polynomials are dense in $(C(T), \|\cdot\|_{\sup})$.*

Proof: We parametrize T by the interval $[0, 2\pi]$ with 0 and 2π identified. So, $C(T)$ is then the space of continuous function $[0, 2\pi] \to \mathbb{C}$, that satisfy $f(0) = f(2\pi)$.

For this proof we introduce the notion of *approximate identity.* We will define a sequence of operators $I_n : C(T) \to P$ such that

$$\lim_{n\to\infty} \|I_n f - f\|_{\sup} = 0, \qquad \text{for all } f \tag{8.2}$$

It is clear that such a sequence of operators approximates the identity operator in the uniform norm on the continuous functions. Obviously, if we can find such a sequence of operators the theorem follows.

We claim that the following I_n satisfy (8.2).

$$I_n f(x) = \int_0^{2\pi} \varphi_n(x - y) f(y)\, dy$$

where

$$\varphi_n(x) = c_n(1 + \cos x)^n$$

where the c_n are positive constants chosen such that

$$\int_0^{2\pi} \varphi_n(x)\, dx = 1, \qquad \text{for all } n \tag{8.3}$$

Note that $\varphi_n \in P$ for all n:

$$\varphi_n(x) = \sum_{k=-n}^{n} a_n e^{inx}$$

First of all, the I_n are well-defined maps $C(T) \to P$. Indeed, for any $f \in C(T)$ we have

$$\begin{aligned} I_n f(x) &= \int_0^{2\pi} \sum_{k=-n}^{n} a_k e^{ik(x-y)} f(y)\, dy \\ &= \sum_{k=-n}^{n} a_k e^{ikx} \int_0^{2\pi} e^{-iky} f(y)\, dy \\ &= \sum_{k=-n}^{n} b_k e^{ikx} \end{aligned}$$

where

$$b_k = a_k \int_0^{2\pi} e^{-iky} f(y)\, dy$$

Next, we have to check the uniform convergence (8.2). This is similar to the proof of Weierstrass' approximation theorem. We use that any continous f on a compact (here T), is bounded and uniformly continuous. We also use (8.3), and note that

$$\int_0^{2\pi} \varphi_n(x-y) f(y)\, dy = \int_0^{2\pi} \varphi_n(y) f(x-y)\, dy$$

Pick $\epsilon > 0$ such that $|x-y| < \epsilon \Rightarrow |f(x) - f(y)| \leq \delta$. Then

$$\begin{aligned} |I_n f(x) - f(x)| &= \left| \int_0^{2\pi} \varphi_n(y)(f(x-y) - f(x))\, dy \right| \\ &= \left| \int_{|y|\leq\epsilon} \varphi_n(y)(f(x-y) - f(x))\, dy \right| \end{aligned}$$

$$+ \left| \int_{|y|>\epsilon} \varphi_n(y)(f(x-y) - f(x))\, dy \right|$$

$$\leq \delta \int_{|y|\leq\epsilon} \varphi_n(y) + 2\|f\|_{\sup} \int_{|y|>\epsilon} \varphi_n(y)\, dy$$

By taking the sup over x, and by using

$$\lim_n \int_{|y|>\epsilon} \varphi_n(y)\, dy = 0, \qquad \text{for all } \epsilon > 0 \tag{8.4}$$

we obtain (8.2). ∎

Exercise 8.2 *Prove (8.4).*

We have proved that for $f \in L^2([0, 2\pi])$

$$\lim_{N\to\infty} \|\frac{1}{2\pi} \sum_{n=-N}^{N} \left[\int_0^{2\pi} e^{-iny} f(y)\, dy\right] e^{inx} - f(x)\|_{L^2} = 0 \tag{8.5}$$

Theorem 7.6 implies that (8.5) is the best possible approximation with a trigonometric polynomial, in the sense of L^2. It is interesting to ask for other kinds of convergence of the partial sums of the series to the function. E.g., it can be proved that they also converge pointwise a.e. to $f(x)$. The proof of this is beyond the scope of this course.

One can also show that for *smooth* f the convergence of the partial sums is *uniform* (see Section 8.2 for a more detailed result. Another interesting question is how the partial sums behave in the neighborhood of a point of discontinuity of f. See H. Dym, and H.P. McKean, *Fourier Series and Integrals*, Section 1.6 for a discussion of this question (the Gibbs Phenomenon).

It is sometimes useful to modify the o.n. basis $\{e_n(x)\}$ a little bit. It is clear that the functions

$$\begin{aligned} e_0(x) &= \frac{1}{\sqrt{2\pi}} \\ \frac{1}{\sqrt{2}}(e_n + e_{-n}) &= \frac{1}{\sqrt{\pi}} \cos(nx), \quad n \geq 1 \\ \frac{1}{i\sqrt{2}}(e_n - e_{-n}) &= \frac{1}{\sqrt{\pi}} \sin(nx), \quad n \geq 1 \end{aligned}$$

together, also form an o.n. basis. This is useful, e.g., because, in this basis, a real-valued function will have real Fourier coeffients. Another useful property of the o.n. basis with cos and sin functions is the following. Consider all functions to be defined on $[-\pi, \pi]$. A function f is called *even* if $f(-x) = f(x)$, for all x, and f is called *odd* if $f(-x) = -f(x)$, for all x. It is then easy to verify that even functions f have an expansion of the form

$$f(x) = \sum_{n=0}^{\infty} a_n \cos(nx)$$

while for odd f

$$f(x) = \sum_{n=1}^{\infty} a_n \sin(nx) \quad .$$

The Fourier coefficients of f are often denoted by $\hat{f}(n)$.

Exercise 8.3 *Prove that the sets $\{e_n \mid n \geq 1\}$ defined by*

$$e_n(x) = \sqrt{\frac{2}{\pi}} \sin nx, \quad 0 \leq x \leq \pi$$

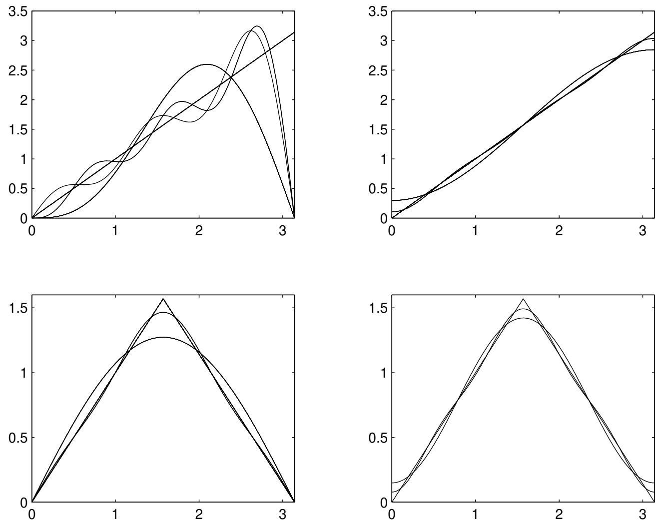

Figure 8.1: Sine and Cosine series for two piecewise linear functions. Note that the difference in the qualtity of the approximations.

and $\{f_n \mid n \geq 0\}$ *defined by*

$$f_0(x) = \sqrt{\frac{1}{\pi}}, \quad f_n(x) = \sqrt{\frac{2}{\pi}} \cos nx, \quad 0 \leq x \leq \pi, n \geq 1,$$

are two o.n. bases of $L^2([0, \pi])$.

8.2 Fourier series of $f \in C^p(T)$

Our aim is to show that if f has p continuous derivatives, with $p = 1, 2, 3, \ldots$, then the partial sums S_N of its Fourier series converge uniformly to f, and to give an estimate of the speed of convergence. We will use the following observation. If $f \in C^1(T)$, we can relate the Fourier coefficients of f' to the ones of f by integration by parts:

$$\begin{aligned} \widehat{f'}(n) &:= \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-inx} f'(x)\, dx \\ &= \frac{1}{\sqrt{2\pi}} [f(2\pi) - f(0)] - \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} (-in) e^{-inx} f(x)\, dx \\ &=: in \hat{f}(n) \end{aligned} \tag{8.6}$$

By induction it follows from from (8.6) that if $f \in C^p(T)$ then $\widehat{f^{(p)}}(n) = (in)^p \hat{f}(n)$.

Note that for relation (8.6) to hold it is, in fact, sufficient that $f' \in L^2(T)$. We will return to this later when we discuss weak derivatives and distributions.

Theorem 8.2 *Let $p \geq 1$. If $f \in C^p(T)$ the partial sums*

$$S_N(x) = \frac{1}{\sqrt{2\pi}} \sum_{n=-N}^{N} \hat{f}(n) e^{inx}$$

satisfy

$$\|S_N - f\|_{\sup} \leq \frac{\|f^{(p)}\|_{L^2}}{\sqrt{2\pi p} N^{p-1/2}}$$

In particular, the partial sums converge uniformly to f.

Proof: Note that $f \in C(T)$ and $S_N \in C(T)$ for all N. Due to the completeness of $C(T)$ for the supremum norm it is, therefore, sufficient to prove that for all $M \geq N$

$$\|S_N - S_M\|_{\sup} \leq \frac{1}{\sqrt{2\pi p}} \frac{1}{N^{p-1/2}} \|f^{(p)}\|_{L^2}$$

For this we use (8.6) and the inequalities of Cauchy-Schwarz and Bessel:

$$\begin{aligned}
\|S_N - S_M\|_{\sup} &\leq \frac{1}{\sqrt{2\pi}} \sum_{N<|n|\leq M} |\hat{f}(n)| \\
&= \frac{1}{\sqrt{2\pi}} \sum_{N<|n|\leq M} |\widehat{f^{(p)}}(n)| \frac{1}{|n|^p} \\
&\leq \frac{1}{\sqrt{2\pi}} \left[\sum_{N<|n|\leq M} |\widehat{f^{(p)}}(n)|^2 \right]^{1/2} \left[\sum_{N<|n|\leq M} \frac{1}{|n|^{2p}} \right]^{1/2} \\
&\leq \frac{1}{\sqrt{2\pi}} \|f^{(p)}\|_{L^2} \times \left[2 \int_N^\infty \frac{1}{r^{2p}}\, dr \right]^{1/2} \\
&\leq \frac{\|f^{(p)}\|_{L^2}}{p\sqrt{2\pi} N^{p-1/2}}
\end{aligned}$$

■

8.3 A Sobolev Embedding Theorem

As an application of the foregoing we discuss the simplest example of a Sobolev Embedding Theorem. Recall the we defined the Sobolev space $H^1(T)$ as the Hilbert space obtained by completion of $C^1(T)$ for the norm derived from the inner product

$$\langle f, g\rangle_{H^1} = \langle f, g\rangle_{L^2} + \langle f', g'\rangle_{L^2}$$

Using the the orthonormality of the Fourier basis and (8.6), we can rewrite this as

$$\langle f, g\rangle_{H^1} = 2\pi \sum_{n=-\infty}^{+\infty} \overline{\hat{f}(n)}\hat{g}(n) + \overline{\widehat{f'}(n)}\widehat{g'}(n)$$

In other words, $H^1(T)$ is also the completion of the trigonometric polynomials for the H^1 norm, and, therefore

$$H^1(T) = \{\sum_{n=-\infty}^{+\infty} a_n e^{inx} \mid \sum_{n=-\infty}^{+\infty} n^2|a_n|^2 < \infty\}$$

More generally, for $p \geq 1$, one defines

$$H^p(T) = \{\sum_{n=-\infty}^{+\infty} a_n e^{inx} \mid \sum_{n=-\infty}^{+\infty} n^{2p}|a_n|^2 < \infty\} \quad .$$

Theorem 8.3 (A Sobolev Embedding Theorem) $H^p(T) \subset C^{p-1}(T)$.

Proof: In the same way as in the proof of Theorem 8.2, one shows that the condition

$$\sum_{n=-\infty}^{+\infty} n^{2p}|a_n|^2 < \infty$$

implies that the partial sums for $f^{(p-1)}$ converge uniformly. And, hence, the limit is continous. ■

8.4 Multiple Fourier series

Let $n \geq 1$, and denote by T^n the $n-$dimensional torus, i.e., the Cartesian product of n circles. The notion of Fourier series is straightforwardly extended to functions of n variables $f : T^n \to \mathbb{C}$, by considering the o.n. basis of $L^2(T^n)$ consisting of the functions

$$e_{\mathbf{n}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{i\mathbf{n}\cdot\mathbf{x}}, \mathbf{n} \in \mathbb{Z}^n$$

Where $\mathbf{x} = (x_1, \ldots, x_n) \in T^n$ and $\mathbf{n} = (n_1, \ldots, n_n) \in \mathbb{Z}^n$, and $\mathbf{n} \cdot \mathbf{x} = n_1 x_1 + \cdots + n_n x_n$.

Exercise 8.4 *Consider $f : T^3 \to \mathbb{C}$ defined by*

$$f(\mathbf{x}) = \sum_{\mathbf{n} \in \mathbb{Z}^3} a_{\mathbf{n}} e^{i\mathbf{n}\cdot\mathbf{x}}$$

Where $\mathbf{x} = (x_1, x_2, x_3)$ and $\mathbf{n} = (n_1, n_2, n_3)$, and $\mathbf{n} \cdot \mathbf{x} = n_1 x_1 + n_2 x_2 + n_3 x_3$. Prove that if, for some $\alpha > 3$,

$$\sum_{\mathbf{n} \in \mathbb{Z}^3} \|n\|^\alpha |a_{\mathbf{n}}|^2 < \infty$$

then, f is continuous.

9 Applications of Fourier series

The applications of Fourier series that we will cover this week are all elementary, but they demonstrate that Fourier series are a very useful tool in the study of a wide variety of concrete problems in engineering and science. There will be some loose ends here and there because we do not discuss conditions for existence and uniqueness of solutions of PDE problems.

9.1 The heat/diffusion equation

This is the original application for which Jean-Baptiste Fourier developed the series expansion named after him, but suggested earlier by Daniel Bernoulli (J.-B. Fourier, *Théorie analytique de la chaleur*, 1822).

9.1.1 Heat conduction in an insulated ring

Consider a thin ring made of a heat conducting material and modelled by the unit circle T. The temperature at time t and position x along the ring is given by $u(x,t)$, which satisfies the *heat equation* given by

$$u_t = u_{xx} \tag{9.1}$$

$$u(x,0) = u_0(x) \tag{9.2}$$

where $u_0(x)$ is a given function describing the initial condition. The same equation also describes diffusion of, e.g., a dye in water or perfume in air. The units of space and time have been chosen such that the thermal conductivity of the material is 1. For physical reasons u_0 is real, bounded and piecewise continuous, and hence $u_0 \in L^2(T)$. From the the heat equation it then follows that $u(\cdot,t) \in L^2(T)$ for all $t \geq 0$. Indeed:

$$\frac{d}{dt}\int_T |u(x,t)|^2\,dx = \int_T 2u(x,t)u_t(x,t)\,dx \tag{9.3}$$

$$= \int_T 2u(x,t)u_{xx}(x,t)\,dx \tag{9.4}$$

$$= -2\int_T |u_x(x,t)|^2\,dx \leq 0 \quad . \tag{9.5}$$

It follows that $u(x,t)$, for all fixed $t \geq 0$, can be expanded as a Fourier series:

$$u(x,t) = \sum_{n=-\infty}^{\infty} u_n(t)e^{inx} \tag{9.6}$$

with

$$u_n(t) = \frac{1}{2\pi}\int_0^{2\pi} u(x,t)e^{-inx}\,dx$$

The coefficients, which still depend on t, can be found by plugging (9.6) into the heat equation:

$$u_t(x,t) = \sum_{n=-\infty}^{\infty} \dot{u}_n\,(t)e^{inx} = \sum_{n=-\infty}^{\infty} -n^2 u_n(t)e^{inx} = u_{xx}(x,t)$$

By taking the inner product of both sides with e^{imt} we find that $u_n(t)$ must satisfy

$$\dot{u}_n = -n^2 u_n \tag{9.7}$$

$$u_n(0) = \frac{1}{2\pi}\int_0^{2\pi} u_0(x)e^{-inx}\,dx \tag{9.8}$$

Thus, the PDE is reduced to an infinite set of decoupled ODE. The solutions are

$$u_n(t) = u_n(0)e^{-n^2 t}$$

Summarizing we obtain that the solution of the heat equation on the unit circle with initial data u_0 is given by

$$u(x,t) = \sum_{n=-\infty}^{\infty} u_n(0)e^{-n^2 t}e^{inx} \tag{9.9}$$

The above arguments are a bit 'formal' if u_0 is not smooth, because the the RHS of the heat equation does not seem to be well-defined in that case. For comfort you can limit the discussion to smooth initial conditions, but this is not necessary, basically due to the first of the following three consequences of (9.9).

Properties the solution $u(x,t)$:

1. *Smoothing property:* For all $t > 0$, $u(x,t) \in C^\infty(T)$. This follows from

$$\sum_{n=-\infty}^{\infty} n^{2p}|u_n(0)e^{-n^2 t}|^2 \le \left[\max_n n^{2p}e^{-n^2 t}\right]\sum_m |u_m(0)|^2 < \infty$$

   for all $p \ge 0$, and the Sobolev Embedding Theorem 8.3.

2. *Irreversibility:* Bad things may happen for $t < 0$, even for C^∞ 'final data'. E.g., $u_0(x) = \sum_n e^{-|n|}e^{inx} \in C^\infty(T)$, but there is no corresponding $u(x,-\epsilon)$ for any $\epsilon > 0$.

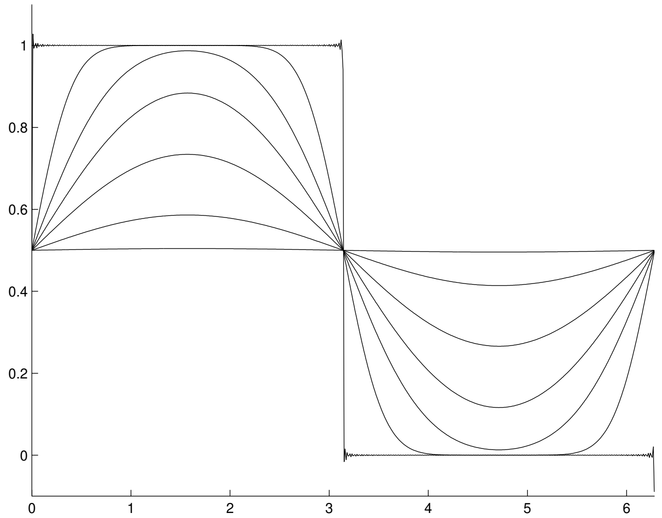

Figure 9.1: Time evolution of the temperature distribution on a ring. The initial distribution is a step function (or rather, a truncated Fourier series approximating a step function). The final distribution is uniform.

3. *Exponential decay and limiting behavior:* The temperature $u(x,t)$ converges exponentially fast to its average value as $t \to \infty$. This is expressed by the following estimate, valid for all $t > \epsilon$, for any $\epsilon > 0$:

$$
\begin{aligned}
& \sup_{x \in T} \left| \sum_{n=-\infty}^{\infty} u_n(0) e^{-n^2 t} e^{inx} - \frac{1}{2\pi} \int_0^{2\pi} u_0(x)\, dx \right| \\
\leq \ & \sum_{n \neq 0} |u_n| e^{-n^2 t} \\
\leq \ & \left[\sum_{n \neq 0} |u_n|^2 \right]^{1/2} \left[2 \sum_{n=1}^{\infty} e^{-2n^2 t} \right]^{1/2} \\
\leq \ & \text{constant} \times \|u_0\|_{L^2} e^{-t}
\end{aligned}
$$

Note that the temperature averaged over the ring is independent of time. This can be seen by computing its derivative with respect to time and using the heat equation:

$$\frac{d}{dt} \int u(x,t)\, dx = \int u_t(x,t)\, dx = \int u_{xx}(x,t)\, dx = 0.$$

9.1.2 The heat equation on an interval

The only difference with the case of a ring is that one needs to specify boundary conditions, e.g., one can fix the temperatures at the end points of the interval.

Exercise 9.1 *Solve the heat equation*

$$\begin{aligned} u_t &= u_{xx} \\ u(x,0) &= u_0(x), \quad 0 \leq x \leq L \\ u(0,t) &= u(L,t) = 0, \quad t > 0 \end{aligned}$$

Hint: Expand the initial data and the solution in a suitable basis of $L^2([0,L])$.

9.1.3 Temperature of the earth as a function of depth

If one is not interested in the daily fluctuations, a crude but reasonable model is the following. Assume that the surface temperature of the earth is a periodic function with the period equal to one year. It is then reasonable to expect that after many years, the temperature at a depth x below the surface will also be a periodic function of time with period one year. For simplicity, choose a time unit such that 1 year $= 2\pi$. Then, for any fixed depth x, the temperature $u(x,t)$ can be expanded in a Fourier series:

$$u(x,t) = \sum_{n=-\infty}^{\infty} u_n(x) e^{int}$$

and, of course, has to satisfy the heat equation. This leads to the following ODE for the coefficients u_n as functions of x ($'$ denotes derivative with respect to x):

$$-inu_n = u_n'' \tag{9.10}$$

Of which the solutions are

$$\begin{aligned} u_n(x) &= u_n(0)e^{\pm\sqrt{n}(1+i)x/\sqrt{2}} \quad \text{if } n \geq 0 \qquad (9.11) \\ &= u_n(0)e^{\pm\sqrt{|n|}(1-i)x/\sqrt{2}} \quad \text{if } n < 0 \qquad (9.12) \end{aligned}$$

The solutions with the $+$ sign in the exponent, should be excluded because they blow up. These solutions can also be removed by imposing any reasonable boundary condition at infinite depth. Note that we are considering the heat equation on the half-line. The general solution is then:

$$u(x,t) = \sum_{n=0}^{\infty} u_n(0)e^{-|n/2|^{1/2}x} e^{i(nt+|n/2|^{1/2}x)} + \sum_{n=-\infty}^{-1} u_n(0)e^{-|n/2|^{1/2}x} e^{i(nt-|n/2|^{1/2}x)} \tag{9.13}$$

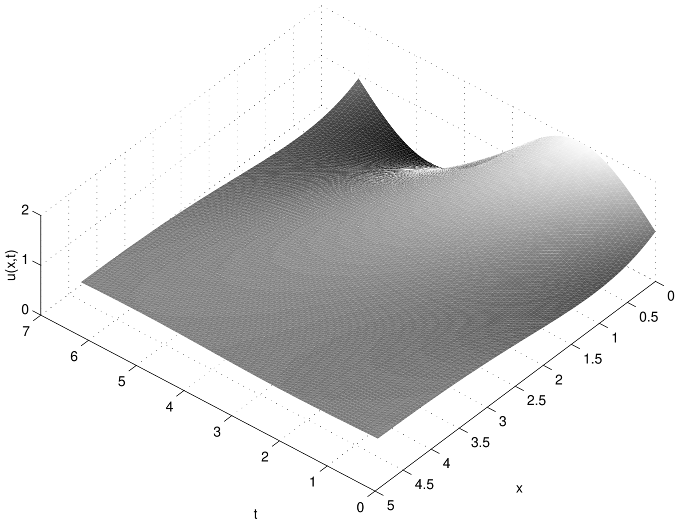

Figure 9.2: The temperature of the earth as a function of time and depth. The time unit is one year. The unti of depth is roughly 1 meter. The temperature unit is arbitary. The maximum and minimum temperature at the surface ($x = 0$) represent the maximum and minimum mean soil temperatures attained during summer and winter respectively.

For concreteness suppose that the surface temperature is given by a simple harmonic function:

$$u(0, t) = A + B \sin t$$

Then, (9.13) becomes

$$u(x, t) = A + Be^{-x/\sqrt{2}} \sin(t - x/\sqrt{2})$$

The exponential damping factor in front of the sine function reduces the fluctuations of the earth temperature below the surface. The argument of the sine function indicates that there is a phase shift for the temperature fluctuations depending on the depth. At a depth x such that $x/\sqrt{2} = \pi$, the fluctuations are reduced by a factor of $e^{-\pi} \sim 0.04$, and are opposite in phase with the surface temperature. For realistic numerical values of the thermal conductivity of the soil this happens for a depth of about 13 feet. So, 13 feet below the surface maximum temperatures are reached in winter and minima in the summer! At the same time the difference between winter and summer temperature is reduced by a factor of about 25 as compared to the average daily temperature at the surface.

9.2 The wave equation: a vibrating string

The one-dimensional wave equation is

$$\frac{1}{c^2}u_{tt} = u_{xx} \tag{9.14}$$

where c is the speed of propagation of the wave (e.g., the speed of sound in a medium, or the speed of light in vacuum).

Note that

$$u(x,t) = \exp(i(kx - \omega t)) \tag{9.15}$$

is a solution of this equation whenever $\omega^2 = c^2k^2$. As this equation is second order in t one should expect that two initial conditions are needed to specify a unique solution. For concreteness, let us consider waves in a vibrating string of length L held fixed at the end points. The following is then a well-posed problem with a unique solution:

$$\frac{1}{c^2}u_{tt} = u_{xx} \tag{9.16}$$

$$u(0,t) = u(L,t) = 0 \tag{9.17}$$

$$u(x,0) = f(x) \tag{9.18}$$

$$u_t(x,0) = g(x) \tag{9.19}$$

where f and g are continuous functions satisfying the boundary conditions. By taking linear combinations of the solutions (9.15) we obtain the following solutions with factorized x and t dependence and which are real for real constants α, β, A, and B:

$$(\alpha \sin kx + \beta \cos kx)\,(A \sin \omega t + B \cos \omega t) \tag{9.20}$$

To satisfy the boundary conditions one needs $\beta = 0$ and

$$k_n = \frac{n\pi}{L}, \quad \omega_n = \frac{cn\pi}{L}, \quad n = 1, 2, 3, \ldots$$

Due to the linearity of the equation all linear combinations of solutions of the form (9.20) will also be solutions. Therefore, we propose a general solution of the form

$$u(x,t) = \sum_{n=1}^{\infty} \sin(k_n x)\,(A_n \sin \omega_n t + B_n \cos \omega_n t)$$

with real constants A_n, B_n.

Exercise 9.2 *Is the set* $\{\sqrt{2/L}\sin(n\pi x/L) \mid n = 1, 2, \ldots\}$ *an o.n. basis for* $L^2([0,L])$*? Why (not)?*

Exercise 9.3 *Find sufficient conditions on the coefficients A_n and B_n such that $u(x,t)$ is continuous as a function of x for all times $t \in \mathbb{R}$.*

Requiring that the solution satisfy the initial conditions (9.18) and (9.19) determines A_n and B_n uniquely:

$$\begin{aligned} B_n &= \frac{2}{L}\int_0^L \sin(n\pi x/L) f(x)\, dx \\ \omega_n A_n &= \frac{2}{L}\int_0^L \sin(n\pi x/L) g(x)\, dx \end{aligned}$$

Note that for the wave equation solutions exist for both $t \geq 0$ and $t < 0$, and that there is no convergence to a stationary solution as $t \to \infty$, as we found for the heat equation.

Exercise 9.4 *Prove the following conservation law for the wave equation:*

$$0 = \frac{d}{dt}\int_0^L |u_t(x,t)|^2 + c^2 |u_x(x,t)|^2\, dx$$

9.3 The Schrödinger equation: a particle in a box

The Schrödinger equation for a particle in a one-dimensional box of size L is a PDE for a complex-valued function $u(x,t)$, $x \in [0,L]$, $t \in \mathbb{R}$, (called the *wavefunction* of the particle, and in the physics literature usually denoted by $\psi(x,t)$). A possible choice of boundary conditions is to require that the wavefunction vanishes at the boundary (*Dirichlet boundary conditions*). So, we are looking for solutions of

$$\begin{aligned} iu_t &= -u_{xx} && (9.21) \\ u(0,t) &= u(L,t) = 0, \quad \text{for all } t && (9.22) \\ u(x,0) &= \psi_0(x) && (9.23) \end{aligned}$$

There are two important conservation laws for this equation. The first one is that the L^2-norm of any solution is independent of time:

$$\begin{aligned} \frac{d}{dt}\int_0^L |u(x,t)|^2\, dx &= \int_0^L \overline{u_t}\, u\, dx + \int_0^L \overline{u}\, u_t\, dx \\ &= \int_0^L \overline{iu_{xx}}\, u\, dx + \int_0^L i\overline{u}\, u_{xx}\, dx \\ &= i\int_0^L \overline{u_x}\, u_x\, dx - i\int_0^L \overline{u_x}\, u_x\, dx = 0 \end{aligned}$$

This property implies that if $\psi_0 \in L^2([0,L])$, $u(\cdot,t) \in L^2([0,L])$, for all t. This property is often called *unitarity*, for reasons that will become clear later. Note that the boundary conditions are essential for this unitarity property.

Exercise 9.5 *Prove unitarity for the Schrödinger equation with Neumann boundary conditions, i.e.,* $u'(0,t) = u'(L,t) = 0$.

The physical interpretation of $|u(x,t)|^2$ is that it is proportional to the probability distribution of the position of the particle at time t. I.e.,

$$\mathrm{prob}\begin{pmatrix}\text{at time } t \text{ the particle is observed}\\ \text{to be in the interval } [a,b]\end{pmatrix} = \frac{\int_a^b |u(x,t)|^2\,dx}{\int_0^L |u(x,t)|^2\,dx}$$

for $0 \le a \le b \le L$. In this interpretation unitarity means conservation of probability, or conservation of the normalization of the probability distribution.

A second important conservation law is conservation of energy given by

$$\frac{d}{dt}\int_0^L |u_x|^2\,dx = 0 \tag{9.24}$$

Exercise 9.6 *Verify (9.24).*

Using Fourier series one can again find the general solutions of (9.21):

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin(k_n x) e^{-i\omega_n t}$$

where

$$k_n = \frac{\pi n}{L}, \quad \omega_n = \left(\frac{\pi n}{L}\right)^2$$

Stationary solutions exhibit the phenomenon called *quantization of energy*. Here, stationary can be defined to mean that $|u(x,t)|^2$ is independent of time, which holds if only one term in (9.3) is non-vanishing. Then, the energy integral equals one of the discrete values ω_n:

$$\frac{\int_0^L |u_x|^2\,dx}{\int_0^L |u|^2\,dx} = \left(\frac{\pi n}{L}\right)^2 = \omega_n$$

The stationary solutions are also the solutions of the following eigenvalue problem:

$$\begin{aligned} -\frac{d^2}{dx^2}u(x) &= \lambda u(x) && (9.25)\\ u(0) &= u(L) = 0 && (9.26)\\ u &\in L^2([0,L]) && (9.27) \end{aligned}$$

More generally, in quantum mechanics stationary solutions are found by solving an (infinite-dimensional) eigenvalue problem of this kind. We will study those in considerable detail later.

9.4 A dispersive wave equation: the linearized KdV equation

In 1895 Korteweg and de Vries introduced a non-linear PDE to describe water waves in shallow channels:

$$u_t = uu_x + u_{xxx} \tag{9.28}$$

There are two remarkable facts about this equation: 1) it has solitary wave solutions (solitons) and 2) in spite of the fact that it is non-linear, it can be solved exactly (by the so-called inverse scattering method poineered by Gardner, Greene, Kruskal, and Miura 1967). This goes beyond simple Fourier series and we do not intend to discuss the inverse scattering method in this course. There is one property, however, that can be illustrated by considering the linearized version of the KdV equation, namely that this equation is *dispersive* (see below). As long as u is small enough, the term uu_x is not expected to influence the solutions very much, and it can be omitted in a first approximation. So, let us consider

$$\begin{aligned} u_t &= u_{xxx} && (9.29)\\ u(x,0) &= u_0(x) && (9.30)\\ u(\cdot,t) &\in L^2(T), \quad \text{for all } t && (9.31) \end{aligned}$$

The solution is

$$u(x,t) = \sum_{n=-\infty}^{\infty} a_n e^{-in^3 t} e^{inx} = \sum_{n=-\infty}^{\infty} a_n e^{in(x-n^2 t)}$$

Notice that the speed of wave propagation depends on n, i.e, on the wavelength. As components in the wave with different wavelengths travel at different speeds, a wavefront is generally spreading out. Hence the name *dispersive* waves. This has to be compared the solutions of

$$\begin{aligned} u_{tt} &= u_{xx} && (9.32)\\ u(x,0) &= u_0(x) && (9.33)\\ u(\cdot,t) &\in L^2(T), \quad \text{for all } t && (9.34) \end{aligned}$$

of which the solution is

$$u(x,t) = \sum_{n=-\infty}^{\infty} a_n^+ e^{in(x+t)} + a_n^- e^{in(x-t)} \tag{9.35}$$

which is non-dispersive.

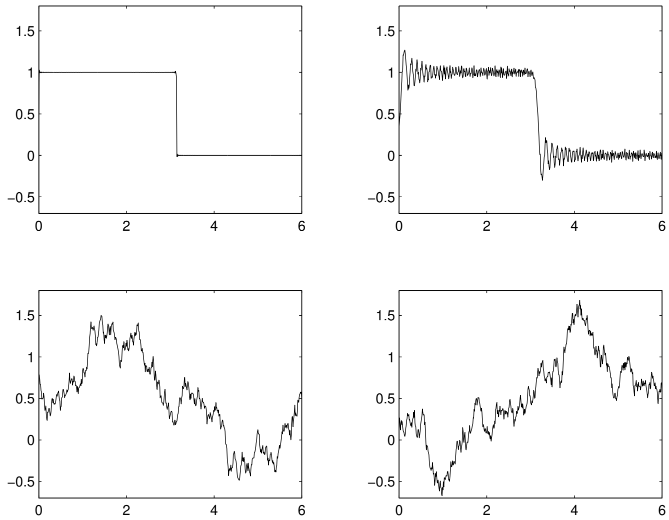

Figure 9.3: The effect of dispersion is illustrated here with the solution (9.35) of the linearized KdV equation on a ring for times $t = 0, e^{-10}, e^{-2}$, and $t = e$. The initial condition is a step function.

9.5 More applications of Fourier Series

9.5.1 The isoperimetric problem

Of all smooth (or piecewise smooth) closed curves in the plane of a given length, say 1, the ones that maximize the area they enclose are the circles. This result is sometimes stated as an inequality: for any smooth closed curve of length L and enclosing an area A one has

$$\frac{A}{L^2} \leq 4\pi \tag{9.36}$$

with equality iff the curve is a circle. This is called (the) an isoperimetric inequality.

A proof using Fourier series is given in the book by Dym and McKean, *Fourier Series and Integrals.* A sketch of the proof is in Naylor and Sell, it Linear Operator Theory in Engineering and Science, Section 5.18, Ex. 6 (pp 329–330).

9.5.2 Ergodicity of arithmetic sequences

Our last application is an ergodic theorem for one of the simplest dynamical systems one can imagine. Let $\gamma \in \mathbb{R}$ and consider the following map on the circle:

$$F : T \to T, \theta \mapsto \theta + 2\pi\gamma \tag{9.37}$$

This map is called the *circle map* or *rotation map.* For every $\theta_0 \in T$ the iterated application of F generates an infinite sequence $(\theta_n = F^n(\theta_0))_{n=0}^{\infty}$ in T.

For any $f \in C(T)$ we can define two averages, a 'time' average, $\langle f\rangle_t$, and a 'phase-space' average, $\langle f\rangle_{ph}$, by

$$\begin{aligned}
\langle f\rangle_t &:= \lim_{N\to\infty} \frac{1}{N+1}\sum_{n=0}^{N} f(\theta_n) \\
\langle f\rangle_{ph} &:= \frac{1}{2\pi}\int_0^{2\pi} f(\theta)\,d\theta
\end{aligned}$$

Theorem 9.1 (Weyl, 1916) *If γ is irrational, then*

$$\langle f\rangle_t = \langle f\rangle_{ph} \tag{9.38}$$

for all $f \in C(T)$ and all $\theta_0 \in T$.

Proof: First, we show the equality (9.38) for the functions $e_m(\theta) = e^{im\theta}$, $m \in \mathbb{Z}$. The case $m = 0$ is trivial: both averages are equal to 1. So, suppose, $m \neq 0$. Then $\langle e_m\rangle_{ph} = 0$. And the time averages can be easily computed as follows:

$$\begin{aligned}
\langle e_m\rangle_t &= \lim_N \frac{1}{N+1}\sum_{n=0}^{N} e^{im(\theta_0+2\pi n\gamma)} \\
&= \lim_N \frac{1}{N+1} e^{im\theta_0} \sum_{n=0}^{N} [e^{2\pi im\gamma)}]^n \\
&= \lim_N \frac{1}{N+1} e^{im\theta_0} \frac{1-[e^{2\pi im\gamma}]^{N+1}}{1-e^{2\pi im\gamma}} \\
&= 0
\end{aligned}$$

where we used the fact that $e^{2\pi im\gamma} \neq 1$ for *irrational* γ.

Both averages are linear in f. So (9.38) also holds for arbitrary linear combinations of the functions e_m, i.e., it holds for the trigonometric polynomials.

Next, we use the fact that the trigonometric polynomials are dense in $C(T)$ with the uniform norm. I.e., for $f \in C(T)$ and any $\epsilon > 0$ there is a trigonometric polynomial P such that $\|P - f\|_{\sup} \leq \epsilon$. Pick such a P. Then

$$\left|\frac{1}{N+1}\sum_{n=0}^{N} f(\theta_n) - \frac{1}{2\pi}\int_0^{2\pi} f(\theta)\,d\theta\right| \leq 2\epsilon + \left|\frac{1}{N+1}\sum_{n=0}^{N} P(\theta_n) - \frac{1}{2\pi}\int_0^{2\pi} P(\theta)\,d\theta\right|$$

By taking $\limsup_N$ we obtain

$$|\langle f\rangle_t - \langle f\rangle_{ph}| \leq 2\epsilon$$

As ϵ is arbitrary this proves (9.38) for all $f \in C(T)$ and $\theta_0 \in T$. ■

Note that for rational γ (9.38) does not hold for most f.

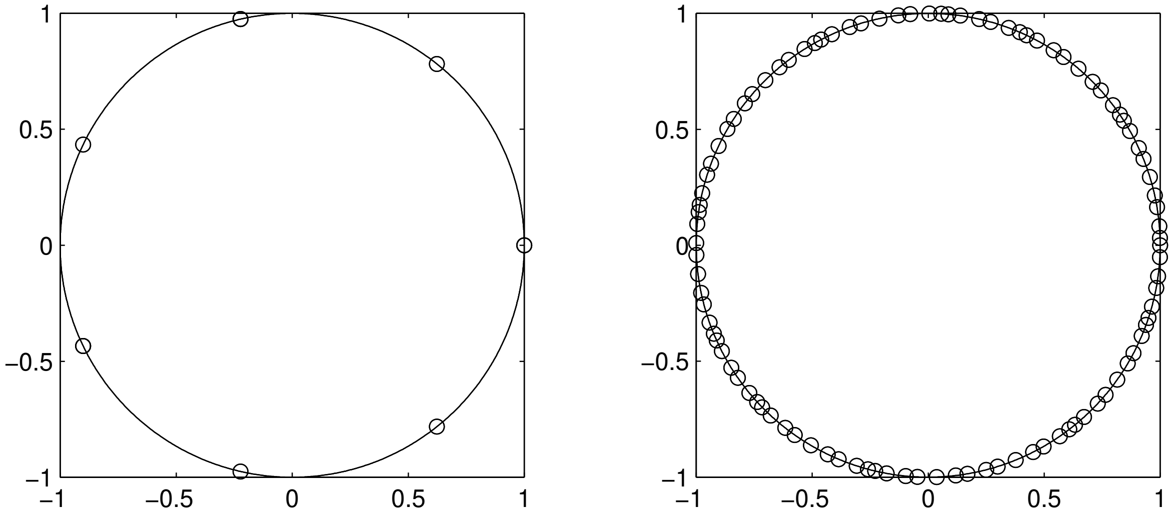

Figure 9.4: The repeated images of the origin under the circle map, $F^n(0)$, $1 \leq 100$. On the left γ is rational: $\gamma = 2/7$. On the right γ is the golden ration, an irational number: $\gamma = (\sqrt{5}-1)/2$.

Corollary 9.2 *For any interval* $[\theta_1, \theta_2] \subset T$*, and irrational* γ*, one has*

$$\lim_{N\to\infty} \frac{1}{N+1} \#\{n \mid 0 \leq n \leq N, \theta_n \in [\theta_1, \theta_2]\} = \frac{|\theta_1 - \theta_2|}{2\pi} \tag{9.39}$$

The relation (9.39) expresses the property that the points of the sequence θ_n are uniformly distributed on the circle.

Exercise 9.7 *As a math writing exercise, write a proof of Corollary 9.2.*

10 Special orthonormal bases 2. Wavelets

In this chapter we will introduce a special class of o.n. bases of $L^2([0,T])$ and $L^2(\mathbb{R})$, called *wavelet* bases. These bases have proved to be very useful in signal analysis, image processing, and data compression. For more information on the mathematics and applications of wavelets, see one or more of the following references:

- For an an elementary introduction, see
  J. Walker, *Fourier Analysis and Wavelet Analysis*, Notices of the AMS, **44** (1997) 658–670, and references therein.

- More examples are in
  G. Strang, *Wavelets and dilation equations: a brief introduction*, SIAM Review, **31** (1989), 614–627.

- This paper has more of the math:
  S. Mallat, *Multiresolution approximations and wavelet orthonormal basis of* $L^2(\mathbb{R})$., Trans. American Math. Soc., **315** (1989), 69–87.

- Read this if you are interested in applications to data compression, image coding etc.
  S. Mallat, *A theory of multiresolution sigma decomposition: the wavelet approximation*, IEEE Trans. Pattern Anal. Machine Intell., **11** (1989), 674–693.

- For a more complete account see the book
  I. Daubechies, *Ten lectures on wavelets*, SIAM, 1992.

- Some routines for numerical implementation are described in
  W.H. Press et al., *Numerical Recipes in C*, Cambridge University Press, Second Edition, 1992

10.1 Fourier series versus wavelets

Fourier series expansions are very effective for representing "nice" periodic functions. For other types of functions, such as functions that vanish on most of their domain, they are not very efficient in the sense that a partial sum with a large number of terms is needed to obtain an good approximation of the function. In signal processing such functions are called *transients*. Recall Parseval's identity, Theorem 7.11 iii):

$$||f||^2 = \sum_n |\hat{f}(n)|^2$$

From this relation we should expect that a large number of terms are needed for a good approximation if the quantity $||f||$ is distributed over a large number of not-so-small coefficients $\hat{f}(n)$. By Theorem 8.2, this must happen, e.g., when f is discontinuous or non-smooth.

Transient signals, supported on a relatively small portion of the relevant time interval, and signals with sharp –almost discontinuous–, transitions are very common. A digitized signal, e.g., may consist of a series of irregularly spaced voltage spikes, or short beeps transmitted by a modem.

Often it is useful to compress a signal before transmission or storage. E.g., to accurately represent a function $f(t), t \in [0, T]$, by storing a finite number of values $f(k\Delta t), k = 0, \ldots, N, N = T/\Delta t$, one needs to choose Δt small enough such that all rapid transitions (or short beeps) can be reconstructed from this list of values. If changes are rapid, Δt will have to be small, and, therefore N will be large. Suppose, however, that we have an o.n. basis of $L^2([0, T])$, with the property that a finite linear combination of basis elements, say with N_0 terms, $N_0 \ll N$, yields a good approximation of the function f. In order to store or transmit the function we can then suffice with N_0 instead of N numbers (the coefficients in the linear combination), without giving up much of the interesting information. Roughly speaking, we have achieved a compression of the data with compression ratio $N : N_0$.

A certain class of o.n. bases, called wavelets, have become very popular for representing signals (functions of time) and images (functions of $x \in I \subset \mathbb{R}^2$) during the last decade, because they allow for large compression ratios. There are different kinds of wavelets, but all share the property that they describe the function at a sequence of different time or length scales. Before we explain precisely what this means, it is instructive to first consider a simple example.

10.2 The Haar wavelets

Consider the Hilbert space $L^2([0, 1])$ with its usual inner product. Later we will also consider $L^2(\mathbb{R})$. Define

$$C_n = \{f \in L^2([0,1]) \mid f \text{ is constant on } [k/2^n, (k+1)/2^n), \text{ for all } k = 0, 1, \ldots, 2^n - 1\} \tag{10.1}$$

Note that for all $n \geq 0$, C_n is a finite-dimensional, and hence closed (Theorem 6.13), linear subspace of $L^2([0, 1])$. Clearly $C_n \subset C_{n+1}$, for all $n \geq 0$. Define $\psi \in L^2([0, 1])$ by

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < 1/2 \\ -1 & \text{if } 1/2 \leq x \leq 1 \end{cases}$$

Let $B_0 = \{1\}$, where 1 denotes the constant function $= 1$. For $n \geq 0$ define

$$B_{n+1} = \{x \mapsto 2^{n/2}\psi(2^n x - k) \mid k = 0, 1, \ldots, 2^n - 1\} \tag{10.2}$$

Note that $B_n \subset C_n$.

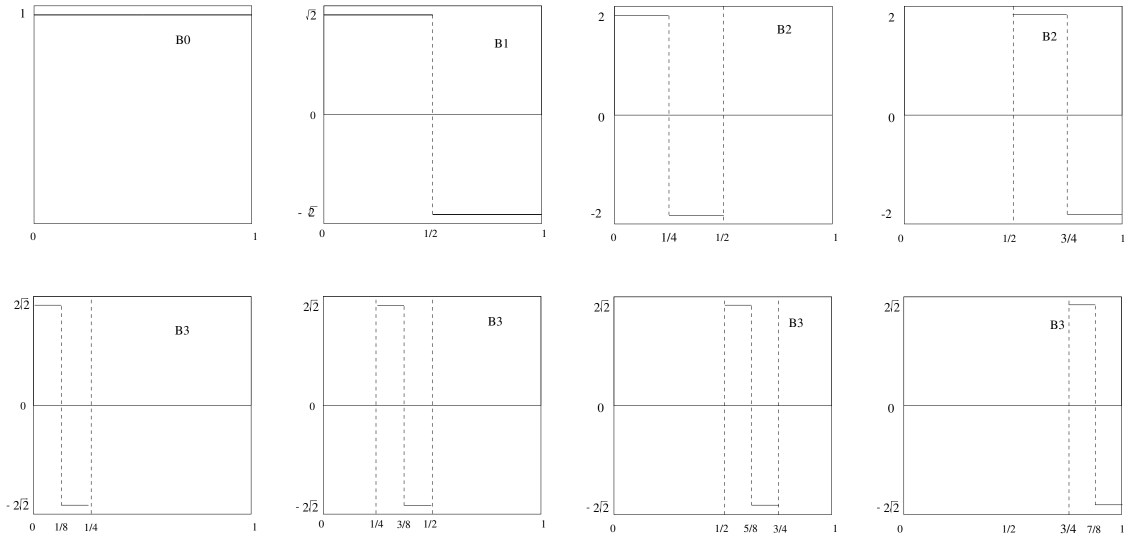

Figure 10.1: Some members of the sets of functions B_n defined in (10.2).

Exercise 10.1 *Prove that* $\cup_{n=0}^{N} B_n$ *is an o.n. basis of* C_N. *(Hint: prove that the set is orthonormal and count its elements.)*

Lemma 10.1 $\mathcal{B} := \cup_{n=0}^{\infty} B_n$ *is an o.n. basis of* $L^2([0,1])$

Proof: From the previous exercise we already know that $\mathcal{B}$ is an o.n. set. As the continuous functions are dense in $L^2([0,1])$, it is sufficient to show that for every $f \in C([0,1])$, and $\epsilon > 0$, there is a finite linear combination of elements in $\mathcal{B}$, i.e., an element g in one of the spaces C_n, such that $\|f - g\|_{L^2} \leq \epsilon$. Actually we will prove the stronger statement that there is $g \in C_N$, for a suitable N, such that $\|f - g\|_{\sup} \leq \epsilon$.

By the uniform continuity of f (Theorem 2.11), there exists N such that for all $x, y \in [0,1]$, with $|x - y| \leq 2^{-N}$, one has $|f(x) - f(y)| \leq \epsilon$. It follows that $g \in C_N$ defined by

$$g(x) = \sum_{k=0}^{2^N - 1} f(k/2^N) \mathrm{Ind}_{[k/2^N, (k+1)/2^N)}(x)$$

satisfies the desired bound. ■

Exercise 10.2 *Suppose* $\mathcal{B} = \{e_n\}_{n=1}^{\infty}$ *is an o.n. basis for* $L^2([0,1])$. *Prove the following:*

a) *For any* $a \in \mathbb{R}$, $\mathcal{B}_a = \{x \mapsto e_n(x - a)\}_{n=1}^{\infty}$ *is an o.n. basis for* $L^2([a, a+1])$.

b) *For any* $c > 0$, $\mathcal{B}^c = \{x \mapsto \sqrt{c} e_n(cx)\}_{n=1}^{\infty}$ *is an o.n. basis for* $L^2([0, c^{-1}])$.

c) *With the convention that functions are extended to a larger than their original domain by putting them equal to 0, prove that* $\mathcal{B} \cup \mathcal{B}_1$ *is an o.n. basis for* $L^2([0,2])$*.*
d) *Prove that* $\cup_{k\in\mathbb{Z}}\mathcal{B}_k$ *is an o.n. basis for* $L^2(\mathbb{R})$*.*

The basis $\mathcal{B} = \cup_{k\in\mathbb{Z}}\mathcal{B}_k$ is called the *Haar wavelet basis* of $L^2(\mathbb{R})$, and its elements are called the *Haar wavelets* . The Haar wavelets are very simple functions, step-functions each taking only three different values. They have compact support and have a simple scaling behavior, as illustrated in the following exercise.

Exercise 10.3 *With* $\mathcal{B}$ *as in Lemma 10.1 and with the notations and conventions of the previous exercise, show that* $\mathcal{B}^{1/2}$ *considered as an o.n. basis of* $L^2([0,2])$ *is equal to* $B \cup B_1$*.*

These properties, compact support and scaling, make the Haar wavelets especially suitable for decomposition of functions with a compact support that is small compared to their domain. But as the Haar wavelets are not continuous, partial sums approximating a continuous function are not continuous. It is often desirable to have continuous approximations of continuous functions, or C^p approximations of C^p functions. This is one motivation for introducing other wavelet basis.

10.3 Multiresolution analysis

In this section we want to illustrate how other wavelet bases can be constructed. We will mention the Daub4 wavelet basis, due to Ingrid Daubechies (1988).

We will consider wavelet bases for $L^2(\mathbb{R})$. Multiresolution analysis refers to the following structure. There is a family of closed linear subspaces of $L^2(\mathbb{R})$, indexed by $n \in \mathbb{Z}$, and denoted by $\{C_n\}$, and a function $\varphi \in L^2(\mathbb{R}) \cap L^1(\mathbb{R})$ such that the following properties hold:

MRA1: Scaling. $x \mapsto f(x) \in C_n$ iff $x \mapsto f(2x) \in C_{n+1}$, for all $n \in \mathbb{Z}$.

MRA2: Inclusion. $C_n \subset C_{n+1}$, for all $n \in \mathbb{Z}$.

MRA3: Density. $\overline{\bigcup_{n\in\mathbb{Z}} C_n} = L^2(\mathbb{R})$.

MRA4: Maximality. $\bigcap_{n\in\mathbb{Z}} V_n = \{0\}$.

MRA5: Basis. $\{\varphi(x-k)\}_{k\in\mathbb{Z}}$ is an o.n. basis of C_0.

These properties are the "axioms of multiresolution analysis". Note that the spaces C_n defined in (10.1) (extend the definition to all integers n), and the function $\varphi(x) = \mathrm{Ind}_{[0,1)}(x)$, satisfy MRA1-5.

Next, we show how a wavelet basis can be obtained from this structure. For the special case described in the previous paragraph the procedure will reproduce the Haar wavelets. But, as we will explain, there are other possibilities. In particular, there we will see how to construct φ that lead to an orthonormal wavelet basis consisting of continuous functions with compact support.

From MRA1 and MRA5 it follows that

$$B_n := \{2^{n/2}\varphi(2^n x - k) \mid k \in \mathbb{Z}\}$$

is an o.n. basis of C_n, for all $n \in \mathbb{Z}$. Of course, $\bigcup_n B_n$ is not a linearly independent set, due to MRA2. For this reason, we define closed linear subspaces of $L^2(\mathbb{R})$ by

$$C_{n+1} = C_n \oplus W_n \tag{10.3}$$

The W_n are called the wavelet subspaces and using MRA2-4, one can easily prove that

$$\bigoplus_{n\in\mathbb{Z}} W_n := \overline{\bigcup_{N\geq 1} \bigoplus_{-N\leq n\leq N} W_n} = L^2(\mathbb{R}) \tag{10.4}$$

Recall that in our discussion of the Haar basis we used the function ψ given by

$$\psi(x) = \varphi(2x) - \varphi(2x-1) \tag{10.5}$$

And one also has

$$\varphi(x) = \varphi(2x) + \varphi(2x-1) \tag{10.6}$$

We will generalize this relation between φ and its rescaled version and between φ and ψ later. For now, just observe the following properties: $x \mapsto \psi(x-k) \in C_1$ and $\perp C_0$; hence, these functions are in W_0. We already know that $\{\varphi(2x-2k) - \varphi(2x-(2k+1)) \mid k \in \mathbb{Z}\} \cup \{\varphi(2x-2k) + \varphi(2x-(2k+1)) \mid k \in \mathbb{Z}\}$ is an o.n. basis of C_1. From (10.6) we then see that $\{\varphi(2x-2k) + \varphi(2x-(2k+1)) \mid k \in \mathbb{Z}\}$ is an o.n. basis of C_0. From this and (10.5) it follows that $\{x \mapsto 2^{n/2}\psi(2^n x - k) \mid k \in \mathbb{Z}\}$ is an o.n. basis of W_0. By combining this with MRA1 we obtain the following lemma.

Lemma 10.2
i) $\{x \mapsto 2^{n/2}\psi(2^n x - k) \mid k \in \mathbb{Z}\}$ *is an o.n. basis for* W_n.
ii) $\{x \mapsto 2^{n/2}\psi(2^n x - k) \mid n, k \in \mathbb{Z}\}$ *is an o.n. basis for* $L^2(\mathbb{R})$.

So far, we have just rederived the Haar wavelet basis. The spaces W_n are also useful for investigating the practical issues involved with the Fast Haar Transform, which is important for numerical applications. See, e.g., the paper by Walker for a discussion.

Next, we explain how the same structure leads to other wavelet bases, in particular the continuous Daub4 wavelets. To this end we will make some further assumptions about the functions φ and ψ (which are satisfied in the Haar case).

Suppose we have real-valued functions φ and ψ with the following properties:

1. Normalization properties

$$\int_{\mathbb{R}} \varphi(x)\, dx = 1 \int_{\mathbb{R}} \varphi(x)^2\, dx = 1, \int_{\mathbb{R}} \psi(x)^2\, dx = 1 \quad .$$

2. The functions $\{x \mapsto \varphi(x-k) \mid k \in \mathbb{Z}\}$ are an orthonormal set.

3. Orthogonality of φ and ψ:

$$\int_{\mathbb{R}} \varphi(x)\psi(x)\, dx = 0$$

4. l vanishing moments of ψ:

$$\int_{\mathbb{R}} x^j \psi(x)\, dx = 0, \quad 0 \leq j \leq l-1$$

Define

$$C_0 = [\{\varphi(\cdot - k)\}_{k\in\mathbb{Z}}]$$

where, as before $[\cdot]$ denotes the closed linear span. Suppose that $\{C_n\}_{n\in\mathbb{Z}}$ and φ satisfy MRA1-5. Then, by MRA1,

$$C_n = [\{x \mapsto 2^{n/2}\varphi(2^n x - k)\}_{k\in\mathbb{Z}}\}]$$

From MRA2 we see that $\varphi \in C_1$, and, therefore, φ can be expanded in the basis of rescaled and shifted functions as follows

$$\varphi(x) = \sum_{k\in\mathbb{Z}} a_k \varphi(2x - k) \tag{10.7}$$

In order to facilitate finding examples of useful pairs of functions φ and ψ, we propose the following form for ψ:

$$\varphi(x) = \sum_{k\in\mathbb{Z}} (-1)^k a_{1-k} \varphi(2x - k) \tag{10.8}$$

From the orthogonality property 3 we then derive the following equation for the coefficients a_k:

$$\sum_{k\in\mathbb{Z}} (-1)^k a_{1-k} a_k = 0$$

The normalization properties 1) translate into

$$\sum_{k\in\mathbb{Z}} a_k = 2; \quad \sum_{k\in\mathbb{Z}} a_k^2 = 2$$

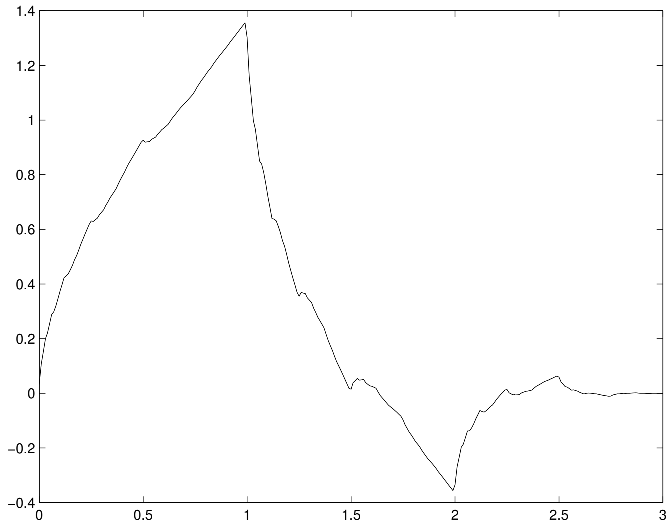

Figure 10.2: The function ψ for the Daub4 wavelets.

Note that the Haar case is a solution with $l = 1, \varphi = \mathrm{Ind}_{[0,1)}$, and $a_0 = a_1 = 1, a_k = 0$, for $k \neq 0, 1$. Daubechies's solution with $l - 2$ is determined by

$$
\begin{aligned}
a_0 &= \frac{1}{4}(1+\sqrt{3}) \quad , \quad a_1 = \frac{1}{4}(3+\sqrt{3}) \\
a_2 &= \frac{1}{4}(3-\sqrt{3}) \quad , \quad a_3 = \frac{1}{4}(1-\sqrt{3}) \\
a_k &= 0 \qquad \text{otherwise}
\end{aligned}
$$

The corresponding φ is determined by the equation (10.7), with the specified values for the coefficients a_k. One can find φ by solving (10.7) iteratively (regard it as a fixed point equation):

$$
\begin{aligned}
\varphi_{n+1}(x) &= \sum_{k\in\mathbb{Z}} a_k \varphi_n(2x-k), n \geq 0 \\
\varphi_0(x) &= \mathrm{Ind}_{[0,1)}(x)
\end{aligned}
$$

In fact, (10.7) is an eigenvalue equation with eigenvalue 1 and a rather complicated linear operator which is combination of translations of the rescaled function. One can show that φ_n converges to a continuous function with support in $[0, 3]$. Clearly, ψ, defined by (10.8), is then also continuous and with support contained in $[0, 3]$. These functions are not differentiable however. Their derivatives have an infinite number of discontinuities, as can be guessed from their graphs (see Figure 10.2).

11 Linear operators on a Hilbert space

Here, we introduce some special classes of bounded linear operators on a Hilbert space $\mathcal{H}$ or from a Hilbert space $\mathcal{H}_1$ to another Hilbert space $\mathcal{H}_2$.

11.1 Unitary operators

Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces. A linear map $U : \mathcal{H}_1 \to \mathcal{H}_2$ is called *unitary* if it is invertible and

$$\langle Ux, Uy\rangle_{\mathcal{H}_2} = \langle x, y\rangle_{\mathcal{H}_1},$$

i.e., U is one-to-one and preserves the inner product. Two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ are said to be *isomorphic* (as Hilbert spaces) if there exists a unitary linear map between them.

Exercise 11.1 a) *Show that a linear operator $U : \mathcal{H}_1 \to \mathcal{H}_2$ is unitary iff it is an isometric isomorphism of normed linear spaces.*
b) *Show that an invertible linear map is unitary iff its inverse is.*

Examples:

1. Let $\{u_\alpha\}$ be an o.n. basis of $\mathcal{H}_1$ and $\{v_\alpha\}$ an o.n. basis of $\mathcal{H}_2$. Define $U : \mathcal{H}_1 \to \mathcal{H}_2$ as follows. Any $x \in \mathcal{H}_1$ can be written uniquely as

$$x = \sum_\alpha c_\alpha u_\alpha$$

   with coefficients $c_\alpha \in \mathbb{C}$. Define Ux by

$$Ux = \sum_\alpha c_\alpha v_\alpha$$

   Then U is a linear unitary map.

2. Let $\{u_\alpha\}$ be an o.n. basis of $\mathcal{H}$ and for all α pick $\lambda_\alpha \in \mathbb{C}$ with $|\lambda_\alpha| = 1$. Then

$$Ux = \sum_\alpha \lambda_\alpha \langle u_\alpha, x\rangle u_\alpha$$

   is unitary.

3. Consider $\mathcal{H}_1 = L^2(T)$ and $\mathcal{H}_2 = l^2(\mathbb{Z})$. Then

$$Uf = (a_n)$$

where (a_n) is the sequence of Fourier coefficients of f, i.e.,

$$a_n = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} e^{-inx} f(x)\, dx,$$

is unitary.

4. For $a \in T$ and $f \in L^2(T)$ define

$$U_a f(x) = f(x-a)$$

Then U_a ia unitary. The set of all such U_a form a representation of the additive group $\mathbb{R}/(2\pi)\mathbb{Z}$:

$$U_{a+b} = U_a U_b$$

5. Let $\mathcal{H} = \mathbb{C}^n$. A complex $n \times n$ matrix represents a unitary linear map $\mathcal{H} \to \mathcal{H}$ iff $U^*U = \mathbb{1}$, where U^* is the hermitian conjugate of U.

6. Let $\mathcal{H} = \mathbb{C}^n$. For any hermitian $n \times n$ matrix A the matrix $U = \exp(iA)$ represents a unitary linear map $\mathcal{H} \to \mathcal{H}$, where, for any matrix B

$$\exp B = \sum_{n=0}^{\infty} \frac{1}{n!} B^n$$

Exercise 11.2 *Let $U : \mathcal{H}_1 \to \mathcal{H}_2$ be a unitary linear map, and $\{u_\alpha\}$ an o.n. basis of $\mathcal{H}_1$. Show that $\{v_\alpha = Uu_\alpha\}$ is an o.n. basis of $\mathcal{H}_2$.*

11.2 Orthogonal Projections

Let $\mathcal{H}$ be a Hilbert space and $\mathcal{H}_1$ and $\mathcal{H}_2$ be two subspaces of $\mathcal{H}$. If $\mathcal{H}_1$ and $\mathcal{H}_2$ are mutually orthogonal and such that for all $x \in \mathcal{H}$, there are unique vectors $x_1 \in \mathcal{H}_1$ and $x_2 \in \mathcal{H}_2$, such that $x = x_1 + x_2$, then we say that $\mathcal{H}$ is the direct sum of $\mathcal{H}_1$ and $\mathcal{H}_2$. The standard notation for this situation is

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2 \tag{11.1}$$

and one writes $x = x_1 \oplus x_2$, where x_1 and x_2 are the unique vectors in $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, for which $x = x_1 + x_2$. It follows that $\mathcal{H}_1^\perp = \mathcal{H}_2$ and $\mathcal{H}_2^\perp = \mathcal{H}_1$, and hence, by Theorem 1.5, $\mathcal{H}_1$ and $\mathcal{H}_2$ are closed. (11.1) is called an *orthogonal decomposition* of $\mathcal{H}$. Similarly, for any two closed and mutually orthogonal subspaces $\mathcal{H}_1$ and $\mathcal{H}_2$ of a Hilbert space $\mathcal{H}$, one has that

$$\mathcal{H}_1 + \mathcal{H}_2 = \mathcal{H}_1 \oplus \mathcal{H}_2$$

is a closed linear subspace of $\mathcal{H}$. The same notation is used for the construction that defines a new Hilbert space $\mathcal{H}_1 \oplus \mathcal{H}_2$ for any two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$. $\mathcal{H}_1 \oplus \mathcal{H}_2$ is then called the *orthogonal sum* or *direct sum* of $\mathcal{H}_1$ and $\mathcal{H}_2$.

For every closed linear subspace $\mathcal{H}_1$ of $\mathcal{H}$ there is an orthogonal decomposition of $\mathcal{H}$. It suffices to take $\mathcal{H}_2 = \mathcal{H}_1^\perp$. This is a consequence of Theorem 7.6.

A linear map $P : \mathcal{H} \to \mathcal{H}$, satisfying $P^2 = P$ and $\langle Px, y\rangle = \langle x, Py\rangle$, is called an *orthogonal projection* on $\mathcal{H}$..

Next, we show that orthogonal projections P are bounded, and that $\|P\| = 1$ except for $P = 0$. If $P \neq 0$, there is $x \in \mathcal{H}$ such that $Px \neq 0$, and we have

$$\begin{aligned} \|Px\| &= \langle Px, Px\rangle/\|Px\| = \langle x, P^2x/\|Px\|\rangle \\ &= \langle x, Px/\|Px\|\rangle \leq \|x\| \left\| \frac{Px}{\|Px\|} \right\| = \|x\| \end{aligned}$$

Therefore, we always have $\|P\| \leq 1$. In order to prove the other inequality we start from $\|P(Px)\| = \|Px\|$. Therefore, $\|P\| \geq 1$. Hence, $\|P\| = 1$, if $P \neq 0$.

For a linear map $A : \mathcal{H} \to \mathcal{H}$ the *range* of A, $\mathrm{ran}A$, is defined by

$$\mathrm{ran}A = \{y \in \mathcal{H} \mid \text{ there exists } x \in \mathcal{H} \text{ such that } y = Ax\}$$

Exercise 11.3 *Prove that if P is an orthogonal projection, then $x \in \mathrm{ran}P$ iff $Px = x$. (Hint: use $P^2 = P$).*

Exercise 11.4 *Check that if P is an orthogonal projection,* $\mathrm{ran}P = (\ker P)^\perp$. *(Hint prove* $\mathrm{ran}P \subset (\ker P)^\perp$ *and* $\ker P \subset (\mathrm{ran}P)^\perp$*.)*

Lemma 11.1 *If P is an orthogonal projection on $\mathcal{H}$, then* $\mathrm{ran}P$ *is a closed linear subspace of $\mathcal{H}$.*

Proof: That $\mathrm{ran}P$ is a linear subspace follows from the linearity of P. That $\mathrm{ran}P$ is closed follows from Exercise 11.4 and Theorem 7.5. ■

Next, we show that there is a one-to-one correspondence between orthogonal projections and closed linear subspaces, or, equivalently, between orthogonal projections and orthogonal decompositions of $\mathcal{H}$.

Theorem 11.2 *i) Let P be any orthogonal projection on $\mathcal{H}$. Then*

$$\mathcal{H} = \mathrm{ran}P \oplus \ker P$$

ii) If $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$, then

$$Px = x_1, \quad \text{for } x = x_1 \oplus x_2$$

defines on orthogonal projection on $\mathcal{H}$ with $\mathrm{ran}P = \mathcal{H}_1$ *and* $\ker P = \mathcal{H}_2$.

Proof:
Proof of i): This is a direct consequence from the fact that $\ker P$ is a closed linear subspace and Exercise 11.4.
Proof of ii): The only non-trivial property to check is that P satisfies $\langle Px, y \rangle = \langle x, Py \rangle$, for all $x, y \in \mathcal{H}$. This follows from the linearity of the inner product and the orthogonality of the decomposition $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$. ■

Examples:

1. One-dimensional projections. For any vector $u \in \mathcal{H}$, with $\|u\| = 1$, P_u defined by
$$P_u x = \langle u, x \rangle u$$
is an orthogonal projection. E.g., let $\mathcal{H} = l^2(\mathbb{Z})$, and $u = e_n$, where for any $n \in \mathbb{Z}$
$$e_n = (\delta_{k,n})_{k=-\infty}^{\infty}$$
Then, for $x = (x_k)$,
$$P_{e_n} x = x_n e_n$$
Physicists like to denote P_u by $|u\rangle\langle u|$.

2. Let $\mathcal{H} = L^2(\mathbb{R})$ and $[a, b] \subset \mathbb{R}$. Then
$$P_{[a,b]} f(x) = \chi_{[a,b]}(x) f(x)$$
is an orthogonal projection. Here $\chi_{[a,b]}$ is the *characteristic function* or *indicator function* of the interval $[a, b]$. For any subset A of a set X, $\chi_A : X \to \{0, 1\}$ defined by
$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

3. If P is an orthogonal projection on $\mathcal{H}$, then so is $Q = \mathbb{1} - P$. Note that $\ker Q = \operatorname{ran} P$ and $\operatorname{ran} Q = \ker P$.

4. Let $\mathcal{H} = L^2(T)$. Then
$$Pf = f - \frac{1}{2\pi} \int_0^{2\pi} f(x)\, dx$$
defines an orthogonal projection on $\mathcal{H}$.

5. If U is unitary and such that $U^2 = 1$, then $P = (\mathbb{1} + U)/2$, is an orthogonal projection.

Exercise 11.5 *Let $\{u_\alpha\}$ be an orthogonal set of vectors in a Hilbert space $\mathcal{H}$. Prove that $\{u_\alpha\}$ is an o.n. basis iff*

$$\sum_\alpha |u_\alpha\rangle\langle u_\alpha| = \mathbb{1}$$

One way in which general Banach spaces differ from Hilbert spaces is that there are, in general, not enough continuous projections. See Section 50 in Simmons for a discussion.

12 Linear functionals on a Hilbert space

12.1 The dual of a Hilbert space

A *linear functional* on a complex Hilbert space $\mathcal{H}$ is a linear map $\mathcal{H} \to \mathbb{C}$. As linear functionals are a special case of linear maps between Banach spaces, we have that a linear functional φ is continuous iff it is bounded, i.e., iff there exists a constant M such that

$$|\varphi(x)| \leq M\|x\|, \quad \text{for all } x \in \mathcal{H} \tag{12.1}$$

The smallest constant for which this inequality holds is then the norm of the functional φ, i.e.,

$$\|\varphi\| = \sup\{|\varphi(x)| \mid x \in \mathcal{H}, \|x\| = 1\} \tag{12.2}$$

Example:
For any $y \in \mathcal{H}$,

$$\varphi_y(x) = \langle y, x\rangle \tag{12.3}$$

defines a bounded linear functional on $\mathcal{H}$ with norm $\|\varphi_y\| = \|y\|$. This equality for the norm is a consequence of the following lemma.

Lemma 12.1 *For any $x \in \mathcal{H}$ one has*

$$\|x\| = \sup\{|\langle x, y\rangle| \mid y \in \mathcal{H}, \|y\| = 1\} \tag{12.4}$$

Proof: For $x = 0$ the equality is obvious. So suppose $x \neq 0$. By taking $y = x/\|x\|$, we see that $\|x\| \leq \sup\{\cdots\}$. On the other hand the inner product satisfies $|\langle x, y\rangle| \leq \|x\|\|y\|$. Hence, by taking $\|y\| = 1$, we also have $\sup\{\cdots\} \leq \|x\|$. This proves (12.4). ■

Example:
Let $\mathcal{H} = L^2([0, \pi])$. Then, for each $n = 1, 2, \ldots$ the maps a_n defined by

$$a_n(f) = \sqrt{2/\pi} \int_0^\pi \sin(nx) f(x)\, dx$$

that associate to any function its n^{th} Fourier coefficient, are continuous linear functionals on $L^2([0, \pi])$.

The space of all bounded ($\equiv$ continuous) linear functionals on a Banach space X (e.g., a Hilbert space), X^*, is again a Banach space with the linear structure determined by

$$(\varphi + \lambda\xi)(x) = \varphi(x) + \lambda\xi(x), \quad \text{for all } \varphi, \eta \in X^*, \lambda \in \mathbb{C}, x \in \mathcal{H}$$

and the norm defined by (12.2) (see Theorem 6.6). X^* is called the *dual* (space) of X.

The following theorem tells us that the examples of bounded linear functionals on a Hilbert space determined by a vector $y \in \mathcal{H}$ as in (12.3), are in fact *all* bounded linear functionals on $\mathcal{H}$.

Theorem 12.2 (The Riesz Representation Theorem) *Let φ be a bounded linear functional on a Hilbert space $\mathcal{H}$, i.e., $\varphi \in \mathcal{H}^*$. Then there is a unique vector $y \in \mathcal{H}$ such that*

$$\varphi(x) = \langle y, x \rangle \tag{12.5}$$

Proof: We already know that $\varphi_y(x) = \langle y, x \rangle$ defines a bounded linear functional on $\mathcal{H}$. It is also easy to see that if $y_1 \neq y_2$, then $\varphi_{y_1} \neq \varphi_{y_2}$. Indeed, if $\varphi_{y_1}(x) = \varphi_{y_2}(x)$, for all $x \in \mathcal{H}$, then $\langle y_1 - y_2, x \rangle = 0$, for all x, and in particular for $x = y_1 - y_2$. This implies $\|y_1 - y_2\| = 0$, and hence, $y_1 = y_2$.

The non-trivial part of the proof of this theorem is to show that any functional φ is of the form (12.5) for some $y \in \mathcal{H}$. If $\varphi = 0$, we can take $y = 0$. So, suppose $\varphi \neq 0$. Then there exists $z \in \mathcal{H}$ such that $z \perp \ker \varphi$. Pick such a vector z. We claim that there is a constant $\lambda \in \mathbb{C}$ such that

$$\varphi(x) = \langle \lambda z, x \rangle$$

λ is then necessarily given by

$$\lambda = \frac{\overline{\varphi(z)}}{\|z\|^2} \quad .$$

In order to prove this consider the linear map P defined by

$$Px = \frac{\varphi(x)}{\varphi(z)}\, z$$

It is then straightforward to check that P is linear, and that $P^2 = P$. Then $\mathrm{ran} P = \{\alpha z \mid \alpha \in \mathbb{C}\}$, and $\ker P = \ker \varphi$, and in particular $\mathrm{ran} P \perp \ker P$, from which it follows that P is an orthogonal projection ($x = Px + (x - Px) \Rightarrow \mathcal{H} = \mathrm{ran} P \oplus \ker P$, and from this it is easy to check that $\langle Px, y \rangle = \langle x, Py \rangle$, for all $x, y \in \mathcal{H}$). So, we have proved that

$$\mathcal{H} = \ker \varphi \oplus \{\alpha z \mid \alpha \in \mathbb{C}\}$$

and for any $x = x_1 + \alpha z$, with $x_1 \in \ker \varphi$ we have

$$\varphi(x) = \alpha \varphi(z) = \langle \frac{\overline{\varphi(z)}}{\|z\|^2} z, z \rangle = \langle \lambda z, x \rangle$$

■

One says that Hilbert spaces are *selfdual*: $\mathcal{H}^*$ is (anti)isomorphic to $\mathcal{H}$; $\mathcal{H}$ and $\mathcal{H}^*$ are isomorphic as Banach spaces, and anti-isomorphic as Hilbert spaces: the map $y \mapsto \varphi_y$ is one-to-one and onto, and it preserves the norm, but it is anti-linear: $\varphi_{\lambda y} = \overline{\lambda}\varphi_y$.

Later we will consider the duals of other Banach spaces (e.g., L^p-spaces, and $C([a, b])$), and we will see that a Banach space is in general quite different from its dual. Hilbert spaces are special in this respect.

Exercise 12.1 *Prove that $\mathcal{H}^*$ is a Hilbert space with the inner product defined by*

$$\langle \varphi_x, \varphi_y \rangle_{\mathcal{H}^*} = \langle y, x \rangle_{\mathcal{H}}$$

Exercise 12.2 *Let $A \subset \mathcal{H}$ be such that*

$$\mathcal{H}_0 = \{x \in \mathcal{H} \mid x \textit{ is a finite linear combination of elements in } A\},$$

is a dense linear subspace of $\mathcal{H}$. Prove that any bounded linear functional on $\mathcal{H}$ is uniquely determined by its values on A. In other words, show that if $\varphi, \xi \in \mathcal{H}^*$ such that $\varphi(a) = \xi(a)$, for all $a \in A$, then $\varphi = \xi$.

Note that that any o.n. basis $\{u_\alpha\}$ of $\mathcal{H}$ has the property required of the set A in the previous exercise. Hence, bounded linear functionals are uniquely determined by their values on an o.n. basis. What is a necessary and sufficient condition on a family of complex numbers c_α for there to be a bounded linear functional φ such that $\varphi(u_\alpha) = c_\alpha$?

Exercise 12.3 *Let $\{u_\alpha\}$ be an o.n. basis of $\mathcal{H}$. prove that $\{\varphi_{u_\alpha}\}$ is an o.n. basis of $\mathcal{H}^*$.*

Example:
A *state* ω of a quantum mechanical system, in general, is a linear functional ω on a space $\mathcal{A}$ of linear operators on a Hilbert space $\mathcal{H}$ ($\mathcal{A}$ represents the *observables* of the system), that satisfies the following two properties:

1. $\omega(A^*A) \geq 0$, for all $A \in \mathcal{A}$ (*positivity*).
2. $\omega(\mathbb{1}) = 1$ (*normalization*).

For simplicity suppose $\mathcal{H} = \mathbb{C}^n$, and $\mathcal{A}$ is the space of all $n \times n$ matrices with complex entries, acting as linear transformations on $\mathcal{H}$. Recall that $\mathcal{A}$ is a Hilbert space with the inner product given by

$$\langle A, B \rangle = \mathrm{Tr} A^* B$$

By the Riesz Representation Theorem 12.2 for each ω there is a unique $\rho \in \mathcal{A}$ such that

$$\omega(A) = \mathrm{Tr}\rho^* A, \qquad \text{for all } A \in \mathcal{A}$$

The conditions 1) and 2) translate into $\rho \geq 0$, and $\mathrm{Tr}\rho = 1$, respectively.

12.2 Adjoint operation on $\mathcal{B}(\mathcal{H})$

The most important application of the Riesz Representation Theorem is probably the definition of the adjoint operation of $\mathcal{B}(\mathcal{H})$. Our aim is to show that for any $A \in \mathcal{B}(\mathcal{H})$, there is a unique $A^* \in \mathcal{B}(\mathcal{H})$, such that

$$\langle x, Ay \rangle = \langle A^* x, y \rangle \tag{12.6}$$

The uniqueness of A^*, if it exists, follows from the following exercise.

Exercise 12.4 **a)** *Suppose $A, B \in \mathcal{B}(\mathcal{H})$ satisfy*

$$\langle x, Ay \rangle = \langle x, By \rangle, \quad \text{for all } x, y \in \mathcal{H} \quad .$$

Prove that it follows that $A = B$.
b) *Use the result of art a) and polarization to prove that if*

$$\langle x, Ax \rangle = \langle x, Bx \rangle, \text{for all } x,$$

then, $A = B$.

To prove that A^* exists means to show that for every $x \in \mathcal{H}$, there is a vector $z \in \mathcal{H}$, depending linearly on x, and such that

$$\langle z, y \rangle = \langle x, Ay \rangle \tag{12.7}$$

Observe that, for x fixed, the map φ_x defined by

$$\varphi_x(y) = \langle x, Ay \rangle$$

is a linear functional on $\mathcal{H}$, and that $\|\varphi_x\|$ is bounded by $\|x\|\|A\|$. By the Riesz Representation Theorem there is then a unique z satisfying (12.7). The linearity follows from the uniqueness in the Riesz Representation Theorem and the linearity of the inner product.

Exercise 12.5 *Prove the following properties of the adjoint: for all $A, B \in \mathcal{B}(\mathcal{H})$,and $\lambda \in \mathbb{C}$ one has*
a) $A^{**} = A$.
b) $(AB)^* = B^* A^*$.
c) $(\lambda A)^* = \overline{\lambda} A^*$.
d) $(A + B)^* = A^* + B^*$.
e) $\|A^*\| = \|A\|$.
f) $\|A^* A\| = \|A\|^2$.

Exercise 12.6 *Show by an example that, in general, $\|AA\| \neq \|A\|^2$.*

12.3 Weak convergence in a Hilbert space

Let $\mathcal{H}$ be a Hilbert space and (x_n) a sequence in $\mathcal{H}$. We say that x_n converges *weakly* to $x \in \mathcal{H}$ iff

$$\lim_{n\to\infty} \langle y, x_n\rangle = \langle y, x\rangle, \qquad \text{for all } y \in \mathcal{H}$$

Weak convergence for sequences in a Hilbert space is a special case of weak convergence in a Banach space. Let X be a Banach space, and (x_n) a sequence in X. We say that x_n converges weakly to $x \in X$ iff $\lim_n \varphi(x_n) = \varphi(x)$ for all $\varphi \in X^*$. In the case that X is a Hilbert space, why is this equivalent to the definition given above for weak convergence in a Hilbert space?

It is easy to see that, in any Banach space, weak convergence is weaker than norm convergence in the sense that if $x_n \to x$ in norm, then $x_n \to x$ weakly. This follows from

$$|\varphi(x_n) - \varphi(x)| = |\varphi(x_n - x)| \le \|\varphi\|\|x_n - x\|$$

The converse is not true whenever the Hilbert space is infinite-dimensional. This follows from the next exercise.

Exercise 12.7 *Let $\{u_n\}$ be a sequence of orthonormal vectors in a Hilbert space. Prove that $u_n \to 0$ weakly.*

Exercise 12.8 *Prove that in a finite-dimensional Hilbert space (e.g., $\mathbb{C}^n$) norm and weak convergence are equivalent.*

Exercise 12.9 *Let $\mathcal{H}$ be a Hilbert space, $\{u_\alpha\}$ an o.n. basis of $\mathcal{H}$, and (x_n) a sequence in $\mathcal{H}$. Suppose there is $x \in \mathcal{H}$ such that*

$$\langle u_\alpha, x_n\rangle \to \langle u_\alpha, x\rangle, \qquad \textit{for all } \alpha$$

Prove or disprove that then necessarily $x_n \to x$ weakly.

Theorem 12.3 *Let (x_n) be a sequence in $\mathcal{H}$ converging weakly to x. If in addition*

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

then $x_n \to x$ in norm.

Proof: We start from

$$\|x_n - x\|^2 = \|x_n\|^2 + \|x\|^2 - 2\mathrm{Re}\langle x_n, x\rangle$$

If $x_n \to x$ weakly we have, in particular,

$$\lim_n \langle x_n, x\rangle = \langle x, x\rangle$$

and, hence, under the hypotheses of the theorem, $\lim_n \|x_n - x\|^2 = 0$. ∎

13 The spectrum of bounded linear operators

13.1 Motivation

We have used Fourier series to solve, e.g., the heat equation on an interval $[0, L]$, with given boundary conditions and inital data. What would it take to express the solution of

$$\begin{aligned} u_t = u_x x + qu \quad & 0 \le x \le L, t \ge 0 \\ u(0,t) = u(L,t) = 0 \quad & \quad t \ge 0 \\ u(x,0) = f(x) \quad & \end{aligned}$$

for given functions f and q, as a series expansion similar to a Fourier series? This equation describes the temperature of a heat conducting bar with a non-uniform heat loss term given by $-q(x)u(x,t)$. If we try separation of variables and assume

$$u(x,t) = \sum_n a_n(t) u_n(x)$$

where $\{u_n\}$ is an o.n. basis of $L^2([0, L])$, and where the $a_n(t)$ are solutions of $\dot{a}_n = \lambda_n a_n$, with constants λ_n, we find that u_n should satisfy

$$\frac{d^2}{dx^2} u_n(x) + q(x) u_n(x) = \lambda_n u_n$$

I.e., the u_n should be eigenvectors of the linear operator D defined by

$$Df(x) = \frac{d^2}{dx^2} f(x) + q(x) f(x)$$

Diagonalization problems of linear maps on infinite-dimensional spaces arise in many other ways. The mathematics of this so-called spectral problem is the main topic of this course for the rest of the quarter. This is again a generalization of something we know well in the finite-dimensional case: diagonalization of matrices. First, we briefly review the finite-dimensional situation.

13.2 Diagonalization of matrices

We consider an $n \times n$ matrix A with complex entries as a (bounded) linear map $\mathbb{C}^n \to \mathbb{C}^n$. A complex number λ is called an *eigenvalue* of A if there is a non-zero vector $u \in \mathbb{C}^n$ such that

$$Au = \lambda u \tag{13.1}$$

A non-zero vector $u \in \mathbb{C}^n$ such that (13.1) holds for $\lambda \in \mathbb{C}$, is called an *eigenvector* of A (belonging to the eigenvalue λ).

One says that A is *diagonalizable* if there is a basis $\{u_1, \dots, u_n\}$ of $\mathbb{C}^n$, consisting of eigenvectors, i.e.,

$$Au_k = \lambda u_k, \quad \lambda_k \in \mathbb{C}, k = 1, \dots, n$$

The set of eigenvalues of A is called the *spectrum* of A, often denoted by $\sigma(A)$.

In this course we will stick with the Hilbert space setting, where the interesting bases are o.n. bases.

When does an $n \times n$ matrix have an o.n. basis of eigenvectors? If there is an o.n. basis $\{u_1, \dots, u_n\}$ such that $Au_k = \lambda_k u_k$, there exists a unitary U defined by

$$Ue_k = u_k$$

where $\{e_k\}$ is the standard o.n. basis of $\mathbb{C}^n$. Clearly one then also has $U^* u_k = U^{-1} u_k = e_k$, $k = 1, \dots n$, and therefore

$$U^* A U e_k = \lambda_k e_k$$

which means that $U^* AU$ is a diagonal matrix with the eigenvalues of A on the diagonal:

$$U^* AU = D := \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Or, equivalently, $A = UDU^*$, with U unitary and D diagonal.

When can a matrix A that can be written in the form UDU^*? If $A = UDU^*$, we also have $A^* = U\overline{D}U^*$, and

$$A^* A = U\overline{D}U^* UDU^* = U\overline{D}DU^* = UD\overline{D}U^* = UDU^* U\overline{D}U^* = AA^*$$

In other words $A^* A = AA^*$, i.e., A commutes with its hermitian conjugate. Matrices with this property are called *normal* matrices (or, more generally, normal operators). So, we found that matrices that have an o.n. basis of eigenvectors are normal. The converse also holds.

Theorem 13.1 *For any normal $n \times n$ matrix there is an o.n. basis of eigenvectors.*

This is a standard theorem of linear algebra.

Special cases of normal matrices are the *hermitian* matrices (satisfying $A^* = A$), and the *unitary* matrices (satisfying $U^* U = \mathbb{1} = UU^*$).

Any normal martrix N can be written as the product of a unitary matrix V and a hermitian matrix A. A can be chosen to be non-negative definite. This follows directly from the diagonal form of N: if $N = UDU^*$, with eigenvalues $\lambda_k = e^{i\varphi_k} |\lambda_k$, we can write $D = \Phi|D|$, where Φ is a diagonal matrix with $e^{i\varphi_k}$ on the diagonal and $|D|$ is diagonal with $|\lambda_k|$ on the diagonal. Then,

$$N = UDU^* = (U\Phi U^*)(U|D|U^*) = VA \tag{13.2}$$

with $V = (U\Phi U^*)$, as a product of unitaries, is unitary, and $A = (U|D|U^*)$ is non-negative definite. It is easy to check that $VA = AV$. (13.2) is called the *polar decomposition* of N.

The converse is also true: if $N = VA$, with V unitary, and A hermitian, and such that $VA = AV$, then N is normal.

The eigenvalues are the roots of the polynomial of degree n given by $P(x) = \det(A - x\mathbb{1})$. And if $P(\lambda) = \det(A - \lambda\mathbb{1}) = 0$ we have that $\ker(A - \lambda\mathbb{1}) \neq \{0\}$, i.e., there is a non-zero vector u such that $Au = \lambda u$. It follows that all matrices have at least one eigenvalue, and that to each eigenvalue there corresponds a non-zero subspace of eigenvectors.

The following countereample shows that not all matrices have a basis of eigenvectors:

$$A = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$$

λ is the only eigenvalue of A and the eigenvectors belonging to λ are multiples of $u = (1, 0)$, which does not contain a basis of $\mathbb{C}^2$.

13.3 The spectrum: definition

It turns out that it is useful to define the spectrum of a linear operator on an infinite-dimensional space a bit more generally than just the set of eigenvalues. As we will see in an example below, in inifinite dimensions, a bounded linear operator does not necessarily have any eigenvalues (and eigenvectors), even if it is self-adjoint ($A^* = A$).

As before, let $\mathcal{H}$ be a Hilbert space, and let $\mathcal{B}(\mathcal{H})$ denote the Banach space of all bounded linear transformations of $\mathcal{H}$.

Definition 13.2 (Resolvent Set, Spectrum) *For $A \in \mathcal{B}(\mathcal{H})$, the resolvent set of A, $\rho(A)$, is the set of all complex numbers λ such that*

$$A - \lambda\mathbb{1}$$

is one-to-one and onto and (hence[1]) has a bounded inverse. The spectrum of A, $\sigma(A)$, is the complement of the resolvent set: $\sigma(A) = \mathbb{C} \setminus \rho(A)$.

[1]Due to the open mapping theorem (see, e.g., M. Reed and B. Simon, *Modern Methods of Mathematical Physics*, Vol I, Theorem III.11.)

Example: If λ is an eigenvalue of A in the sense that there is a non-zero vector $u \in \mathcal{H}$ such that $Au = \lambda u$, then $\ker(A - \lambda\mathbb{1}) \neq \{0\}$, and hence $A - \lambda\mathbb{1}$ is not one-to-one. It follows that $\lambda \in \sigma(A)$.

This example, however, does not cover all the possible reasons for which a complex number λ may belong to the spectrum:

- $A - \lambda\mathbb{1}$ is not one-to-one. In this case λ is called an *eigenvalue*. The set of all eigenvalues of A is called the *point spectrum* of A.

- $A - \lambda\mathbb{1}$ is one-to-one, but, although $\operatorname{ran}(A - \lambda\mathbb{1})$ is dense in $\mathcal{H}$, $A - \lambda\mathbb{1}$ is not onto, i.e., $A - \lambda\mathbb{1}$ does not have a *bounded* inverse. In this case λ is said to belong to the *continuous spectrum* of A.

- It is also possible that $A - \lambda\mathbb{1}$ is one-to-one, but $\operatorname{ran}(A - \lambda\mathbb{1})$ is not dense in $\mathcal{H}$, in which case we say that λ belongs to the *residual spectrum* of A.

Here is an example of an operator without any eigenvalues. This is to be contrasted with the finite-dimensional situation where there always is at least one eigenvalue.

Example: Let $\mathcal{H} = L^2([0,1])$, and define $X \in \mathcal{B}(\mathcal{H})$ by

$$Xf(x) = xf(x)$$

It is easy to see that X is bounded and that $\|X\| = 1$. For any $\lambda \in \mathbb{C}$, suppose there is a function f such that $Xf = \lambda f$. Then $f(x) = 0$, for all $x \neq \lambda$. Hence $f = 0 \in L^2$. So, X has no eigenvalues.

On the other hand, $\operatorname{ran}(X - \lambda\mathbb{1})$ is dense for all $\lambda \in \mathbb{C}$. In fact for $\lambda \in \mathbb{C} \setminus [0,1]$, $\operatorname{ran}(X - \lambda\mathbb{1}) = L^2$: as $x \mapsto (x-\lambda)^{-1} \in L^2([0,1])$, is bounded and continuous we have that for any $f \in L^2([0,1])$, $(x-\lambda)^{-1}f(x) \in L^2([0,1])$, and $(X - \lambda\mathbb{1})(x-\lambda)^{-1}f(x) = f(x)$. As f is arbitrary, this proves that $\operatorname{ran}(X - \lambda\mathbb{1}) = L^2([0,1])$.

For $\lambda \in [0,1]$, the non-zero constant functions $f(x) = c$ are not in the range of $X - \lambda\mathbb{1}$, because $c(x-\lambda)^{-1} \notin L^2([0,1])$. However $\operatorname{ran}(X - \lambda\mathbb{1})$ is dense also in this case:

$$\lim_n (X - \lambda\mathbb{1})\mathrm{Ind}[|x-\lambda| \geq 1/n]\frac{f}{x-\lambda} = f, \quad f \in L^2([0,1])$$

where the limit is in the sense of L^2.

We have proved that $\sigma(X) = [0,1]$, and that all $\lambda \in [0,1]$ belong to the continuous spectrum of X.

Note that if a Kronecker delta $\delta_{x,\lambda}$, $0 \leq \lambda \leq 1$, (or a "delta function") were a non-zero element of the space on which X acts, it would be an eigenvector with eigenvalue λ. In some sense it is still true that the elements of the spectrum are similar to eigenvalues.

13.4 Elementary properties of the spectrum

Recall that the resolvent set $\rho(A)$ of a linear operator A is the set of all $\lambda \in \mathbb{C}$ such that $A - \lambda\mathbb{1}$ has an everywhere defined, bounded inverse. It is conventional to denote the inverse of $\lambda\mathbb{1} - A$ by R_λ. R_λ is called the *resolvent* (or *resolvent operator*) of A (at λ). The following proposition and its proof show that R_λ can be considered as an (operator-valued) analytic function defined on the open domain $\rho(A)$.

Proposition 13.3 *For any $A \in \mathcal{B}(\mathcal{H}), \rho(A)$ is an open subset of $\mathbb{C}$.*

Proof: We have to show that for any $\lambda_0 \in \rho(A)$, there is an open disk in the complex plane with center λ_0, and entirely contained in $\rho(A)$. We start from

$$\lambda\mathbb{1} - A = (\lambda_0\mathbb{1} - A)\left[\mathbb{1} - (\lambda_0 - \lambda)(\lambda_0\mathbb{1} - A)^{-1}\right]$$

If $|\lambda_0 - \lambda| < \|(\lambda_0\mathbb{1} - A)^{-1}\|^{-1}$ we can invert the RHS (and hence the LHS) by the Neumann series . ∎

The Neumann series also shows that

$$\sigma(A) \subset \{z \in \mathbb{C} \mid |z| \leq \|A\|\}$$

The *spectral radius* of A, $r(A)$, is defined as the radius of the smallest disk with center 0 that contains $\sigma(A)$:

$$r(A) := \sup\{|\lambda| \mid \lambda \in \sigma(A)\}$$

One can show that for self-adjoint operators A, $r(A) = \|A\|$. But, in general, one only has $r(A) \leq \|A\|$. E.g., consider $\mathcal{H} = \mathbb{C}^2$ and

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

Then $r(A) = 0$, and $\|A\| = 1$.

Proposition 13.4 *For any $A \in \mathcal{B}(\mathcal{H})$, $\sigma(A)$ is not empty.*

Proof: For sufficiently large $|\lambda|$, i.e., $|\lambda| > \|A\|$, the resolvent $R_\lambda = (\lambda\mathbb{1} - A)^{-1}$ is bounded and given by a norm-convergent series:

$$R_\lambda = \frac{1}{\lambda}\left[\sum_{n=0}^{\infty}\left(\frac{A}{\lambda}\right)^n\right]$$

It follows that for any $x, y \in \mathcal{H}$, the function f defined by

$$f(\lambda) := \langle x, R_\lambda y\rangle \tag{13.3}$$

is given by a convergent series expansion as well, and is analytic and uniformly bounded for $|\lambda| > \|A\|$. Its domain of analyticity is $\rho(A)$. Indeed, f is analytic because it is given by a power series on $\rho(A)$ (previous proposition). (Note also that R_λ cannot be analytic on a domain containing points of the spectrum, because then, as in the proof of the previous proposition, $\lambda\mathbb{1}-A$ would be invertible for those λ.) So, if $\sigma(A)$ were empty, f would be a bounded entire function. By Liouville's Theorem, this implies that f is a constant function, and as $\lim_{\lambda\to\infty} f(\lambda) = 0$, this means $f = 0$. $f = 0$ for all choices if $x, y \in H$, in (13.3), would imply $R_\lambda = 0$ for all λ, which is impossible. Therefore, we must conclude that $\sigma(A) \neq \emptyset$ (and thus alow for singularites in the functions f). ■

The following lemma gives an important relation between an operator and its adjoint.

Lemma 13.5 *For any $A \in \mathcal{B}(\mathcal{H})$ the following relations hold:*

$$(\ker A)^\perp = \overline{\mathrm{ran} A^*} \tag{13.4}$$

$$(\mathrm{ran} A)^\perp = \ker A^* \tag{13.5}$$

Proof: If $x \in \mathrm{ran} A^*$, there exists $y \in \mathcal{H}$ such that $x = A^* y$. For any $z \in \ker A$, we then have

$$\langle x, z\rangle = \langle A^* y, z\rangle = \langle y, Az\rangle = 0$$

This proves that $\mathrm{ran} A^* \subset (\ker A)^\perp$. As $(\ker A)^\perp$ is closed, this implies $\overline{\mathrm{ran} A^*} \subset (\ker A)^\perp$.

On the other hand, if $x \in (\mathrm{ran} A^*)^\perp$, we have, for all $y \in \mathcal{H}$,

$$0 = \langle A^* y, x\rangle = \langle y, Ax\rangle$$

and therefore, $Ax = 0$. This means $(\mathrm{ran} A^*)^\perp \subset \ker A$. By taking the $\perp$ complement of both sides of this relation we get $(\ker A)^\perp \subset (\mathrm{ran} A^*)^{\perp\perp} = \overline{\mathrm{ran} A^*}$. This concludes the proof of the first relation.

The second relation follows from the first by applying it to A^*, instead of A, and by using $A^{**} = A$, and taking orthogonal complements. ■

Proposition 13.6 *For any $A \in \mathcal{B}(\mathcal{H})$, if λ is in the residual spectrum of A, then $\overline{\lambda}$ is an eigenvalue of A^*.*

Proof: If λ belongs to the residual spectrum of A, we have that $\mathrm{ran}(\lambda\mathbb{1} - A)$ is not dense in $\mathcal{H}$. By the previous lemma, this means that there exists $0 \neq x \in (\mathrm{ran}(\lambda\mathbb{1} - A))^\perp = \ker(\lambda\mathbb{1} - A)^* = \ker(\overline{\lambda}\mathbb{1} - A^*)$. ■

Exercise 13.1 *Prove that $\rho(A^*) = \overline{\rho(A)}$, where $\overline{\rho(A)}$ is the set $\{\lambda \in \mathbb{C} \mid \overline{\lambda} \in \rho(A)\}$.*

Exercise 13.2 *If λ is an eigenvalue of A, $\overline{\lambda}$ is in the spectrum of A^*. What can you say about the type of spectrum $\overline{\lambda}$ belongs to?*

Lemma 13.7 *If $A = A^* \in \mathcal{B}(\mathcal{H})$, we have that $\sigma(A) \subset [-\|A\|, \|A\|] \subset \mathbb{R}$.*

Proof: We already know that $r(A) \leq \|A\|$. So we only have to prove that the spectrum of a self-adjoint A consists only of real numbers.

Write $\lambda = a + ib$, with $a, b \in \mathbb{R}$. Then, for all $x \in \mathcal{H}$

$$\begin{aligned} \|(A - \lambda\mathbb{1})x\|^2 &= \langle (A - \lambda\mathbb{1})x, (A - \lambda\mathbb{1})x \rangle \\ &= \langle (A - a\mathbb{1})x, (A - a\mathbb{1})x \rangle + \langle (-ib)x, (-ib)x \rangle \\ &\quad + \langle Ax, (-ib)x \rangle + \langle (-ib)x, Ax \rangle \\ &= \|(A - a\mathbb{1})x\|^2 + b^2\|x\|^2 \geq b^2\|x\|^2 \end{aligned}$$

Suppose $\lambda \in \sigma(A)$ and $b \neq 0$. Then, λ cannot be an eigenvalue because the inequality shows that $\ker(A - \lambda) = 0$. Also, as $b \neq 0$, the inequality above implies that $(A - \lambda\mathbb{1})$ has a bounded inverse on its range (see Exercise 13.3 below). It is then easy to check that $\operatorname{ran}(A - \lambda\mathbb{1})$ is closed. If $\operatorname{ran}(A - \lambda\mathbb{1}) \neq \mathcal{H}$, λ would therefore be in the residual spectrum of A, and by Proposition 13.6, $\overline{\lambda} = a - ib$, would then be an eigenvalue of $A^* = A$. But this would contradict the inequality given above applied to $\overline{\lambda}$. Therefore, we must conclude that $\lambda \in \rho(A)$ if $b \neq 0$. ∎

Corollary 13.8 *If A is self-adjoint the residual spectrum of A is empty.*

Proof: As the point spectrum and the residual spectrum are subsets of $\mathbb{R}$ (previous lemma) with empty intersection, this follows immediately from Proposition 13.6. ∎

Exercise 13.3 *Let X and Y be normed linear spaces and $T : X \to Y$ a linear map.*
a) *Suppose there exists a constant $c > 0$ such that*

$$\|Tx\| \geq c\|x\| \quad . \tag{13.6}$$

Prove that T has a bounded inverse defined on $\operatorname{ran}(T)$, with $\|T^{-1}\| \leq c^{-1}$.
b) *If T has a bounded inverse defined on its range and X is a Banach space, then $\operatorname{ran}(T)$ is closed.*
c) *Prove the converse of part a): if T has a bounded inverse defined on its range, then there exists a constant $c > 0$ such that (13.6).*

Exercise 13.4 *Prove that the spectrum of an orthogonal projection P is either $\{0\}$, in which case $P = 0$, or $\{1\}$, in which case $P = \mathbb{1}$, or else $\{0, 1\}$.*

Exercise 13.5 *Let $A \in \mathcal{B}(\mathcal{H})$ be such that $\langle x, Ax\rangle \geq 0$, for all $x \in \mathcal{H}$. Prove that $\sigma(A) \subset [0, +\infty)$.*

The following set of exercises provide excellent examples of the different kinds of spectrum. The shift operators appearing in these exercises can be used, e.g., to define difference equations, which, in turn, show up in numerical approximation schemes for ODE and PDE. See Exercise 13.10 for an example.

Exercise 13.6 *Let $\mathcal{H} = l^2(\mathbb{Z})$. The right shift S on $\mathcal{H}$ is defined by*

$$S(x)_k = x_{k-1}, \qquad \text{for all } k \in \mathbb{Z} \tag{13.7}$$

where $x = (x_k)_{k\in\mathbb{Z}} \in l^2(\mathbb{Z})$.

a) *Prove that the point spectrum of S is empty, i.e., S has no eigenvalues, or, equivalently, $\lambda \mathbb{1} - S$ is one-to-one for all $\lambda \in \mathbb{C}$.*

b) *Prove that* $\mathrm{ran}(\lambda \mathbb{1} - S) = l^2(\mathbb{Z})$, *for all $\lambda \in \mathbb{C}$ with $|\lambda| > 1$. In other words, prove that the spectrum of S lies entirely in the unit disk.*

c) *Prove that* $\mathrm{ran}(\lambda \mathbb{1} - S) = l^2(\mathbb{Z})$, *for all $\lambda \in \mathbb{C}$ with $|\lambda| < 1$. In other words, in combination with the previous part, prove that the spectrum of S lies entirely in the unit circle.*

d) *Prove that all $\lambda \in \mathbb{C}, |\lambda| = 1$, belong to the continuous spectrum of S.*

It is interesting to compare the results of the previous exercise to the following two, which concern the spectrum of similar shift operators on $l^2(\mathbb{N})$.

Exercise 13.7 *Define the right shift R on $l^2 := l^2(\{0, 1, 2, 3, \ldots\})$ by*

$$R(x) = (0, x_0, x_2, \ldots) \text{ for all } x = (x_0, x_1, x_2, \ldots) \in l^2 \quad . \tag{13.8}$$

a) *Prove that* $\mathrm{ran}(\lambda \mathbb{1} - R) = l^2(\mathbb{Z})$, *for all $\lambda \in \mathbb{C}$ with $|\lambda| > 1$. In other words, prove that the spectrum of R lies entirely in the unit disk.*

b) *Prove that all $\lambda \in \mathbb{C}, |\lambda| = 1$, belong to the continuous spectrum of R.*

c) *Prove that* $\mathrm{ran}(\lambda \mathbb{1} - R)$ *is not dense in l^2, for all $\lambda \in \mathbb{C}$ with $|\lambda| < 1$. In other words, prove that the interior of the unit disk is part of the residual spectrum of R.*

Exercise 13.8 *Define the left shift L on $l^2 := l^2(\{0, 1, 2, 3, \ldots\})$ by*

$$L(x) = (x_1, x_2, \ldots) \text{ for all } x = (x_0, x_1, x_2, \ldots) \in l^2 \quad . \tag{13.9}$$

Prove the following:

a) *The resolvent set of L consists of all $\lambda \in \mathbb{C}, |\lambda| > 1$.*

b) *The continuous spectrum of L is the unit circle.*

c) *The point spectrum of L is given by the interior of the unit disc.*

d) *The residual spectrum of L is empty.*

Exercise 13.9 *Compute the adjoints S^*, R^*, and L^* of the operators S, R, and L defined in the previous exercises.*

Exercise 13.10 *Define an operator Δ on $l^2(\mathbb{Z})$ by*

$$(\Delta x)_k = x_{k-1} - 2x_k + x_{k+1}, \text{ for all } x = (x_k)_{k\in\mathbb{Z}} \in l^2(\mathbb{Z}) \ . \tag{13.10}$$

Prove that the spectrum of Δ is entirely continuous, real, and given by the interval $[-4, 0]$. (Hint: Note that $\Delta = S + S^* - 2\mathbb{1}$.)

14 The spectral theorem for compact self-adjoint operators

Let X and Y be Banach spaces (e.g., $X = Y = \mathcal{H}$, a Hilbert space).

Definition 14.1 *A linear operator $T : X \to Y$ is called compact if for every bounded set $B \subset X$, the set $T(B)$ is pre-compact in Y.*

An equivalent formulation is: T is compact if the image of any bounded sequence $x_n \in X$, has a convergent subsequence $Tx_{n_k} \to y \in Y$.

Example:
An orthogonal projection on a Hilbert space is compact iff its range is finite-dimensional.

Proposition 14.2 *Let X, Y, and Z be Banach spaces. Compact linear maps have the following properties:*

1. *If $C : X \to Y$ is compact then C is bounded.*
2. *If $C_1, C_2 : X \to Y$ are compact then $C_1 + C_2$ is compact.*
3. *If $C_n \to C$ in norm, and $C_n : X \to Y$ is compact for all n, then C is compact.*
4. *If $A : X \to Y$ and $B : Z \to X$ are bounded linear maps, and $C : Y \to Z$ is compact, then CA and BC are compact.*

It follows from this proposition that the compact linear operators $X \to Y$ form a Banach subspace of $\mathcal{B}(X, Y)$. And that the compact operators $X \to X$ are a two-sided ideal of $\mathcal{B}(X)$.

Exercise 14.1 *Prove Proposition 14.2.*

We now focus on the problem of diagonalizing compact self-adjoint operators on a Hilbert space $\mathcal{H}$. We will give two versions of the spectral theorem for compact self-adjoint operators, one for general Hilbert spaces (Theorem 14.4), and one for separable Hilbert spaces (Theorem 14.3). This is merely for convenience of the user. In fact, the two versions have the same proof (see p. 152)

Theorem 14.3 (Spectral Theorem, version 1) *Let $\mathcal{H}$ be a separable Hilbert space and suppose $A = A^* \in \mathcal{B}(\mathcal{H})$ is compact. Then there is an o.n. basis $\{e_n\}$ of eigenvectors $\mathcal{H}$ with real eigenvalues λ_n, such that*

$$A = \sum_n \lambda_n \langle e_n, \cdot \rangle e_n \tag{14.1}$$

The multiplicity of each non-zero eigenvalue is finite, and, if $\mathcal{H}$ is infinite-dimensional, $\lim_n \lambda_n = 0$.

Note that if $\mathcal{H}$ is not separable, and C is compact and self-adjoint, we necessarily have that $\mathcal{H}_0 := (\ker C)^\perp$ is a separable sub-Hilbert space of $\mathcal{H}$ (why ?), and $\mathcal{H}_0$ is an invariant subspace for C. We can therefore use the previous theorem to diagonalize C restricted to $\mathcal{H}_0$, and, as $C = 0$ on $\mathcal{H}_0^\perp$ this also diagonalizes C on the whole of $\mathcal{H}$. So, the condition of separability in the theorem is not really a restriction and the theorem can also be formulated as follows:

Theorem 14.4 (Spectral Theorem, version 2) *Let $\mathcal{H}$ be a (not necessarily separable) Hilbert space and suppose $A = A^* \in \mathcal{B}(\mathcal{H})$ is compact. Then A has at most a countable number of non-zero eigenvalues $\lambda_n \in \mathbb{R}$, with a corresponding orthonormal set $\{e_n\}$ of eigenvectors, such that*

$$A = \sum_n \lambda_n \langle e_n, \cdot\rangle e_n \tag{14.2}$$

The multiplicity of each non-zero eigenvalue is finite, and 0 is the only possible accumulation point of $\{\lambda_n\}$ (and 0 is an accumulation point if the set of non-zero eigenvalues is infinite).

The relation (14.1) means that the sequence of operators defined by the partial sums,

$$S_N x = \sum_{n \le N} \lambda_n \langle e_n, x\rangle e_n, \qquad \text{for all } x \in \mathcal{H}$$

converges to A in the operator norm.

It is easy to see why the non-zero eignvalues of a compact self-adjoint A have finite multiplicity, and that if there are an infinite number of those, one must have that $\lim_n \lambda_n = 0$. Indeed if $\lambda \neq 0$ would have infinite multiplicity, the corresponding o.n. eigenvectors e_n would form a bounded sequence, but $Ae_n = \lambda e_n$ does not have a convergent subsequence contradicting the assumption that A is compact. If A has an infinite number of non-zero eigenvalues λ_n, then, as the eigenvalues are all bounded by $\|A\|$, there are convergent subsequences $\lambda_{n_k} \to \lambda$. If $\lambda \neq 0$, the sequence of eigenvectors $\lambda_n^{-1} e_n$, is bounded, but again, $A\lambda_n^{-1} e_n = e_n$ does not have a convergent subsequence, which is impossible since A is compact.

Before we embark on the proof of this theorem we will first prove some lemmas that we will need in the proof, but that have other uses as well.

After the proof of the theorem we will address the question how one can verify that a bounded linear operator is compact. We will also see how the theorem can be used to diagonalize certain *unbounded* (hence non-compact) linear operators.

Lemma 14.5 *If $A = A^* \in \mathcal{B}(\mathcal{H})$, we have that*

$$\|A\| = \sup_{x, \|x\|=1} |\langle x, Ax\rangle|$$

Proof: One inequality is obvious:

$$|\langle x, Ax\rangle| \leq \|Ax\|\,\|x\| \leq \|A\|\,\|x\|^2$$

and therefore,

$$a := \sup_{x,\|x\|=1} |\langle x, Ax\rangle| \leq \|A\|$$

To prove the second inequality we use Lemma 14.5 to express $\|A\|$ as follows:

$$\|A\| = \sup_{x,\|x\|=1} \|Ax\| = \sup_{x,y;\|x\|=\|y\|=1} |\langle y, Ax\rangle|$$

and then polarization to rewrite the RHS as

$$\langle y, Ax\rangle = \frac{1}{4}\times$$
$$[\langle x+y, A(x+y)\rangle - \langle x-y, A(x-y)\rangle + i\langle x+iy, A(x+iy)\rangle - i\langle x-iy, A(x-iy)\rangle]$$

As $A^* = A$, the first two terms are real, and the last two are imaginary. By replacing y by $e^{i\varphi}y$, with a suitable $\varphi \in \mathbb{R}$, we can make the imaginary terms vanish, i.e., make $\langle e^{i\varphi}y, Ax\rangle$ real. This leads to the bound

$$\begin{aligned} 16|\langle y, Ax\rangle|^2 &\leq |\langle x+y, A(x+y)\rangle - \langle x-y, A(x-y)\rangle|^2 \\ &\leq a^2(\|x+y\|^2 + \|x-y\|^2)^2 \\ &= a^2(2\|x\|^2 + 2\|y\|^2)^2 = 16a^2 \end{aligned}$$

where we have used the definition of a, the Parallelogram Law, and the fact that $\|x\| = \|y\| = 1$. ■

Lemma 14.6 *Let $A = A^* \in \mathcal{B}(\mathcal{H})$, and suppose that λ and μ are two disctinct eigenvalues of A with eigenvectors u and v respectively. Then, $u \perp v$.*

Proof: By Lemma 13.7 we know that λ and μ are real. Therefore,

$$\lambda\langle u, v\rangle = \langle Au, v\rangle = \langle u, Av\rangle = \mu\langle u, v\rangle$$

If $\lambda \neq \mu$, this implies $\langle u, v\rangle = 0$. ■

Let A be a linear operator on $\mathcal{H}$ and suppose $\mathcal{H}_0$ is a linear subspace of $\mathcal{H}$, satisfying $Ax \in \mathcal{H}_0$, for all $x \in \mathcal{H}_0$. $\mathcal{H}_0$ is then called an invariant subspace for A. If A is bounded the closure of any invariant subspace is still an invariant subspace (why?).

Lemma 14.7 *Let $A = A^* \in \mathcal{H}$, and let $\mathcal{H}_0$ be a closed invariant subspace for A. Then $\mathcal{H}_0^\perp$ is also an invariant subspace for A (and, as it is an orthogonal complement, it is automatically closed).*

Proof: Let $x \in \mathcal{H}_0^\perp$, and $y \in \mathcal{H}_0$. Then, as $A = A^*$,

$$\langle y, Ax\rangle = \langle Ay, x\rangle = 0$$

where we have used that $Ay \in \mathcal{H}_0$. Therefore, $Ax \in \mathcal{H}_0^\perp$. ∎

In the situation of the lemma $\mathcal{H}$ has an orthogonal decomposition:

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^\perp$$

and the lemma implies that A also decomposes into $A = A_0 + A_1$, where $A_0(x_0 \oplus x_1) = Ax_0$, and $A_1(x_0 \oplus x_1) = Ax_1$, for all $x_0 \in \mathcal{H}_0$, and $x_1 \in \mathcal{H}_0^\perp$.

We are now ready to prove the spectral theorem for compact self-adjoint linear operators.

Proof: (of Theorems 14.3 and 14.4)

Step 1.

First we will prove that if $A = A^*$ is compact, then either $\lambda = \|A\|$, or $\lambda = -\|A\|$, or both, is an eigenvalue of A. As there is nothing to prove if $A = 0$, suppose $A \neq 0$.

We start from the following expression for $\|A\|$ (Lemma 14.5):

$$\|A\| = \sup_{x, \|x\|=1} |\langle x, Ax\rangle|$$

which implies that we can find a sequence (x_n), with $\|x_n\| = 1$, for all n, and such that

$$\lim_n |\langle x_n, Ax_n\rangle| = \|A\|$$

Since $A = A^*$, $\langle x_n, Ax_n\rangle$ is real for all n, there must be a subsequence of (x_n), say (y_n), such that

$$\lim_n \langle y_n, Ay_n\rangle = \lambda \tag{14.3}$$

with $\lambda = \|A\|$, or $\lambda = -\|A\|$. (y_n) is a sequence of unit vectors and, therefore, bounded. The compactness of A then guarantees the existence of a subsequence, which we denote again by (y_n), such that $\lim_n Ay_n = z$. We claim that z is an eigenvector of A with eigenvector λ. $z \neq 0$, since otherwise from (14.3) it would follow that $\lambda = 0$ and hence $\|A\| = |\lambda| = 0$, which is not the case since we assumed $A \neq 0$. That z is an eigenvector can be checked as follows:

$$\begin{aligned}
\|(A-\lambda)z\|^2 &= \lim_n \|(A-\lambda)Ay_n\|^2 = \|A(A-\lambda)y_n\|^2 \\
&\leq \lim_n \|A\|^2 \|(A-\lambda)y_n\|^2
\end{aligned}$$

$$\begin{aligned} &= \lim_n \|A\|^2 \left[\|Ay_n\|^2 + \lambda^2\|y_n\|^2 - 2\lambda\langle y_n, Ay_n\rangle\right] \\ &\leq \lim_n \|A\|^2 \left[\|A\|^2\|y_n\|^2 + \lambda^2\|y_n\|^2 - 2\lambda\langle y_n, Ay_n\rangle\right] \\ &= \|A\|^2 \left[\lambda^2 + \lambda^2 - 2\lambda^2\right] = 0 \end{aligned}$$

Step 2.

The idea is now to iterate Step 1 as follows. Denote the eigenvector found in Step 1 by e_0, and the corresponding eigenvalue by λ_0. Let $\mathcal{H}_0$ be the one-dimensional subspace of $\mathcal{H}$ generated by e_0, and consider the orthogonal decomposition of $\mathcal{H}$:

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^\perp$$

It is obvious that $\mathcal{H}_0$ is an invariant subspace of A. As explained after Lemma 14.7, A decomposes into $A = A_0 + A_1$, where A_1 is again compact (as the difference of two compact operators, Proposition 14.2). We also have that $\|A_1\| \leq \|A\|$:

$$\|A_1\| = \sup_{x\in\mathcal{H}_0^\perp, \|x\|=1} \|Ax\| \leq \sup_{x\in\mathcal{H}, \|x\|=1} \|Ax\| = \|A\|$$

Now, we can apply Step 1 to A_1 and find an eigenvector e_1 with eigenvalue λ_1, $|\lambda_1| = \|A_1\| \leq \|A\| = |\lambda_0|$. As $e_1 \in \mathcal{H}_0^\perp$ we have $A_0 e_1 = 0$, and

$$Ae_1 = A_0e_1 + A_1e_1 = \lambda_1 e_1$$

So, we have found a second eigenvalue of A, or, if $\lambda_1 = \lambda_0$, a second, independent, eigenvector belonging to λ_0.

We can now repeat this procedure : define A_2 to be the restriction of A to the orthogonal complement of the two-dimensional subspace generated by e_0 and e_1. Then, A_2 is again compact and $\|A_2\| \leq \|A_1\|$. So, there will be an eigenvalue λ_2, $|\lambda_2| \leq |\lambda_1| \leq |\lambda_0|$. etc.

After n iterations of this procedure, we will have found $n+1$ eigenvalues $\lambda_0, \ldots, \lambda_n$, in order of decreasing magnitude. If $A_{N+1} = 0$, for some n, we are done, and can write

$$A = \sum_{n=0}^{N} \lambda_n \langle e_n, \cdot\rangle e_n$$

The orthonormal set $\{e_n\}$ can be completed to an o.n. basis of $\mathcal{H}$, and all the other basis vectors will be eigenvectors of A with eigenvalue 0.

If each $A_n \neq 0$, for all n, we obtain an infinite sequence of non-zero eigenvalues. We already know that the compactness of A implies that non-zero eigenvalues have finite multiplicities, and also that

$$\lim_n \lambda_n = 0 \tag{14.4}$$

For any N we have

$$A = \sum_{n=0}^{N} \lambda_n \langle e_n, \cdot \rangle e_n + A_{N+1}$$

where A_{N+1} is 0 on the subspace generated by $e_0, \ldots, e_N$, and $\|A_{N+1}\| = |\lambda_{N+1}|$. By (14.4) we then have

$$\lim_N \|A - \sum_{n=0}^{N} \lambda_n \langle e_n, \cdot \rangle e_n\| = \lim_N |\lambda_{N+1}| = 0$$

In other words

$$A = \sum_{n=0}^{\infty} \lambda_n \langle e_n, \cdot \rangle e_n$$

where the sum converges in the operator norm. The vectors e_n are o.n., but need not be complete. Denote their closed linear span by $\mathcal{H}'$, i.e.,

$$\mathcal{H}' = \{\sum_n c_n e_n \mid c_n \in \mathbb{C}, \sum_n |c_n|^2 < \infty\}$$

It is easy to see that any vector $x \in (\mathcal{H}')^\perp$, $Ax = A_N x$ for all N. Therefore,

$$\|Ax\| \leq \|A_N\| \, \|x\|, \quad \text{for all } N$$

As $\|A_N\| = |\lambda_n| \to 0$, this implies $Ax = 0$. So, $(\mathcal{H}')^\perp = \ker A$, and if so desired, we can complete $\{e_n\}$ to an o.n. basis by adding a basis of $\ker A$, consisting of zero-eigenvectors. ∎

15 Examples of compact linear operators

Let $\mathcal{H}$ be a Hilbert space and $A \in \mathcal{B}(\mathcal{H})$. Before one can apply the spectral theorem in the form we have derived, one has to check that A is compact. There are many ways of verifying that A is compact. The following is a partial list:

1. Check the definition (Definition 14.1 directly. See Section 15.1).
2. Show that A is Hilbert-Schmidt (see Section 15.2).
3. Show that A is the norm limit of a sequence of compact (e.g., finite-rank) operators (Proposition 14.2).
4. Show that there is a self-adjoint compact operator B, and a function $f : \mathbb{R} \to \mathbb{C}$, continuous at 0, such that $A = f(B)$ (see Section 16.1).

15.1 Compactness in Hilbert space

One way of proving that A is compact would be by direclty verifying the definition: prove that for any bounded set $B \subset \mathcal{H}$, the set $A(B) = \{Ax \mid x \in B\}$, is precompact. A characterization of precompact sets in a Hilbert space is given by the following theorem.

Theorem 15.1 *Let $\mathcal{H}$ be a Hilbert space, and let C be a subset of $\mathcal{H}$.*
i) If C is precompact one has that for any o.n. set $\{e_n\}$ the following property holds: for all $\epsilon > 0$, there exists N such that

$$\sum_{n \geq N} |\langle e_n, x\rangle|^2 < \epsilon, \qquad \text{for all } x \in C \tag{15.1}$$

ii) If C is bounded and there is an o.n. basis $\{e_n\}$ such that for all $\epsilon > 0$, there exists N such that (15.1) holds, then C is pre-compact.

Proof: *Proof of i):* It is easy to show that if C is pre-compact, then it is necessarily bounded (as in any normed space). Suppose C is bounded and such that (15.1) does not hold. We will prove that C cannot be precompact.

If (15.1) does not hold, there is $\epsilon > 0$, such that for all N, there is $x_N \in C$ such that

$$\sum_{n > N} |\langle e_n, x_N\rangle|^2 \geq \epsilon \tag{15.2}$$

We construct a subsequence of (x_N) as follows: let $n_1 = 1$. Then, pick n_2 such that

$$\sum_{n > n_2} |\langle e_n, x_{n_1}\rangle|^2 \leq \epsilon/4$$

etc. Given n_k, pick n_{k+1} such that

$$\sum_{n>n_{k+1}} |\langle e_n, x_{n_k}\rangle|^2 \leq \epsilon/4 \tag{15.3}$$

For any $n \geq 1$ define the orthogonal projection

$$P_n = \sum_{j\leq n} \langle e_j, \cdot\rangle e_j$$

The sequence (x_{n_k}) does not have any Cauchy subsequence: for $k > l$ we have

$$\begin{aligned}
\|x_{n_k} - x_{n_l}\|^2 &\geq \|(\mathbb{1} - P_{n_k})(x_{n_k} - x_{n_l})\|^2 \\
&\geq [\|(\mathbb{1} - P_{n_k})x_{n_k}\| - \|(\mathbb{1} - P_{n_k})x_{n_l})\|]^2 \\
&= \left[\sqrt{\sum_{n>n_k} |\langle e_n, x_{n_k}\rangle|^2} - \sqrt{\sum_{n>n_k} |\langle e_n, x_{n_l}\rangle|^2}\right]^2 \\
&\geq (\epsilon^{1/2} - (\epsilon/4)^{1/2})^2 \geq \frac{\epsilon}{4}
\end{aligned}$$

where we have used (15.2), $n_k \geq n_{l+1}$, and (15.3). It follows that (x_{n_k}) does not have any convergent subsequence, contradicting the hypothesis that C is precompact.
Proof of ii): Let $\{e_n\}$ be an o.n. basis with the stated property, and let (x_n) be any sequence in C. We will use the diagonalization trick to construct a convergent subsequence, thus proving that C is precompact. Without loss of generality we may assume that $\|x\| \leq 1$, for all $x \in C$. Start with $n_1 = 1$. We have that

$$\|(\mathbb{1} - P_{n_1})x_n\| \leq 1, \quad \text{for all } n$$

As all $P_{n_1}x_n$ are in the finite-dimensional Hilbert subspace spanned by $e_1, \ldots, e_{n_1}$, there is subsequence $(x_{1,k})$ such that that $P_{n_1}x_{1,k}$ converges. Therefore, we can pick the subsequence such that (see Excercise 15.1)

$$\|P_{n_1}(x_{1,k} - x_{1,l})\|^2 \leq \frac{1}{k}$$

for all $l \geq k$. Next pick n_2 such that

$$\|(\mathbb{1} - P_{n_2})x_{1,k}\|^2 \leq \frac{1}{2}, \quad \text{for all } k$$

This is possible due to the assumption (15.1). Then pick a subsequence $x_{2,k}$ of $x_{1,k}$, such that $P_{n_2}x_{2,k}$ is Cauchy and such that

$$\|P_{n_2}(x_{2,k} - x_{2,l})\|^2 \leq \frac{1}{k}$$

for all $l \geq k$. Etc: take n_l such that

$$\|(\mathbb{1} - P_{n_l})x_{l-1,k}\|^2 \leq \frac{1}{l}, \qquad \text{for all } k$$

and then pick a subsequence $x_{l,k}$ of $x_{l-1,k}$, such that $P_{n_l}x_{l,k}$ satisfies:

$$\|P_{n_l}(x_{l,k} - x_{l,r})\|^2 \leq \frac{1}{k}$$

for all $r \geq k$. The diagonal sequence $x_{k,k}$ is then Cauchy:

$$\|x_{m,m} - x_{n,n}\|^2 = \|P_k(x_{m,m} - x_{n,n})\|^2 + \|(\mathbb{1} - P_k)x_{m,m} - x_{n,n}\|^2 \leq \frac{2}{k}$$

for all $m, n \geq k$. ∎

In the proof above we have used the following excercise. If it is not entirely obvious to you, you should write a proof of this statement:

Exercise 15.1 *Let (X, d) be a metric space, and let (x_n) be a sequence in X. Prove that if (x_n) has a Cauchy subsequence, then, for any decreasing sequence of positive $\epsilon_k \to 0$, (x_n) has a subsequence (x_{n_k}) such that*

$$d(x_{n_k}, x_{n_l}) \leq \epsilon_k, \qquad \textit{for all } l > k$$

15.2 Hilbert-Schmidt operators

Any finite-rank operator ($\dim \operatorname{ran} A < \infty$), is compact. From Proposition 14.2 we know that uniform limits of compact operators are compact. Therefore we can prove that A is compact by showing that there is a sequence of finite-rank operators A_n that converges to A in norm.

E.g., this way we can prove that any operator on $L^2(D)$, $D \subset \mathbb{R}^n$, of the form

$$Kf(x) = \int_D k(x, y)f(y)\, dy, \qquad \text{for all } f \in L^2(D) \tag{15.4}$$

with $k \in \mathcal{L}^2(D \times D)$, is compact. These operators are called *Hilbert-Schmidt* operators.

Exercise 15.2 *Consider the operator K defined in (15.4).*

a) *Prove that K is compact if $k \in \mathcal{L}^2(D \times D)$.*

b) *Show that K^* is also given as an integral operator of the form*

$$Kf(x) = \int_D l(x, y)f(y)\, dy$$

and find its kernel l in terms of k.

More generally, suppose $\mathcal{H}$ is a separable Hilbert space, and $A \in \mathcal{B}(\mathcal{H})$. If there is an o.n. basis $\{e_n\}$ such that

$$\sum_n |Ae_n\|^2 < +\infty$$

A is said to be Hilbert-Schmidt, and the quantity

$$\|A\|_{HS} = \sqrt{\sum_n \|Ae_n\|^2} \tag{15.5}$$

is the *Hilbert-Schmidt norm* of A.

Exercise 15.3 *Let $\mathcal{H}$ be a separable Hilbert space with an o.n. basis $\{e_n\}$, and $A \in \mathcal{B}(\mathcal{H})$ such that*

$$\sum_n \|Ae_n\|^2 < +\infty$$

a) *Prove that the Hilbert-Schmidt norm defined in (15.5) is independent of the basis. I.e., show that for any other o.n. basis $\{f_n\}$ one has*

$$\sum_n \|Af_n\|^2 = \sum_n \|Ae_n\|^2$$

b) *Prove that*

$$\|A\|_{HS} = \|A^*\|_{HS}$$

15.3 Sturm-Liouville problems

When we studied Fourier series we effectively diagonalized the unbounded operator Δ (called the *Laplacian*, or *Laplace operator*) given by

$$\Delta f = \frac{d^2}{dx^2} f$$

Indeed the o.n. basis of $L^2([0, 2\pi])$, defined by $e_n(x) = e^{inx}$, $n \in \mathbb{Z}$, is a basis of eigenvectors of Δ:

$$\Delta e^{inx} = -n^2 e^{inx}$$

So, the eigenvalues are $-n^2$ (which are unbounded, showing that Δ is not a bounded linear operator). The inverse operator of Δ is not well-defined as the constant functions are in the kernel of Δ. The inverse of $\Delta - \lambda$, with $\lambda \neq -n^2$, for any $n \in \mathbb{Z}$, is defined by

$$(\Delta - \lambda)^{-1} = -\sum_{n \in \mathbb{Z}} (n^2 + \lambda)^{-1} \langle e_n, \cdot \rangle e_n$$

This is a special case of a more general situation.

Let p, q, and w be real-valued functions defined on an interval $[a, b]$ and satisfying: $q, w \in C([a, b])$, $p \in C^1([a, b])$, $w(x) > 0, p(x) > 0$, for all $x \in [a, b]$.

Let H be the Hilbert space of functions f satisfying

$$\int_a^b |f(x)|w(x)dx < \infty$$

with inner product

$$\langle f, g\rangle = \int_a^b \overline{f(x)}g(x)w(x)dx$$

The differential operator given by

$$Lf = \frac{1}{w}[-(pf')' + qf)] \tag{15.6}$$

for f such that $Lf \in H$ is well-defined is called a *Sturm-Liouville* operator. Next quarter, we will be more precise about the definition of unbounded operators such as L. In particular, we will then explain why (15.6) should be supplemented by boundary conditions such as

$$Af(a) + A'f'(a) = 0 \tag{15.7}$$

$$Bf(b) + B'f'(b) = 0 \tag{15.8}$$

where A, A', B, B' are real constants such that A and A' are not both 0, and B and B' are not both 0.

Theorem 15.2 *Let L be defined as in (15.6) with the boundary conditions (15.8) and assume* 0 *is not an eigenvalue of L. Then there exists a continuous function $g : [a, b] \times [a, b] \to \mathbb{R}$, satisfying $g(x, y) = g(y, x)$, such that*

$$L^{-1}f(x) = \int_a^b g(x, y)f(y)dy \tag{15.9}$$

defines the inverse operator.

For a proof of this theorem and more details on Sturm-Liouville operators, see MAT203C next quarter. Note that Proposition 5.2 of MAT203A is a special case of the above theorem.

It follows from (15.9) that L^{-1} is Hilbert-Schmidt and hence compact. And as g is real and symmetric, it is self-adjoint. Therefore, there exists an o.n. basis of eigenvectors $\{e_n\}$ of L^{-1}. $\{e_n\}$ is then also an o.n. basis of eigenvectors of L (with the inverse eigenvalues).

Note that if 0 is an eigenvalue of L, one can apply the theorem to the operator $L - \lambda \mathbb{1}$, with $\lambda \in \mathbb{R}$ chosen such that 0 is not an eigenvalue of $L - \lambda \mathbb{1}$. The relation between L and $L - \lambda \mathbb{1}$ is a redefinition of the function q:

$$(L - \lambda) f = \frac{1}{w} \left[-(pf')' + (q - \lambda w) f) \right]$$

16 Applications of the spectral theorem

16.1 Functions of operators

The theory of functions of operators is called *operational calculus*. There are several ways of defining a function of a linear operator. E.g., for any $A \in \mathcal{B}(\mathcal{H})$ (or for any element in a Banach algebra for that matter) we can define $\exp A$ by the following norm convergent series:

$$\exp A = \sum_{n=0}^{\infty} \frac{A^n}{n!}$$

Here, we focus on functions of self-adjoint compact operators on a Hilbert space $\mathcal{H}$. If A is self-adjoint and compact, we can diagonalize it:

$$A = \sum_{n=0}^{\infty} \lambda_n P_n$$

where, $\{e_n\}$ is an o.n. basis of $\mathcal{H}$, $P_n = \langle e_n, \cdot \rangle e_n$, $\lambda_n \in [-\|A\|, \|A\|]$, $Ae_n = \lambda_n e_n$, and if $\dim \mathcal{H} = \infty$, $\lim_n \lambda_n = 0$.

Exercise 16.1 *Suppose* $\dim \mathcal{H} = \infty$. *If* $A \in \mathcal{B}(\mathcal{H})$ *is self-adjoint and compact and* $\{\lambda_n\}$ *is the set of eigenvalues of* A*, then the spectrum of* A*,* $\sigma(A)$*, is given by*

$$\sigma(A) = \overline{\{\lambda_n\}} = \{\lambda_n\} \cup \{0\}$$

It is obvious how to define $Q(A)$, for a polynomial of degree d given by $Q(x) = \sum_{n=0}^{d} Q_n x^n$:

$$Q(A) = \sum_{n=0}^{d} Q_n A^n \tag{16.1}$$

The spectral theorem, however, provides an alternative definition that gives $Q(A)$ in diagonal form:

$$Q(A) = \sum_{n=0}^{\infty} Q(\lambda_n) P_n \tag{16.2}$$

Exercise 16.2 a) *Suppose* $\dim H = \infty$. *Prove that the series (16.2) is convergent in norm iff* $Q(0) = 0$.

b) *In general, i.e., for an arbitrary value* $Q(0)$*, prove that the series (16.2) is convergent in the strong operator topology.*

Exercise 16.3 *Prove that (16.1) and (16.2) indeed define the same operator* $Q(A)$.

For a continuous function $f \in C([-\|A\|, \|A\|])$, we also have two options to define $f(A)$. We can make use of a uniformly convergent sequence of polynomials $Q_N \to f$, or we can directly extend (16.2) to continuous functions:

$$f(A) = \sum_{n=0}^{\infty} f(\lambda_n) P_n \tag{16.3}$$

Exercise 16.4 *Prove that the series in (16.3) is convergent in the strong operator topology for any $f \in C([-\|A\|, \|A\|])$, and that it converges uniformly if in addition $f(0) = 0$.*

To prove that the first option also works one needs the following property:

Exercise 16.5 *Let $f \in C([-\|A\|, \|A\|])$, and consider $f(A)$ defined by (16.3). Prove that*

$$\|f(A)\| = \sup\{|f(\lambda_n)| \mid n\}$$

Let (Q_N) be a sequence of polynomials of degree N, converging uniformly to f on $[-\|A\|, \|A\|]$. The existence of such a sequence is a consequence of the Weierstrass Approximation Theorem.

Exercise 16.6 *Prove that $(Q_N(A))$ converges in norm, and that its limit equals $f(A)$ as defined in (16.3).*

Exercise 16.7 *Let A be a self-adjoint compact operator on a Hilbert space $\mathcal{H}$, and let $f : [-\|A\|, \|A\|] \to \mathbb{C}$ be a continuous function. When exactly is $f(A)$ compact?*

Exercise 16.8 *Let A be a self-adjoint compact operator on a Hilbert space $\mathcal{H}$. Prove that*

$$e^{iA} = \cos A + i \sin A$$

Exercise 16.9 *Let A be a non-negative self-adjoint compact operator on a Hilbert space $\mathcal{H}$, i.e., for all $x \in \mathcal{H}$*

$$\langle x, Ax \rangle \geq 0$$

a) *Prove that there is a unique non-negative self-adjoint operator $B \in \mathcal{B}(\mathcal{H})$ such that $B^2 = A$, i.e., $B = A^{1/2}$.*
b) *Prove that $A^{1/2}$ is also compact.*

16.2 Matched filter

The following is a common variational problem on a Hilbert space $\mathcal{H}$: given a vector $0 \neq h \in \mathcal{H}$, and a positive operator W, i.e., $\langle x, Wx \rangle > 0$, for all $0 \neq x \in \mathcal{H}$, find all $0 \neq g^* \in \mathcal{H}$ such that

$$\frac{|\langle g^*, h \rangle|^2}{\langle g^*, Wg^* \rangle} = \sup_{0 \neq g \in \mathcal{H}} \frac{|\langle g, h \rangle|^2}{\langle g, Wg \rangle} \tag{16.4}$$

In the case of compact W we can solve this problen using the spectral theorem for s.a. compact operators. But let us first give an example of a concrete instance where this kind of variational problem arises.

Consider a *filter* L, acting on *signals* $S(t)$ as follows:

$$(LS)(t) = \int_0^t g(s) S(t-s)\, ds$$

Unfortunately, the input to the filter is expected to be contaminated by random *noise* $N(\omega, t)$, where ω takes values in a probability space Ω, with a probability measure $P(d\omega)$. Assume that both signal and noise are given by elements of $L^2([0,T])$. The standard definition of the *signal-to-noise ratio* Q, is

$$Q = \frac{\Sigma^2}{\sigma^2}$$

where

$$\Sigma^2 := (LS)(T)^2 = \left| \int_0^T g(s) S(t-s)\, ds \right|^2$$

is the square of the signal observed through the filter L at time T, and

$$\sigma^2 := \mathbb{E}[|(LN)(T)|^2]$$

is the same quantity for the noise averaged over the probability measure ($\mathbb{E}[\cdot]$ denotes the expectation with respect to the probability measure $P(d\omega)$), i.e.,

$$\begin{aligned}
\sigma^2 &= \int_\Omega P(d\omega) \int_0^T ds_1 \int_0^T ds_2\, g(s_1) g(s_2) \overline{N(\omega, T - s_1)} N(\omega, T - s_2) \\
&= \int_0^T ds_1 \int_0^T ds_2\, g(s_1) g(s_2) w(s_1, s_2)
\end{aligned}$$

with

$$w(s_1, s_2) := \int_\Omega P(d\omega) \overline{N(\omega, T - s_1)} N(\omega, T - s_2)$$

Q is a measure of the contrast between signal and noise (a good filter has a high Q, i.e., amplifies the signal and/or reduces the noise). Alternatively, ω could have the interpretation of a frequency.

Define an operator W on $L^2([0,T])$ by

$$Wg(t) = \int_0^T w(t,s)g(s)\,ds \tag{16.5}$$

Exercise 16.10 *Prove that W defined in (16.5) is a non-negative, self-adjoint Hilbert-Schmidt operator, if $N(\cdot,\cdot) \in L^2(\Omega \times [0,T])$.*

The signal-to-noise ratio can now be expressed as

$$Q = \frac{|\langle g, S(T-\cdot)\rangle|^2}{\langle g, Wg\rangle}$$

To find a filter (determined by $g \in L^2([0,T])$) that maximizes Q we need to solve a variational problem of the type (16.4).

We assume that W is invertible (does not have 0 as an eigenvalue). Then, $W^{1/2}$ is invertible too, and we can vary over $f = W^{1/2}g$ instead of g.

$$\frac{|\langle g,h\rangle|^2}{\langle g, Wg\rangle} = \frac{|\langle W^{1/2}g, W^{-1/2}h\rangle|^2}{\langle W^{1/2}g, W^{1/2}g\rangle} = \frac{|\langle f,k\rangle|^2}{\langle f,f\rangle}$$

where $k = W^{-1/2}h$ (which we assume to be well-defined). So we have to find f^* such that

$$\frac{|\langle f^*,k\rangle|^2}{\langle f^*,f^*\rangle} = \sup_{0\neq f\in\mathcal{H}} \frac{|\langle f,k\rangle|^2}{\langle f,f\rangle}$$

The Cauchy-Schwarz inequality gives

$$\frac{|\langle f,k\rangle|^2}{\langle f,f\rangle} \le \langle k,k\rangle$$

for all $f \in L^2([0,T])$, with equality iff $f = ck$, $c \in \mathbb{C}$. Therefore, the solutions of (16.4) are given by

$$g^* = W^{-1/2}f^* = cW^{-1/2}k = cW^{-1}h$$

If one knows the eigensystem of W, g^* can be given more explicitly as:

$$g^* = c\sum_n (\lambda_n)^{-1}\langle e_n, h\rangle e_n$$

16.3 The mean ergodic theorem

The following theorem, due to von Neumann, is not an application of the spectral theorem, but it is another good example of Hilbert space methods.

Theorem 16.1 *Let U be a unitary operator on a Hilbert space $\mathcal{H}$, and let P be the orthogonal projection on the vectors $x \in \mathcal{H}$, invariant under U, i.e., such that $Ux = x$. Then, for all $x \in \mathcal{H}$ one has*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n x = Px \tag{16.6}$$

i.e., the average of the U^n converges strongly to P.

Proof: Consider the orthogonal decomposition of $\mathcal{H}$ determined by P:

$$\mathcal{H} = \ker P \oplus \mathrm{ran} P$$

(16.6) trivially holds for $x \in \mathrm{ran}P$.

We can complete the proof by showing that (16.6) also holds for $x \in \ker P$. Note that $\mathrm{ran}P = \ker(\mathbb{1} - U)$, and hence $\ker P = \ker(\mathbb{1} - U)^\perp = \ker(\mathbb{1} - U^*)^\perp = \overline{\mathrm{ran}(\mathbb{1} - U)}$, where we have used the fact that for a unitary U, $U^{-1} = U^*$ (which implies that $Ux = x \Leftrightarrow U^*x = x$), and also Lemma 13.5. So, any $x \in \ker P$, can be approximated by vectors of the form $(\mathbb{1} - U)y$, $y \in \mathcal{H}$. For such $x = (\mathbb{1} - U)y$, we have

$$\begin{aligned}
\left\| \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n x \right\| &= \left\| \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n (\mathbb{1} - U) y \right\| \\
&= \lim_{N\to\infty} \frac{1}{N+1} \| y - U^{N+1} y \| = 0
\end{aligned}$$

If $x_k \to x \in \ker P$, with $x_k \in \mathrm{ran}(\mathbb{1} - U)$, for all k, we then have

$$\begin{aligned}
\left\| \lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} U^n x \right\| &\le \limsup_{N\to\infty} \left\| \frac{1}{N+1} \sum_{n=0}^{N} U^n (x - x_k) \right\| \\
&\quad + \limsup_{N\to\infty} \left\| \frac{1}{N+1} \sum_{n=0}^{N} U^n x_k \right\| \\
&\le \| x - x_k \| + 0
\end{aligned}$$

As k is arbitrary and $x_k \to x$,this proves that (16.6) holds for any $x \in \ker P$. It follows easily then, that (16.6) holds for all $x \in \mathcal{H}$, by using the orthogonal decomposition $x = x_0 + x_1$, with $x_0 \in \ker P$, and $x_1 \in \mathrm{ran}P$. ∎

A class of unitaries for which one applies this theorem is given in terms of bijective maps T of a probability space $(\Omega, P(d\omega))$ into itself, that satisfy

$$\int_A P(d\omega) = \int_{T^{-1}(A)} P(d\omega)$$

where $T^{-1}(A) := \{\omega \in \Omega \mid T(\omega) \in A\}$. Such T are called *measure preserving*. An operator $U : L^2(\Omega, P(d\omega)) \to L^2(\Omega, P(d\omega))$ is then defined by

$$Uf(\omega) = f(T(\omega)) \tag{16.7}$$

and one has

$$\int_\Omega Uf(\omega)P(d\omega) = \int_\Omega f(\omega)P(d\omega)$$

for all $f \in L^2(\Omega, P(d\omega))$.

Exercise 16.11 *Prove that U, defined in (16.7), is unitary.*

Definition 16.2 *The measure preserving bijection T is called ergodic if the only functions $f \in L^2(\Omega, P(d\omega))$, satisfying $f(\cdot) = f(T(\cdot))$, are the constant functions.*

For ergodic T time averages and phase-space averages coincide:

Theorem 16.3 *If T is ergodic then, for all $f \in L^2(\Omega, P(d\omega))$*

$$\lim_{N\to\infty} \frac{1}{N+1} \sum_{n=0}^{N} f(T^n(\cdot)) = \int_\Omega f(\omega)P(d\omega) \tag{16.8}$$

where the convergence is in the L^2 norm.

Proof: The proof is a direct consequence of Theorem 16.1 applied to the unitary U defined in (16.7). ∎

Compare this theorem with Theorem 9.1. There is another ergodic theorem (Birkhoff's ergodic theorem) stating that the averages on the LHS of (16.8) converge $P-$*almost everywhere* to the constant on the RHS for all $f \in L^1(\Omega, P(d\omega))$. This theorem is a bit harder to prove however.

Note that for (16.8) to hold for all f it is also necessary that T is ergodic.

17 Unbounded linear operators and Green's functions

We have noted before that differential operators on a (non-trivial) normed function space cannot be bounded. Clearly, differential operators are important for the study of differential equations and we would like to study them in detail in spite of their lack of continuity. There are two main approaches to this problem, which are the main themes of this quarter. One is to consider differential operators as acting on topological function spaces (necessarily not normed) designed such that the relevant operators become continuous. The other approach, which we will follow in this chapter, is to consider special classes of unbounded operators (e.g., self-adjoints) defined on a dense linear subspace of a Hilbert space. Often, the important differential operators will have bounded inverses (defined in terms of their "Green's function") and this will help in the analysis.

17.1 Definitions: domain, adjoint, selfadjoint

We will restrict ourselves to linear operators defined on a dense linear subspace of a Hilbert space $\mathcal{H}$, although unbounded operators on a Banach space can be considered along the same lines.

The definition of a (not necessarily bounded) linear operator A on $\mathcal{H}$ includes the definition of its *domain* $\mathcal{D}(A)$, which is assumed to be a linear subspace of $\mathcal{H}$. So, a linear operator is a linear map $A : \mathcal{D}(A) \to \mathcal{H}$. If $\mathcal{D}(A)$ is *dense* in $\mathcal{H}$, we say that A is *densely defined*. We will usually assume that $\mathcal{D}(A)$ is dense. If it is not, we can always define A to be 0 on $\mathcal{D}(A)^{\perp}$. So, this is without loss of generality.
Example: Let L be the differential operator with domain $\mathcal{D}(L) = C^2([0,1]) \subset L^2([0,1])$, and $Lf = f''$, for all $f \in \mathcal{D}(L)$. As we will see lateron, there are other, more useful, choices of the domain on which to define the second derivative as an operator on L^2.

An operator B is called an *extension* of A if $\mathcal{D}(A) \subset \mathcal{D}(B)$, and $Bx = Ax$, for all $x \in \mathcal{D}(A)$.

Theorem 17.1 *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces, and let A be a linear map $\mathcal{D}(A) \to \mathcal{H}_2$ with $\mathcal{D}(A)$ a linear subspace of $\mathcal{H}_1$. If A is bounded with norm (least bound) $\|A\|$, then there is a unique bounded extension $\tilde{A}$ of A with $\mathcal{D}(\tilde{A}) = \overline{\mathcal{D}(A)}$, and one has $\|\tilde{A}\| = \|A\|$.*

Note: For this reason if $\mathcal{D}(A)$ is dense and A is bounded we can equally well define A on all of $\mathcal{H}_1$, i.e., A is a bounded linear transformation $\mathcal{H}_1 \to \mathcal{H}_2$.

The *adjoint* of a linear operator A with dense domain is defined as follows. For any $y \in \mathcal{H}$ consider the linear functional on $\mathcal{D}(A)$ defined by

$$\varphi_y(x) = \langle y \mid Ax \rangle$$

As $\mathcal{D}(A)$ is assumed to be dense, by the previous theorem, if φ_y is bounded it can be uniquely extend to a bounded linear functional on $\mathcal{H}$. By the Riesz Representation Theorem we then have a unique vector $z \in \mathcal{H}$ satisfying

$$\langle y, Ax \rangle = \langle z, x \rangle, \qquad \text{for all } x \in \mathcal{D}(A)$$

The adjoint of A, denoted by A^*, is then the linear operator with

$$\mathcal{D}(A^*) = \{ y \in \mathcal{H} \mid \varphi_y \text{ is bounded } \}$$

and $A^* y = z$ for all $y \in \mathcal{D}(A^*)$.

As before, we call an operator A *self-adjoint* if $A^* = A$, but now this includes the equality of their domains: $A^* = A$ iff $\mathcal{D}(A^*) = \mathcal{D}(A)$, and $A^* x = Ax$, for all $x \in \mathcal{D}(A)$.

Note that it is possible that A^* exists but one cannot define A^{**} because $\mathcal{D}(A^*)$ is not dense.

Exercise 17.1 *Prove that if A^{**} exists, then it is an extension of A.*

Examples:
1) Take $\mathcal{H} = l^2(\mathbb{N})$, and define an operator A as follows:

$$\mathcal{D}(A) = \{ (z_n)_{n=1}^{\infty} \in l^2(\mathbb{N}) \mid \text{ only finitely many of the } z_n \neq 0 \}$$

and

$$A(z_1, z_2, z_3, \ldots) = (z_1, 2z_2, 3z_3, \ldots)$$

Then A^* is an extension of A with domain

$$\mathcal{D}(A^*) = \{ (z_n)_{n=1}^{\infty} \in l^2(\mathbb{N}) \mid \sum_{n=1}^{\infty} n^2 |z_n|^2 < \infty \}$$

A^* is self-adjoint, i.e., $A^{**} = A^*$.
2) Take $\mathcal{H} = L^2(T^1)$. Define an operator A with domain

$$\mathcal{D}(A) = H^2(T^1) = \{ f(x) = \sum_{n \in \mathbb{Z}} a_n e^{inx} \mid \sum_{n \in \mathbb{Z}} n^4 |a_n|^2 < \infty \}$$

and, for any $f(x) = \sum_{n \in \mathbb{Z}} a_n e^{inx}$,

$$Af(x) = \sum_{n \in \mathbb{Z}} n^2 a_n e^{inx}$$

A is just one of the possible definitions of $-d^2/dx^2$ as an operator on $L^2(T^1)$. We now verify that A is self-adjoint. For any $g \in \mathcal{D}(A^*)$, $g(x) = \sum_n b_n e^{inx}$, we should have

$$\langle g, Af \rangle = \sum_n \overline{b_n} n^2 a_n$$

which is a bounded linear functional of f iff $(n^2 b_n) \in l^2(\mathbb{Z})$, due to the Riesz Representation Theorem, i.e., iff $g \in \mathcal{D}(A)$. This proves $\mathcal{D}(A^*) = \mathcal{D}(A)$. Clealry we also have $A^* g = Ag$ for all $g \in \mathcal{D}(A)$.

Note that A is also an extension of $-d^2/dx^2$ with domain $C^2(T^1)$.

We will come back to the problem of finding self-adjoint extensions (restrictions) of more general differential operators.

17.2 Inverses

One is often interested in solving linear equations of the form

$$Ax = y \tag{17.1}$$

for given y.

If $\dim \mathcal{H} < \infty$ the conditions for unique solvability of (17.1) are given by the so-called *Fredholm Alternative*:

A. $\ker A^* = \{0\}$. This is equivalent to $\ker A = \{0\}$ (finite dimension! See Exercise 17.3), in which case A is invertible, i.e., there is a unique linear map A^{-1} such that $AA^{-1} = \mathbb{1} = A^{-1}A$, and, hence, (17.1) is solved by $x = A^{-1}y$, for all $y \in \mathcal{H}$.

B. $\ker A^* \neq \{0\}$. (17.1) can be solved iff $y \in \mathrm{ran} A$, by the definition of $\mathrm{ran} A$. As $\mathrm{ran} A = (\ker A^*)^\perp$, this can be expressed as $\langle z, y \rangle = 0$ for all z such that $A^* z = 0$. In this case, however, the solution is not unique: if x_0 is a solution, i.e, $Ax_0 = y$, then all $x \in x_0 + \ker A$ are solutions, too (and these are exactly all solutions).

Exercise 17.2 *Prove that, if* $\dim \mathcal{H} < \infty$, $\dim \ker A + \dim \mathrm{ran} A = \dim \mathcal{H}$.

Exercise 17.3 *Prove that, if* $\dim \mathcal{H} < \infty$, $\dim \ker A = \dim \ker A^*$. *In particular,* $\ker A = \{0\}$ *iff* $\ker A^* = \{0\}$.

If $\dim \mathcal{H} = \infty$, things are not so simple. We study this problem in detail for differential operators in the next section.

As a matter of illustration, do the following exercise:

Exercise 17.4 *Consider the right shift S on $\mathcal{H} = l^2(\mathbb{N})$. Prove that* $\ker S = \{0\}$, *but* $\mathrm{ran} S$ *is not dense in $\mathcal{H}$. Hence, there cannot be a densely defined inverse S^{-1} such that $SS^{-1} = \mathbb{1}$.*

17.3 Differential operators

A linear differential operator of order n is a linear map L defined on a suitable function space that acts on $f \in C^n$ by

$$Lf = \sum_{j=0}^{n} a_j(x) f^{(j)}(x)$$

Here, $f^{(j)}$ denotes the jth derivative of f, and the $a_j(x)$ are complex-valued functions. For concreteness we assume that all functions are defined of the interval $[0, 1]$. If the coefficients a_j are polynomials, then L maps polynomials into polynomials. If the coefficients are C^∞ functions, L transforms C^∞ into itself. But it is often necessary to consider L as an operator defined on various other dense domains in L^2.

Our goal is to study ODE's of the form

$$Lu = f \tag{17.2}$$

where we want to find the solutions u for given f. First, we will look for conditions for unique solvability of (17.2) (conditions for the existence of a unique Green's function). Then we will learn how to find, i.e., construct, the Green's function and the solutions.

17.3.1 The Green's function for boundary value problems

In this section, we will limit ourselves to 2nd order differential operators L acting on function $u : [0, 1] \to \mathbb{C}$:

$$Lu = au'' + bu' + cu \tag{17.3}$$

Assume that $a(x) > 0$ for all $x \in [0, 1]$, and that a, b, and c are continuous. Then, the ODE

$$Lu = 0$$

has a two-dimensional space of C^2 solutions

$$u = \alpha_1 u_1 + \alpha_2 u_2, \quad \alpha_1, \alpha_2 \in \mathbb{C}$$

Obviously, if we do not impose extra conditions on the solutions, (17.3) will have at least a two-dimensional space of solutions if it has a solution at all. Unique solutions are obtained by imposing boundary conditions, e.g.,

Dirichlet:

$$u(0) = u(1) = 0,$$

Neumann:

$$u'(0) = u'(1) = 0,$$

periodic:

$$\begin{aligned} u(0) &= u(1) \\ u'(0) &= u'(1), \end{aligned}$$

mixed:

$$\begin{gathered} \alpha_0 u(0) + \beta_0 u'(0) = 0 \\ \alpha_1 u(1) + \beta_1 u'(1) = 0. \end{gathered}$$

where $\alpha_0, \alpha_1, \beta_0, \beta_1$ are complex constants.

Instead of imposing conditions at both endpoints, one can also impose two conditions at one of the endpoints. E.g., $u(0) = u'(0) = 0$. This is then usually called *initial conditions*.

The boundary conditions given above are *homogeneous* boundary conditions. For linear ODE, non-homogeneous boundary conditions can be reduced to homogeneous ones by defining $v(x) = u(x) - u_0 - (u_1 - u_0)x$. If v is then a solution of the homogeneous problem

$$\begin{aligned} Lv &= f \\ v(0) &= 0 \\ v(1) &= 0 \end{aligned}$$

u will solve the non-homogeneous system

$$\begin{aligned} Lu &= f \\ u(0) &= u_0 \\ u(1) &= u_1 \end{aligned}$$

Before we commence the rigorous discussion of Green's functions of differential operators + boundary conditions, we first explain the basic idea behind it. For concreteness, let us consider Dirichlet boundary condtions, i.e., we want to solve

$$\begin{aligned} Lu &= f \\ u(0) &= 0 \\ u(1) &= 0 \end{aligned} \tag{17.4}$$

As the inverse operator of taking the derivative can be expressed as an integral operator with kernel function $\mathrm{Ind}[y \le x]$, it is natural to try to find the inverse of a differential operator in the form of an integral operator. I.e., we look for a function $g : [0,1] \times [0,1] \to \mathbb{C}$ such that the operator G defined by

$$Gf(x) = \int_0^1 g(x,y)f(y)\,dy \tag{17.5}$$

is the inverse of L, i.e.,

$$LG = \mathbb{1}, \qquad \text{on } L^2([0,1]) \tag{17.6}$$

(and $GL = \mathbb{1}$ on $\mathcal{D}(L)$).

Using (17.5) and (17.6) we obtain the following condition on g:

$$L\int_0^1 g(x,y)f(y)\,dy = f(x), \qquad \text{for all } f \tag{17.7}$$

In order to have Gf satsify the Dirichlet boundary conditions we also require that, for all $y \in [0,1]$, $g(0,y) = g(1,y) = 0$. Then, for any f, the solution u of (17.4) can be written as

$$u(x) = \int_0^1 g(x,y)f(y)\,dy$$

There are several ways of rewriting the condition (17.7). One is to make use of the notion of delta function, and write a differential equation for g in distributional sense:

$$Lg(x,y) = \delta(x-y)$$

Here we consider L to be acting on g as a function of its first variable.

We will study distributions later, but in any case, at first, it is preferable to formulate an ordinary ODE for g.

Proposition 17.2 *Let L be given by (17.3) with $a \in C^1([0,1])$, $a(x) > 0$, for all $x \in [0,1]$, and $b, c \in C([0,1])$. Suppose $g : [0,1] \times [0,1] \to \mathbb{C}$ is continuous and, for all $y \in [0,1]$, g is C^1 in x on $x < y$ and $x > y$, and such that*

$$Lg = 0, \qquad \textit{for all } x \neq y \tag{17.8}$$

$$a(y)\lim_{\epsilon \downarrow 0}\left[g_x(x, x+\epsilon) - g_x(x, x-\epsilon)\right] = 1 \tag{17.9}$$

$$g(0,y) = g(1,y) = 0 \tag{17.10}$$

Then G given by (17.5) satisfies $LG = \mathbb{1}$ on $L^2([0,1])$, and for any $f \in C([0,1])$, Gf satisfies the Dirichlet boundary conditions and is a solution of (17.4).

Proof: The proof is by direct computation. The only non-trivial part is to check that

$$u(x) := \int_0^1 g(x,y)f(y)\,dy$$

satisfies $Lu = f$. From the ODE, and the fact that a is continuous and striclty positive, it is easy to see that g will be C^2 where it is C^1. $g_x x$ is therefore a.e. well-defined, and, ignoring the arguments where it is not, $g_x x$ is integrable as a function of two variables.

$$\begin{aligned} Lu &= L\int_0^1 g(x,y)f(y)\,dy \\ &= \left[a(x)\frac{d^2}{dx^2} + b(x)\frac{d}{dx} + c(x)\right]\left[\int_0^x g(x,y)f(y)\,dy + \int_x^1 g(x,y)f(y)\,dy\right] \end{aligned}$$

To compute the derivatives we use Lemma 17.3 given below. The result is:

$$\int_0^1 Lg(x,y)f(y)\,dy + a(x)[g_x(x,x-0)f(x-0) - g_x(x,x+0)]f(x+0)$$
$$= 0 + a(x)[g_x(x+0,x) - g_x(x-0)]f(x) = f(x)$$

by the assumptions. ■

Lemma 17.3 *If $h(x,y)$ is a continuous function of two variables, and has an integrable almost everywhere defined partial derivative with repsect to x, and a and b are C^1, then*

$$\frac{d}{dx}\int_{a(x)}^{b(x)} h(x,y)\,dy = b'(x)h(x,b(x)) - a'(x)h(x,a(x)) + \int_{a(x)}^{b(x)} \frac{\partial}{\partial x}h(x,y)\,dy$$

Exercise 17.5 *Prove Lemma 17.3.*

17.3.2 Construction of the Green's function

We start from some basic facts about the solutions of linear 2nd order ODE. When a, b, c, f are continuous functions and $a(x) \neq 0$ on an interval (a, b) containing x_0, the homogeneous ODE

$$au'' + bu' + cu = 0$$

has a two-dimensional space of solutions, i.e., spanned by two linearly independent solutions. As long as $a(x)$ does not vanish the Wronskian of two independent solutions does not vanish either:

$$W(x) = \det\begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_1'(x) \end{pmatrix} \neq 0$$

Also, under the same assumptions, the ODE with b.c.

$$au'' + bu' + cu = f \tag{17.11}$$
$$u(x_0) = U, u'(x_0) = V \tag{17.12}$$

has a unique solution $u(x)$ on the interval (a, b).

In order to construct g satisfying the conditions of Proposition 17.2 we start from a pair of linearly independent solutions v_0 and v_1 such that

$$v_0(0) = 0, \text{ and } v_1(1) = 0$$

Such a pair exists iff the homogeneous Dirichlet problem, i.e., (17.11-17.12) with $f = U = V = 0$, has only the solution $u \equiv 0$.

Exercise 17.6 *Prove the preceding statement.*

The Green's function can now be found in the following form:

$$g(x, y) = C(y) \times \begin{cases} v_0(x)v_1(y) & \text{if } x < y \\ v_0(y)v_1(x) & \text{if } x > y \end{cases} \tag{17.13}$$

The function $C(y)$ is determined by (17.9):

$$C(y) = \frac{1}{a(y)W(y)}$$

It is then clear that g is continuous, satisfies $Lg = 0$ whenever $x \neq y$, and $g(0, y) = g(1, y) = 0$.

To summarize, we found that, if the homogeneous Dirichlet problem has only the zero solution, g defined by has all the properties required in Proposition 17.2.

Example: The stationary temperature distribution in a rod of unit length, with both ends kept at a constant temperature of 0 degrees, with heat loss through its surface given by $\alpha^2 u$, and subject to a non-uniform heat source desrbed by f, is the solution of

$$-u'' + \alpha^2 u = f \tag{17.14}$$

For the construction of the Green's function we need two linearly independent solutions of (17.14) with $f = 0$, v_0 and v_1, satisfying $v_0(0) = 0$, and $v_1(1) = 0$. The general solution of the homogeneous equation is of the form

$$c_+ e^{\alpha x} + c_- e^{-\alpha x}$$

For v_0 and v_1 we choose

$$\begin{aligned} v_0(x) &= \frac{e^{\alpha x} - e^{-\alpha x}}{2} = \sinh \alpha x \\ v_1(x) &= \frac{-e^{-\alpha} e^{\alpha x} + e^{\alpha} e^{-\alpha x}}{2} = \sinh \alpha(1 - x) \end{aligned}$$

the Wronskian turns out to be a constant $W \equiv \alpha \sinh \alpha$. The Green's function is then

$$g(x, y) = \begin{cases} \frac{\sinh \alpha x \sinh \alpha(1-y)}{\alpha \sinh \alpha} & \text{if } 0 \leq x \leq y \leq 1 \\ \frac{\sinh \alpha y \sinh \alpha(1-x)}{\alpha \sinh \alpha} & \text{if } 0 \leq y \leq x \leq 1 \end{cases}$$

17.3.3 Self-adoint differential operators and Sturm-Liouville operators

In this section we will consider self-adjoint differential operators and their Green's functions in some more detail. We start with a formula due to Green for the action of the adjoint L^* of a differential operator L on sufficiently smooth functions (C^n functions if n is the order of the differential operator).

Lemma 17.4 *Let L be a linear operator such that for all $f \in \mathcal{D}(L) \cap C^n([0,1])$, $Lf = (d/dx)^n f$. Then, for all $f, g \in \mathcal{D}(L) \cap C^n([0,1])$,*

$$\langle g, Lf\rangle = (-1)^n \langle Lg, f\rangle + \left[g f^{(n-1)} - g^{(1)} f^{(n-2)} + \cdots + (-1)^{n-1} g^{(n-1)} f\right]_0^1$$

where, for any function h, $h^{(k)}$ denotes the kth derivative of h.

Proof: The proof is nothing more than integration by parts. ■

In particular the operator $D = d/dx$ on the domain

$$\mathcal{D}(D) = \{f \in C^1([0,1]) \mid f(0) = f(1) = 0\}$$

satisfies $\mathcal{D}(D) \subset \mathcal{D}(D^*)$ and for all $f \in \mathcal{D}(D)$

$$D^* f = -Df$$

And similarly for $D^2 = (d/dx)^2$ on the domain

$$\mathcal{D}(D^2) = \{f \in C^2([0,1]) \mid f(0) = f(1) = 0\}$$

satisfies $\mathcal{D}(D^2) \subset \mathcal{D}((D^2)^*)$ and for all $f \in \mathcal{D}(D^2)$

$$(D^2)^* f = D^2 f$$

By combining these relations with the following exercises we can compute the action of the adjoint of any 2nd order differential operator with sufficiently smooth coefficients on sufficiently smooth functions.

Exercise 17.7 *Let A and B be two linear operator on a Hilbert space $\mathcal{H}$ with domains $\mathcal{D}(A)$ and $\mathcal{D}(B)$ respectively, and assume $\mathcal{D}(A) \cap \mathcal{D}(B) \neq \emptyset$. Define an operator C by $\mathcal{D}(C) = \mathcal{D}(A) \cap \mathcal{D}(B)$ and $Cx = Ax + Bx$, for all $x \in \mathcal{D}(C)$. Prove that C^* is an extension of $A^* + B^*$.*

Exercise 17.8 *Let A and B be two linear operator on a Hilbert space $\mathcal{H}$ with domains $\mathcal{D}(A)$ and $\mathcal{D}(B)$ respectively. Define $\mathcal{D}(AB)$ and $\mathcal{D}(B^*A^*)$ by*

$$\begin{aligned}\mathcal{D}(AB) &= \{x \in \mathcal{D}(B) \mid Bx \in \mathcal{D}(A)\} \\ \mathcal{D}(B^*A^*) &= \{x \in \mathcal{D}(A^*) \mid A^*x \in \mathcal{D}(B^*)\}\end{aligned}$$

and assume that $\mathcal{D}(AB)$ and $\mathcal{D}(B^*A^*)$ are non-empty. Define operators AB and B^*A^* on their respective domains in the obvious way. Prove that $(AB)^*$ is an extension of B^*A^*.

Proposition 17.5 (Green's formula) *Define $\mathcal{D}(L)$ by*

$$\mathcal{D}(L) = \{u \in C^2([0,1]) \mid u(0) = u(1) = 0\}$$

Let $a \in C^2([0,1]), b \in C^1([0,1]), c \in C([0,1])$ and define

$$Lu = au'' + bu' + cu, \qquad \text{for all } u \in \mathcal{D}(L).$$

Then $\mathcal{D}(L) \subset \mathcal{D}(L^*)$ and

$$L^*u = (u\overline{a})'' - (u\overline{b})' + u\overline{c}, \qquad \text{for all } u \in \mathcal{D}(L).$$

Proof: By Exercise 17.7 we can compute the action of L^* term by term. E.g., for the first term, apply Exercise 17.8 with $A = (d/dx)^2$ and B is multiplication by the complex-valued function a. B^* is then multiplication by the complex-valued function $\overline{a}$, and the result follows. ∎

Next, we address the question which 2nd order differential operators are self-adjoint. Recall that $L^* = L$ means $\mathcal{D}(L^*) = \mathcal{D}(L)$ and $L^*u = Lu$, for all $u \in \mathcal{D}(L)$. First, we investigate the conseqences of the latter requirement:

$$\begin{aligned} L^*u &= (u\overline{a})'' - (u\overline{b})' + u\overline{c} \\ &= \overline{a}u'' + (2\overline{a}' - \overline{b})u' + (\overline{a}'' - \overline{b}' + \overline{c})u \\ &= au'' + bu' + cu \end{aligned}$$

As u, u', and u'', are in general linearly independent, we must therefore have

$$\begin{aligned} a &= \overline{a} \\ b &= 2\overline{a}' - \overline{b} \\ c &= \overline{a}'' - \overline{b}' + \overline{c} \end{aligned}$$

These relations are easily reduced to the following conditions:

$$\begin{aligned} &a \text{ is real} \\ &\mathrm{Re}\ b = a' \\ &\mathrm{Im}\ c = -\frac{1}{2}\mathrm{Im}\ b \end{aligned}$$

where, Re z and Im z denote the real and imaginary parts of a complex number z, respectively. So, the coefficients of a self-adjoint L are determined by three real functions: a, Im b, and Re c. If one restricts to real coefficients, there are only two real functions as independent parameters, traditional called p and q:

$$\begin{aligned} a &= -p \\ b &= -p' \\ c &= q \end{aligned}$$

In this way we arrive at the standard form for a *Sturm-Liouville* operator:

$$Lu = -(pu')' + qu$$

17.4 The spectral Theorem for Sturm-Liouville operators

The standard Sturm-Liouville problem we considered so far is determined by two real functions, p and q, about which we assume that $p \in C^1([0,1])$, $q \in C([0,1])$, $p(x) > 0$ for all $x \in [0,1]$. The corresponding eigenvalue problem is then:

$$\begin{aligned} Lu := -(pu')' + qu &= \lambda u \\ u \in \mathcal{D}(L) &= \{u \in H^2([0,1]) \mid u(0) = u(1) = 0\} \end{aligned} \tag{17.15}$$

If λ_0 is not an eigenvalue, i.e., the above problem with $\lambda = \lambda_0$ has only the $u \equiv 0$ solution, then by the previous paragraphs, the Green's function g of $L - \lambda_0 \mathbb{1}$ exists and is unique, and we can define

$$Gu(x) = \int_0^1 g(x,y)u(y)\,dy; \qquad G : L^2 \to \mathcal{D}(L)$$

As g is continuous, real, and symmetric, G is self-adjoint and compact, and hence has a o.n. basis $\{u_n\}$ of eigenvectors, with eigenvalues g_n. As $LG = \mathbb{1}$ we must have that $u_n \in \mathcal{D}(L)$, and u_n is and eigenvector of L with eigenvalue λ_n satisfying $\lambda_n - \lambda_0 = g_n^{-1}$. It is also easy to see that (17.15) cannot have other linearly independent solutions. Thus we have proved a spectral theorem for Sturm-Liouville operators.

17.5 Singular Sturm-Liouville problems

The Sturm-Liouville operators considered above are called *regular*. In the construction of the Green's function we have used that p (or a) does not vanish. If p does vanish somewhere in the bounded interval one is considering, or if the interval is unbounded, the Sturm-Liouville problem is said to be *singular*. For singular problems it can be more difficult to prove that G is compact, or even it may not be true. Here, we limit ourselves to a few examples.

Example 1: $p(x) = (x^2 - 1)$ vanishes at the boundaries of the interval $[-1,1]$. The corresponding S-L eigenvalue problem has the Legendre polynomials as its solutions.

Example 2: An example of a problem on the whole of $\mathbb{R}$ with a compact G is the quantum harmonic oscillator:

$$\begin{aligned} Lu &= -u'' + x^2 \\ \mathcal{D}(L) &= \{u \in H^2(\mathbb{R}) \mid \int_{\mathbb{R}} x^4 |u(x)|^2\,dx < \infty\} \end{aligned}$$

Its eigenvectors are the Hermite functions. As the corresponding G is self-adjoint and compact we obtain for free that the Hermite functions as an orthormal set in $L^2(\mathbb{R})$ are complete.

Example 3: Finally, here is a Sturm-Liouville operator on $L^2(\mathbb{R})$ with a non-compact inverse:

$$\begin{aligned} Lu &= -u'' + u \\ \mathcal{D}(L) &= H^2(\mathbb{R}) \end{aligned}$$

Its Green's operator is given by

$$Gu(x) = \int_{\mathbb{R}} e^{-|x-y|} u(y)\, dy$$

In fact it is not hard to see that the spectrum of L (and G) is $[1, +\infty)$.

17.6 Applications of Sturm-Liouville problems

.

We briefly mention some applications of Green's functions and Sturm-Liouville problems. We refer to the literature for many more applications in the theory of PDE.

17.6.1 PDE: initial value problems

This is the application we mentioned at the beginning of this chapter as our main motivation for the study of the eigenvalue problem for unbounded operators. A PDE initial value problem of the form

$$\begin{aligned} u_t &= Lu \\ u(x, 0) &= f(x) \end{aligned}$$

$x \in [a, b]$, $t \in \mathbb{R}$, has solutions of the form

$$u(x,t) = \sum_n e^{\lambda_n t} \langle u_n, f \rangle e_n(x)$$

for f and t for which this defines a differentiable function.

17.6.2 An application in higher dimensions: elliptic boundary value problems

Higher-dimensional problems can often be reduced to lower-dimensional ones by separation of variables and the use of symmetry. But Green's functions can also be constructed for higher-dimensional problems directly. In such case we are strictly speaking no longer delaing with Sturm-Liouville problems. The main issues, however, are the same: self-adjointness, compactness of the inverse operator etc..

We limit ourselves to a particular example: the Dirichlet problem on a bounded domain Ω with smooth boundary in $\mathbb{R}^n$:

$$\begin{aligned} Lu := -\Delta u &= \lambda u && (17.16)\\ \mathcal{D}(L) &= \{u \in H^2(\Omega) \mid u(x) = 0, x \in \partial\Omega\} && (17.17) \end{aligned}$$

where $\partial\Omega$ denotes the boundary of Ω (i.e., the difference between the smallest closed set in $\mathbb{R}^n$ containing it, and the largest open set contained in it). We will return to this example later and prove that indeed the inverse of L is given by a compact self-adjoint Green's operator. As a consequence there is an o.n. basis of $L^2(\Omega)$ consisting of eigenvectors of the Laplacian and satisfying Dirichlet boundary conditions. Other boundary conditions and more general operators can be considered as well.

18 Distributions

Our next two topics are *distributions* and the *Fourier transform*. Both will allow us to study, compute, and use Green's functions more easily. In addition, both topics have many other applications.

18.1 The Schwartz space $\mathcal{S}$

Recall the definition of *seminorm*: a function $p : X \to \mathbb{R}$, where X is a linear space over $\mathbb{C}$, is a seminorm on X if it has the following three properties:

- $p(x) \geq 0$, for all $x \in X$.
- $p(x+y) \leq p(x) + p(y)$, for all $x, y \in X$.
- $p\lambda x) = |\lambda| p(x)$, for all $x \in X$, $\lambda \in \mathbb{C}$.

Let n be a positive integer. We will use the following notations: for $\alpha = (\alpha_1, \ldots, \alpha_n), \beta = (\beta_1, \ldots, \beta_n) \in \mathbb{Z}^n$, $\alpha_i, \beta_i \geq 0$, for $i = 1, \ldots, n$

$$\partial^\alpha := \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_n} \tag{18.1}$$

$$x^\alpha := \prod_{i=1}^n x_i^{\alpha_i} \tag{18.2}$$

$$\alpha + \beta := (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n) \tag{18.3}$$

$$|\alpha| := \sum_{i=1}^n |\alpha_i| = \sum_{i=1}^n \alpha_i \tag{18.4}$$

$$\alpha! := \prod_{i=1}^n \alpha_i! \tag{18.5}$$

$$\alpha \geq \beta \quad \text{iff} \quad \alpha_i \geq \beta_i, i = 1, \ldots, n \tag{18.6}$$

With these notations Taylor's remainder theorem and Leibniz' rule can be written in the following compact form: for any $f \in C^k(\mathbb{R}^n)$, $x_0 \in \mathbb{R}^n$

$$f(x) = \sum_{0 \leq |\alpha| \leq k} \partial^\alpha f(x_0) \frac{(x - x_0)^\alpha}{\alpha!} + R_k(x) \tag{18.7}$$

$$\text{with} \lim_{x \to x_0} \frac{R_k(x)}{\|x - x_0\|^k} = 0 \tag{18.8}$$

$$\partial^\alpha (fg) = \sum_{\beta + \gamma = \alpha} \frac{\alpha!}{\beta! \gamma!} \left(\partial^\beta f\right) \left(\partial^\gamma g\right) \tag{18.9}$$

The seminorms we will use in this section are defined on C^∞ functions f. For all $n \geq 0, \alpha \in (\mathbb{Z}_+)^n$, define

$$p_{n,\alpha}(f) = \sup_{x\in\mathbb{R}^n} (1 + \|x\|)^n \left|\partial^\alpha f(x)\right| \tag{18.10}$$

The space of C^∞ functions for which all the seminorms $p_{n,\alpha}$ are finite is called *Schwartz space*, and is conventionally denoted by $\mathcal{S}(\mathbb{R}^n)$. We will often simply write $\mathcal{S}$ instead of $\mathcal{S}(\mathbb{R}^n)$, and $\|f\|_{n,\alpha}$ instead of $p_{n,\alpha}$. The elements of $\mathcal{S}$ are often called *test functions*, which is less specific than *Schwartz functions.*

Exercise 18.1 *Prove that the seminorms $p_{n,\alpha}$ are in fact norms on $\mathcal{S}$.*

There are plenty of functions in $\mathcal{S}$. E.g., every functions of the form $Q(x)\exp(-c\|x-x_0\|^2)$, with Q a polynomial, $c > 0$, and $x_0 \in \mathbb{R}$, is a Schwartz function.

Exercise 18.2 *Check that $\mathcal{S}$ is an algebra for the usual operations of addition and multiplication of functions.*

In general, if $\{p_a\}_{a\in A}$ is family of seminorms on a linear space X, indexed by A, X becomes a topological vector space with the following definition of a base of open neighborhoods $\mathcal{N}$:

$$\begin{aligned}
\mathcal{N} &= \{N_{x,\{a_i\},\epsilon} \mid x \in X, \{a_i\} \subset A \text{ finite}, \epsilon > 0\} \\
N_{x,\{a_i\},\epsilon} &= \{y \in X \mid p_a(x-y) < \epsilon, \text{ for all } a \in \{a_i\}\}
\end{aligned}$$

The following three exercises are general properties topological vector spaces defined using a family of seminorms that we will need for our study of distributions.

Exercise 18.3 *Prove the following:*

a) *The addition of vectors in X and the multiplication by a scalar are continuous.*

b) *A topology defined by a family of seminorms as above has a base of convex open neighborhoods. Such a topological space is called* locally convex.

c) *If for all $x \in X$ there exists $a \in A$ such that $p_a(x) > 0$, then the topology defined by $\{p_a\}_{a\in A}$ is Haussdorff.*

If the family of seminorms is countable, the topology defined by it is *metric*, meaning that there is a metric d on X that defines the same open sets. One also says that X (or its topology) is *metrizable.*

Exercise 18.4 *Suppose that A is countable, and let $p_1, p_2, p_3, \ldots$ be an enumeration of the $\{p_a\}$.*

a) *Prove that*

$$d(x,y) = \sum_{n=1}^{\infty} 2^{-n} \frac{p_n(x-y)}{1+p_n(x-y)} \tag{18.11}$$

defines a metric on X.

b) *Prove that the topology defined by d coincides with the one defined by $\{p_a\}_{a\in A}$.*

Note that the set of seminorms $\{p_{n,\alpha}\}$ on $\mathcal{S}$ is countable. Hence, by the previous exercise, $\mathcal{S}$ is a metric space. The convergence of a sequence (x_n) to a limit $x \in X$ can be expressed in terms of the metric, but it is often more convenient to express it in terms of the seminorms, i.e., $x_n \to x$ iff for all $a \in A$, $p_a(x_n - x) \to 0$.

Exercise 18.5 *Let (x_n) be a sequence in X. Prove that (x_n) is a Cauchy sequence for the metric d defined in (18.11) iff for all $a \in A$, and $\epsilon > 0$ there exists N such that $p_a(x_n - x_m) < \epsilon$ for all $n, m \geq N$.*

In the following proposition we prove that $\mathcal{S}$ is a *complete* metric space.

Proposition 18.1 *$\mathcal{S}$ with the topology defined above is a complete metric space.*

Proof: Let (f_n) be a Cauchy sequence in $\mathcal{S}$. We have to prove that (f_n) converges for the topology on $\mathcal{S}$ to a function $f \in \mathcal{S}$. We start from the fact that the bounded continuous functions with the sup norm are complete. From the previous exercise we know that (f_n) Cauchy in the topology of $\mathcal{S}$ implies that it is a Cauchy sequence for the sup norm ($p_{0,0}$ is the sup norm). Therefore, there exists a bounded continuous function f such that $f_n \to f$ in the sup norm. Clearly, f is the candidate limit for f_n in $\mathcal{S}$. Two questions remain: is f in $\mathcal{S}$, and does (f_n) converge to f in the topology of $\mathcal{S}$?

Again by the previous exercise we have that for all α the sequence $\partial^\alpha f_n$ is also Cauchy for the sup norm, and hence converges uniformly to a function g_α. We claim that

$$g_\alpha = \partial^\alpha f, \qquad \text{for all } \alpha. \tag{18.12}$$

Then, by using the fact that f_n is Cauchy for all the norms $p_{n,\alpha}$, we see that $f \in \mathcal{S}$ and $f_n \to f$ in $\mathcal{S}$.

We prove (18.12) for all α, by induction on $|\alpha|$. It holds for $|\alpha| = 0$. Suppose we have proved (18.12) for all α with $|\alpha| \leq m$. Then, if $|\beta| = m+1$, there exist α, $|\alpha| = m$, and j such that $\beta = \alpha + e_j$, where e_j is the jth standard basis vector of $\mathbb{Z}^n$, and we have by the Fundamental Theorem of Calculus that

$$\partial^\alpha f_n(x + te_j) - \partial^\alpha f_n(x) = \int_0^t \partial^{e_j} \partial^\alpha f_n(x + se_j)\, ds$$

Clearly, $\partial^{e_j}\partial^\alpha = \partial^\beta$. By taking $n \to \infty$ we obtain

$$\partial^\alpha f(x + te_j) - \partial^\alpha f(x) = \int_0^t g_\beta(x + se_j)\, ds$$

Next, take $\lim_{t\downarrow 0}$ of $(1/t)$ times this expression and using the definition of derivative and the continuity of g_β you find

$$\partial^\beta f(x) = g_\beta(x)$$

■

Summarizing, we have found that $\mathcal{S}$ with its topology defined by the seminorms $p_{n,\alpha}$, is a locally convex metrizable complete Haussdorff space. A topological linear space with these properties is called a *Fréchet space.*

Exercise 18.6 *Let $f \in C^\infty(\mathbb{R}^n)$. Then, $f \in \mathcal{S}$ iff $x^\alpha \partial^\beta f$ is bounded for all α, β iff $\partial^\beta(x^\alpha f)$ is bounded for all α, β.*

Exercise 18.7 *Let $k \in \mathcal{S}$ and define*

$$Kf(x) = \int k(x - y) f(y)\, dy, \qquad \textit{for all } f \in \mathcal{S}$$

Prove that $K : \mathcal{S} \to \mathcal{S}$, which is called a convolution operator, is a continuous linear operator for the topology of $\mathcal{S}$.

Exercise 18.8 *For every $a \in \mathbb{R}^n$ define a linear transformation $T_a : \mathcal{S} \to \mathcal{S}$ by $T_a(f)(x) = f(x - a)$. These are the translation operators on $\mathcal{S}$.*

a) *Prove that for all $a \in \mathbb{R}^d$, T_a is continuous for the topology of $\mathcal{S}$.*

b) *Prove that for all $f \in \mathcal{S}$, the map $a \mapsto T_a f$ is continuous from $\mathbb{R}^d$ to $\mathcal{S}$. (Hint: First, prove that for $f \in C(\mathbb{R}^n)$ one has $\lim_{a\to 0} \|T_a f - f\|_{\sup} = 0$ iff f is uniformly continuous. Also, note that it is sufficient to prove continuity in $a = 0$, due to the group property of T_a.)*

18.2 The space $\mathcal{S}^*$, the tempered distributions

The space of all continuous linear functionals on $\mathcal{S}$ is denoted by $\mathcal{S}^*$ (or $\mathcal{S}'$). It is the topological dual of $\mathcal{S}$. The elements of $\mathcal{S}^*$ are called *tempered distributions.*

As $\mathcal{S}$ is a metric space, we can prove that a functional $\varphi : \mathcal{S} \to \mathbb{C}$ is continuous by verifying that for every convergent sequence $f_n \to f$ in $\mathcal{S}$ we have

$$\lim_n \varphi(f_n) = \varphi(f)$$

Example:
The prime example is the so-called delta-function (which is a bit of a misnomer because it is not a function on $\mathbb{R}^n$, but a functional on $\mathcal{S}$). For every $x_0 \in \mathbb{R}^n$ define a linear functional δ_{x_0} on $\mathcal{S}$ by

$$\delta_{x_0}(f) = f(x_0), \quad f \in \mathcal{S}$$

If $f_n \to f$ in $\mathcal{S}$, then, in particular we know that $f_n \to f$ uniformly. Therefore, $f_n(x_0) \to f(x_0)$, which proves that δ_{x_0} is continuous. The linearity of δ_{x_0} is trivial. We conclude that $\delta_{x_0} \in \mathcal{S}^*$.

The action of the elements of the dual X^* on X is often represented by a *pairing*, which can be viewed as a generalization of an inner product.

$$\langle \cdot, \cdot \rangle : \mathcal{S}^* \times \mathcal{S} : \langle \varphi, f \rangle = \varphi(f)$$

Note that, due to the Riesz Representation Theorem, in the case where $X = \mathcal{H}$ is a Hilbert space the pairing on $\mathcal{H}^* \times \mathcal{H}$ can be indentified with the inner product on $\mathcal{H}$.

A popular notation for the pairing on $\mathcal{S}^* \times \mathcal{S}$ is given by

$$\varphi(f) = \langle \varphi, f \rangle = \int \varphi(x) f(x)\, dx$$

as if $\mathcal{S}^*$ were a function space. This notation is justified by the fact that the pairing on $\mathcal{S}^* \times \mathcal{S}$ shares a number of properties with inner products defined through an integral. For the same reason it often is a convenient notation in computations, but keep in mind that it is just a notation for continuous linear functionals. The distribution δ_{x_0} is then written as

$$\delta_{x_0}(f) = \int \delta(x - x_0) f(x)\, dx = f(x_0)$$

See the representation theorem 18.2 later for further justification of this notation.
More examples:

1. Let F be a continous function on $\mathbb{R}^n$ of at most polynomial growth, i.e., there exists p and $C > 0$ such that

   $$|F(x)| \leq C(1 + \|x\|)^p$$

   Define $\varphi_F \in \mathcal{S}^*$ by

   $$\varphi_F(f) = \int_{\mathbb{R}^n} F(x) f(x)\, dx$$

2. If $K : \mathcal{S} \to \mathcal{S}$ is a continuous linear transformation of $\mathcal{S}$, and $\varphi \in \mathcal{S}^*$, then $\varphi \circ K \in \mathcal{S}^*$.

3. An important case of the previous example is $K = \partial^\alpha$. So, e.g.,

$$f \mapsto \delta_{x_0}(\partial^\alpha f) = \partial^\alpha f(x_0)$$

is a functional in $\mathcal{S}^*$. The following is a standard notation inspired by the Integration by Parts formula for true integrals:

$$\begin{aligned}
\delta_{x_0}(\partial^\alpha f) &= \int \delta(x - x_0)\partial^\alpha f(x)\, dx \\
&= (-1)^{|\alpha|} \int (\partial^\alpha \delta)(x - x_0) f(x)\, dx \\
&= (-1)^{|\alpha|} \int \delta^{(\alpha)}(x - x_0) f(x)\, dx
\end{aligned}$$

So, in one dimension ($n = 1$), δ'_{x_0} denotes the functional on $\mathcal{S}(\mathbb{R})$ defined by

$$\delta'_{x_0}(f) = -f'(x_0)$$

and similarly

$$\delta''_{x_0}(f) = f''(x_0)$$

This notation can be generalized to arbitrary tempered distributions: one defined the parial derivative $\partial^\alpha \varphi$, for any $\varphi \in \mathcal{S}^*$ by

$$\partial^\alpha \varphi(f) = (-1)^{|\alpha|} \varphi(\partial^\alpha f)$$

4. For any $g \in L^1(\mathbb{R}^d)$, φ_g defined by

$$\varphi_g(f) = \int g(x) f(x)\, dx, \quad f \in \mathcal{S}$$

is a tempered distribution. Moreover, φ_g is uniquely determined by $g \in L^1$. This can be seen as follows. Suppose $g_1, g_2 \in L^1(\mathbb{R}^d)$. Then $h = g_1 - g_2 \in L^1(\mathbb{R}^d)$. We have to show that if $\varphi_{g_1} = \varphi_{g_2}$, then $h = 0$. The latter can be proved by showing that

$$\int_A h(x)\, dx = 0 \tag{18.13}$$

for all sets $A \subset \mathbb{R}^d$ of the form $[a_1, b_1] \times \cdots [a_d, b_d]$. (18.13) follows from $\varphi_{g_1} = \varphi_{g_2}$ by approximating the characteristic function χ_A with a sequence of test functions f_n:

$$\int_A h(x)\, dx = \lim_{n\to\infty} \int h(x) f_n(x) = \lim_{n\to\infty} \varphi_{g_1}(f_n) - \varphi_{g_2}(f_n) = 0$$

Exercise 18.9 *Let $\theta : \mathbb{R} \to \mathbb{R}$ denote the Heaviside function defined by*

$$\theta(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

For any $x_0 \in \mathbb{R}$, let Θ be the tempered distribution defined by

$$\Theta_{x_0}(f) = \int \theta(x - x_0) f(x0 \, dx$$

Prove that $\Theta'_{x_0} = \delta_{x_0}$.

18.3 Convergence of tempered distributions

Let (φ_n) be a sequence in $\mathcal{S}^*$. We will say that (φ_n) converges to $\varphi \in \mathcal{S}^*$ iff

$$\lim_{n\to\infty} \varphi_n(f), \qquad \text{for all } f \in \mathcal{S} \tag{18.14}$$

Note that for each $f \in \mathcal{S}$ the map

$$\varphi \mapsto \varphi(f) \tag{18.15}$$

is a linear functional on $\mathcal{S}$. The notion of convergence of distributions we just introduced corresponds to a natural topology on the tempered distributions defined as the weakest topology for which all the functionals of the form (18.15) are continuous. This topology is called the *weak-* topology*, often abreviated as w^*-topology. It can also be viewed as a topology defined by the family of seminorms $\{p_f \mid f \in \mathcal{S}\}$, where
,

$$p_f(\varphi) = |\varphi(f)|$$

Exercise 18.10 *Prove that p_f is a seminorm on $\mathcal{S}^*$.*

A particularly important kind of sequences of distributions are the ones that converge to δ_{x_0}. They are called *delta sequences.* One is usually interested in approximating δ_{x_0} by distributions of the form

$$\varphi_n(f) = \int \delta_n(x - x_0) f(x) \, dx \tag{18.16}$$

where δ_n is a true function with nice properties, e.g., $\delta_n \in \mathcal{S}$.
Examples of delta sequences:

1. By translation a sequence converging to δ_{x_0} can be obtained from a sequence converging to δ_0. Therefore we will just give examples for delta sequences converging to δ_0. You already encountered several delta sequences without thinking of them in terms of distributions. E.g., in one dimension,

$$\varphi_n(f) = \frac{n}{2}\int_{-\frac{1}{n}}^{\frac{1}{n}} f(x)\,dx \to \delta_0(f) = f(0)$$

which is of the form (18.16) with

$$\delta_n(x) = \begin{cases} n/2 & \text{if } -1/n \leq x \leq 1/n \\ 0 & \text{elsewhere} \end{cases}$$

2. The approximation by Gaussians is another example we are already familiar with:

$$\delta_n(x) = \sqrt{\frac{n}{\pi}}e^{-nx^2} \tag{18.17}$$

3. The following is also a useful delta sequence:

$$\delta_n(x) = \frac{\sin nx}{\pi x} \tag{18.18}$$

Exercise 18.11 a) *Prove that*

$$\int_{-\infty}^{+\infty} \frac{\sin x}{x}\,dx = \pi$$

(Hint: use Cauchy's formula applied to the function $\exp(iz)$*.)*
b) *Prove that the* δ_n *of (18.18) indeed defines a delta sequence.*

It is obvious that $\mathcal{S}$ can be continuously embedded in $\mathcal{S}^*$, and it is also clear that this embedding will not be onto. Therefore, we cannot expect to have a strict analogue of the Riesz Representation theorem for the continuous linear functionals on a Hilbert space. But there is an analogous theorem that we state here without proof.

Theorem 18.2 *For every* $\varphi \in \mathcal{S}^*$ *there is a continuous function* $F : \mathbb{R}^n :\to \mathbb{C}$ *of at most polynomial growth and a multi-index* α *such that for all* $f \in \mathcal{S}$

$$\varphi(f) = \int F(x)\partial^\alpha f(x)\,dx$$

18.4 Convolution

By Exercise 18.7 we know that for $f, g \in \mathcal{S}$, the function $f * g$ defined by

$$x \mapsto f * g(x) := \int_{\mathbb{R}^d} f(x-y)g(y)\,dy \tag{18.19}$$

is in $\mathcal{S}$ and depends continuously on g (and f). $f * g$ is called the *convolution* of f and g. Using the translation operator defined in Exercise 18.8, and the reflection operator R defined by

$$(Rf)(x) = f(-x)$$

this can also be written as

$$f * g(x) := \int_{\mathbb{R}^d} (RT_x f)(y)g(y)\,dy$$

The following properties are easy to prove.

Proposition 18.3 *For any $f, g, h \in \mathcal{S}$ one has*
i) $f * g = g * f$.
ii) $(f * g) * h = f * (g * h)$.
iii) $T_a(f * g) = (T_a f) * g = f * (T_a g)$, for all $a \in \mathbb{R}^d$.

It is clear from (18.19) that the definition of convolution can be extended from test-functions to more general functions as long the integrals exist. E.g. the convolution of two L^1 functions is well-defined and is also L^1. It is also easy to see that if $f \in L^1$ and $g \in C^k$ with all derivatives up to order k bounded, then $f * g \in C^k$.

One can even define the convolution of $\varphi \in \mathcal{S}^*$ and $f \in \mathcal{S}$ by

$$\varphi * f(x) = \varphi(RT_x f)$$

One can prove that $\varphi * f \in C^\infty(\mathbb{R}^d)$, and that $\varphi * f$ is of at most polynomial growth.

We state without proof the following inequality for the L^p norm of convolutions. In particular, this inequality shows that convolution is a continuous operation when defined on an appropriate choice of Banach spaces.

Proposition 18.4 (Young's inequality) *Suppose $f \in L^p(\mathbb{R}^d), g \in L^q(\mathbb{R}^d)$, and $1 \leq, p, q, r \leq \infty$ such that $p^{-1} + q^{-1} = 1 + r^{-1}$. Then, $f * g \in L^r(\mathbb{R}^d)$, and*

$$\|f * g\|_{L^r} \leq \|f\|_{L^p}\,\|g\|_{L^q}$$

19 The Fourier Transform

19.1 The Fourier transform of test functions

Proposition 19.1 *For any $f \in \mathcal{S}(\mathbb{R}^d)$, the function $\hat{f} : \mathbb{R}^d \to \mathbb{C}$ defined by*

$$\hat{f}(k) = \frac{1}{(2\pi)^{d/2}} \int e^{-ik\cdot x} f(x)\,dx$$

is an element of $\mathcal{S}(\mathbb{R}^d)$.

Proof: $\partial^\alpha \hat{f}$ is bounded for all $\alpha \in (\mathbb{Z}^+)^d$. This can be seen as follows:

$$\begin{aligned}
|\partial^\alpha \hat{f}(k)| &= \frac{1}{(2\pi)^{d/2}} \left| \int (-ix)^\alpha e^{-ik\cdot x} f(x)\,dx \right| \\
&\leq \frac{1}{(2\pi)^{d/2}} \int |x^\alpha f(x)|\,dx \\
&\leq \frac{1}{(2\pi)^{d/2}} \int (1+\|x\|)^{|\alpha|+d+1} |f(x)| \frac{1}{(1+\|x\|)^{d+1}}\,dx \\
&\leq \frac{1}{(2\pi)^{d/2}} \|f\|_{|\alpha|+d+1,0} \int_{\mathbb{R}^d} \frac{1}{(1+\|x\|)^{d+1}}\,dx
\end{aligned}$$

It then also follows that $\partial^\alpha \hat{f}$ is continuous for all α.

To prove that $\|\hat{f}\|_{n,\alpha} < \infty$ for all n, α, we use Exercise 18.6 and repeat the above argument combined with integration by parts: for any α, β

$$\begin{aligned}
|k^\beta \partial^\alpha \hat{f}(k)| &= \frac{1}{(2\pi)^{d/2}} \left| \int (-ix)^\alpha (\partial^\beta e^{-ik\cdot x}) f(x)\,dx \right| \\
&\leq \frac{1}{(2\pi)^{d/2}} \int |\partial^\beta (x^\alpha f(x))|\,dx
\end{aligned}$$

It is then clear that $|k^\beta \partial^\alpha \hat{f}(k)|$ can be bounded by a finite sum of the form

$$\sum_{0\leq m\leq|\alpha|, 0\leq\gamma\leq\beta} C(m,\gamma)\|f\|_{m,\gamma}$$

∎

From the proof of the previous proposition it also follows that the map $\mathcal{F} : \mathcal{S} \to \mathcal{S} : f \mapsto \hat{f}$ is a *continuous* linear transformation from $\mathcal{S}$ to itself with its standard topology.

Next, we prove that $\mathcal{F}$ is invertible with a continuous inverse.

Proposition 19.2 (Fourier Inversion Formula) *Define* $\mathcal{G} : \mathcal{S} \to \mathcal{S}$ *by*

$$\mathcal{G}f(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{ik\cdot x} f(k)\, dk$$

Then $\mathcal{G}$ *is a continuous linear transformation from* $\mathcal{S}$ *to itself and* $\mathcal{F}\mathcal{G} = \mathcal{G}\mathcal{F} = \mathrm{id}$, *i.e.,* $\mathcal{G} = \mathcal{F}^{-1}$.

Proof: First, note that

$$\mathcal{G} = R \circ \mathcal{F} \tag{19.1}$$

where R is the reflection defined by $Rf(x) = f(-x)$. Clearly, R is linear and continuous. Therefore, (19.1) and the fact that implies that $\mathcal{F}$ is a continuous linear transformation implies that $\mathcal{G}$ is linear and continuous, too.

To prove that $\mathcal{G}$ is the inverse of $\mathcal{F}$ we will compute, for any $f \in \mathcal{S}$, $\mathcal{G}\mathcal{F}f = \mathcal{G}\hat{f}$. At first, suppose $d = 1$.

$$\begin{aligned} \mathcal{G}\hat{f}(x) &= \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{ik\cdot x} \left[\int_{-\infty}^{+\infty} e^{-ik\cdot y} f(y)\, dy \right] dk \\ &= \lim_{n\to\infty} \frac{1}{2\pi} \int_{-n}^{+n} \left[\int_{\infty}^{+\infty} e^{ik\cdot(x-y)} f(y)\, dy \right] dk \\ &= \lim_{n\to\infty} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left[\int_{-n}^{+n} e^{ik\cdot(x-y)}\, dk \right] f(y)\, dy \\ &= \lim_{n\to\infty} \frac{1}{2\pi} \int \left[\frac{2\sin n(x-y)}{x-y}\, dx \right] f(y)\, dy \end{aligned}$$

Note that by introducing the "ultraviolet cut-off" in the integral over k, we were able to interchange the order of integration. Now, we use the fact that $(\sin nx)/\pi x$ is a delta sequence (see the previous section). We conclude that $\mathcal{G}\hat{f}(x) = f(x)$, i.e., $\mathcal{G}\mathcal{F} = \mathrm{id}$. Because the relation (19.1) $\mathcal{F}\mathcal{G} = \mathrm{id}$ then easily follows.

For arbitrary dimensions d it suffices to observe that the multidimensional Fourier transform is nothing but the compositions of the one-dimensional Fourier transforms in each of the components x_i of $x \in \mathbb{R}^d$, $i = 1, \ldots, d$. ■

It is convenient to denote the inverse Fourier transform $\mathcal{F}^{-1}f$ by $\tilde{f}$. So, we have

$$\tilde{\hat{f}} = \hat{\tilde{f}} = f, \qquad \text{for all } f \in \mathcal{S} \tag{19.2}$$

Example: Fourier transform of a Gaussian: suppose $a > 0$, and let $f(x) = \exp -a(\frac{1}{2}x \cdot x)$. Then

$$\hat{f}(k) = \frac{1}{a^{d/2}} \exp(-\frac{1}{a}(\frac{1}{2}k \cdot k)) \tag{19.3}$$

To prove this formula it suffices to consider the case $a = 1$. The formula for all $a > 0$ then follows by a change of variables $x \mapsto \sqrt{a}x$. We can also reduce the case of arbitrary d to the the one-dimensional case because the Fourier transform of the multi-dimensional Gaussian factorizes into a product of one-dimensional integrals. So, we have to prove that

$$\frac{1}{\sqrt{2\pi}} \int e^{-ikx} e^{-\frac{1}{2}x^2}\, dx = e^{-\frac{1}{2}k^2}$$

The LHS can be written as

$$\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x+ik)^2}\, dx \times e^{-\frac{1}{2}k^2}$$

So, the point is to prove that

$$\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x+ik)^2}\, dx = 1$$

This integral in independent of k:

$$\begin{aligned} \frac{d}{dk}\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x+ik)^2}\, dx &= -i\frac{1}{\sqrt{2\pi}} \int (x+ik) e^{-\frac{1}{2}(x+ik)^2}\, dx \\ &= i\frac{1}{\sqrt{2\pi}} \int \frac{d}{dx} e^{-\frac{1}{2}(x+ik)^2}\, dx \\ &= i\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x+ik)^2}\Big|_{-\infty}^{+\infty} = 0 \end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}(x+ik)^2}\, dx = \frac{1}{\sqrt{2\pi}} \int e^{-\frac{1}{2}x^2}\, dx = 1$$

Exercise 19.1 (Behavior under translations) *Recall the action of the translation over $a \in \mathbb{R}^d$ on $\mathcal{S}$: $T_a f(x) = f(x-a)$. Prove that*

$$\widehat{T_a f}(k) = e^{-ik\cdot a} \hat{f}(k) \tag{19.4}$$

Exercise 19.2 (Fourier transform of derivative) *Prove that for any multi-index α*

$$\widehat{\partial^\alpha f}(k) = (-ik)^\alpha \hat{f}(k) \tag{19.5}$$

Exercise 19.3 (Fourier transform of convolutions) *For all $f, g \in \mathcal{S}$ we have*

$$\frac{1}{(2\pi)^{d/2}} \widehat{f * g}(k) = \hat{f}(k)\hat{g}(k) \tag{19.6}$$

19.2 The Fourier transformation of tempered distributions

Suppose $\varphi \in \mathcal{S}^*$. We define the Fourier transform $\hat{\varphi}$ of φ by

$$\hat{\varphi}(f) = \varphi(\tilde{f}), \qquad \text{for all } f \in \mathcal{S} \tag{19.7}$$

The tilde in this formula is not a misprint. That $\hat{\varphi}$ is a tempered distribution follows directly from the fact that $\mathcal{F}^{-1}$ is a continuous linear transformation of $\mathcal{S}$. In fact, the map $\varphi \mapsto \hat{\varphi}$ is a continuous transformation of $\mathcal{S}^*$ to itself with its standard topology. It is also clear that the Fourier transformation on $\mathcal{S}^*$ is also invertible with a continuous inverse given by $\varphi \mapsto \tilde{\varphi}$, $\tilde{\varphi}(f) = \varphi(\hat{f})$.

Examples:

- For φ that are defined in terms of a function $g \in \mathcal{S}$ by

  $$\varphi(f) = \int \overline{g(x)} f(x)\, dx$$

  we have that

  $$\begin{aligned} \hat{\varphi}(f) &= \int \overline{g(x)} \tilde{f}(x)\, dx \\ &= \int \overline{g(x)} \frac{1}{\sqrt{2\pi}} \left[\int e^{ik\cdot x} f(k)\, dk \right] dx \\ &= \int \overline{\left[\frac{1}{\sqrt{2\pi}} \int g(x) e^{-ik\cdot x}\, dx \right]} f(k)\, dk \\ &= \int \overline{\hat{g}(x)} f(x)\, dx \end{aligned}$$

  So, by making the pairing sesquilinear (i.e., by introducing the complex conjugation), the Fourier transform of a distribution determined by a test function coincides with the distribution determined by the Fourier transform of the test function.

- The Fourier transform of the delta distribution at the origin:

  $$\hat{\delta}_0(f) = \delta_0(\tilde{f}) = \tilde{f}(0) = \frac{1}{\sqrt{2\pi}} \int f(x)\, dx$$

  which can be written with the pairing as

  $$\int \hat{\delta}(x) f(x)\, dx = \int \frac{1}{\sqrt{2\pi}} f(x)\, dx$$

  If you wish, you can think of $\hat{\delta}$ as the constant function $\frac{1}{\sqrt{2\pi}}$.

From the Fourier inversion formula (19.2) we immediately see that

$$\hat{\varphi}(\hat{f}) = \varphi(f) = \tilde{\varphi}(\tilde{f}) \tag{19.8}$$

This is a version of *Plancherel's formula.* If we write φ (formally or literally) as

$$\varphi(f) = \int \overline{\varphi(x)} f(x)\, dx$$

Plancherel's formula becomes

$$\int \overline{\varphi(x)} f(x)\, dx = \int \overline{\hat{\varphi}(x)} \hat{f}(x)\, dx = \int \overline{\tilde{\varphi}(x)} \tilde{f}(x) \tag{19.9}$$

Remark: The definition of the Fourier tranform of distributions in terms of the Fourier transform of test functions is a special case of the more general procedure of obtaining the *adjoint* of a (continuous) linear transformation $T : X \to X$, where X is a (topological) linear space, as a (continuous) linear transformation on the (topological) dual X^*:

$$T^* : X^* \to X^* : \xi \mapsto \xi \circ T$$

T^* is then called the *adjoint* of T. In our case the Fourier transform on $\mathcal{S}^*$ is defined as the adjoint of the *inverse* Fourier transform on $\mathcal{S}$.

19.3 The Fourier transform on L^2

There are several ways of defining the Fourier transform on $L^2(\mathbb{R}^d)$. For us the easiest method starts from the observation that $\mathcal{S}$ is a linear subspace of L^2, which is dense for the L^2-norm. The Plancherel formula (19.9) shows that $\mathcal{F}$ is isometric, and *a fortiori*, bounded. Therefore it has a unique extension to an isometric transformation of L^2 into itself. The same is true for $\mathcal{F}^{-1}$. Therefore, $\mathcal{F}$ is isometric and invertible, hence unitary.

The main drawback is that it does not give us a formula to compute the Fourier transform of arbitrary elements of L^2, represented by a function. In fact, the latter is not possible. Of course, we can always resort to an approximation argument. Any L^2 function can be approximated by a sequence of test functions. By continuity of the Fourier transform, the Fourier transform of the limit of any convergent sequence is then the limit of the sequence of the Fourier transforms of the elements in the sequence.

Examples:

1. Let $f = \chi_{-N,N}$ be the characteristic function of the interval $[-N, N]$. Then,

$$\hat{f}(k) = \frac{1}{\sqrt{\pi}} \frac{\sin Nk}{k}$$

2. For $a > 0$, define $f(x) = \exp(-a|x|)$. Then $f \in L^2(\mathbb{R})$ and

$$\hat{f}(k) = \sqrt{\frac{2}{\pi}}\frac{a}{a^2 + k^2}$$

We defined the Fourier transform as a unitary operator on $L^2(\mathbb{R}^d)$. What is its spectrum? What are its eigenvectors if any? It is sufficient to consider this question for $d = 1$ (Why?).

As before denote the Fourier transform by $\mathcal{F}$. As $R\mathcal{F}^{-1} = \mathcal{F}$, we have that $\mathcal{F}^2 = R$, and therefore $\mathcal{F}^4 = \mathbb{1}$. It follows that if $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{F}$, then $\lambda \in \{1, i, -1, -i\}$. It turns out that the spectrum of $\mathcal{F}$ consist entirely of eigenvalues, more precisely, $\sigma(\mathcal{F}) = \{1, i, -1, -i\}$, each of which is an infinitely degenerated eigenvalue. The eigenvectors form an o.n. basis and can be taken to be the Hermite functions defined by

$$H_n(x) = \frac{(-1)^n}{n!} e^{\pi x^2} \frac{d^n}{dx^n} e^{-2\pi x^2}, \quad n = 0, 1, 2, \ldots \tag{19.10}$$

The H_n are not normalized. Their L^2-norm is given by

$$||H_n||^2_{L^2} = \frac{(4\pi)^n}{\sqrt{2}n!} \tag{19.11}$$

One can then prove that

$$\mathcal{F}H_n = (-i)^n H_n$$

Suppose we have proved that the normalized Hermite functions are an o.n. basis of $L^2(\mathbb{R}^2)$. Then, we can define the Fourier transform on $L^2(\mathbb{R})$ in the following alternative way. For any $f \in L^2(\mathbb{R})$, f can be written as

$$f = \sum_{n=0}^{\infty} a_n H_n$$

with

$$a_n = \frac{1}{||H_n||^2_{L^2}} \int H_n(x) f(x)\, dx$$

Then

$$\mathcal{F}f = \sum_{n=0}^{\infty} (-i)^n a_n H_n$$

19.4 The Fourier transform on L^1

For any $f \in \mathcal{L}^1(\mathbb{R}^d)$, and $k \in \mathbb{R}^d$, the function $x \mapsto e^{-ik\cdot x} f(x)$ is also L^1. Therefore, we can define the Fourier transform of $f \in L^1$ simply by

$$\hat{f}(k) = \int e^{-ik\cdot x} f(x)\, dx$$

So, no appromation argument is needed for the definition of the Fourier transform on L^1. Approximations arguments are useful, however, to prove certain properties of the Fourier transform of L^1. E.g., in the proof of the following result we will use that $\mathcal{S}$ is dense in L^1 with the L^1-norm. Denote by $C_0(\mathbb{R}^d)$ the space of continuous functions tending to zero at infinity, i.e., $f \in C_0(\mathbb{R}^d)$ is continous and $\lim_{x\to\infty} f(x) = 0$. Recall that $C_0(\mathbb{R}^d)$ equipped with the supremum norm is a Banach space.

Theorem 19.3 (The Riemann-Lebesgue Lemma) *For all $f \in L^1(\mathbb{R}^d)$, $\hat{f} \in C_0(\mathbb{R}^d)$.*

Proof: Let (f_n) be a sequence in $\mathcal{S}$ converging to f in the L^1-norm. We claim that $(\hat{f}_n)$ converges uniformly to $\hat{f}$. The theorem then follows as any uniform limit of continuous functions vanishing at ∞ is a continuous function vanishing at ∞. We now prove the claim.

First, we show that $\hat{f}_n$ is a Cauchy sequence for the supremum norm:

$$\begin{aligned}
|\hat{f}_n(k) - \hat{f}_m(k)| &= \left|\frac{1}{\sqrt{2\pi}} \int e^{-ik\cdot x}(f_n(x) - f_m(x))\, dx\right| \\
&\le \frac{1}{\sqrt{2\pi}} \int |f_n(x) - f_m(x)|\, dx \\
&= \|f_n - f_m\|_{L^1}
\end{aligned}$$

As we assumed that $f_n \to f$ in the L^1 norm, this estimate shows that $(\hat{f}_n)$ is Cauchy for the supremum norm.

As $C_0(\mathbb{R}^d)$ is complete for the supremum norm, there is $g \in C_0(\mathbb{R}^d)$, such that $\hat{f}_n \to g$ uniformly. It remains to show that $g = \hat{f}$. But this is quite obvious:

$$\begin{aligned}
&|g(k) - \hat{f}(k)| \\
&\quad \left|\lim_{n\to\infty} \hat{f}_n(k) - \lim_{n\to\infty}\left[\frac{1}{\sqrt{2\pi}} \int e^{-ik\cdot x} f_n(x)\, dx - \frac{1}{\sqrt{2\pi}} \int e^{-ik\cdot x}(f_n(x) - f(x))\, dx\right]\right| \\
&\quad \le \lim_{n\to\infty} \frac{1}{\sqrt{2\pi}} \int |f_n(x) - f(x)|\, dx = \lim_{n\to\infty} \|f_n - f\|_{L^1} = 0
\end{aligned}$$

■

Note that the Riemann-Lebesgue Lemma does not hold in general for functions in L^2 (give a counterexample).

The following obvious inequality shows that the Fourier transform is a bounded (hence continuous) linear map L^1 into C_0 equipped with their respective norm topologies. The following properties show that the Fourier transform is an isomorphism between the algebra L^1 considered with convolution (up to a factor $1/(2\pi)^{d/2}$) as the product and its image $\mathcal{F}(L^1)$ considered as a subalgebra of C_0 with the ordinary pointwise product of functions. It is known that $\mathcal{F}(L^1) \neq C_0$, but a precise description of it is lacking. In other words, $\mathcal{F} : L^1(\mathbb{R}^d) \to C_0(\mathbb{R}^d)$ is ono-to-one but not onto.

Proposition 19.4 *The Fourier transform on $L^1(\mathbb{R}^d)$ has the following two properties:*
i) For any two functions $f, g \in L^1(\mathbb{R}^d)$, $f * g \in L^1(\mathbb{R}^d)$, and

$$\frac{1}{(2\pi)^{d/2}} \widehat{f * g} = \hat{f}\hat{g}$$

ii) $\hat{f} = 0$ iff $f = 0$.

Proof: The proof of i) just involves the interchange of order of integration, which is justified because the integral involved exist:

$$\begin{aligned}
\frac{1}{(2\pi)^{d/2}} \widehat{f * g}(k) &= \frac{1}{(2\pi)^d} \int e^{-ik\cdot x} \left[\int f(x-y) g(y)\, dy \right] dx \\
&= \frac{1}{(2\pi)^d} \int e^{-ik\cdot y} \left[\int e^{-ik\cdot(x-y)} f(x-y)\, dx \right] g(y)\, dx \\
&= \frac{1}{(2\pi)^{d/2}} \int e^{-ik\cdot z} f(z)\, dz \times \frac{1}{(2\pi)^{d/2}} \int e^{-ik\cdot y} g(y)\, dy
\end{aligned}$$

To prove ii) we consider $f \in L^1$ as a tempered distribution. Plancherel's formula then reads

$$\int f(x) g(x)\, dx = \int \hat{f}(k) \hat{g}(k)\, dk, \qquad \text{for all } g \in \mathcal{S}$$

from which we see that $\hat{f} = 0$ iff the tempered distribution determined by f vanishes. As the tempered distribution is uniquely determined by f as an L^1 function, this proves the statement. ∎

Exercise 19.4 *Let $f \in L^1(\mathbb{R}^3)$ be a rotation invariant function in the sense that there is a function $g : \mathbb{R}^+ \to \mathbb{C}$ such that*

$$f(x) = g(\|x\|)$$

Prove that the Fourier transform of f is a continuous function $\hat{f}$ that is also rotation invariant and more specifically

$$\hat{f}(k) = h(\|k\|)$$

with

$$h(k) = \frac{1}{k}\sqrt{\frac{2}{\pi}} \int_0^{+\infty} r \sin(kr) g(r) \, dr$$

20 Applications of the Fourier Transform

20.1 Translation Invariance (Symmetry)

Consider an integral operator K on $L^2(\mathbb{R})$ (L^2, for short), defined by

$$Kf(x) = \int_{-\infty}^{+\infty} k(x,y) f(y)\, dy \tag{20.1}$$

Suppose there is a function $g \in L^2 \cap L^1$, satisfying $g(-x) = \overline{g(x)}$, and such that $k(x,y) = g(x-y)$. It follows that K is a self-adjoint bounded linear operator on L^2 and that it commutes with translations, which are represented by the unitary operators T_a:

$$KT_a = T_a K, \quad \text{for all } a \in \mathbb{R}$$

If K and T_a were two finite-dimensional matrices, we would conclude that there is a basis of common eigenvectors of T_a and K. Note that K of the form (20.1) cannot be Hilbert-Schmidt. The Fourier transform, considered as a unitary $\mathcal{F} : L^2 \to L^2$, allows us the directly compute the spectrum of K. Define $\hat{K} = \mathcal{F}^{-1} K \mathcal{F}$. Then

$$\hat{K}\hat{f}(k) = \sqrt{2\pi}\hat{g}(k)\hat{f}(k)$$

As $\hat{f}$ is an arbitrary element of L^2, this means that $\hat{K}$ is the multiplication operator by the function $\sqrt{2\pi}\hat{g}$. As we assumed that $g \in L^1$, $\hat{g} \in C_0$, by the Riemann-Lebesgue Lemma (Theorem 19.3). Just as with the operator of multiplication by x, on $L^2([0,1])$, that we studied last quarter, the spectrum of $\hat{K}$ is exactly the closure of the range of values taken by the function $\hat{g}$. As K and $\hat{K}$ are unitarily equivalent ($\hat{K} = \mathcal{F}^{-1} K \mathcal{F}$), this is then also the spectrum of K.

Exercise 20.1 *Prove that if $g \in L^2$ satisfies $g(-x) = \overline{g(x)}$, then $\hat{g}$ is real-valued.*

Exercise 20.2 *Let G be a multiplication operator on $L^2(\mathbb{R})$, defined by*

$$Gf(x) = g(x)f(x)$$

where g is continuous and bounded. Prove that G is a bounded linear operator on $L^2(\mathbb{R})$ and that its spectrum is given by

$$\sigma(G) = \overline{\{g(x) \mid x \in \mathbb{R}\}}$$

Exercise 20.3 *Can an operator G of the form given in the previous exercise have any eigenvalues? Give a (counter)example.*

For concreteness consider the following example:

$$Kf(x) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-y)^2} f(y)\, dy$$

The kernel of K is called the *heat kernel.* Then

$$\hat{K}\hat{f}(k) = e^{-\frac{1}{2}k^2}\hat{f}(k)$$

By the discussion above we then have $\sigma(K) = \sigma(\hat{K}) = [0, 1]$. It is not hard to prove that this is entirely continuous spectrum. So, there are no eigenvectors. There are, however, generalized eigenvectors, in the sense of distributions: for any $l \in \mathbb{R}$

$$\hat{K}\delta_l = e^{-\frac{1}{2}l^2}\delta_l$$

By taking Fourier transforms we find for K that

$$Ke^{-ilx} = e^{-\frac{1}{2}l^2}e^{-ilx}$$

where the functions are to be interpreted as the distributions they define via the pairing. As one expects from the fact that K and T_a commute, these are also "eigendistributions" of T_a:

$$T_a e^{-lx} = e^{-il(x-a)} = e^{ial}e^{-ilx}$$

The corresponding "eigenvalues" of the unitary T_a are e^{ial}. So, if $a \neq 0$, $\sigma(T_a) = S^1$, and it is entirely continuous spectrum of T_a, considered as an operator on L^2. For $a = 0$, we have $T_0 = \mathbb{1}$, whose spectrum contains just the eigenvalue 1.

A concrete application of the integral operator K is in the study of the heat equation on the real line (an "infinite rod"):

$$\begin{aligned}
u_t &= \frac{1}{2}u_{xx}, && \text{for all } x \in \mathbb{R}, t > 0 \\
\lim_{x\to\pm\infty} u(x,t) &= 0, && \text{for all } t > 0 \\
\lim_{t\downarrow 0} u(x,t) &= u_0(x), && \text{for all } x \in \mathbb{R}
\end{aligned}$$

Here, u_0 is the initial condition, which we can allow to be a distribution satisfying certain conditions. The solution is then given by

$$u(x,t) = \int_{-\infty}^{+\infty} \frac{e^{-\frac{1}{2t}(x-y)^2}}{\sqrt{2\pi t}} u_0(y)\, dy$$

One immediately sees that for any initial condition $u_0 \in L^1$, the solution is C^∞ in both t and x, for $t > 0$ and $x \in \mathbb{R}$. This is the well-known *smoothing property* of the heat equation.

One can also consider a half-infinite rod, i.e., the heat equation on $[0, \infty)$. The solution is then given by (see old homework)

$$u(x,t) = \int_0^{+\infty} \frac{e^{-\frac{1}{2t}(x-y)^2} - e^{-\frac{1}{2t}(x+y)^2}}{\sqrt{2\pi t}} u_0(y)\, dy \tag{20.2}$$

Exercise 20.4 *Check that (20.2) is indeed a solution of the heat equation on* $[0, \infty)$ *with initial condition* u_0. *(Hint: use the "method of images".)*

In the homework you will learn about an application of the Fourier transform in the study of the wave equation.

20.2 The Poisson Summation Formula

The Poisson Summation Formula states that for a large class of functions $f : \mathbb{R} \to \mathbb{C}$ one has the followig intriguing identity

$$\sum_{n=-\infty}^{+\infty} f(2\pi n) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{+\infty} \hat{f}(n) \tag{20.3}$$

You can easily imagine that this formula can be useful to derive identies between infinite series, or even permit one to explicitly sum a series. In a slight variant, the Euler-MacLaurin formula, one can use it to find approcimate values for the sum of series with a an error term that can be easily estimated see, e.g., H. Dym and H.P. McKean, *Fourier series and integrals*, Academic Press, New York - London, 1972, p 112.). Note that the presence or absence of factors 2π in (20.3) depends on the chosen normalization for the Fourier transform.

Theorem 20.1 (Poisson Summation Formula) *Let* $f \in C^1(\mathbb{R})$ *such that there exists a constant* $C > 0$, *and* $\epsilon > 0$, *such that*

$$|(1+|x|)^{1+\epsilon} f(x)| \leq C, \quad |(1+|x|)^{1+\epsilon} f'(x)| \leq C, \quad \text{for all } x \in \mathbb{R}$$

Then one has the identity

$$\sum_{n=-\infty}^{+\infty} f(2\pi n) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{+\infty} \hat{f}(n)$$

Proof: First, note that, under the conditions of the theorem, the following sum

$$g(x) = \sum_{n=-\infty}^{+\infty} f(x + 2\pi n) \tag{20.4}$$

converges uniformly and defines a function $g \in C^1(S^1)$, i.e., a periodic continuously differentiable function on $\mathbb{R}$ with period 2π (see Exercise 20.5). Therefore, the Fourier series for g converges uniformly (Theorem 8.2):

$$g(x) = \frac{1}{2\pi} \sum_{n=-\infty}^{+\infty} e^{inx} \int_0^{2\pi} e^{-iny} g(y)\, dy$$

As g is related to f by (20.4), we can rewrite this as follows:

$$\sum_{n=-\infty}^{+\infty} f(x + 2\pi n) = \frac{1}{\sqrt{2\pi}} \sum_{n=-\infty}^{+\infty} e^{inx} \hat{f}(n)$$

which for $x = 0$ reduces to the Poisson Summation Formula we set out to prove. ■

Exercise 20.5 *Suppose $(S_N)_{N=1}^{\infty}$ is a sequence of functions in $C^1(\mathbb{R})$ such that $S_N \to g$ and $S_N' \to h$, both uniformly. Prove that $g \in C^1(\mathbb{R})$, and $g\prime = h$.*

A standard application of the Poisson Summation Formula is an identity for the Jacobi theta function defined, for $t > 0$, by

$$\theta(t) = \sum_{n=-\infty}^{+\infty} e^{-\pi n^2 t}$$

Jacobi's formula is then a symmetry property of this function expressed by

$$\theta(t) = \frac{1}{\sqrt{t}} \theta(1/t) \tag{20.5}$$

Exercise 20.6 *Prove (20.5).*

Exercise 20.7 *Prove the following identity for all $a > 0$:*

$$\sum_{n=-\infty}^{+\infty} \frac{1}{n^2 + a^2} = \frac{\pi}{a} \frac{1 + e^{-2\pi a}}{1 - e^{-2\pi a}}$$

Note that by considering the limit $a \downarrow 0$ of a rearrangement of this identity you can obtain

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$$

which was a homework problem in MAT203B.

20.3 The Central Limit Theorem

Let X be a real random variable, with mean 0 and variance 1, and with distribution f. This means that f is a non-negative function in $L^1(\mathbb{R})$, satisfying

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1, \quad \int_{-\infty}^{+\infty} xf(x)\,dx = 0, \quad \int_{-\infty}^{+\infty} x^2 f(x)\,dx = 1 \tag{20.6}$$

The interpretation is that, for any $a \leq b$, the probability that X has a value between a and b is given by

$$\text{Prob}\,(a \leq X \leq b) = \int_a^b f(x)\,dx$$

If X has a finite mean and variance, the normalization of the mean to 0 and the variance to 1 can be achieved by replacing X by Y defined by

$$Y = A + BX$$

with suitable real constants A and B (if this is not familiar to you, find A and B explicitly in terms of the mean and the variance of X). Therefore, the assumption that the mean vanishes and that the variance is 1 just amounts to assuming that these quantities are finite.

In study of random events described by such a random variable X, and in applications, it is often important to consider the sum of a large number of independent observations (or measurements) of X. Indeed, it is a standard procedure to reduce certain types of error in measurements of physical quantities in every experimental science, by measuring the quantity many times and taking the average. The central limit theorem explains how this error reduction works. In other applications the sum of many independent copies of X is the quantity of interest itself. E.g., a (discrete-time) *Random Walk* in one dimension is defined as the sequence of real numbers

$$(x_0, x_1, x_2, \ldots)$$

where $x_{n+1} = x_n + X_n$, i.e., at each time step the value of a random variable X_n is added to the previous position. E.g., we can take every X_n to be an independent copy of X. Independent means independent of the outcomes of X_m, for $m \neq n$. The random walker is then walking up and down the real line with steps that are independent of each other and that are chosen at random with distribution f. A natural question to ask is then what the position of the random walker is after N steps and assuming, e.g., that she started at 0.

For $N = 1$, the distribution of the position of the walker is obviously given by f. And if the position of the walker after N steps is x_N, the distribution of her position after $N + 1$ steps is clearly given by $f(x - x_N)$, because this is the probability that the $N+1$st step will be of magnitude $x - x_N$. Of course, x_N is random itself, say with distribution f_N. Then, the distribution at time step $N+1$ will be given by

$$f_{N+1}(x_{N+1}) = \int f(x_{N+1} - x_N) f_N(x_N)\, dx_N = f * f_N(x_{N+1})$$

It is then clear that the distribution of the position after N steps, starting from 0, i.e., $S_N = X_1 + X_2 + \cdots + X_N$, is given by

$$f_N = \underbrace{f * f * \cdots * f}_{N \text{ times}}$$

The Central Limit Theorem is a statement about the behavior of f_N in the limit of large N. As we should expect that typically values of S_N increase with N, the limit is better described by a properly normalized quantity. It turns out that the correct normalization is to consider

$$Y_N = \frac{1}{\sqrt{N}} S_N \tag{20.7}$$

Denote by g_N the distribution of Y_N. Then

$$g_N(x) = \sqrt{N} f_N(\sqrt{N} x)$$

The Central Theorem shows that, in the limit $N \to \infty$, g_N converges to a standard Gaussian distribution. The sense in which this convergence occurs, depends on the properties of f. Here we will prove that, for any $h \in S$,

$$\lim_{N\to\infty} \int g_N(x) h(x)\, dx = \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} h(x)\, dx$$

This means that the convergence is in the topology of $\mathcal{S}^*$. In probability theory this is called *weak convergence.*

By approximating the indicator function of the interval $[a, b]$ by test functions one can then also show that, for any $a < b$,

$$\lim_{N\to\infty} \text{Prob}\left(a \le \frac{S_N}{\sqrt{N}} \le b \right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}x^2}\, dx$$

Theorem 20.2 (Central Limit Theorem)

$$\lim_{N\to\infty} g_N(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{20.8}$$

in the sense of distributions.

Proof: As the Fourier transform on $\mathcal{S}$ is a continuous transformation with continuous inverse, instead of (20.8), we can as well prove

$$\lim_{N\to\infty}\int g_N(k)\hat{h}(k) = \frac{1}{\sqrt{2\pi}}\int e^{-\frac{1}{2}k^2}\hat{h}(k)\,dk, \qquad \text{for all } h\in\mathcal{S}$$

Due to Plancherel's formula (19.9), and the formula for the Fourier transform of a Gaussian (19.3), this is then equivalent with

$$\lim_{N\to\infty}\int \widetilde{g_N}(k)h(k) = \frac{1}{\sqrt{2\pi}}\int e^{-\frac{1}{2}k^2}h(k)\,dk, \qquad \text{for all } h\in\mathcal{S} \tag{20.9}$$

As $g_N(x) = \sqrt{N}f_N(\sqrt{N}x)$, and f_N is the $N-$fold convolution of f with itself, we have that

$$\widehat{g_N}(k) = \widehat{f_N}(\frac{k}{\sqrt{N}})$$

As f_N is the $N-$fold convolution of f, we have, by Exercise 19.3,

$$\widehat{f_N}(\frac{k}{\sqrt{N}}) = (2\pi)^{(N-1)/2}(\hat{f}(\frac{k}{\sqrt{N}}))^N$$

The limit of the latter quantity can be most easily found by expanding e^{-iky} as follows

$$e^{-iky} = 1 - iky - \frac{1}{2}k^2y^2(1+r(ky))$$

where $r(y)$ is a continuous function of $y\in\mathbb{R}$, vanishing at $y=0$. By using the assumptions (20.6), we find the following expression for $\hat{f}(\frac{k}{\sqrt{N}})$:

$$\begin{aligned}\hat{f}(\frac{k}{\sqrt{N}}) &= \frac{1}{\sqrt{2\pi}}\int\left[1-\frac{ikx}{\sqrt{N}}-\frac{k^2x^2}{2N}(1+r(kx/\sqrt{N}))\right]f(x)\,dx\\ &= \frac{1}{\sqrt{2\pi}}\left[1-\frac{k^2}{2N}(1+R_N)\right]\end{aligned}$$

with $\lim_{N\to\infty}R_N = 0$ (by the dominated convergence theorem, see later). To compute the Nth power is now easy and we obtain

$$\lim_{N\to\infty}(2\pi)^{(N-1)/2}(\hat{f}(\frac{k}{\sqrt{N}}))^N = \lim_{N\to\infty}\frac{1}{\sqrt{2\pi}}\left[1-\frac{k^2}{2N}(1+R_N)\right]^N = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}k^2}$$

As the Fourier transform of the standard Gaussian distribution is the standard Gaussian distribution itself, we have proved that

$$g_n(x)\to\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$$

as a tempered distribution. ∎

21 Measure and Integration

This section supplies a minimal introduction to measure and integration theory. It is not an overview or a summary of measure theory. We will not explicitly construct the Lebesgue measure on the real line and the most important results for applications will be stated without proofs. We believe that by introducing a few fundamental notions of the theory we will make it possible to understand and use the concepts and theorems in measure and integration theory, as we will need them for our subsequent study of L^p spaces. For a more complete but concise introduction to this subject we recommend E.H. Lieb and M. Loss, *Analysis*, Graduate Studies in Mathematics volume 14, American Mathematical Society, 1997. A standard reference for the basics of measure theory is P.R. Halmos, *Measure Theory*, Graduate Texts in Mathematics volume 18, Springer Verlag, 1974.

21.1 Measures

In order to be able to define integrals of real-valued functions defined on a domain Ω, we need the notion of measure. Indeed, intuitively, we would like the integral of a characteristic function $\chi_A : \Omega \to \mathbb{R}$, for some subset $A \subset \Omega$, to be equal to the size of A, given by the *measure*, $\mu(A)$. A measure on Ω is a function defined on subsets of Ω, and for us its values will always be non-negative.

It turns out that it is not a good idea to try to define the size (measure) of arbitrary subsets of say $\Omega = \mathbb{R}$. Some sets are just too wild to be *measurable*. So, μ will be defined on a subset of all subsets of Ω. Because of the interpretation of $\mu(A)$ as the size of A, μ will be defined on a *sigma-algebra* Σ, which is defined as follows.

Definition 21.1 (sigma-algebra) *A set of subsets Σ of a set Ω is a sigma-algebra if it satisfies the following properties:*
i) $\emptyset \in \Sigma$,
ii) $A \in \Sigma \Rightarrow A^c := \Omega \setminus A \in \Sigma$,
iii) For any countable family $\{A_i\}$ of elements in Σ, $\bigcup_i A_i \in \Sigma$.

In other words, a sigma-algebra is a set of subsets of Ω that contains the empty set and that is closed under taking complements and countable unions. It follows that a sigma-algebra also contains Ω itself, and that it is closed under countable intersections as well.

Just as in the case of defining a topology (as a set of open sets satisfying the appropriate axioms) the most common way of defining a sigma-algebra is to specify a set of subsets, say $\mathcal{A}$, and consider the sigma-algebra $\Sigma(\mathcal{A})$ generated by it, i.e., $\Sigma(\mathcal{A})$ is the smallest sigma-algebra of subsets of Ω that contains $\mathcal{A}$. One can show that there is a unique such sigma-algebra.

Examples:

1. As any sigma-algebra should contain $\emptyset$, the smallest possible sigma-algebra of subsets of Ω is $\Sigma(\emptyset) = \{\emptyset, \Omega\}$.

2. The largest sigma-algebra of subsets of Ω is just the collection of *all* subsets.

3. The previous two examples are not particularly interesting. For most Ω, e.g., $\Omega = \mathbb{R}$, the first one is too small to be usefull and the second one is too large to be practical. For any topological space $(X, \mathcal{T})$, such as $\mathbb{R}^n$ with its standard topology, there is a natural sigma-algebra, called the *Borel sigma-algebra*, often denoted by $\mathcal{B}(X)$, defined as the smallest sigma-algebra of subsets of X that contains all the open subsets of X :
$$\mathcal{B}(X) = \Sigma(\mathcal{T})$$
E.g.,
$$\mathcal{B}(\mathbb{R}) = \Sigma(\{(a, b) \mid a < b\})$$
The elements of a Borel sigma-algebra are called *Borel sets.* Of course $\{(a, b) \mid a < b\}$ is only a subset of all open subsets of $\mathbb{R}$, but it is easy to see that this set is enough to generate all of the Borel sets.

Exercise 21.1 *Prove that*

$$\mathcal{B}(\mathbb{R}) = \Sigma(\{(a, b] \mid a < b\}) = \Sigma(\{[a, b) \mid a < b\}) = \Sigma(\{[a, b] \mid a < b\}) = \Sigma(\{(a, +\infty) \mid a \in \mathbb{R}\})$$

Definition 21.2 (measure) *Let Σ be a sigma-algebra of subsets of Ω. A measure on Ω is a function*
$$\mu : \Sigma \to [0, +\infty]$$
with the property of countable additivity, which, by definition, means that for any countable family $\{A_i\}$ of mutually disjoint elements of Σ, i.e., $A_i \cap A_j = \emptyset$, whenever $i \neq j$, we have
$$\mu(\bigcup_i A_i) = \sum_i \mu(A_i) \tag{21.1}$$

Note that (21.1) makes sense because Σ is closed under taking countable unions. This property is sometimes relaxed to just finite disjoint unions, but we will always assume countable additivity of any measure we consider. The elements of Σ are called *measurable sets*, for obvious reasons.

If $\mu(\Omega) = 1$, μ is called a *probability measure*, and the triple (Ω, Σ, μ) is referred to as a *probability space.* In general, a *measure space* is a triple (Ω, Σ, μ), where Σ is a sigma-algebra of subsets of Ω, and μ is a measure on Ω.

Examples:

1. Consider an arbitrary set Ω, e.g., $\Omega = \mathbb{N}$, with the sigma-algebra consisting of all subsets of Ω. The *counting measure* on Ω is defined by

$$\mu(A) = \text{ the number of elements of } A$$

   with the convention that if A is an infinite set, $\mu(A) = +\infty$. Check that this is a measure.

2. Let $\Omega = \{1, 2, \ldots, n\}$ be a finite set, and take again Σ to be the set of all subsets of Ω. All the probability measures on Ω are then determined by n non-negative numbers, $p_i \geq 0, i = 1, \ldots, n$, satisfying $p_1 + \cdots + p_n = 1$, and the measure is determined by

$$\mu(\{i\}) = p_i$$

In other contexts *signed measures*, which take arbitrary values in the reals, or *complex-valued measures*, are useful to consider. Here, we will always assume the measure is non-negative. Note that some sets A, may have infinite measure, i.e, $\mu(A) = +\infty$.

Exercise 21.2 *Let Σ be a sigma-algebra of subsets of Ω, and suppose μ is a measure on Ω. Prove the following properties:*
a) *If $A, B \in \Sigma$, then $A \setminus B \in \Sigma$.*
b) *If $A, B \in \Sigma$, and $A \subset B$, then $\mu(A) \leq \mu(B)$.*
c) *If $A, B \in \Sigma$, then $\mu(A \cup B) \leq \mu(A) + \mu(B)$.*

Another property that we will always assume is σ-finiteness: μ is *σ-finite* if there exists an at most countable family $\{A_i\}$ of measurable subsets of Ω such that

$$\bigcup_i A_i = \Omega, \qquad \text{and } \mu(A_i) < \infty$$

Exercise 21.3 *Is the counting measure on $[0,1]$ sigma-finite? Prove or disprove.*

Theorem 21.3 *Suppose $\Sigma = \Sigma(A)$ is the sigma-algebra generated by the elements of $\mathcal{A}$, and suppose μ_1 and μ_2 are two measures such that*

$$\mu_1(A) = \mu_2(A), \qquad \textit{for all } A \in \mathcal{A}$$

and suppose there is an at most countable family of sets A_i in $\mathcal{A}$, such that $\bigcup_i A_i = \Omega$, and $\mu_1(A_i) < \infty$, then $\mu_1 = \mu_2$.

The above theorem gives a useful sufficient condition for a measure to be uniquely determined on all of $\Sigma(\mathcal{A})$ by its values on $\mathcal{A}$.

Examples:

1. Let $f : \mathbb{R} \to \mathbb{R}$ be a non-negative Riemann-integrable function. There is then a unique measure μ on the Borel sigma-algebra of $\mathbb{R}$ satisfying

$$\mu((a,b)) = \int_a^b f(x)\,dx$$

2. By choosing $f \equiv 1$ in the previous example we obtain a unique measure on $\mathbb{R}$ with the Borel sigma-algebra such that

$$\mu((a,b)) = |b-a|$$

Here, we have not explained why such a measure should exists at all, but the theorem above implies there can be at most one. One can prove the existence of this measure in Example 2) above by construction but it is not easy. It is called the *Lebesgue measure* on $\mathbb{R}$, or more precisely, it is the restriction of the Lebesgue measure to the Borel sigma-algebra. This is because it is convenient to have the property that any subset B of a measurable set A with $\mu(A) = 0$, is measurable and, then, of course, $\mu(B) = 0$. The sigma-algebra which contains all Borel sets, and all subsets of sets of measure zero is the standard sigma-algebra of subsets of $\mathbb{R}$ on which the Lebesgue measure is defined. If you are interested in the details, check out an analysis textbook, e.g., W. Rudin, *Real and Complex Analysis*, McGraw Hill, 3rd edition, 1987, or, G.B. Folland, *Real Analysis*, John Wiley & Sons, 1984.

21.2 Integration

Assuming that you are familiar with the *Riemann integral*, we will now define a general notion of integration of real- and complex-valued functions, called the *Lebesgue integral*, which includes the Lebesgue integral of functions on $\mathbb{R}$ as a special case.

Suppose Σ is a sigma-algebra of subsets of Ω, and $\mu : \Sigma \to [0, +\infty]$ is a measure. A function $f : \Omega \to \mathbb{R}$, is called *measurable* if all the sets $S_f(t)$, $t \in \mathbb{R}$, defined by

$$S_f(t) = \{x \in \Omega \mid f(x) > t\}$$

are measurable, i.e., $S_f(t) \in \Sigma$, for all $t \in \mathbb{R}$. The sets $S_f(t)$ are called the *level sets* of f.

Example:
Consider $f : \mathbb{R} \to \mathbb{R}$, defined by

$$f(x) = \frac{1}{1+x^2}$$

Then

$$\begin{array}{rcl} S_f(t) = & \emptyset & \text{if } t \geq 1 \\ S_f(t) = & (-\sqrt{\frac{1}{t}-1}, +\sqrt{\frac{1}{t}-1}) & \text{if } 0 < t < 1 \\ S_f(t) = & \mathbb{R} & \text{if } t \leq 0 \end{array}$$

Let, at first, $f : \Omega \to \mathbb{R}$, be a non-negative measurable function. Following Lieb & Loss, we will define $\int f\, \mu(dx)$ as the Riemann integral of the function $F(t)$ defined by

$$F(t) = \mu(S_f(t)) \tag{21.2}$$

F is well-defined because f is measurable. From the properties of Exercise 21.2, it is obvious that $F(t)$ is a non-negative, non-increasing monotone function of t. Therefore, its Riemann integral is well-defined (allowing for the value $+\infty$ for the integral). Hence, we can define the *Lebesgue integral* of f by

$$\int_\Omega f(x)\, \mu(dx) := \int_0^{+\infty} F(t)\, dt \tag{21.3}$$

Give this formula a minute thought and you will realize that this is a very natural definition of the integral. One can prove that if f is itself Riemann integrable, its Riemann integral and its Lebesgue integral as defined in (21.3) conincide.

In the case of $\Omega = \mathbb{R}$, and with μ the Lebesgue measure, we have defined the standard Lebesgue integral of any non-negative Lebesgue-measurable function on $\mathbb{R}$.

Example:
For $f(x) = 1/(1+x^2)$, $F(t) = 2\sqrt{\frac{1}{t}-1}$. So, in this case, $F = 2f^{-1}$, and it is then obvious that

$$\int_{-\infty}^{+\infty} f(x)\, dx = \int_0^{+\infty} F(t)\, dt$$

A non-negative function f is called *integrable* (or *summable*), if f is measurable and

$$\int f(x)\mu(dx) < \infty$$

A function $\Omega \to \mathbb{C}$ is defined to be measurable if the four non-negative functions $f_1, \dots, f_4$ appearing in the decomposition

$$f = f_1 - f_2 + if_3 - if_4$$

are all measurable. If $f_1, \dots, f_4$ are all integrable, f is called integrable and the integral of f is then defined by linear extension:

$$\int_\Omega f\,\mu(dx) := \int_\Omega f_1\,\mu(dx) - \int_\Omega f_2\,\mu(dx) + i\int_\Omega f_3\,\mu(dx) - i\int_\Omega f_4\,\mu(dx)$$

Note that, for measurable f, f is integrable iff $|f|$ is integrable.

21.3 Sets of measure zero, almost everywhere, and all that

We have introduced the Borel sigma-algebra $\mathcal{B}(\mathbb{R})$ of subsets of $\mathbb{R}$. It can be shown that there is a unique measure λ defined on $\mathcal{B}(\mathbb{R})$, with the following two properties:

1. For every measurable set A, and $a \in \mathbb{R}$ one has $\lambda(A + a) = \lambda(A)$ (translation invariance of λ);
2. $\lambda([0, 1]) = 1$

The standard Lebesgue measure on $\mathbb{R}$ is an extension of λ to a sigma-algebra Σ that contains $\mathcal{B}(\mathbb{R})$, but that also has the property that if $A \in \Sigma$, $\lambda(A) = 0$, and $B \subset A$, then $B \in \Sigma$ and $\lambda(B) = 0$. A measurable set A with the property $\lambda(A) = 0$, is called a *set of measure zero*. So, for the Lebesgue measure any subset of a set of measure zero is a set of measure zero. This is a convenient property to have. The Borel sigma-algebra is too small in the sense that it does not contain necessaritly all subsets of a set of measure zero, i.e., not all subsets of a set of measure zero are necessarily measurable. We will use the same symbol λ for the measure discussed above extended to Σ, and called it the *(standard) Lebesgue measure.*

Examples:

1. Any finite subset $A \subset \mathbb{R}$ has measure zero. Indeed, for any $a, b \in \mathbb{R}$, $\lambda(\{a\} = \lambda(\{b\})$, by the translation invariance of the measure. Then, for any n we can find $a_1 < a_2 < \cdots a_n \in [0, 1]$, for which we then have

$$n\lambda(\{a_1\}) = \sum_{i=1}^{n} \lambda(\{a_i\} = \lambda(\{a_1, \dots, a_n\} \leq \lambda([0, 1]) = 1$$

where we have used the elementary properties of measures and the translation invariance of λ. As, n is arbitrary, it follows that $\lambda(\{a_1\}) = 0$, and hence $\lambda(\{a\} = 0$, for all $a \in \mathbb{R}$.

2. Any countable subset of $\mathbb{R}$ has measure zero. This is a consequence of the previous example and the countable additivity of measures. If $A = \{a_i\}_{i=1}^{\infty}$, we have

$$\lambda(A) = \lambda(\bigcup_{i=1}^{\infty}\{a_i\}) = \sum_{i=1}^{\infty}\lambda(\{a_i\}) = 0$$

 E.g., the set $\mathbb{Q}$ of all rational numbers has measure zero. It follows that the irrational numbers between 0 and 1 have measure 1.

3. There are also sets of measure zero that are uncountably infinite. An example is the Cantor set.

The expression *almost everywhere*, abbreviated as *a.e.*, means "with possible exception of a set of measure zero". E.g for two functions f and g, $f = g$ a.e., means that there is a set of measure zero E, such that $f(x) = g(x)$, for all $x \in \Omega \setminus E$. If necessary the measure μ for which this exceptional set has measure zero may be indicated by writing μ-a.e. A sequence of measurable funtions (f_n) converges pointwise μ-a.e. to f, if there is a measurable set E with $\mu(E) = 0$, and such that

$$\lim_{n\to\infty} f_n(x) = f(x), \qquad \text{for all } x \in \Omega \setminus E$$

It can be show that the a.e. pointwise limit of a sequence of measurable functions is measurable. Strictly speaking, for this to make sense one has to (re)define f on E by hand in a suitable way, e.g., put $f(x) = 0$ for all $x \in E$.

From the definition of the integral it is easy to see that if f and g are integrable functions and $f = g$ μ-a.e., then $\int f(x)\,\mu(dx) = \int g(x)\,\mu(dx)$.

In probability theory almost everywhere is ususally called *almost surely.*

A related concept is the *essential supremum* of a set of real numbers $A \subset \mathbb{R}$: indexsupremum!essential

$$\text{ess sup}A := \inf\{C > 0 \mid |x| \leq C, \text{ for a.e. } x \in A\}$$

21.4 The Monotone and Dominated Convergence Theorems

By now, you know from experience that convergent sequences of functions are at the core of the majority of the arguments in analysis. Therefore, the following theorems about integrals of sequences of functions are extremely important. In this section the notions of measurability, integrability etc. will refer to an arbitrary measure space (Ω, Σ, μ), and we will implicitly assume that μ is sigma-finite. Functions are $\Omega \to \mathbb{C}$ unless otherwise mentioned.

The easiest to prove of the convergence theorems is the Monotone Convergence Theorem.

Theorem 21.4 (Monotone Convergence Theorem) *Let (f_n) be a sequence of non-negative functions, converging pointwise and in a monotone increasing way to a function f, i.e.,*

$$f_n(x) \leq f_{n+1}(x), \quad \text{for all } n, \text{ and } \lim_{n\to\infty} f_n(x) = f(x), \quad \text{for a.e. } x$$

Then

$$\lim_{n\to\infty} \int f_n(x)\,\mu(dx) = \int f(x)\,\mu(dx) \tag{21.4}$$

Equality also holds when the LHS is $+\infty$. In this case it means that f is measurable and non-negative but has an infinite integral in the sense of (21.3).

While it is obvious that the same result holds for a decreasing sequence of non-positive functions converging pointwise, (21.4) is generally false for arbitrary pointwise converging sequences or even decreasing sequences of non-negative functions.

Exercise 21.4 *Give an example of a sequence of non-negative integrable functions converging pointwise and monotonically to a function f, but such that the equality (21.4) does not hold.*

Exercise 21.5 *Prove Theorem 21.4 by reducing (21.4) to the case of Riemann integrable functions. Hint: first, show that the sequence of functions (F_n), where F_n is related to f_n as in (21.2), is monotone increasing. Use the countable additivity of the measure to show that $F_n \to F$, where F is also given by (21.2).*

For arbitrary pointwise converging sequences of non-negative integrable functions there is a inequality instead of the equality (21.4). This is a consequence of Fatou's Lemma.

Theorem 21.5 (Fatou's Lemma) *Let (f_n) be a sequence of non-negative integrable functions. Then $f(x) := \liminf_{n\to\infty} f_n(x)$ defines a measurable function and*

$$\liminf_{n\to\infty} \int f_n(x)\,\mu(dx) \geq \int f(x)\,\mu(dx)$$

The most useful of the convergence theorems is the Dominated Convergence Theorem. It does not require the functions to be non-negative nor the convergence to be monotone.

Theorem 21.6 (Dominated Convergence Theorem) *Let (f_n) be a sequence of integrable functions, converging to f pointwise a.e.. Suppose there exists an integrable function g such that $|f_n(x)| \leq g(x)$, μ-a.e., then*

$$\lim_{n\to\infty} \int f_n(x)\,\mu(dx) = \int f(x)\,\mu(dx)$$

Proof: First, show that it is sufficient to prove the statement for non-negative functions f_n. Then, apply Fatou's Lemma 21.5 twice, once to the sequence (f_n), and once to the sequence $(g - f_n)$. ∎

Exercise 21.6 *Fill in the details of the proof of the Dominated Convergence Theorem.*

21.5 Fubini's Theorem

The Lebesgue measure on $\mathbb{R}^n$ defined on the Borel sigma-algebra of $\mathbb{R}^n$ (or, again, an extension of it with lots of sets of measure zero), is the unique measure with the properties of translation and rotation invariance and such that the measure of the unit cube is 1. We denote it again by λ. It is the product of the Lebesgue measure on n copies of $\mathbb{R}$ in the sense that, if $A_i \in \mathcal{B}(\mathbb{R})$, then $A = A_1 \times \cdots \times A_n \in \mathcal{B}(\mathbb{R}^n)$ and

$$\lambda(A) = \lambda(A_1)\lambda(A_2)\cdots\lambda(A_n)$$

This is a particular case of a general construction of a measure $\mu = \mu_1 \times \mu_2$ on $\Omega = \Omega_1 \times \Omega_2$, given any two measure spaces $(\Omega_1, \Sigma_1, \mu_1)$ and $(\Omega_2, \Sigma_2, \mu_2)$. Take for Σ the sigma-algebra of subsets of Ω generated by the subsets of the form $A_1 \times A_2$, where $A_1 \in \Sigma_1$ and $A_2 \in \Sigma_2$, and put

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$$

It can be shown that μ extends uniquely to a measure on all of Σ. It is called the *product measure* of μ_1 and μ_2.

Example:
Let μ_1 and μ_2 be two probability measures on Ω_1 and Ω_2 respectively. In the case of probability measures, the measurable subsets are called *events*. So, if A_1 is a measurable subset of Ω_1 and A_2 is a measurable subset of Ω_2, then $\mu_1(A_1)$ and $\mu_2(A_2)$ are the probabilities with which these events occur. One says that A_1 and A_2 are *independent* if the probability that A_1 and A_2 both happen is the product of their respective probabilities. If all events in Ω_1 are independent of all events in Ω_2, the measure on $\Omega = \Omega_1 \times \Omega_2$ describing their joint probabilities must then satisfy

$$\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2),$$

i.e., μ is the product measure $\mu_1 \times \mu_2$.

We are now faced with an apparent ambiguity when we integrate functions of more then one variable, say $f : \mathbb{R}^2 \to \mathbb{C}$. Do we use the Lebesgue measure on $\mathbb{R}^2$ to integrate $f(x, y)$, or do we first fix one of the variables, say x and integrate over y using the Lebesgue measure on $\mathbb{R}$, which yields a function of x that we can then integrate by using the Lebesgue measure on $\mathbb{R}$ again? Or the other way around, first integrate over x, then over y? Fortunately there is a theorem that makes this question irrelevant, as it should be.

Theorem 21.7 (Fubini's Theorem) *Suppose μ_1 and μ_2 are two sigma-finite measures on Ω_1 and Ω_2 respectively. Let f be a measurable function defined on $\Omega_1 \times \Omega_2$. If f is integrable for $\mu = \mu_1 \times \mu_2$, i.e.,*

$$\int |f(x,y)| \, \mu(dx\, dy) < \infty$$

then

$$\begin{aligned} & \int f(x,y) \, \mu(dx\, dy) \\ = & \int \left[\int f(x,y) \, \mu_1(dx) \right] \mu_2(dy) \\ = & \int \left[\int f(x,y) \, \mu_2(dy) \right] \mu_1(dx) \end{aligned}$$

The equality also holds if f is a non-negative measurable function with an infinite integral.

22 L^p spaces

For this chapter, *Analysis* by Lieb & Loss is recommended reading.

22.1 Definition

Let (Ω, Σ, μ) be a measure space. We are mainly interested in Ω that are a measurable subset of $\mathbb{R}^n$. Most of the results, however, are valid for an arbitrary measure space. We will introduce the spaces $L^p(\Omega, \mu)$, for $1 \leq p < +\infty$ and $p = +\infty$ separately, as they require a different approach.

Fix $1 \leq p < +\infty$. $L^p(\Omega)$, or $L^p(\Omega, \mu(dx))$ if it is desirably to express the measure μ explicitly, is the linear space of all measurable functions $f : \Omega \to \mathbb{C}$, such that $|f|^p$ is integrable, i.e.,

$$\int_\Omega |f(x)|^p \, \mu(dx) < +\infty \tag{22.1}$$

We will often write L^p instead of $L^p(\Omega)$ if this does not lead to confusion. Note that the Lebesgue measure λ, or $\lambda(dx)$, is often written as dx. So, $L^p([a, b], dx)$ means the L^p-space of complex-valued functions on the interval $[a, b]$ with the standard Lebesgue measure.

Exercise 22.1 *Prove that for all $1 \leq p < +\infty$, the space of measurable functions such that (22.1) holds is a linear space for the standard linear operations on functions: pointwise addition and multiplication by a scalar.*

The space $L^\infty(\Omega, \mu)$ consists of all functions $f : \Omega \to \mathbb{C}$ such that there exists a finite constant M such that

$$|f(x)| \leq M, \quad \mu - \text{a.e.}$$

For $1 \leq p < +\infty$, the L^p norm is defined by

$$\|f\|_p := \|f\|_{L^p} := \left[\int_\Omega |f(x)|^p \, \mu(dx)\right]^{1/p}, \quad \text{for all } f \in L^p(\Omega, \mu) \tag{22.2}$$

In fact, this is not a norm defined on the L^p-spaces as we have described them so far. This is because if $f = g$ μ-a.e., then $\|f - g\|_p = 0$. So, in order to make L^p into a normed space we need to consider all functions that are equal to f μ-a.e., as one element, i.e., L^p is really a space of equivalence classes of functions. We will prove that L^p, thus defined, and equipped with the L^p-norm, is a Banach space. Recall that in MAT203A we defined $L^p([0, 1])$ as the completion of the space of all continuous functions with the topology defined by the L^p-norm. The present definition has the advantage of giving us immediately a realization of L^p as a space of equivalence classes of functions, instead of a space of equivalence classes of Cauchy sequences of functions. These two definitions are equivalent however, i.e., they lead to the same Banach space.

Exercise 22.2 *Prove that* $f \sim g$ *iff* $f = g$ μ-a.e., *indeed defines an equivalence relation on the space of all measurable functions.*

Exercise 22.3 *Prove that* $\|\cdot\|_p$ *defined by (22.2) is a norm on* $L^p(\Omega)$. *Hint: use Minkowski's inequality (see the next subsection).*

Theorem 22.1 *For any measure space* (Ω, Σ, μ), *and* $1 \leq p < +\infty$, $L^p(\Omega, \mu)$, *with the* L^p*-norm, is a Banach space.*

Proof: We will abreviate $L^p(\Omega, \mu)$ by L^p and $\|\cdot\|_p$ by $\|\cdot\|$. Recall that for normed linear spaces completeness can be proved by showing that every absolutely convergent series converges. So, let $f_n \in L^p$, $n = 1, 2, \ldots$, be such that

$$\sum_{n=1}^{\infty} \|f_n\| = M < +\infty \tag{22.3}$$

We need to show that there exists $f \in L^p$ such that

$$\lim_{N\to\infty} \|f - \sum_{n=1}^{N} f_n\| = 0 \tag{22.4}$$

To this end, we first consider the sequence of functions g_N defind by

$$g_N(x) = \sum_{n=1}^{N} |f_n(x)| \quad .$$

Clearly, $g_N(x) \geq 0$, and pointwise convergence of (g_N) to a function $g : \Omega \to [0, +\infty]$, is guaranteed by the monotonicity of the sequence. We also have that

$$\|g_N\| \leq \sum_{n=1}^{N} \|f_n\| \leq M \quad .$$

Hence,

$$\int |g_N|^p d\mu = \|g_N\|^p \leq M$$

and, by the monotone convergence theorem (Theorem 21.4), this implies that

$$\|g\|^p = \lim_{N\to\infty} \|g_N\| \leq\leq M$$

In particular, $\|g\| < +\infty$, i.e., $g \in L^p$, and this implies that $g(x) < +\infty \mu -$ a.e., which, by the definition of g, means that $\sum_n f_n(x)$ is absolutely convergent $\mu -$ a.e., and, therefore, the series converges pointwise $\mu -$ a.e. and we can define

$$f(x) = \sum_{n=1}^{\infty} f_n(x)$$

Note that the partial sums of this series satisfy

$$\left| \sum_{n=1}^{N} f_n(x) \right|^p \leq g(x)^p$$

As, $g \in \mathcal{L}^p$, and $|f(x)| \leq g(x)$, we mist have that $f \in L^p$. Furthermore, this also means that $g^p \in L^1$, and we have

$$\left| f(x) - \sum_{n=1}^{N} f_n(x) \right|^p \leq (2g(x))^p \quad .$$

Therefore, we can apply the dominated convergence theorem (Theorem 21.6) to show the convergence of the series in the L^p-norm:

$$\lim_{N \to \infty} \int \left| f - \sum_{n=1}^{N} f_n \right|^p \mu(dx) = \int 0 \mu(dx) = 0.$$

■

Note that, in the proof above, we used the following trivial consequence of the dominated convergence theorem.

Exercise 22.4 *Let $f_n : \Omega \to \mathbb{C}$ be a sequence of measurable functions converging to f pointwise μ-a.e.. Suppose there exists $g \in L^p(\Omega, \mu)$ such that $|f_n| \leq g$, μ-a.e.. Then $f_n \to f$ in the L^p-norm.*

There is a semi-converse to Exercise 22.4: it is also true that if $f_n \to f$ in L^p, then there exists a subsequence of (f_n) converging to f pointwise a.e.. We do not prove this here.

On $L^\infty(\Omega, \mu)$ the L^∞-norm is defined as follows:

$$\|f\|_\infty := \|f\|_{L^\infty} := \inf\{C \mid |f(x)| \leq C, \ \mu - \text{a.e.}\}$$

Clearly, if f is continuous, $\|f\|_\infty = \|f\|_{\text{sup}}$. In general, for $f \in L^\infty$, $\|f\|_\infty$ is the essential supremum of $|f|$, i.e.

$$\|f\|_\infty = \text{ess sup}\{|f(x)| \mid x \in \Omega\}$$

Note that $|f|$ may be greater than its essential supremum on a set of measure zero (but this is not essential!).

An important fact that we will not prove here is that $L^\infty(\Omega, \mu)$ with the L^∞-norm is also a Banach space, when considered as a space of equivalence classes of functions.

22.2 The basic inequalities

It has been said that Analysis is the art of estimating, i.e., proving inequalities. Inequalities are proved using other inequalities. Here are the Ur-inequalities of L^p theory.

22.2.1 Jensen's inequality

If (Ω, Σ, μ) is a measure space with $\mu(\Omega) < \infty$, we can define the *average* of any $f \in L^1(\Omega, \mu)$ by

$$\langle f \rangle_\mu := \frac{1}{\mu(\Omega)} \int_\Omega f(x)\, \mu(dx) \tag{22.5}$$

Jensen's inequality says that the average of a convex function of f is greater then or equal to the value of the convex function at the average of f.

Before we state and proof the theorem, we recall a few properties of convex functions. Let C be a *convex subset* of a linear space, i.e., for every pair $x, y \in C$, and any $t \in [0, 1]$, we have $tx + (1-t)y \in C$. Geometrically this means that the line segment connecting x and y is entirely in C. By definition $J : C \to \mathbb{R}$ is *convex* iff for every $x, y \in C$, and $t \in [0, 1]$,

$$J(tx + (1-t)y) \leq tJ(x) + (1-t)J(y) \tag{22.6}$$

which means that the subset of $C \times \mathbb{R}$ bounded from below by the graph (this is called the *epi-graph*) of J is convex.

Exercise 22.5 *Let $J : \mathbb{R} \to \mathbb{R}$ be a convex function. Prove the following properties:*
a) *For all $x \in \mathbb{R}$ the left- and right-derivatives, $J'_-(x)$ and $J'_+(x)$ exist and satisfy*

$$J'_-(x) \leq J'_+(x)$$

b) *J is continuous on $\mathbb{R}$.*
c) *For all $x \in \mathbb{R}$, there exists a constant $c \in \mathbb{R}$ such that*

$$J(y) \geq J(x) + c(y - x), \qquad \textit{for all } y \in \mathbb{R} \tag{22.7}$$

The graph of the function $y \mapsto c(y - x)$ satisfying (22.7) is called a *support line* of J at x.

Exercise 22.6 *Prove that a function $J : \mathbb{R} \to \mathbb{R}$ is convex iff J has at least one support line at each $x \in \mathbb{R}$.*

Theorem 22.2 (Jensen's inequality) *Let (Ω, Σ, μ) be a measure space with $\mu(\Omega) < \infty$. For any convex function $J : \mathbb{R} \to \mathbb{R}$ and any real-valued $f \in L^1(\Omega, \mu)$, one has the inequality*

$$\langle J \circ f \rangle_\mu \geq J(\langle f \rangle_\mu) \tag{22.8}$$

Proof: From Exercise 22.5 c) we know that there exists a constant $c \in \mathbb{R}$ such that, for all $y \in \mathbb{R}$,

$$J(y) \geq J(\langle f \rangle_\mu) - c(y - \langle f \rangle_\mu) \tag{22.9}$$

By putting $y = f(x)$ in this inequality and integrate both sides, we obtain

$$\begin{aligned} \int_\Omega J \circ f(x)\, \mu(dx) &\geq J(\langle f \rangle_\mu) \int_\Omega \mu(dx) + c\left(\int_\Omega f(x)\, \mu(dx) - \langle f \rangle_\mu \int_\Omega \mu(dx)\right) \\ &= J(\langle f \rangle_\mu) \int_\Omega \mu(dx) \end{aligned}$$

where we have used that $\int_\Omega \mu(dx) = \mu(\Omega) < \infty$ and (22.5). By dividing by $\mu(\Omega)$ one obtains Jensen's inequality. ∎

Note that the LHS of Jensen's inequality may be infinite. Although $J \circ f$ need not be in L^1, its negative part is due to (22.9). Therefore, its integral is well-defined.

One can also show that equality in Jensen's inequality can only occur in the following two circumstances: either $J(x)$ is not *strictly convex* at $x = \langle f \rangle_\mu$, meaning that J has a linear piece in its graph, or f is a constant function.

22.2.2 Hölder's inequality

The Hölder inequalities, and the special case of the Cauchy-Schwarz inequality, are probably the most important ones for proving estimates in L^p-spaces. Hölder's inequality is formulated for a pair of functions, one in L^p and one in L^q, where p and q satisfy: $1 \leq p, q \leq +\infty$, and

$$\frac{1}{p} + \frac{1}{q} = 1$$

with the convention that $1/(+\infty) = 0$. Such p and q are called *dual indices* or *dual exponents.*

Theorem 22.3 (Hölder's inequality) *Let $1 \leq p, q \leq +\infty$ and such that $1/p + 1/q = 1$. Then, for all $f \in L^p(\Omega, \mu)$, and $g \in L^q(\Omega, \mu)$, one has*

$$\left| \int f(x) g(x)\, \mu(dx) \right| \leq \|f\|_p \|g\|_q$$

Proof: For the dual pair of indices $\{p, q\} = \{1, +\infty\}$, Hölder's inequality reduces to the well-known inequality

$$\left| \int_\Omega f(x) g(x)\, \mu(dx) \right| \leq \|g\|_\infty \int_\Omega |f(x)| \mu(dx)$$

So, we may as well assume that $1 < p, q < \infty$.

Next, we observe that

$$\Omega_1 := \{x \in \Omega \mid g(x) \neq 0\}$$

is a measurable set and that

$$\left| \int_\Omega f(x) g(x)\, \mu(dx) \right| = \left| \int_{\Omega_1} f(x) g(x)\, \mu(dx) \right| \leq \int_{\Omega_1} |f(x)|\, |g(x)|\, \mu(dx)$$

And for the RHS of Hölder's inequality we have

$$\|f\|_p^p = \int_\Omega |f(x)|^p\, \mu(dx) \geq \int_{\Omega_1} |f(x)|^p\, \mu(dx)$$

Therefore, without loss of generality, we may assume that $f(x) \geq 0$, and $g(x) > 0$, for all $x \in \Omega$.

With these assumptions in mind we can define a new measure μ on Ω, by

$$\nu(dx) := g(x)^q\, \mu(dx)$$

Then, we apply Jensen's inequality (Theorem 22.2), with the convex function $J(x) = |x|^p$ $(1 \leq p < \infty\, !)$, and the function

$$h(x) = \frac{f(x)}{g(x)^{q/p}}$$

Note that $h \in L^1(\Omega, \nu)$. We obtain

$$\begin{aligned}
\frac{1}{\int_\Omega g(x)^q\, \mu(dx)} \int_\Omega f(x)^p\, \mu(dx) &= \frac{1}{\int_\Omega \nu(dx)} \int_\Omega \frac{f(x)^p}{g(x)^q}\, \nu(dx) \\
&= \langle J \circ h \rangle_\nu \\
&\geq J(\langle h \rangle_\nu) \\
&= \left| \frac{1}{\int_\Omega \nu(dx)} \int_\Omega \frac{f(x)}{g(x)^{q/p}}\, \nu(dx) \right|^p \\
&= \frac{1}{\left| \int_\Omega g(x)^q\, \mu(dx) \right|^p} \left| \int_\Omega f(x) g(x)^{q - q/p}\, \mu(dx) \right|^p \\
&= \left| \frac{\int_\Omega f(x) g(x)\, \mu(dx)}{\int_\Omega g(x)^q\, \mu(dx)} \right|^p
\end{aligned}$$

Note that we have used the fact that p and q are dual exponents for the last equality. By taking the pth root, rearranging and using again that p and q are dual exponents, we see that the inequality above is equivalent to Hölder's inequality. ■

Note that the Hölder inequalities also imply that the LHS is finite if the RHS is.

We refer to Lieb & Loss for a discussion of the case of equality in the Hölder inequalities.

In the special case $p = q = 2$, Hölder's inequality is the Cauchy-Schwarz inequality for L^2-spaces.

Exercise 22.7 *Prove the following generalization of Hölder's inequality: for n functions, $f_i \in L^{p_i}(\Omega, \mu)$, with $1 \leq p_i \geq +\infty$, such that $\sum_{i=1}^n 1/p_i = 1$, one has*

$$\left| \int f_1(x) \cdots f_n(x)\, \mu(dx) \right| \leq \|f_1\|_{p_1} \cdots \|f_n\|_{p_n}$$

As a simple application of Hölder's inequality we can show that, if $\mu(\Omega_1) < \infty$, $L^p(\Omega, \mu) \subset L^{p'}(\Omega, \mu)$, whenever $1 \leq p' \leq p \leq +\infty$. Indeed, define $q \geq 1$ such that

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{p'}$$

p/p' and q/p' are then a pair of dual indices and therefore we can apply Hölder's inequality (with $f \to |f|^{p'}$, and $g \to 1$):

$$\begin{aligned} \|f\|_{p'}^{p'} &= \int_\Omega |f(x)|^{p'}\, \mu(dx) \\ &\leq \left[\int_\Omega (|f(x)|^{p'})^{p/p'}\, \mu(dx) \right]^{p'/p} \times \left[\int_\Omega \mu(dx) \right]^{p'/q} \\ &= (\mu(\Omega)^{p/(p-p')} \|f\|_p^{p'} \end{aligned}$$

This inequality between the norms implies that $\|f\|_p$ is finite for every $f \in L^p$, which is equivalent to the claimed inclusion.

Remark: It is crucial for these inclusions to hold that $\mu(\Omega) < \infty$. This is illustrated in the following exercise.

Exercise 22.8 *Prove that for any pair of distinct exponents $1 \leq p, p' \leq +\infty$, we have $L^p(\mathbb{R}, dx) \not\subset L^{p'}(\mathbb{R}, dx)$.*

22.2.3 Minkowski's inequality

Theorem 22.4 (Minkowski's inequality) *Let $(\Omega_1, \Sigma_1, \mu_1)$ and $(\Omega_2, \Sigma_2, \mu_2)$ be two sigma-finite measure spaces, and suppose $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is a non-negative, $\mu_1 \times \mu_2$ measurable function. Then*

$$\int_{\Omega_2} \left[\int_{\Omega_1} f(x,y)^p \, \mu_1(dx) \right]^{1/p} \mu_2(dy) \geq \left(\int_{\Omega_1} \left[\int_{\Omega_2} f(x,y) \, \mu_2(dy) \right]^p \mu_1(dx) \right)^{1/p}$$

We will omit the proof of this theorem, although it is a nice application of Hölder's inequality that also uses the Monotone Convergence Theorem and Fubini's Theorem. See Lieb & Loss if you are interested.

The following corollary is a special case of the previous theorem. It is also often referred to as Minkowski's inequality. It is the triangle inequality for the L^p-norm.

Corollary 22.5 *For $1 \leq p \leq +\infty$, and any $f, g \in L^p(\Omega, \mu)$ we have*

$$\|f + g\|_p \leq \|f\|_p + \|g\|_p$$

22.2.4 Chebyshev's inequality

Chebyshev's inequality is an elementary but very useful inequality in probability theory.

Theorem 22.6 (Chebyshev's inequality) *For all $f \in L^p(\Omega, \mu)$, $0 < p < \infty$, and $\epsilon > 0$ one has*

$$\mu(\{x \in \Omega \mid |f(x)| > \alpha\}) \leq \frac{1}{\alpha^p} \|f\|_p^p$$

Proof: Define $S_\alpha \in \Sigma$ by

$$S_\alpha = \{x \in \Omega \mid |f(x)| > \alpha\}$$

Then

$$\begin{aligned} \|f\|_p^p &= \int_\Omega |f(x)|^p \, \mu(dx) \\ &\geq \int_{E_\alpha} |f(x)|^p \, \mu(dx) \\ &\geq \alpha^p \int_{E_\alpha} 1 \, \mu(dx) \\ &= \alpha^p \mu(E_\alpha) \end{aligned}$$

which is what we had to prove. ∎

Note that Chebyshev's inequality holds for all $p \in (0, \infty)$.

22.3 Continuous linear functionals; weak convergence

Let $L^p(\Omega, \mu)^*$ denote the space of all continuous (i.e., bounded) linear functionals on $L^p(\Omega, \mu)$.

Example:

Suppose $g \in L^q(\Omega, \mu)$, with q the exponent dual to p, i.e., $1/p + 1/q = 1$. Define φ_g by

$$\varphi_g(f) = \int_\Omega g(x) f(x) \, \mu(dx) \tag{22.10}$$

Clearly, φ_g is a linear functional where it is well-defined. Hölder's inequality shows that φ_g is in fact well-defined on all of L^p, and bounded, with an explicit bound on its norm given by

$$||\varphi_g||_{(L^p)^*} \leq ||g||_q$$

In one of the following subsection we will see that, if $1 \leq p < +\infty$ (note the strict inequality!), the linear functionals of the form φ_g, with $g \in L^q$, are in fact all of $(L^p)^*$, but for the moment this is immaterial.

As we have done several times before, the continuous linear functionals can be used to define a new topology on L^p, the topology of *weak (L^p) convergence*. Let f_n be a sequence in L^p. If there exists $f \in L^p$ such that

$$\lim_n \varphi(f_n) = \varphi(f), \quad \text{for all } \in (L^p)^*$$

we say that f_n converges to f, weakly in L^p. As expected, $f_n \to f$, weakly, does not imply $f_n \to f$, strongly, i.e., for the L^p-norm.

In order to understand how f_n may converge to f weakly but not strongly it is sufficient to consider the case $f = 0$. There are essentially three ways in which this can happen. These three are were fittingly baptized by Lieb & and Loss as

1. f_n *oscillates to death*: e.g. $f_n(x) = \sin nx$ in $L^p([0, \pi], dx)$, for any $1 \leq p < \infty$.
2. f_n *goes up the spout*: e.g. $f_n(x) = n^{1/p} g(nx)$, for a fixed $g \in L^p(\mathbb{R})$.
3. f_n *wanders off to infinity*: e.g. $f_n(x) = g(x - n)$, for a fixed $g \in L^p(\mathbb{R})$.

Note that, in the case of the Hilbert space $L^2(\Omega, \mu)$, all three of these possibilities lead to expansion coefficients of f_n in any o.n. basis that wander off to infinity as the bases index tends to infinity.

Exercise 22.9 *Prove that in each of the three examples above one has $f_n \to 0$ weakly, but not strongly.*

22.3.1 The Hahn-Banach Theorem

The Hahn-Banach theorem is a fundamental result in functional analysis, i.e., the theory of topological vector spaces (see. e.g., Rudin ***). We will prove it here for the special case of the L^p spaces. In words the Hahn-Banach Theorem says that *continuous linear functionals separate*, meaning that, if $f \neq g$, then there is a continuous linear functional φ, such that $\varphi(f) \neq \varphi(g)$. Due to linearity this can be expressed as follows.

Theorem 22.7 (Hahn-Banach Theorem) *Suppose $f \in L^p(\Omega, \mu)$ satisfies*

$$\varphi(f) = 0, \qquad \textit{for all } \varphi \in L^p(\Omega, \mu)^*$$

Then, $f = 0$.

Proof: The idea of the proof is to show that for every $f \in L^p$ there is an element $g \in L^q$, with q the exponent dual to p such that the functional φ_g defined in (22.10) satisfies

$$\varphi_g(f) = \|f\|_p^p$$

Clearly, the existence of such a functional for each f implies the statement of the theorem.

First, we treat the case $1 < p < \infty$. Simply put

$$g(x) = \begin{cases} |f(x)|^{p-2}\overline{f(x)} & \text{if } f(x) \neq 0 \\ 0 & \text{if } f(x) = 0 \end{cases}$$

Then $\varphi_g(f) = \|f\|_p^p$. We prove that φ_g is a continuous linear functional by showing that $g \in L^q$, with $q = (1 - 1/p)^{-1}$. But this is clear as

$$|g(x)|^q = |f(x)|^{(p-1)/(1-1/p)} = |f(x)|^p$$

This argument can be adapted to the special cases $p = 1$ and $p = \infty$ as follows.

If $p = 1$, we define

$$g(x) = \begin{cases} \frac{\overline{f(x)}}{|f(x)|} & \text{if } f(x) \neq 0 \\ 0 & \text{if } f(x) = 0 \end{cases} \tag{22.11}$$

Then, $|g(x)| \leq 1$, and hence $g \in L^\infty$. Again, $\varphi_g(f) = \|f\|_1$.

If $p = \infty$, we modify our strategy slightly as follows: for every $\epsilon > 0$ the level set

$$E_\epsilon = \{x \in \Omega \mid |f(x)| \geq \|f\|_\infty - \epsilon\}.$$

has non-vanishing measure. Then, from the sigma-finiteness and the sigma-additivity of μ it follows that there is $A \in \Sigma$ such that

$$\mu(A) < \infty, \quad |f(x)| \geq \|f\|_\infty - \epsilon$$

If $f \neq 0$, we have $\|f\|_\infty \neq 0$. So, we can take $0 < \epsilon < \|f\|_\infty$, and we can define

$$g(x) = \begin{cases} \frac{\overline{f(x)}}{\mu(A)|f(x)|} & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Then, $g \in L^1$ and

$$\varphi_g(f) = \frac{1}{\mu(A)} \int_A |f(x)|\, \mu(dx) \geq \|f\|_\infty - \epsilon > 0$$

and this concludes the proof. ■

Note that the sigma-finiteness of the measure was used only in the proof of the case $p = \infty$. Even in that case one can dispose of it, but then the proof is no longer constructive.

This theorem implies, among other things, that the weak topology is Hausdorff, e.g., a sequence in L^p can converge to at most one limit.

22.3.2 The Uniform Boundedness Principle

The following theorem is also a special case of a more general theorem (with the same name) in functional analysis.

Theorem 22.8 (Uniform Boundedness Principle for L^p-spaces) *Let $1 \leq p \leq +\infty$ and let f_n be a sequence in $L^p(\Omega, \mu)$, such that for all $\varphi \in (L^p)^*$, there exists $M_\varphi \geq 0$ such that*

$$|\varphi(f_n)| \leq M_\varphi, \qquad \text{for all } n$$

Then f_n is a bounded sequence for the L^p-norm, i.e., there exists a constant $M \geq 0$ such that

$$\|f_n\|_p \leq M$$

Proof: Following Lieb & Loss, we work out the proof in detail for the case $1 < p < \infty$, and leave it to the reader to make the necessary modifications to treat the cases $p = 1$ and $p = \infty$, in the same spirit as in the proof of Theorem 22.7

Suppose that $\|f_n\|_p$ is *not* bounded. We will then construct a functional $\varphi \in (L^p)^*$ such that the sequence $\varphi(f_n)$ is unbounded as well.

First, we observe that for any sequence (f_n) in L^p such that $(\|f_n\|_p)$ is unbounded, we can find a sequence (g_n) in L^p, such that

$$\|g_n\| = 4^n, \qquad \text{and } |\varphi(g_n)| \leq |\varphi(f_{k_n})|$$

where (f_{k_n}) is a subsequence of (f_n). It suffices to choose (f_{k_n}) such that $\|f_{k_n}\|_p \geq 4^n$ and to define

$$g_n = \frac{4^n}{\|f_{k_n}\|_p} f_{k_n}$$

From this observation is follows that it is sufficient to show that for every sequence (g_n) in L^p such that $\|g_n\|_p = 4^n$, we can find a $\varphi \in (L^P)^*$ such that $\varphi(g_n)$ is unbounded. The proof of Theorem 22.7 suggests how to do this.

Define functions h_n by

$$h_n(x) = \frac{|g_n(x)|^{p-2}\overline{g_n(x)}}{\|g_n\|_p^{p-1}}$$

One has

$$\|h_n\|_q = 1 \tag{22.12}$$

where q is the exponent dual to p.

Next, we recursively choose complex numbers of modulus one, $e^{i\theta_n}$, $\theta_n \in \mathbb{R}$, such that

$$\begin{aligned}
\theta_1 &= 0 \\
\theta_2 &= \arg\left(\int_\Omega h_1(x)g_2(x)\,\mu(dx)\right) \\
&\vdots \\
\theta_n &= \arg\left(\sum_{k=1}^{n-1}\frac{e^{i\theta_k}}{3^k}\int_\Omega h_k(x)g_n(x)\,\mu(dx)\right) - \arg\left(\int_\Omega h_n(x)g_n(x)\,\mu(dx)\right) \\
&\vdots
\end{aligned}$$

As a result of this choice we have that, for all $n \geq 1$,

$$\int_\Omega h_n(x)g_n(x)\,\mu(dx)$$

and

$$\sum_{k=1}^{n-1}\frac{e^{i\theta_k}}{3^k}\int_\Omega h_k(x)g_n(x)\,\mu(dx)$$

have the same argument as complex numbers and, hence,

$$\begin{aligned}
\left|\sum_{k=1}^{n}\frac{e^{i\theta_k}}{3^k}\int_\Omega h_k(x)g_n(x)\,\mu(dx)\right| &\geq \frac{1}{3^n}\int_\Omega h_n(x)g_n(x)\,\mu(dx) \\
&= \frac{\|g_n\|_p}{3^n} \\
&= \left(\frac{4}{3}\right)^n
\end{aligned} \tag{22.13}$$

Now, we can define a functional φ by

$$\varphi(f) = \sum_{n=1}^{\infty} \frac{e^{i\theta_n}}{3^n} \int_\Omega h_n(x) f(x)\, \mu(dx)$$

By using Hölder's inequality and (22.12) we can bound φ as follows:

$$|\varphi(f)| \leq \sum_{n=1}^{\infty} \frac{1}{3^n} \|f\|_p$$

and, therefore, $\varphi \in (L^p)^*$.

On the other hand, using (22.13), we find

$$\begin{aligned} |\varphi(g_n)| &\geq \left| \sum_{k=1}^{n} \frac{e^{i\theta_k}}{3^k} \int_\Omega h_k(x) g_n(x)\, \mu(dx) \right| - 4^n \sum_{k=n+1}^{\infty} \frac{1}{3^n} \\ &\geq \left(\frac{4}{3}\right)^n - 4^n \frac{3^{-(n+1)}}{1-(1/3)} \\ &= \frac{1}{2}\left(\frac{4}{3}\right)^n \end{aligned}$$

showing that $\varphi(g_n)$ is unbounded. ∎

22.3.3 L^p-L^q duality

We state without proof the Riesz Representation Theorem for L^p-spaces.

Theorem 22.9 (Riesz Representation Theorem) *If $1 \leq p < \infty$, every $\varphi \in (L^p(\Omega,\mu))^*$ is of the form*

$$\varphi(f) = \int_\Omega g(x) f(x)\, \mu(dx$$

for some $g \in L^q(\Omega,\mu)$, with q such that $1/p + 1/q = 1$.

The topological dual of L^∞ contains L^1 as a strict subspace. Its full description is somewhat involved and rarely useful, so we do not give it here. If you are interested see e.g., K. Yosida, *Functional Analysis*, Springer Verlag, 5th edition (1980), section IV.9, Example 5.

For $1 < p < \infty$, we have $(L^p)^{**} = L^p$. One says that these spaces are *reflexive*.

It follows from this theorem that L^2 is its own dual, but, that $(L^p)^* \neq L^p$, if $p \neq q$ (except in the trivial case where these spaces are finite-dimensional). Therefore, L^p, with $p \neq 2$, is not a Hilbert space. It is clear that $\|\cdot\|_p$ does not satisfy the parallelogram identity if $p \neq 2$, and if the space is at least two-dimensional. The fact that the infinite-dimensional L^p spaces with $p \neq 2$ are not self-dual, , implies the stronger statement that the L^p topology cannot be derived from an equivalent norm that satisfies the parallelogram identity, i.e., one cannot find a norm such that it becomes a Hilbert space.

22.4 Approximations of L^p-functions

In this section we will suppose that Ω is a measurable subset of $\mathbb{R}^n$ (e.g. $\mathbb{R}^n$ itself). As we have seen many times, it is often useful to approximate an arbitrary element in some space as the limit of a sequence of elements with special properties. For lack of time, we will not give all proofs in this section.

22.4.1 Approximation by C^∞ functions

For $\alpha > 0$, let φ_α be the d-dimensional Gaussian given by

$$\varphi_\alpha(x) = \frac{1}{(2\pi\alpha)^{d/2}} e^{-\alpha\frac{1}{2}\|x\|^2}$$

For, $1 \leq p \leq \infty$, and $f \in L^p(\mathbb{R}^d)$, define f_α by

$$f_\alpha(x) = \varphi_\alpha * f(x) := \int_{\mathbb{R}^d} \varphi_\alpha(x-y) f(y)\, \mu(dy)$$

It is not hard to prove that, for all $\alpha > 0$, $f_\alpha \in C^\infty(\mathbb{R}^d) \cap L^p(\mathbb{R}^d, \mu)$ (See Exercise 22.10).

Exercise 22.10 *For, $\alpha > 0$, $1 \leq p \leq \infty$, $f \in L^p(\mathbb{R}^d)$, and f_α defined above, prove that*
a) $f_\alpha \in C^\infty(\mathbb{R}^d)$. *(Hint: use that $\varphi_\alpha \in \mathcal{S}(\mathbb{R}^d)$)*
b) $f_\alpha \in L^p(\mathbb{R}^d, \mu)$. *(Hint: use Young's inequality, Proposition 18.4)*

Theorem 22.10 (Approximation by C^∞-functions) *If $1 \leq p < \infty$, $C^\infty(\mathbb{R}^d) \cap L^p(\mathbb{R}^d, \mu)$ is a dense subspace of $L^p(\mathbb{R}^d, \mu)$ with the L^p-norm. In particular, the functions f_α are C^∞ approximations of f in the sense of L^p:*

$$\lim_{\alpha \downarrow 0} \|f_\alpha - f\|_p = 0$$

Exercise 22.11 *As a corollary of Theorem 22.10, prove that C_c^∞, the C^∞ functions with compact support, are also dense in L^p.*

Exercise 22.12 *Are the C^∞ functions dense in L^∞? Or the continuous functions?*

22.4.2 Approximation by simple functions

Approximation by simple functions is another common tool in L^p theory. Simple functions are measurable functions $f : \Omega \to \mathbb{C}$ that take only finitely many values, i.e., the range of f is a finite set of complex numbers. Together with measurability this means that there are a finite number n of measurable sets $A_i \in \Sigma$, and complex numbers $c_i \in \mathbb{C}$, such that

$$f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x)$$

where χ_A denotes the characteristic function of the set A.

Using the previous result on approximations by C_c^∞ functions, it is quite straigtforward to prove that the simple functions are also dense in $L^p(\mathbb{R}^d)$, for all $1 \le p < \infty$.

Theorem 22.11 (Approximation by simple functions) *Let $1 \le p < \infty$, and Ω a measurable subset of $\mathbb{R}^d$. For all $f \in L^p(\Omega, dx)$, there exists a sequence (f_n) of simple functions such that*

$$\lim_{n\to\infty} \|f - f_n\|_p = 0$$

Exercise 22.13 *Prove Theorem 22.11.*

As an application of approximation by simple functions we can now prove that $L^p(\mathbb{R})$ is separable as a metric space. This extends a property that we already proved for $L^2(\mathbb{R}^d)$ in the context of Hilbert spaces.

Theorem 22.12 (Separability of L^p) *For $1 \le p < \infty$, $L^p(\mathbb{R}^d)$ is a separable metric space.*

Proof: We have to show that $L^p(\mathbb{R}^d)$ contains a countable dense subset. The set of all simple functions that take values in the "rational complex numbers", i.e., the complex numbers with both real and imaginary part in $\mathbb{Q}$, is obviously dense in the space of all simple functions and, hence, due to Theorem 22.11, also dense in $L^p(\mathbb{R}^d, \mu)$. Of course, it is not countable, because there are far to many measurable sets. The proof of the theorem consists in showing that one can approximate any L^p function by a simple function of the form

$$f(x) = \sum_{i=1}^{n} c_i \chi_{A_i}(x)$$

with all A_i chosen from a suitable countable family of measurable sets. E.g., one can take cubes of the form $[a_1, b_1] \times \cdots \times [a_d, b_d]$,m where $a_1, \ldots, a_d, b_1, \ldots, b_d \in \mathbb{Q}$. To prove this, is straightforward but somewhat tedious. $\blacksquare$

Note that this result extends immediately to $L^p(\Omega, \mu)$, with $1 \le p < \infty$, and Ω an arbitrary measurable subset of $\mathbb{R}^d$.

22.5 Compactness in L^p-spaces

22.5.1 The Banach-Alaoglu Theorem

We will prove that every bounded sequence in L^p, with $1 < p < \infty$, has a *weakly* convergent subsequence. This usually proved as an application of the more general Banach-Alaoglu Theorem (see. e.g., W. Rudin, *Functional Analysis*, McGraw Hill, 2nd edition, 1991). Our proof will be based on the separability of the dual of L^p, i.e., L^q. The general Banach-Alaoglu Theorem does not use separability.

Theorem 22.13 (Banach-Alaoglu for L^p) *Let Ω be a measurable subset of $\mathbb{R}^d$, and (f_n) a norm-bounded sequence in $L^p(\Omega, \mu)$, with $1 < p < \infty$. Then, there exists a subsequence (f_{n_k}) weakly converging to $f \in L^p(\Omega)$.*

Proof: By the Riesz Representation Theorem 22.9 f_{n_k} converges weakly to f in L^p iff for all $g \in L^q$, with $1/p + 1/q = 1$,

$$\int_\Omega g(x) f_{n_k}(x)\, dx \to \int_\Omega g(x) f(x)\, dx \tag{22.14}$$

As L^q is separable (Theorem 22.12), there is a countable dense set of $g_i \in L^q$. This means that for any $g \in L^q$, there is a subsequence (g_{i_l}) of (g_i) such that

$$\lim_{l\to\infty} \|g_{i_l} - g\|_q = 0$$

It will be sufficient to prove (22.14) for g in that countable set, i.e., for $g = g_i$, $i = 1, 2, \ldots$ (see below). This can be done using Cantor's diagonal argument.

The sequence $(\int g_1(x) f_n(x)\, dx)$ is bounded. This follows from Hölders inequality and the boundedness of f_n in L^p. Therefore, it has a convergent subsequence, with limit L_1. This subsequence determines a corresponding subsequence of (f_n). Denote this subsequence by $(f_{1,n})$. Next, consider the sequence $(\int g_2(x) f_{1,n}(x)\, dx)$, which is again bounded and, therefore, has a convergent subsequence, with limit L_2, and denote the corresponding subsequence of $(f_{1,n})$ by $(f_{2,n})$. By repeating this argument we obtain a double sequence $(f_{m,n})$ with the following properties:

1. $(f_{1,n})_{n=1}^\infty$ is a subsequence of (f_n).
2. $(f_{m,n})_{n=1}^\infty$ is a subsequence of $(f_{m-1,n})_{n=1}^\infty$, for all $m \geq 2$.
3. $\lim_{n\to\infty} \int g_k(x) f_{m,n}(x)\, dx$ exists and equals L_k, for all $1 \leq k \leq m$.

The argument of Cantor's then shows that the subsequence $(f_{m,m})_{m=1}^\infty$ of $(f_n)_{n=1}^\infty$ satisfies

$$\lim_{m\to\infty} \int g_i(x) f_{m,m}(x)\, dx = L_i, \qquad \text{for all } i = 1, 2, \ldots$$

As the set $\{g_i\}$ is dense in L^q, and using the boundedness of (f_n), we see that $g_i \mapsto L_i$ defines a bounded linear functional on a dense subset of L^q. By Theorem 6.7 this functional extends uniquely to a bounded linear functional on all of L^q. Then, again by $L^p - L^q$ duality (Theorem 22.9), there is a element $f \in L^p$ such that

$$\lim_{m\to\infty} g(x) f_{m,m}(x)\, dx = \int g(x) f(x), dx$$

showing that the subsequence $(f_{m,m})$ converges weakly to f. ∎

Another way of expressing Theorem 22.13 is to say that bounded sets in L^p, with $1 < p < \infty$ are weakly compact.

Exercise 22.14 *Prove that the unit ball in $L^p([a,b], dx)$, $1 \le p \le +\infty$, is not strongly compact (strongly refers to the norm topology).*

Exercise 22.15 *Prove that Theorem 22.13 does not hold for $p = 1$. I.e., give an example of a bounded sequence in L^1 that does not have a weakly convergent subsequence.*

In the study of variantional problems the following theorem is sometimes useful. It shows that although a weakly convergent sequence is in general not strongly convergent, one can construct a strongly convergent sequence from it by considering convex combinations of the elements in the sequence. We say that g is a *convex combination* of $f_1, \ldots, f_n$, if there exists non-negative constants t_i, $i = 1, \ldots, n$, such that

$$\sum_{i=1}^{n} t_i = 1, \quad \text{and} \quad g = \sum_{i=1}^{n} t_i f_i \tag{22.15}$$

This is sometimes good enough, e.g., if one is looking for a minimizer of a convex functional. We skip the proof of this theorem.

Theorem 22.14 (Mazur's Theorem) *Let $1 < p < \infty$, and let (f_n) be a sequence in an arbitrary L^p space $L^p(\Omega, \mu)$ that converges weakly to $f \in L^p(\Omega, \mu)$. Then, there exists a sequence (g_n), where g_n is a convex combination of $f_1, \ldots, f_n$, such that (g_n) converges strongly to f.*

22.5.2 Strong compactness in L^p

Finally, we give a sufficient condition for a subset of $L^p(\mathbb{R}^d)$ to be compact in the norm topology. Note the similarity with the Arzelá-Ascoli Theorem, which characterizes compact sets of continuous functions with the sup-norm topology. We state the criterion for pre-compactness. Recall that a pre-compact set is a set with compact closure. So, a pre-compact set is compact iff it is closed.

Theorem 22.15 (Strong compactness in L^p spaces) *Let $1 \leq p < \infty$, and $\Omega \subset \mathbb{R}^d$. A set $A \subset L^p(\Omega, dx)$ is precompact iff the following three conditions are satisfied:*
i) A is bounded, i.e., there exists $M \geq 0$ such that

$$\|f\|_p \leq M, \quad \text{for all } f \in A$$

ii) For all $\epsilon > 0$ there exists $\Delta > 0$ such that for all $a \in \mathbb{R}^d$,with $\|a\| < \delta$, one has

$$\|T_a f - f\|_p < \epsilon, \quad \text{for all } f \in A$$

Here, T_a is the translation over a: $T_a f(x) = f(x - a)$, for all $x \in \Omega$.
iii) For all $\epsilon > 0$, there exists $R > 0$ such that

$$\int_{\|x\| \geq R} |f(x)|^p \, dx < \epsilon, \quad \text{for all } f \in A$$

The second condition is called *L^p-equicontinuity.* Note that the third condition is automatically satisfies if Ω is a bounded set, i.e., there exists $R > 0$ such that Ω is contained in a ball of radius R.

23 Sobolev Spaces

Many problems in applied analysis involve differentiable functions. Roughly speaking, *Sobolev spaces* are Banach spaces of functions with derivatives in some L^p space. We encountered a special case in the chapter on Fourier series. Here, we give the general definition and highlight some of the main results.

23.1 Definitions

In this section we will consider $\mathbb{R}^d$ with the d-dimensional Lebesgue measure. Let Ω be a measurable subset of $\mathbb{R}^d$. If $\Omega_0 \subset \Omega$, and $f : \Omega \to \mathbb{C}$, we will use the same symbol f to denote its restriction to Ω_0. Therefore, it makes sense to ask whether an element $f \in L^p(\Omega)$, is in $L^q(\Omega_0)$ or not, for any $1 \leq p, q \leq \infty$, and measurable $\Omega_0 \subset \Omega$.

It is useful to introduce *local* L^p spaces: a measurable function f is in $L^p_{\text{loc}}(\Omega)$ if $f \in L^p(K)$, for all *compact* subsets $K \subset \Omega$. Clearly, $L^p \subset L^p_{\text{loc}}$. A sequence (f_n) is called *weakly (strongly) convergent in* $L^p_{\text{loc}}(\Omega)$ if (f_n) is weakly (strongly) convergent in $L^p(K)$, for every compact $K \subset \Omega$.

For $1 \leq p \leq \infty$, any $f \in L^p(\mathbb{R}^d)$,

$$\varphi_f(g) := \int_\Omega f(x)g(x)\,\mu(dx), \qquad \text{for all } g \in \mathcal{S}(\mathbb{R}^d)$$

defines a tempered distribution. With this identification of f and φ_f implicitly understood, the distributional derivatives of f are well-defined. If there is $g \in L^q(\Omega)$ such that

$$\partial^\alpha \varphi_f = \varphi_g$$

we say that the *weak (or distributional) derivative* of f is in $L^q(\Omega)$, and write $\partial^\alpha f \in L^q(\Omega)$. A similar meaning is given to $\partial^\alpha f \in L^q_{\text{loc}}(\Omega)$, for any multi-index α.

Definition 23.1 (Sobolev Spaces) *With the notations and conventions of above we define the Sobolev space $W^{1,p}(\Omega)$ (and its local variant $W^{1,p}_{\text{loc}}(\Omega)$) by*

$$W^{1,p}_{(\text{loc})}(\Omega) = \{f \in L^p_{(\text{loc})}(\Omega) \mid \partial^i f \in L^p_{(\text{loc})}(\Omega), i = 1, \ldots, d\}$$

$W^{1,p}(\Omega)$ is Banach space for the norm

$$\|f\|_{W^{1,p}(\Omega)} = \left[\|f\|^p_{L^p(\Omega)} + \sum_{i=1}^n \|\partial^i f\|^p_{L^p(\Omega)}\right]^{1/p}$$

Similarly, one defines $W^{n,p}(\Omega)$ and $W^{n,p}_{\text{loc}}(\Omega)$ as the spaces of L^p functions with all partial derivatives of order $\leq n$, in L^p. And the norm on $W^{n,p}(\Omega)$ is given by

$$\|f\|_{W^{n,p}(\Omega)} = \left[\sum_{0\leq|\alpha|\leq n} \|\partial^\alpha f\|^p_{L^p(\Omega)}\right]^{1/p}$$

A sequence (f_n) is *weakly (strongly) convergent in the Sobolev space* $W^{n,p}(\Omega)$ iff the sequences $(\partial^\alpha f_n)$ are weakly (strongly) convergent in $L^p(\Omega)$, for all $\alpha, 0 \leq |\alpha| \leq n$. And similarly for weak and strong convergence in the local Sobolev spaces.

Example:
Recall the definition of the Cantor set $C \subset [0,1]$. The complement of the Cantor set $D = [0,1] \setminus C$, is a countable union of closed intervals $I_{n,i}$, where $|I_{n,i}| = 3^{-n}$, and there are 2^n intervals of length 3^{-n}. The *Cantor function*, f_C, or *Devil's Staircase* vanishes on C, and is constant on the intervals $I_{n,i}$. The values on these intervals are such that on each interval $I_{n+1,i}$, f_C takes the average value of the adjacent intervals of the preceding level, i.e., $I_{n,j}$ and $I_{n,j+1}$ for some j, and $f_C(0) = 0$ and $f_C(1) = 1$.

Clearly the f'_C exists and vanishes on the interior of D, which is a set of Lebesgue measure 1. So, as an element of $L^p([0,1])$, $1 \leq p \leq \infty$, f_C is a.e.differentiable and its derivative is equivalent to the zero function, although f_C is far from being a constant function. However, this is *not* the weak derivative of f_C!

Exercise 23.1 *Let $1 \leq p \leq \infty$. Prove that if $f \in C^1(0,1])$, its weak derivative coincides with the usual pointwise definition of derivative.*

Exercise 23.2 *Let $1 \leq p \leq \infty$. Prove that if $f \in L^p([0,1], dx)$, and its weak derivative is identically zero, then there exists $c \in \mathbb{C}$, such that $f(x) = c$* a.e..

If $p = 2$ the Sobolev spaces are Hilbert spaces. To remind us of this fact $W^{n,p}(\Omega)$ is usually denoted by $H^n(\Omega)$.

Exercise 23.3 *Prove that $H^n(\Omega)$ is a Hilbert space for $n \geq 1$. I.e., define the appropriate inner product and prove that $H^n(\Omega)$ is complete (Hint: use that $L^2(\Omega)$ is a Hilbert space).*

Exercise 23.4 *Of the following functions in $L^2([-1,1], dx)$, which are in $H^1([-1,1], dx)$?*
i) $x \mapsto |x|$;
ii) $x \mapsto \text{sign}x$;
iii) $x \mapsto \sum_{n=1}^\infty (1/n)^{3/2} \sin(2\pi nx)$.

Exercise 23.5 *For $\alpha = e_1, \ldots, e_d$, let ∂^α denote the linear operators defined by taking the weak derivatives of functions in the their domain $H^1(\Omega, dx) \subset L^2(\Omega, dx)$. Prove that the unbounded linear operators ∂^α are closed.*

23.2 Fourier characterization

Theorem 23.2 *Let $\hat{f}$ denote the Fourier transform of $f \in L^2(\mathbb{R}^d, dx)$. Then, $f \in H^1(\mathbb{R}^d, dx)$ iff $k \mapsto |k|\hat{f}(k)$ is in $L^2(\mathbb{R}^d)dx)$.*

23.3 The Sobolev Embedding Theorem

As an application of Fourier Series in Section 7.4.3 we proved a Sobolev Embedding Theorem for functions defined on the circle. Here, we will state a more general result. But first we need the notion of *regular domain*. In fact, there are many notions of regularity that one requires to be satisfied for the domain Ω, depending on the context. The differences among those are sometimes quite subtle. Any domain bounded by a smooth hypersurface (i.e., locally parametrically defined by smooth functions), will satisfy the necessary regularity for all the results mentioned in this chapter. Domains with outward pointing cups or needle-shaped protrusions are typical examples of domains with insufficient regularity and often the desired property does not hold for such domains. For more details on Sobolev Embedding Theorems and the regularity conditions they require, see, e.g., R.A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

It is customary to formulate the Sobolev Embedding Theorem in terms of a embedding $j : H^n(\Omega) \to C(\overline{\Omega})$ for regular, bounded open domains $\Omega \subset \mathbb{R}^d$, or $j : H^n(\mathbb{R}^d) \to C_0(\mathbb{R}^d)$, in the case of $\Omega = \mathbb{R}^d$. Part of the theorem is that such an injective map can be defined if $n > d/2$, i.e., H^n functions are continuous if $n > d/2$. The theorem also says that the natural embedding j is a continuous linear transformation, i.e., mapping bounded sets into bounded sets, when each space is considered with its own norm.

Theorem 23.3 (Sobolev Embedding Theorem) *Suppose $n > d/2 + r$, $r \geq 0$. Then*
i) $H^n(\mathbb{R}^d) \subset C_0^r(\mathbb{R}^d)$;
ii) For every regular bounded open domain Ω

$$H^n(\Omega) \subset C^r(\overline{\Omega})$$

iii) For $\Omega = \mathbb{R}^d$, or Ω regular, bounded, and open, there exists a constant M, depending on d and n, such that

$$\|f\|_{\sup} \leq M\|f\|_{H^n}, \quad \text{for all } f \in H^n(\Omega)$$

23.4 Rellich's Theorem

Rellich's theorem is a common tool to prove that a set of functions in a Sobolev space is compact. We will state it without proof.

It is obvious from the definition that, if $m \geq n+1$, $H^m(\Omega) \subset H^n(\Omega)$. The corresponding embedding $j_{m,n}$ is again a bounded linear map. Rellich's Theorem states that $j_{m,n}$ is a *compact embedding* if $m \geq n+1$, i.e., $j_{m,n}$ maps bounded subsets of $H^m(\Omega)$ into pre-compact subsets of $H^n(\Omega)$, when each space is equipped with its own norm.

Theorem 23.4 (Rellich's Theorem) *Suppose $\Omega \subset \mathbb{R}^d$ is a regular bounded open domain, and suppose $m \geq n+1$. If A be a bounded subset of $H^m(\Omega)$, then A is pre-compact as a subset of $H^n(\Omega)$.*

Exercise 23.6 *Compare Rellich's Theorem with the Arzelá-Ascoli Theorem. (Hint: replace the condition of equicontinuity in the latter theorem by the stronger condition of a uniform bound on the first-order partial derivatives.)*

24 Differential Calculus

24.1 Introduction

This course is mostly about function spaces and linear operators acting on them. Linear operators on function spaces is the natural and quite satisfactory framework for the majority of the linear equations that arise in applied mathematics. Many interesting problems, however, lead to *nonlinear* equations, formulated in terms of a nonlinear map $F : X \to Y$, between two topological spaces X and Y. There is no general theory of nonlinear maps that is as powerful as, e.g., spectral theory for linear operators. But there is a procedure, called *linearization* , which allows one to address at some of the interesting questions about nonlinear problems in terms of a linear operator or a set of linear operators. This is very similar to looking for extrema of nonlinear functions by studying its differential, which can be considered as a linear map.

For concreteness, we first mention a couple of examples. Unless otherwise indicated F is a map from X to Y, where X and Y are toplogical spaces.

1. For fixed $y_0 \in Y$, we are interested in the solutions $x \in X$ of a nonlinear equation of the form
$$F(x) = y_0 \tag{24.1}$$
A common problem is to determine the *stability* of a given solution x_0 under small perturbations of y_0. Let us assume that X and Y are topological linear spaces, and ϵ is to be thought of as a small real parameter. If y is close to y_0 we can write y as
$$y = y_0 + \epsilon y_1,$$
and ask whether (24.1) has a solution $x(\epsilon)$ close to x_0. E.g., we look for
$$x(\epsilon) = x_0 + \epsilon x_1 + \mathbf{o}(\epsilon^2)$$
such that
$$F(x(\epsilon)) = y_0 + \epsilon y_1 \quad .$$
The form (1) for the solution implies that $x(\epsilon)$ is differentiable, where the derivative of a funtion $x : \mathbb{R} \to X$, where X is a topological linear space is defined by the standard formula:
$$\frac{d}{d\epsilon}x(\epsilon) = \lim_{h\to 0} \frac{x(\epsilon + h) - x(\epsilon)}{h},$$
where the limit is in the topology of X. If F is also differentibale in the sense that there is a map $L : X \to Y$ such that
$$F(x_0 + \epsilon x_1) = F(x_0) + \epsilon L x_1 + \mathbf{o}(\epsilon^2),$$

we can expect to have solutions of the form (1) with

$$x_1 = L^{-1} y_1$$

at least of L is invertible. In the finite-dimensional situation the fact that this procedure works is the content of the *inverse function theorem* (see, e.g., Marsden, *Elementary classical analysis*). We will see how one can define L in more general situations.

2. Many evolution equations are of the form

$$u_t = F(u)$$

where $X = Y$ is the state space of the dynamical system, which we assume to be a topological linear space, and F is a map $X \to X$ (in some situations F may only be defined of a subset of X). u_t denotes the derivative with respect to t of $u : \mathbb{R} \to X$ (sometimes solutions are only for $t \geq 0$, or up to some time T), and usually one has specified an initial condition: $u(0) = u_0 \in X$. Partial differential equations can be put into this form. One important type of problem concerns the stability of stationary solutions, i.e., $u^* \in X$ such that

$$F(u^*) = 0$$

Again we may try to linearize the nonlinear equation as follows:

$$F(u^* + \epsilon v) = F(u^*) + \epsilon L v + \mathbf{o}(\epsilon^2)$$

where L is a linear transformation on X, and we look for solutions of the form $u(t) = u^* + \epsilon v(t)$. Then, v satisfies

$$v_t = Lv$$

which is a linear equation with (formal) solution

$$v(t) = e^{tL} v(0) \quad .$$

$e^{tL} v(0)$ will be well-defined, e.g., by the spectral theorem, if L is bounded above. The stationary point u^* is called *linearly (asymptotically) stable* if

$$\lim_{t \to +\infty} u^* + e^{tL} v(0) = u^*$$

which happens if there exists $\gamma > 0$ such that for all $\lambda \in \sigma(L)$ (the spectrum of L), Re $\lambda \leq -\gamma$. If Re $\lambda = 0$ for some $\lambda \in \sigma(L)$, and Re $\lambda \leq 0$ for all $\lambda \in \sigma(L)$, the stability cannot be decided from the linearization. Purely imaginary eigenvalues correspond to periodic solutions of the linear problem, which do not decay nor diverge. If Re $\lambda > 0$ for some $\lambda \in \sigma(L)$, the soltuions of the linear problem blow up exponentially. u^* is then *linearly unstable* .

3. Another class of problems are formulated in terms of the minimization of a functional $F : X \to \mathbb{R}$. We are then looking for $x^* \in X$ such that

$$F(x^*) = \inf_{x \in X} F(x)$$

A maybe naive but sometimes useful appraoch may be to look for critical points, i.e., x^* such that

$$\frac{d}{dt} F(x^* + ty)\Big|_{t=0} = 0, \qquad \text{for all } y \in X$$

or for all y in some suitable subset.

In the above examples we tried to demonstrate the need for a notion of derivative for maps between topological linear spaces. Once again our problem is to generalize a notion which is well-understood in the finite-dimensional case (the differential of a function of finitely many variables) to the infinite-dimensional situation (differential calculus for nonlinear transformations on function spaces). We will introduce the Fréchet and Gâteaux derivatives. The Fréchet derivative generalizes the notion of differential of a vector-valued function of several variables. The Gâteaux derivative generalizes the notion of gradient or directional derivative of a scalar function on a linear space or a manifold.

24.2 The Fréchet derivative

Let X and Y be Banach spaces, and let $U \subset X$ be an open set. Consider a map $F : U \to Y$, which, in general is nonlinear.

Definition 24.1 (Fréchet derivative) *If there exists a dense linear subspace $\mathcal{D}(L) \subset X$, and a linear operator $L : \mathcal{D}(L) \to Y$, such that*

$$\lim_{x \to 0} \frac{\|F(x_0 + x) - F(x_0) - Lx\|}{\|x\|} = 0$$

L is called the Fréchet derivative *of F at x_0. If $\mathcal{D}(L) = X$ and L is bounded, F is called* Fréchet differentiable *at x_0.*

Note that if X and Y are finite-dimensional, then $F : X \to Y$ is differentiable in the standard sense iff F is Fréchet differentiable and the linear map L is determined by the matrix of the partial derivatives of the components of F evaluated in the point x_0.

Here is an infinite-dimensional example:

Example:
Let $X = H^1(\Omega)$ and $Y = L^1(\Omega)$, with Ω a regular domain $\Omega \subset \mathbb{R}^d$, and define $F : X \to Y$ by

$$F(u) = |\nabla u|^2 + |u|^2$$

In order to identify the Fréchet derivative of F at $u \in X$, consider functions $u + \epsilon v$. Then

$$\begin{aligned} F(u+\epsilon v) &= (\nabla u + \epsilon \nabla v)^2 + (u + \epsilon v)^2 \\ &= |\nabla u|^2 + 2\epsilon \nabla u \cdot \nabla v + \mathbf{o}(\epsilon^2) + |u|^2 + 2\epsilon uv + \mathbf{o}(\epsilon^2) \end{aligned}$$

Hence, if it exists, the Fréchet derivative at u is given by

$$Lv = (2\nabla u \cdot \nabla + 2u)v$$

By using the properties of the Sobolev norm, one can show that $L : H^1 \to L^1$, is a bounded linear map, and that L is the Fréchet derivative of F.

In the following exercise you are asked to generalize the property that the derivative of linear functions are constant.

Exercise 24.1 *Prove that if $F : X \to Y$ is a bounded linear map itself, then F is Fréchet differentiable and its Fréchet derivative is independent of the point $x_0 \in X$, and given by F itself.*

Exercise 24.2 *Suppose F and G are two Fréchet differentiable maps $X \to Y$. Show that $F + G$ is Fréchet differentiable with the Fréchet derivative given by the sum of the Fréchet derivatives of F and G.*

24.3 Linearization of nonlinear PDE by means of the Fréchet derivative

We will discuss linearization of nonlinear PDE by considering a class of important examples: *reaction-diffusion equations*. . In general this is a set of nonlinear PDE describing the dynamics of space-dependent chemical concentrations and temperature in the presence of chemical reactions and diffusion of reactants and heat. Suppose there are n such variables. Then, the system is described by a function $u : \Omega \times \mathbb{R} \to \mathbb{R}^n$, where $\Omega \subset \mathbb{R}^d$ is the spatial domain we wish to consider. The diffusion is described by the Laplacian and a $n \times n$ diffsuion matrix D. The reactions are described by a sources-and-sinks term $f : \mathbb{R}^n \to \mathbb{R}^n$.

$$\begin{aligned} &u_t = D\Delta u + \lambda f(u) \\ &u(x,t) = 0, \text{ for } x \in \partial\Omega \\ &u(\cdot, 0) = u_0 \end{aligned}$$

where, for concreteness, we have chosen Dirichlet boundary conditions, and u_0 is the initial condition. The nonlinearity is due to the function f which is often a polynomial in the chemical concentrations, with coefficients which may depend on temperature as exponential functions. In order to elucidate the role of the nonlinearity we introduced a real parameter λ. For simplicity of our discussion, assume $D = \mathbb{1}$. As we discussed before, an important problem is to determine the stability of a stationary solution u^* satisfying

$$\begin{aligned} &\Delta u^* + \lambda f(u^*) = 0 \\ &u^*(x) = 0, \text{ for } x \in \partial\Omega \end{aligned}$$

To study the linear stability of u^*, we have to compute the Fréchet derivative of the map $u \mapsto \Delta u + \lambda f(u)$. Due to the exercises of the previous paragraph we only need to compute the derivative of $u \mapsto f(u)$. The following lemma gives the result as well as sufficient conditions on f.

Lemma 24.2 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded regular open domain, and $f : \mathbb{R}^n \to \mathbb{R}^n$ is twice continuously differentiable with differential $Df : \mathbb{R}^n \to \mathbb{R}^{n\times n}$. Then, $u \mapsto f(u)$ as a map from $H^k(\Omega) \to L^2(\Omega)$, for any $k > d/2$, is Fréchet differentiable and its derivative at $u \in H^k(\Omega)$, is given by $v \mapsto Df(u)v$.*

Proof: Let $|u|$ denote the Euclidean norm of $u \in \mathbb{R}^n$. As f is twice continuously differentiable, by Taylor's theorem we have the following inequality: for all $R > 0$, there is constant $C(R)$ such that

$$|f(u+v) - f(u) - Df(u)v| \le C(R)|v|^2, \quad \text{for all} u, v \in \mathbb{R}^n, |u| \le R, |u+v| \le R$$

By the Sobolev embedding theorem (Theorem 23.3) and the assumptions, there is a constant M such that

$$\|u\|_\infty \le M\|u\|_{H^n}$$

By combining these inequalities and intergating over Ω, we obtain

$$\begin{aligned} \|f(u+v) - f(u) - Df(u)v\|_{L^2} &\le (C(\|u\|_\infty + \|v\|_\infty))^2\|v^2\|_{L^2} \\ &\le C'\|v\|_\infty\|v\|_{L^2} \\ &\le C''\|v\|_{H^n}^2 \end{aligned}$$

where C', and C'' are constants depending on the bounded domain Ω and u, but not on $\|v\|_\infty$, as long as $\|v\|_{H^n} \le 1$. The result follows by dividing by $\|v\|_{H^n}$ and taking the limit $\|v\|_{H^n} \to 0$. ∎

In conclusion, we have found that the Fréchet derivative of $u \mapsto \Delta u + \lambda f(u)$, in the point u, is given by

$$Lv := \Delta v + \lambda Df(u)v \quad .$$

If the functions are defined on a bounded domain Ω with suitable boundary conditions, and $Df(u)$ satisfies some reasonable conditions (cfr. Sturm-Liouville operators), L will have a normal compact inverse, and therefore, L will have a spectrum consisting entirely of eigenvalues, which will depend on the parameter λ. E.g., if $Df(u)$ is a symmetric matrix with constant entries, the eigenvalues will depend linearly on λ, and it is then easy to decide the linear stability properties of u in terms of the eigenvalues of the Laplacian on Ω.

24.4 The Gâteaux derivative

The notion of *Gâteaux derivative* or *directional derivative* generalizes the directional derivative of functions on $\mathbb{R}^n$. If the function is not differentiable, the directional derivative may exists in certain directions, but not in others, and/or it may not depend linearly on the vector specifying the direction.

Definition 24.3 *Let X be a linear space and Y a Banach space. A map $F : \mathcal{D}(F) \to Y$ is called* Gâteaux differentiable *at $x_0 \in X$, in the direction $x \in X$, if the following limit exists:*

$$\lim_{t\to 0} \frac{F(x_0 + tx - F(x_0)}{t} =: F'_x(x_0) \quad .$$

F is called Gâteaux differentiable *or* weakly differentiable *at x_0 if this limit exists for all $x \in X$.*

Note that if F is Fréchet differentiable then F is Gâteaux differentiable, and the map

$$x \mapsto \lim_{t\to 0} \frac{F(x_0 + tx - F(x_0)}{t}$$

is linear and bounded and equal to the Fréchet derivative. However, th converse is not generally true. If F is Gâteaux differentiable, and the map 24.4 is linear and bounded, one can still not conclude that F is Fréchjet differentiable (in x_0). Observe that the limit required for the Fréchet derivative is as $\|x\| \to 0$, which implies that the limits $t \to 0$ in the Gâteaux derivative are uniform in the direction specified by x.

Note that in the case $X = \mathbb{R}^d, Y = \mathbb{R}$, and F differentiable, the directional derivative is given by the gradient of F:

$$F'_{x_0}(x) = \nabla F(x_0) \cdot x$$

As an example in the infinite-dimensional case we will compute the Gâteaux derivative of the L^p norm.

Example:
Let $X = L^p(\Omega, \mu)$, $1 \leq p < +\infty$, and consider $F : X \to \mathbb{R} : f \mapsto \|f\|_p^p$. Then

$$F'_f(g) = \frac{d}{dt} \int |f + tg|^p$$

First, we show that we can interchange the derivative and the integral by the dominated convergence theorem (Theorem 21.6).

$$\frac{d}{dt}|f(x) + tg(x)|^p = \lim_{t\to 0} \frac{1}{t} \left(|f(x) + tg(x)|^p - |f(x)|^p\right)$$

As the function $x \mapsto |x|^p$, for $p \geq 1$, is convex, we have, for $0 < t \leq 1$:

$$|f(x) + tg(x)|^p \leq t|f(x) + g(x)|^p + (1-t)|f(x)|^p$$

and hence

$$\frac{1}{t} \left(|f(x) + tg(x)|^p - |f(x)|^p\right) \leq |f(x) + g(x)|^p - |f(x)|^p$$

and similarly

$$|f(x)|^p - |f(x) - g(x)|^p \leq \frac{1}{t} \left(|f(x) + tg(x)|^p - |f(x)|^p\right)$$

These two inequalities imply

$$\frac{1}{t}\left||f(x) + tg(x)|^p - |f(x)|^p\right| \leq |f(x)|^p + |f(x) + g(x)|^p + |f(x) - g(x)|^p$$

The three terms in the RHS are in L^1. Therefore, we can apply the dominated convergence theorem if the integrand converges pointwise a.e. as $t \to 0$, which is the case:

$$\frac{d}{dt}|f(x) + tg(x)|^p\Big|_{t=0} = p|f(x)|^{p-2}\frac{1}{2}(\overline{f(x)}g(x) + f(x)\overline{g(x)})$$

Note that we were careful and wrote $|z|^p$ as $|z\overline{z}|^{p/2}$, before differentiating. Hence

$$F'_f(g) = \frac{p}{2} \int |f|^{p-2}(\overline{f}g + f\overline{g}) \tag{24.2}$$

This is a bounded real-linear functional on L^p due to Hölder's inequality: $|f|^{p-1} \in L^q$, with q the index dual to p.

We have proved that the L^p norm is Gâteaux differentiable on L^p and the Gâteaux derivative is given by the bounded real-linear functional given in (24.2).

Exercise 24.3 *Is the L^p norm Fréchet differentiable on L^p?*

We already noted that the notion of Gâteaux derivative is weaker than the Fréchet derivative. Its mere existence does not imply that is is continuous in the direction x or even linear in x, two necessary properties for Fréchet differentiability. One can prove that if the Gâteaux derivative $F'_x(x_1)$ exists as a bounded linear operator for all x_1 in an open neighborhood of x_0 and is, as a function $x_1 \to (\mathcal{B}(X,Y), \|\cdot\|)$, it is continuous at x_0, the F is Fréchet differentiable at x_0. This result generalize the theorem relating continuity of all first order partial derivatives to differentiability. As the proof of this statement requires the general Hahn-Banach theorem for Banach spaces, we will not prove this here. We also warn the student that there is little consistency in the literature in the usage of the words Gâteaux derivative, differential, differentiable etc..

Finally, we mention that is is the stronger notion of Fréchet derivative that is needed to generalize the basic theorems of calculus (such as the fundamental theorem of calculus and Taylord' theorem) to the infinite dimensional situation. See R.D. Milne, *Applied Functional Analysis*, for a more detailed account and applications. For other applications the Gâteaux derivative is sufficient. E.g., if the $f : X \to \mathbb{R}$, and F has a minimun or a maximum at x_0, then all Gâteaux derivatives of F at x_0 that exist, must vanish.

Exercise 24.4 *Let $g : \mathbb{R} \times [0,1] \to \mathbb{R}$ be a continuous function such that for each $t \in [0,1]$, the function $g(\cdot, t) \in C^1(\mathbb{R})$. Define a functional $\varphi : C([0,1]) \to \mathbb{R}$ by*

$$\varphi(f) = \int_0^1 g(f(t), t)dt$$

Determine the differentiability properties of φ.

24.5 Functional derivatives

The *functional derivative* of a functional on a function space, also called the *variational derivative*, can be regarded as a special case of the Gâteaux derivative. We limit ourselves to an example that illustrates the most important aspects of this notion.

***

24.6 The Euler-Lagrange equations

Langrangian mechanics can be formulated as a variational principle of the following type. One looks for a differentiable function $x : [0,1] \to \mathbb{R}^n$, satisfying boundary conditions, say $x(0) = x(1) = 0$, for simplicity. x must be a critical point of a Lagrangian functional $\mathcal{L}$ of the following form:

$$\mathcal{L}(x) = \int_0^1 L(x, \dot{x}, t)dt$$

where L is a function $\mathbb{R}^n \times \mathbb{R}^n \times [0,1] \to \mathbb{R}$, and $\dot{x}$ denotes the derivative of x with respect to t. If L has continuous partial derivatives one can show that the vanishing of the Gâteaux derivatives in all directions (or sufficiently many directions, e.g., taken from a test function space), implies that x satisfies -the so-called *Euler-Lagrange equations*:

$$\frac{d}{dt} L_{\dot{x}}(x, \dot{x}, t) - L_x(x, \dot{x}, t) = 0$$

24.7 Appendix: the inverse and implicit function theorems

Recall that we say that $f : D \subset \mathbb{R}^n \to \mathbb{R}^m$ is C^1 iff its *differential*

$$Df = (\frac{\partial f_i}{\partial x_j})$$

exists and is continuous. In the case $n = m$, Df is a square matrix and its determinant is called the *Jacobian* of f.

Theorem 24.4 (Inverse Function Theorem) *Suppose $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$, where D is an open set, is C^1 on D. Let $x_0 \in D$ be such that $\det Df(x_0) \neq 0$. Then, there exists an open neighborhood $V \subset D$ of x_0, and an open neighborhood $W \subset \mathbb{R}^n$ of $f(x_0)$, such that, f has an inverse on U, say f^{-1}, which is C^1 and one-to-one and onto $W \to V$.*

Under the same conditions one can also show that the differential of the inverse function is the inverse matrix of the differential of f, i.e., for every $y \in W$, and $x = f^{-1}(y)$, we have

$$Df^{-1}(y) = (Df(x))^{-1}$$

One can also prove that if f is C^p, $p \geq 1$, then so is f^{-1}.

Theorem 24.5 (Implicit Function Theorem) *Let $D \subset \mathbb{R}^n \times \mathbb{R}^m$ be an open set, and suppose $F : D \to \mathbb{R}^m$ is C^p, $p \geq 1$. Let $(x_0, y_0) \in D$ be a point such that*

$$\det(\frac{\partial}{\partial y_j} F_i(x, y))\Big|_{(x_0, y_0)} \neq 0$$

The, there exists an open neighborhood $V \subset \mathbb{R}^n$ of x_0, and an open neighborhood $W \subset \mathbb{R}^m$ of y_0, such that there is a unique function $V \to W$ satisfying

$$F(x, f(x)) = 0, \text{ for all } x \in V,$$

and f is C^p.

The proof of the inverse function theorem uses the contraction mapping principle (Theorem 4.2). The implicit function theorem can be proved by using the inverse function theorem. For the proofs and applications see Marsden & Hoffman, *Elemantary Classical Analysis.*

Index