Assorted notes on functional analysis

Harald Hanche-Olsen

harald.hanche-olsen@ntnu.no

Abstract. These are supplementary notes for a course on functional analysis. The notes were first made for the course in 2004. For 2005, those notes were worked into a single document and some more material has been added. Only minor changes have been made since then.

The basic text for the course was Kreyszig's *Functional analysis*. These notes are only intended to fill in some material that is not in Kreyszig's book, or to present a different exposition.

Version 2022-02-24

Chapter 1

Transfinite induction

Chapter abstract.

Transfinite induction is like ordinary induction, only more so. The salient feature of transfinite induction is that it works by not only moving beyond the natural numbers, but even works in uncountable settings.

Wellordering

A *strict partial order* on a set S is a binary relation, typically written as $<$ or $\prec$ or some similar looking symbol (let us pick $\prec$ for this definition), which is *transitive* in the sense that, if $x \prec y$ and $y \prec z$, then $x \prec z$, and *antireflexive* in the sense that $x \prec x$ *never* holds. The order is called *total* if, for every $x, y \in S$, either $x \prec y$, $x = y$, or $y \prec x$. We write $x \succ y$ if $y \prec x$.

Furthermore, we write $x \preccurlyeq y$ if $x \prec y$ or $x = y$. When $\prec$ is a partial order then $\preccurlyeq$ is also a transitive relation. Furthermore, $\preccurlyeq$ is *reflexive*, i.e., $x \preccurlyeq x$ always holds, and if $x \preccurlyeq y$ and $y \preccurlyeq x$ both hold, then $x = y$. If a relation $\preccurlyeq$ satisfies these three conditions (transitivity, reflexivity, and the final condition) then we can define $\prec$ by saying $x \prec y$ if and only if $x \preccurlyeq y$ and $x \neq y$. This relation is then a partial order (exercise: prove this). We will call a relation of the form $\preccurlyeq$ a *nonstrict partial order*.

There is clearly a one-to-one correspondence between strict and nonstrict partial orders. Thus we often use the term *partial order* about one or the other. We rely on context as well as the shape of the symbol used (whether it includes something vaguely looking like an equality sign) to tell us which kind is meant.

An obvious example of a total order is the usual order on the real numbers, written $<$ (strict) or $\leq$ (nonstrict).

A much less obvious example is the *lexicographic* order on the set $\mathbb{R}^{\mathbb{N}}$ of sequences $x = (x_1, x_2, \ldots)$ of real numbers: $x < y$ if and only if, for some n, $x_i = y_i$ when $i < n$ while $x_n < y_n$. Exercise: Show that this defines a total order.

An example of a partially ordered set is the set of all real functions on the real line, ordered by $f \leq g$ if and only if $f(x) \leq g(x)$ for all x. This set is not totally ordered. For example, the functions $x \mapsto x^2$ and $x \mapsto 1$ are not comparable in this order.

Another example is the set $\mathscr{P}(S)$ of subsets of a given set S, partially ordered by inclusion $\subset$. This order is not total if S has at least two elements.

A *wellorder* on a set S is a total order $\prec$ so that every nonempty subset $A \subseteq S$ has a smallest element. That is, there is some $m \in A$ so that $m \preccurlyeq a$ for every $a \in A$.

One example of a wellordered set is the set of natural numbers $\{1, 2, \ldots\}$ with the usual order.

Morover, every subset of a wellordered set is wellordered in the inherited order.

1 Proposition. (Principle of induction) *Let S be a wellordered set, and $A \subseteq S$. Assume for every $x \in S$ that, if $y \in A$ for every $y \prec x$, then $x \in A$. Then $A = S$.*

Proof: Let $B = S \setminus A$. If $A \neq S$ then $B \neq \varnothing$. Let x be the smallest element of B. But then, whenever $y \prec x$ then $y \in A$. It follows from the assumption that $x \in A$. This is a contradiction which completes the proof. ∎

An *initial segment* of a partially ordered set S is a subset $A \subseteq S$ so that, if $a \in A$ and $x \prec a$, then $x \in A$. Two obvious examples are $\{x \in S \colon x \prec m\}$ and $\{x \in S \colon x \preccurlyeq m\}$ where $m \in S$. An initial segment is called *proper* if it is not all of S.

Exercise: Show that every proper initial segment of a wellordered set S is of the form $\{x \in S \colon x \prec m\}$ where $m \in S$.

A map $f \colon S \to T$ between partially ordered sets S and T is called *order preserving* if $x \prec y$ implies $f(x) \prec f(y)$. It is called an *order isomorphism* if it has an inverse, and both f and f^{-1} are order preserving. Two partially ordered sets are called *order isomorphic* if there exists an order isomorphism between them.

Usually, there can be many order isomorphisms between order isomorphic sets. However, this is not so for wellordered sets:

2 Lemma. *If S and T are wellordered and $f \colon S \to T$ is an order isomorphism of S to an initial segment of T, then for each $s \in S$, $f(s)$ is the smallest $t \in T$ greater than every $f(x)$ where $x \prec s$.*

Proof: Let S' be the initial segment of T so that f is an order isomorphism of S onto S', and let $s \in S$ be arbitrary. Let

$$z = \min\{t \in T \colon t \succ f(x) \text{ for all } x \prec s\}.$$

Such a z exists, for the set on the righthand side contains $f(s)$, and so is nonempty. In particular, $z \preccurlyeq f(s)$. If $z \prec f(s)$ then $f(x) \neq z$ for all $x \in S$: For if $x \prec s$ then $f(x) \prec z$ by the definition of z, and if $x \succcurlyeq s$ then $f(x) \succcurlyeq f(s) \succ z$. But then S' is not an initial segment of T, since $z \neq S'$ but $z \prec f(s) \in S'$. Thus $z = f(s)$ and the proof is complete. ∎

3 Proposition. *There can be only one order isomorphism from one wellordered set to an initial segment of another.*

Proof: Let S and T be wellordered, and f, g two order isomorphisms from S to initial segments of T. We shall prove by induction that $f(x) = g(x)$ for all $x \in S$. We do this by applying Proposition 1 to the set of all $s \in S$ for which $f(s) = g(s)$.

Assume, therefore, that $s \in S$ and that $f(x) = g(x)$ for all $x \prec s$.

By Lemma 2, then

$$\begin{aligned} f(s) &= \min\{t \in T\colon\ t \succ f(x) \text{ for all } x \prec s\} \\ &= \min\{t \in T\colon\ t \succ g(x) \text{ for all } x \prec s\} = g(s), \end{aligned}$$

and so $f = g$ by induction. ∎

4 Proposition. *Given two wellordered sets, one of them is order isomorphic to an initial segment of the other (which may be all of the other set).*

Proof: Let S and T be wellordered sets, and assume that T is not order isomorphic to any initial segment of S. We shall prove that S is order isomorphic to an initial segment of T.

Let W be the set of $w \in S$ so that $\{y \in S\colon\ y \preccurlyeq w\}$ is order isomorphic to an initial segment of T.

Clearly, W is an initial segment of S. In fact, if $w_1 \in S$ and $w_2 \in W$ with $w_1 \prec w_2$ and we restrict the order isomorphism of $\{y \in S\colon\ y \preccurlyeq w_2\}$ to the set $\{y \in S\colon\ y \preccurlyeq w_1\}$, we obtain an order isomorphism of the latter set to an initial segment of T. By using Lemma 2, we conclude that the union of all these mappings is an order isomorphism f of W to an initial segment of T. Since T is not order isomorphic to an initial segment of S, $f[W] \neq T$.

Assume that $W \neq S$. Let m be the smallest element of $S \setminus W$. Extend f by letting $f(m)$ be the smallest element of $T \setminus f[W]$. Then the extended map is an order isomorphism, so that $m \in W$. This is a contradiction.

Hence $W = S$, and the proof is complete. ∎

It should be noted that if S and T are wellordered and each is order isomorphic to an initial segment of the other, then S and T are in fact order isomorphic. For otherwise, S is order isomorphic to a proper initial segment of itself, and that is impossible (the isomorphism would have to be the identity mapping).

Thus we have (almost) a total order on all wellordered sets. Given two wellordered sets S and T, precisely one of the following conditions holds: Either S is order isomorphic to a proper initial segment of T, or T is order isomorphic to a proper initial segment of S, or S and T are order isomorphic.

Later we shall see how *ordinal numbers* can be used to keep track of the isomorphism classes of wellordered sets and their order relation.

5 Theorem. (The wellordering principle) *Every set can be wellordered.*

Proof: The proof relies heavily on the axiom of choice. Let S be any set, and pick a "complementary" choice function $c\colon \mathscr{P}(S)\setminus\{S\}\to S$.

More precisely, $\mathscr{P}(S)$ is the set of all subsets of S, and so c is to be defined on all subsets of S with *nonempty complement*. We require that $c(A)\in S\setminus A$ for each A. This is why we call it a complementary choice function: It chooses an element of each nonempty complement for subsets of S.

We consider subsets G of S. If G is provided with a wellorder, then G (and the wellorder on it) is called *good* if

$$c\big(\{x\in G\colon x\prec g\}\big)=g$$

for all $g\in G$.

The idea is simple, if its execution is less so: In a good wellorder, the smallest element must be $x_0=c(\varnothing)$. Then the next smallest must be $x_1=c(\{x_0\})$, then comes $x_2=c(\{x_0,x_1\})$, and so forth. We now turn to the formal proof.

If G_1 and G_2 are good subsets (with good wellorders $\prec_1$ and $\prec_2$) then one of these sets is order isomorphic to an initial segment of the other. It is easily proved by induction that the order isomorphism must be the identity map. Thus, one of the two sets is contained in the other, and in fact one of them *is* an initial segment of the other. Let G be the union of *all* good subsets of S. Then G is itself a good subset, with an order defined by extending the order on all good subsets. In other words, G is the largest good subset of S.

Assume $G\neq S$. Then let $G'=G\cup\{c(G)\}$, and extend the order on G to G' by making $c(G)$ greater than all elements of G. This is a wellorder, and makes G' good. This contradicts the construction of G as the largest good subset of S, and proves therefore that $G=S$. ∎

Ordinal numbers. We can define the natural numbers (including 0) in terms of sets, by picking one set of n elements to stand for each natural number n. This implies of course

$$0=\varnothing,$$

so that will be our starting point. But how to define 1? There is one obvious item to use as the element of 1, so we define

$$1=\{0\}.$$

Now, the continuation becomes obvious:

$$2=\{0,1\},\quad 3=\{0,1,2\},\dots$$

In general, given a number n, we let its *successor* be

$$n^+ = n \cup \{n\}.$$

We define n to be an *ordinal number* if every element of n is in fact also a *subset* of n, and the relation $\in$ wellorders n.

Obviously, 0 is an ordinal number. Perhaps less obviously, if n is an ordinal number then so is n^+. Any element of an ordinal number is itself an ordinal number, and each element is in fact the set of all smaller elements.

On the other hand, you may verify that, e.g., $\{0,1,2,4\}$ is *not* an ordinal number, for though it is wellordered by $\in$, 4 is not a subset of the given set.

If m and n are ordinal numbers, then either $m = n$, $m \in n$, or $n \in m$. For one of them is order isomorphic to an initial segment of the other, and an induction proof shows that this order isomorphism must be the identity map.

For the proof of our next result, we are going to need the concept of *definition by induction*. This means to define a function f on a wellordered set S by defining $f(x)$ in terms of the values $f(z)$ for $z \prec x$. This works by letting A be the subset of S consisting of those $a \in S$ for which there exists a unique function on $\{x \in S\colon x \preccurlyeq a\}$ satisfying the definition for all $x \preccurlyeq a$, and then using transfinite induction to show that $A = S$. In the end we have a collection of functions, each defined on an initial segment of S, all of which extend each other. The union of all these functions is the desired function. We skip the details here.

6 Proposition. *Every wellordered set is order isomorphic to a unique ordinal number.*

Proof: The uniqueness part follows from the previous paragraph. We show existence.

Let S be wellordered. Define by induction

$$f(x) = \{f(z)\colon z \prec x\}.$$

In particular, this means that $f(m) = \varnothing = 0$ where m is the smallest element of S. The second smallest element of S is mapped to $\{f(m)\} = \{0\} = 1$, the next one after that to $\{0,1\} = 2$, etc.

Let

$$n = \{f(s)\colon s \in S\}.$$

Then every element of n is a subset of n. Also n is ordered by $\in$, and f is an order isomorphism. Since S is wellordered, then so is n, so n is an ordinal number. ∎

An ordinal number which is not 0, and is not the successor n^+ of another ordinal number n, is called a *limit ordinal*.

We call an ordinal number *finite* if neither it nor any of its members is a limit ordinal. Clearly, 0 is finite, and the successor of any finite ordinal is finite. Let ω be the set of all finite ordinals. Then ω is itself an ordinal number. Intuitively, $\omega = \{0, 1, 2, 3, \ldots\}$. ω is a limit ordinal, and is in fact the smallest limit ordinal.

There exist uncountable ordinals too; just wellorder any uncountable set, and pick an order isomorphic ordinal number. There is a smallest uncountable ordinal, which is called Ω. It is the set of all countable ordinals, and is a rich source of counterexamples in topology.

Arithmetic for ordinals can be tricky. If m and n are ordinals, let A and B be wellordered sets order isomorphic to m and n, with $A \cap B = \varnothing$. Order $A \cup B$ by placing all elements of B after those of A. Then $m + n$ is the ordinal number order isomorphic to $A \cup B$ ordered in this way. You may verify that $0 + n = n + 0 = n$ and $n^+ = n + 1$. However, addition is not commutative on infinite ordinals: In fact $1 + n = n$ whenever n is an infinite ordinal. (This is most easily verified for $n = \omega$.) You may also define mn by ordering the cross product $m \times n$ lexicographically. Or rather, the convention calls for reverse lexicographic order, in which $(a, b) < (c, d)$ means either $b < d$ or $b = d$ and $a < c$. For example, $0n = n0 = 0$ and $1n = n1 = n$, but $\omega 2 = \omega + \omega$ while $2\omega = \omega$:

$$\omega \times 2 \text{ is ordered } (0,0), (1,0), (2,0), \ldots, (0,1), (1,1), (2,1), \ldots,$$
$$2 \times \omega \text{ is ordered } (0,0), (1,0), (0,1), (1,1), (0,2), (1,2), \ldots$$

Zorn's lemma and the Hausdorff maximality principle

As powerful as the wellordering principle may be, perhaps the most useful method for doing transfinite induction is by Zorn's lemma. We need some definitions.

A *chain* in a partially ordered set is a subset which is totally ordered.

7 Lemma. *If $\mathscr{C}$ is a collection of chains in a partially ordered set S, and if $\mathscr{C}$ is itself a chain with respect to set inclusion, then its union $\bigcup \mathscr{C}$ is a chain in S.*

Proof: Let $a, b \in \bigcup \mathscr{C}$. There are $A, B \in \mathscr{C}$ so that $a \in A$ and $b \in B$. Since $\mathscr{C}$ is a chain, either $A \subseteq B$ or $B \subseteq A$. Assume the former. Since now $a, b \in B$ and B is a chain, a and b are comparable. We have shown that any two elements of $\bigcup \mathscr{C}$ are comparable, and so $\bigcup \mathscr{C}$ is a chain. ∎

An element of a partially ordered set is called *maximal* if there is no element of the set greater than the given element.

8 Theorem. (Hausdorff's maximality principle)
Any partially ordered set contains a maximal chain.

The maximality of the chain is with respect to set inclusion.

Proof: Denote the given, inductive, order on the given set S by $\prec$. Let $<$ be a wellorder of S.

We shall define a chain (whenever we say *chain* in this proof, we mean a chain with respect to $\prec$) on S by using induction on S. We do this by going through the elements of S one by one, adding each element to the growing chain if it can be done without destroying its chain-ness.

We shall define $f(s)$ so that it becomes a chain built from a subset of $\{x\colon x \leq s\}$ for each s. Define $f\colon S \to \mathscr{P}(S)$ by induction as follows. The *induction hypothesis* shall be that each $f(s) \subset S$ is a chain, with $f(t) \subseteq f(s)$ when $t < s$.

When $s \in S$ and $f(t)$ has been defined for all $t < s$ so that the induction hypothesis holds, let $F(s) = \bigcup_{t<s} f(t)$. Then $F(s)$ is a chain by the induction hypothesis plus the previous lemma.

Let $f(s) = F(s) \cup \{s\}$ if this set is a chain, i.e., if s is comparable with every element of $F(s)$; otherwise, let $f(s) = F(s)$. In either case, $f(s)$ is a chain, and whenever $t < s$ then $f(t) \subseteq F(s) \subseteq f(s)$, so the induction hypothesis remains true.

Finally, let $C = \bigcup_{s\in S} f(s)$. Again by the lemma, C is a chain. To show that C is maximal, assume not, so that there is some $s \in S \setminus C$ which is comparable with every element of C. But since $F(s) \subseteq C$, then s is comparable with every element of $F(s)$. Thus by definition $s \in f(s)$, and so $s \in C$ – a contradiction. ∎

A partially ordered set is called *inductively ordered* if, for every chain, there is an element which is greater than or equal to any element of the chain.

9 Theorem. (Zorn's lemma) *Every inductively ordered set contains a maximal element.*

Proof: Let S be inductively ordered, and let C be a maximal chain in S. Let $m \in S$ be greater than or equal to every element of C. Then m is maximal, for if $s \succ m$ then s is also greater than or equal to every element of C, and so $s \in C$ because C is a *maximal* chain. ∎

Hausdorff's maximality principle and Zorn's lemma can usually be used interchangably. Some people seem to prefer one, some the other.

The wellordering principle, Zorn's lemma, and Hausdorff's maximality lemma are all equivalent to the axiom of choice. To see this, in light of all we have done so far, we only need to prove the axiom of choice from Zorn's lemma.

To this end, let a set S be given, and let a function f be defined on S, so that $f(s)$ is a nonempty set for each $s \in S$. We define a partial choice function to be a function c defined on a set $D_c \subseteq S$, so that $c(x) \in f(x)$ for each $x \in D_c$.

We create a partial order on the set $\mathscr{C}$ of such choice function by saying $c \preccurlyeq c'$ if $D_c \subseteq D_{c'}$ and c' extends c. It is not hard to show that $\mathscr{C}$ is inductively ordered. Thus it contains a maximal element c, by Zorn's lemma. If c is not defined on all of S, we can extend c by picking some $s \in S \setminus D_c$, some $t \in f(s)$, and letting $c'(s) = t$, $c'(x) = c(x)$ whenever $x \in D_c$. This contradicts the maximality of c. Hence $D_c = S$, and we have proved the axiom of choice.

We end this note with an application of Zorn's lemma. A *filter* on a set X is a set $\mathscr{F}$ of subsets of X so that

- $\varnothing \notin \mathscr{F}$, $X \in \mathscr{F}$,
- $A \cap B \in \mathscr{F}$ whenever $A \in \mathscr{F}$ and $B \in \mathscr{F}$,
- $B \in \mathscr{F}$ whenever $A \in \mathscr{F}$ and $A \subseteq B \subseteq X$.

A filter $\mathscr{F}_1$ is called *finer* than another filter $\mathscr{F}_2$ if $\mathscr{F}_1 \supseteq \mathscr{F}_2$. An *ultrafilter* is a filter $\mathscr{U}$ so that no other filter is finer than $\mathscr{U}$.

Exercise: Show that a filter $\mathscr{F}$ on a set X is an ultrafilter if, and only if, for every $A \subseteq X$, either $A \in \mathscr{F}$ or $X \setminus A \in \mathscr{F}$. (Hint: If neither A nor $X \setminus A$ belongs to $\mathscr{F}$, create a finer filter consisting of all sets $A \cap F$ where $F \in \mathscr{F}$ and their supersets.)

10 Proposition. *For every filter there exists at least one finer ultrafilter.*

Proof: The whole point is to prove that the set of all filters on X is inductively ordered by inclusion $\subseteq$. Take a chain $\mathscr{C}$ of filters, that is a set of filters totally ordered by inclusion. Let $\mathscr{F} = \bigcup \mathscr{C}$ be the union of all these filters. We show the second of the filter properties for $\mathscr{F}$, leaving the other two as an exercise.

So assume $A \in \mathscr{F}$ and $B \in \mathscr{F}$. By definition of the union, $A \in \mathscr{F}_1$ and $B \in \mathscr{F}_2$ where $\mathscr{F}_1, \mathscr{F}_2 \in \mathscr{C}$. But since $\mathscr{C}$ is a chain, we either have $\mathscr{F}_1 \subseteq \mathscr{F}_2$ or vice versa. In the former case, both $A \in \mathscr{F}_2$ and $B \in \mathscr{F}_2$. Since $\mathscr{F}_2$ is a filter, $A \cap B \in \mathscr{F}_2$. Thus $A \cap B \in \mathscr{F}$. ∎

Ultrafilters can be quite strange. There are some obvious ones: For any $x \in X$, $\{A \subseteq X : x \in A\}$ is an ultrafilter. Any ultrafilter that is not of this kind, is called *free*. It can be proved that no explicit example of a free ultrafilter can be given, since there are models for set theory without the axiom of choice in which no free ultrafilters exist. Yet, if the axiom of choice is taken for granted, there must exist free ultrafilters: On any infinite set X, one can construct a filter $\mathscr{F}$ consisting of precisely the *cofinite* subsets of X, i.e., the sets with a finite complement. Any ultrafilter finer than this must be free.

Let $\mathscr{U}$ be a free ultrafilter on $\mathbb{N}$. Then

$$U = \Big\{ \sum_{k \in A} 2^{-k} : A \in \mathscr{U} \Big\}$$

is a non-measurable subset of $[0,1]$. The idea of the proof is as follows: First, show that when $k \in \mathbb{N}$ and $A \subseteq \mathbb{N}$, then $A \cup \{k\} \in \mathscr{U}$ if and only if $A \in \mathscr{U}$. Thus the question of

membership $x \in U$ is essentially independent of any single bit of x: Whether you turn the bit on or off, the answer to $x \in U$ is the same. In particular (using this principle on the first bit), the map $x \mapsto x + \frac{1}{2}$ maps $(0, \frac{1}{2}) \cap U$ onto $(\frac{1}{2}, 1) \cap U$. In particular, assuming U is measurable, these two sets will have the same measure. But the map $x \mapsto 1 - x$ inverts all the bits of x, and so maps U onto its complement. It will follow that $(0, \frac{1}{2}) \cap U$ and $(\frac{1}{2}, 1) \cap U$ must each have measure $\frac{1}{4}$. Apply the same reasoning to intervals of length $\frac{1}{4}$, $\frac{1}{8}$, etc. to arrive at a similar conclusion. In the end one must have $|A \cap U| = \frac{1}{2}|A|$ for every measurable set A, where $|A|$ denotes Lebesgue measure. But no set U with this property can exist: Set $A = U$ to get $|U| = 0$. But we also have $|U| = |U \cap [0,1]| = \frac{1}{2}$, a contradiction.

(One of the details we have skipped in the above proof sketch concerns the dyadically rational numbers, i.e., numbers of the form $n/2^k$ for integers n and k, which have two different binary representations. In fact every dyadically rational number belongs to U (consider the binary representation ending in all ones), and so our statement that $x \mapsto 1 - x$ maps U to its complement is only true insofar as we ignore the dyadically rational numbers. However, there are only a countable number of these, so they have measure zero, and hence don't really matter to the argument.)

The existence of maximal ideals of a ring is proved in essentially the same way as the existence of ultrafilters. In fact, the existence of ultrafilters is a special case of the existence of maximal ideals: The set $\mathscr{P}(X)$ of subsets of X is a ring with addition being symmetric difference and multiplication being intersection of subsets. If $\mathscr{F}$ is a filter, then $\{X \setminus A\colon A \in \mathscr{F}\}$ is an ideal, and similarly the set of complements of sets in an ideal form a filter.

Finally we should mention that the axiom of choice has many unexpected consequences, the most famous being the Banach–Tarski paradox: One can divide a sphere into a finite number of pieces, move the pieces around, and assemble them into two similar spheres.

Further reading

A bit of *axiomatic set theory* is really needed to give these results a firm footing. A quite readable account can be found on the Wikepedia:
`http://en.wikipedia.org/wiki/Axiomatic_set_theory`

Chapter 2

Some Banach space results

Uniform boundedness

The purpose of this section is to present an alternative proof of the uniform boundedness theorem, without the need for the Baire category theorem.

11 Lemma. *Let (X,d) be a complete, nonempty, metric space, and let F be a set of real, continuous functions on X. Assume that F is pointwise bounded from above, in the following sense: For any $x \in X$ there is some $c \in \mathbb{R}$ so that $f(x) \le c$ for all $f \in F$. Then F is uniformly bounded from above on some nonempty open subset $V \subseteq X$, in the sense that there is some $M \in \mathbb{R}$ so that $f(x) \le M$ for all $f \in F$ and all $x \in V$.*

Proof: Assume, on the contrary, that no such open subset exists.

That is, for every nonempty open subset $V \subseteq X$ and every $M \in \mathbb{R}$, there exists some $f \in F$ and $x \in V$ with $f(x) > M$.

In particular (starting with $V = X$), there exists some $f_1 \in F$ and $x_1 \in X$ with $f_1(x_1) > 1$. Because f_1 is continuous, there exists some $\varepsilon_1 > 0$ so that $f_1(z) \ge 1$ for all $z \in \overline{B_{\varepsilon_1}(x_1)}$.

We proceed by induction. For $k = 2,3,\ldots$, find some $f_k \in F$ and $x_k \in B_{\varepsilon_{k-1}}(x_{k-1})$ so that $f_k(x_k) > k$. Again, since f_k is continuous, we can find some $\varepsilon_k > 0$ so that $f_k(z) \ge k$ for all $z \in \overline{B_{\varepsilon_k}(x_k)}$. In addition, we require that $B_{\varepsilon_k}(x_k) \subseteq B_{\varepsilon_{k-1}}(x_{k-1})$, and also $\varepsilon_k < k^{-1}$.

Now we have a descending sequence of nonempty closed subsets

$$X \supseteq \overline{B_{\varepsilon_1}(x_1)} \supseteq \overline{B_{\varepsilon_2}(x_2)} \supseteq \overline{B_{\varepsilon_3}(x_3)} \supseteq \cdots,$$

and the diameter of $B_{\varepsilon_k}(x_k)$ converges to zero as $k \to \infty$. Since X is complete, the intersection $\bigcap_k \overline{B_{\varepsilon_k}(x_k)}$ is nonempty; in fact, $(x_k)_k$ is a Cauchy sequence converging to the single element x of this intersection.

But now $f_k(x) \ge k$ for every k, because $x \in \overline{B_{\varepsilon_k}(x_k)}$. However that contradicts the upper boundedness of F at x, and this contradiction completes the proof. ∎

12 Theorem. (Banach–Steinhaus) *Let X be a Banach space and Y a normed space. Let $\Phi \subseteq B(X,Y)$ be a set of bounded operators from X to Y which is pointwise bounded, in the sense that, for each $x \in X$ there is some $c \in \mathbb{R}$ so that $\|Tx\| \le c$ for all $T \in \Phi$. Then Φ is uniformly bounded: There is some constant C with $\|T\| \le C$ for all $T \in \Phi$.*

Proof: Apply Lemma 11 to the set of functions $x \mapsto \|Tx\|$ where $T \in \Phi$. Thus, there is an open set $V \subseteq X$ and a constant C so that $\|Tx\| \le C$ for all $T \in \Phi$ and all $x \in V$.

Pick some $z \in V$ and $\varepsilon > 0$ so that $\overline{B_\varepsilon(z)} \subseteq V$. Also fix $c \in \mathbb{R}$ with $\|Tx\| \le c$ whenever $T \in \Phi$. Now, if $\|x\| \le 1$ then $z + \varepsilon x \in V$, and so for any $T \in \Phi$ we get

$$\|Tx\| = \|\varepsilon^{-1}\big(T(z+\varepsilon x) - Tz\big)\| \le \varepsilon^{-1}\big(\|T(z+\varepsilon x)\| + \|Tz\|\big) \le \varepsilon^{-1}(M+c).$$

Thus $\|T\| \le \varepsilon^{-1}(M+c)$ for any $T \in \Phi$. ∎

I found the above proof in Emmanuele DiBenedetto: *Real Analysis*. DiBenedetto refers to an article by W. F. Osgood: Nonuniform convergence and the integration of series term by term, *Amer. J. Math.*, **19**, 155–190 (1897). Indeed, the basic idea of the proof seems to be present in that paper, although the setting considered there is much less general: It is concerned with sequences of functions on a real interval.

I rewrote the proof a bit, splitting off the hardest bit as lemma 11.

Chapter 3

Sequence spaces and L^p spaces

Sequence spaces

A *sequence space* is a subspace of the set of all sequences $x = (x_1, x_2, \ldots) = (x_k)_{k=1}^{\infty}$. For the sake of brevity, we shall simply write $x = (x_k)_k$.

We shall be interested in *normed* sequence spaces. We shall usually consider sequences of complex numbers, though almost everything we shall say works equally well if we restrict our attention to real sequences.

All the sequence spaces we shall be concerned with in this note consist of *bounded* sequences, i.e., those for which

$$\|x\|_\infty = \sup_k |x_k| < \infty.$$

We write ℓ^∞ for the space of bounded sequences, equipped with the norm $\|\cdot\|_\infty$.

13 Proposition. *ℓ^∞ is complete.*

Proof: Consider a Cauchy sequence $(x_n)_{n=1}^{\infty}$ in ℓ^∞. Note carefully that each x_n is itself a sequence. Write $x_n = (x_{nk})_k = (x_{n1}, x_{n2}, \ldots)$. If we fix some k, then the sequence $(x_{nk})_{n=1}^{\infty}$ is a Cauchy sequence of complex numbers, because $|x_{mk} - x_{nk}| \le \|x_m - x_n\|_\infty$. Since $\mathbb{C}$ is complete this sequence has a limit, which we shall call y_k. We shall show that the limit sequence $y = (y_k)_k$ is bounded, and $\|y - x_n\|_\infty \to 0$ when $n \to \infty$.

In fact, given $\varepsilon > 0$, let N be so that $\|x_m - x_n\|_\infty < \varepsilon$ whenever $m, n \ge N$.

Then, in particular, $|x_{mk} - x_{nk}| \le \|x_m - x_n\|_\infty < \varepsilon$ for any k. Let $m \to \infty$ to get $|y_k - x_{nk}| \le \varepsilon$. As this holds for every k and $n \ge N$, we get $\|y - x_n\|_\infty \le \varepsilon$ for every $n \ge N$. Thus y is in fact bounded, and we have proved the desired convergence. ∎

Two interesting subspaces are $c_0 \subset c \subset \ell^\infty$, where c is the set of all *convergent* sequences and c_0 is the set of all sequences in c with limit zero.

14 Proposition. *c and c_0 are closed subspaces of ℓ^∞.*

Proof: We first show that c is closed. So let $x_n \in c$ for $n = 1, 2, \ldots$, and assume $x_n \to y$ with $y \in \ell^\infty$. We need to show that $y \in c$. For this, it is enough to show that y is Cauchy. Let $\varepsilon > 0$ and pick some n so that $\|y - x_n\|_\infty < \varepsilon$. Since (x_n) is convergent, it is Cauchy, so there exists some N so that $j, k \geq N \Rightarrow |x_{nj} - x_{nk}| < \varepsilon$. Then if $j, k \geq N$:

$$|y_j - y_k| \leq |y_j - x_{nj}| + |x_{nj} - x_{nk}| + |x_{nk} - y_k| < 3\varepsilon,$$

so y is indeed Cauchy.

Next, we show that c_0 is closed. To this end, define the linear functional f_∞ on c by

$$f_\infty(x) = \lim_{k\to\infty} x_k \qquad (x \in c).$$

We note that f_∞ is in fact bounded, with norm 1. Hence it is continuous, so its null space c_0 is closed. ∎

Of course c_0 and c, being closed subspaces of a Banach space ℓ^∞, are themselves Banach spaces. We shall want to identify their *dual spaces* next.

ℓ^1 is the space of *absolutely summable* sequences, i.e., the space of sequences x for which

$$\|x\|_1 = \sum_{k=1}^{\infty} |x_k| < \infty.$$

15 Proposition. *Whenever $x \in \ell^1$ and $y \in \ell^\infty$ then*

$$\sum_{k=1}^{\infty} |x_k y_k| \leq \|x\|_1 \, \|y\|_\infty.$$

Thus the sum $\sum_{k=1}^{\infty} x_k y_k$ is absolutely convergent, and

$$\left| \sum_{k=1}^{\infty} x_k y_k \right| \leq \|x\|_1 \, \|y\|_\infty.$$

In particular, any $x \in \ell^1$ defines a bounded linear functional $\tilde{x}$ on ℓ^∞, and any $y \in \ell^\infty$ defines a bounded linear functional $\tilde{y}$ on ℓ^1 by

$$\tilde{x}(y) = \tilde{y}(x) = \sum_{k=1}^{\infty} x_k y_k.$$

We have, in fact,

$$\|\tilde{x}\| = \|x\|_1 \quad \textit{and} \quad \|\tilde{y}\| = \|y\|_\infty.$$

Proof: We find, using $|y_k| \le \|y\|_\infty$,

$$\sum_{k=1}^{\infty} |x_k y_k| \le \sum_{k=1}^{\infty} |x_k| \, \|y\|_\infty = \|x\|_1 \, \|y\|_\infty,$$

which proves the first inequality. The second follows immediately from the triangle inequality for infinite sums, and the bounds

$$\|\tilde{x}\| \le \|x\|_1 \quad \text{and} \quad \|\tilde{y}\| \le \|y\|_\infty$$

are also immediate.

If $x \in \ell^1$, let $y_k = \overline{\operatorname{sgn}} \, x_k$. [1] Then $y \in \ell^\infty$, in fact $\|y\|_\infty = 1$, and

$$\tilde{x}(y) = \sum_{k=1}^{\infty} x_k y_k = \sum_{k=1}^{\infty} x_k \, \overline{\operatorname{sgn}} \, x_k = \sum_{k=1}^{\infty} |x_k| = \|x\|_1,$$

so that $\|\tilde{x}\| \ge \|x\|_1$.

Similarly, if $y \in \ell^\infty$, for any k let e_k be the sequence defined by

$$e_{kj} = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{if } j \ne k. \end{cases}$$

Then $e_k \in \ell^1$, $\|e_k\|_1 = 1$, and $\tilde{y}(e_k) = y_k$. Thus $\|\tilde{y}\| \ge |y_k|$ for every k, and taking the supremum over all k we get $\|\tilde{y}\| \ge \|y\|_\infty$. ∎

16 Proposition. *Every bounded linear functional on ℓ^1 is of the form*

$$\tilde{y}(x) = \sum_{k=1}^{\infty} x_k y_k$$

for some $y \in \ell^\infty$.

Proof: Let f be a bounded linear functional on ℓ^1. Define $e_k \in \ell^1$ as above, and let $y_k = f(e_k)$. Then $|y_k| \le \|f\| \, \|e_k\|_1 = \|f\|$, so $y \in \ell^\infty$.

Now f and $\tilde{y}$ take the same value on every vector e_k. But if $x \in \ell^1$ then

$$x = \sum_{k=1}^{\infty} x_k e_k, \tag{3.1}$$

[1] In this note, we use $\operatorname{sgn} z$ for the *complex sign* of a complex number z: $\operatorname{sgn} z = z/|z|$ if $z \ne 0$, while $\operatorname{sgn} 0 = 0$. In all cases, $z = |z| \operatorname{sgn} z$. We also write $\overline{\operatorname{sgn}} \, z = \overline{\operatorname{sgn} z}$ for the complex conjugate of the complex sign, so that in all cases $|z| = z \, \overline{\operatorname{sgn}} \, z$. (That we use overlines both for complex conjugates and for closures of sets should cause no confusion.)

the sum being convergent in ℓ^1, and so because f is bounded,

$$f(x) = \sum_{k=1}^{\infty} x_k f(e_k) = \sum_{k=1}^{\infty} x_k y_k = \tilde{y}(x).$$

It remains to prove (3.1). For any partial sum $s_k = \sum_{j=1}^{k} x_j e_j$, the vector $x - s_k$ has j-component 0 for $j \le k$, while the other components are those of x itself. So

$$\|x - s_k\|_1 = \sum_{j=k+1}^{\infty} |x_j| \to 0 \qquad (k \to \infty)$$

because $\sum_{j=1}^{\infty} |x_j| < \infty$. ∎

In brief, we state the above result by saying that *the dual space of ℓ^1 is ℓ^∞*.

17 Proposition. *Every bounded linear functional on c_0 is of the form*

$$\tilde{y}(x) = \sum_{k=1}^{\infty} x_k y_k$$

for some $y \in \ell^1$. Moreover, $\|\tilde{y}\| = \|y\|_1$.

Proof: Just like in the preceding proof, define y by $y_k = f(e_k)$. Note that (3.1) holds in c_0 as well, with convergence in c_0 (i.e., in the norm $\|\cdot\|_\infty$) – although for a very different reason, namely that $x_k \to 0$ when $k \to \infty$ and $x \in c_0$. (You should work out the details for yourself.)

Then the same argument shows that $f(x) = \sum_{k=1}^{\infty} x_k y_k$ for every $x \in c_0$.

It only remains to show that $y \in \ell^1$ and $\|y\|_1 = \|f\|$. For each k, let

$$x_{kj} = \begin{cases} \overline{\operatorname{sgn} y_j}, & \text{if } j \le k, \\ 0 & \text{otherwise.} \end{cases}$$

Then $x_k \in c_0$ and $\|x_k\|_\infty = 1$, and $f(x_k) = \sum_{j=1}^{k} |y_j|$. Thus $\sum_{j=1}^{k} |y_j| \le \|f\|$. Letting $k \to \infty$, we get $y \in \ell^1$ and $\|y\|_1 \le \|f\|$.

On the other hand,

$$|f(x)| = \Big| \sum_{k=1}^{\infty} x_k y_k \Big| \le \sum_{k=1}^{\infty} |x_k y_k| \le \sum_{k=1}^{\infty} \|x\|_\infty |y_k| = \|x\|_\infty \|y\|_1$$

proves the opposite inequality $\|f\| \le \|y\|_1$. ∎

18 Proposition. *Every bounded linear functional on c is of the form*

$$\tilde{y}(x) + \alpha f_\infty(x) = \sum_{k=1}^{\infty} x_k y_k + \alpha \lim_{k\to\infty} x_k$$

for some $y \in \ell^1$ and scalar α. Moreover,

$$\|f\| = \|y\|_1 + |\alpha|.$$

This means that, if we write $z_1 = \alpha$ and put $z_{k+1} = y_k$ for $k = 1, 2, \ldots$, then $z \in \ell^1$ and $\|f\| = \|z\|_1$. Thus, the dual of c is *also* ℓ^1, so this is an example of distinct spaces having the same dual.

Proof: Let f be a bounded linear functional on c. The restriction of f to c_0 equals $\tilde{y}$ for some $y \in \ell^1$ according to our previous result. Let $e = (1, 1, 1, \ldots) \in c$. We find that $x - f_\infty(x)e \in c_0$ whenever $x \in c$, so that

$$f\big(x - f_\infty(x)e\big) = \tilde{y}\big(x - f_\infty(x)e\big) = \sum_{k=1}^{\infty} \big(x_k - f_\infty(x)\big) y_k$$

and hence

$$f(x) = \sum_{k=1}^{\infty} x_k y_k + \alpha f_\infty(x), \qquad \alpha = f(e) - \sum_{k=1}^{\infty} y_k.$$

The estimate $\|f\| \le \|u\|_1 + |\alpha|$ is immediate. To prove the opposite inequality, for each k define $x_k \in c$ by setting

$$x_{kj} = \begin{cases} \overline{\operatorname{sgn}}\, y_j, & \text{if } j \le k, \\ \overline{\operatorname{sgn}}\, \alpha, & \text{if } j > k. \end{cases}$$

Then $\|x_{kj}\| = 1$, $f_\infty(x_k) = \overline{\operatorname{sgn}}\, \alpha$, and

$$f(x_k) = \sum_{j=1}^{k} |y_j| + \sum_{j=k+1}^{\infty} \overline{\operatorname{sgn}}\, \alpha\, y_j + |\alpha|,$$

so that we get

$$\|f\| \ge |f(x_k)| \ge \sum_{j=1}^{k} |y_j| - \sum_{j=k+1}^{\infty} |y_j| + |\alpha|.$$

Now let $k \to \infty$ to get $\|f\| \ge \|y\|_1 + |\alpha|$. ∎

L^p spaces

In this section μ is a positive, σ-finite measure on a measure space Ω.[2] [3]

Whenever we talk about functions on Ω, we shall only consider *measurable* complex functions. For any function u and real number $p > 0$, define

$$\|u\|_p = \left(\int_\Omega |u|^p \, d\mu\right)^{1/p}.$$

We also define

$$\|u\|_\infty = \operatorname*{ess.sup}_{t\in\Omega} |u(t)| = \min\{M\colon |u(t)| \le M \text{ for a.e. } t \in \Omega\}.$$

(The latter equality is the definition of the *essential supremum*. In this definition, one should first replace the minimum by an infimum, then use a bit of measure theory to show that the infimum is in fact attained, so that the minimum is defined.)

We put $\|u\|_\infty = \infty$ if there is no real number M so that $|u| \le M$ almost everywhere. To sum up:

$$|u| \le \|u\|_\infty \text{ a.e.}, \quad \mu\{t \in \Omega\colon |u(t)| > M\} > 0 \text{ if } M < \|u\|_\infty.$$

Exercise: Prove that

$$\lim_{p\to\infty} \|u\|_p = \|u\|_\infty \qquad \text{if } \mu(\Omega) < \infty.$$

When $0 < p < 1$, $\|\cdot\|_p$ is *not* a norm (the triangle inequality is not satisfied). This case is just too strange in many ways, though it is sometimes encountered. In all cases, the homogeneity $\|\alpha u\|_p = |\alpha| \|u\|_p$ is obvious when $\alpha \in \mathbb{C}$, but the triangle inequality is harder to prove. The triangle inequality for $\|\cdot\|_p$ is called *Minkowski's inequality*, and will be proved later. However, as an easy exercise you are encouraged to give direct proofs for the cases $p = 1$ and $p = \infty$.

For $0 < p \le \infty$, we define L^p to be the set of measurable functions u on Ω so that $\|u\|_p < \infty$. You should verify that, as soon as the triangle inequality is proved, it follows that L^p is a vector space. (In fact, this is true even for $0 < p < 1$, even though the triangle inequality does not hold in this case.)

We shall say that real numbers p and q are *conjugate exponents* if any (hence all) of the following equivalent conditions hold:

$$\frac{1}{p} + \frac{1}{q} = 1, \quad p + q = pq, \quad (p-1)(q-1) = 1.$$

[2] We do not bother to name the σ-algebra, but simply talk about measurable sets when we do need them.

[3] We may be able to get away with less than σ-finiteness: The most important property is that there are no atomic sets of infinite measure. An *atomic set* is a measurable subset $A \subseteq \Omega$ so that, whenever $B \subseteq A$ is measurable, then either $\mu(B) = 0$ or $\mu(B) = \mu(A)$.

In addition to these, we allow as special cases $p = 1$ and $q = \infty$, or $p = \infty$ and $q = 1$.

19 Lemma. (Young's inequality) *For $a \geq 0$, $b \geq 0$, and conjugate exponents p, q with $1 < p < \infty$,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Equality holds if and only if $a^p = b^q$.

For complex numbers a, b, and p, q as above we have

$$\operatorname{Re}(ab) \leq \frac{|a|^p}{p} + \frac{|b|^q}{q}$$

with equality if and only if $|a|^p \operatorname{sgn} a = |b|^q \overline{\operatorname{sgn}} b$.

Proof: First, assume $a > 0$ and $b > 0$. Write $a = e^{x/p}$ and $b = e^{y/q}$, and also $t = 1/p$ and $1 - t = 1/q$. Then the desired inequality is $e^{tx+(1-t)y} \leq te^x + (1-t)e^y$, which follows from the strict convexity of the exponential function. Moreover, the inequality is strict unless $x = y$, which is equivalent to $a^p = b^q$.

The case where $a \geq 0$, $b \geq 0$ and $ab = 0$ is of course obvious, and so the first part is proved.

The second part follows from the first part applied to $|a|$ and $|b|$ instead of a and b, and the fact that $\operatorname{Re} ab \leq |ab|$ with equality precisely when $ab = 0$ or $\operatorname{sgn} a = \overline{\operatorname{sgn}} b$. ∎

20 Proposition. (Hölder's inequality) *Let p, q be conjugate exponents with $1 \leq p \leq \infty$. Then*

$$\int_\Omega |uv|\, d\mu \leq \|u\|_p \, \|v\|_q.$$

for any two measurable functions u and v. In particular, when the righthand side is finite then uv is integrable, and

$$\left| \int_\Omega uv\, d\mu \right| \leq \|u\|_p \, \|v\|_q.$$

If $0 < \|u\|_p \, \|v\|_q < \infty$ and $1 < p < \infty$, equality holds in the latter inequality if and only if there is a scalar γ so that $|u|^p \operatorname{sgn} u = \gamma |v|^q \overline{\operatorname{sgn}} v$ almost everywhere.

Proof: The cases $p = 1$ and $p = \infty$ are easy and left to the reader. So we assume $1 < p < \infty$. Moreover, we may assume that $\|u\|_p < \infty$ and $\|v\|_q < \infty$, since otherwise there is nothing to prove. (The case where one norm is infinite and the other is zero is easy.) Since nothing in the statement of the proposition changes

when u and v are replaced by scalar multiples of themselves, we may even assume that $\|u\|_p = \|v\|_q = 1$.

Now we apply Young's inequality and integrate:

$$\int_\Omega |uv|\,d\mu \le \int_\Omega \Big(\frac{|u|^p}{p} + \frac{|v|^q}{q}\Big)\,d\mu = \frac{\|u\|_p^p}{p} + \frac{\|v\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1,$$

which proves the first inequality.

The integrability of uv and the second inequality follow immediately from the first and the definition of integrability together with the triangle inequality for the integral.

In order to find the proper condition for equality in the second inequality we may replace v by a scalar multiple of itself so that $\int_\Omega uv\,d\mu \ge 0$. Then we can use Young's inequality again:

$$\int_\Omega \mathrm{Re}(uv)\,d\mu \le \int_\Omega \Big(\frac{|u|^p}{p} + \frac{|v|^q}{q}\Big)\,d\mu = \frac{\|u\|_p^p}{p} + \frac{\|v\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1,$$

with equality if and only if $|u|^p \operatorname{sgn} u = |v|^q \overline{\operatorname{sgn}}\, v$ almost everywhere. The factor γ appears because of the change in v above, and because of our normalizing of u and v. ∎

21 Corollary. *Let p and q be conjugate exponents, $1 \le p \le \infty$. For any measurable function u,*

$$\|u\|_p = \sup_{\|v\|_q = 1} \int_\Omega |uv|\,d\mu.$$

If $1 \le p < \infty$, and $\|u\|_p < \infty$, there is some v with $\|v\|_q = 1$ and

$$\int_\Omega uv\,d\mu = \|u\|_p.$$

You may wonder why I call this a corollary when the proof is so long. The reason is that the proof, though lengthy, contains no deep or difficult ideas.

Proof: We prove the final part first; this will take care of most cases for the first statement.

For the case $p = 1$, if $\|u\|_1 < \infty$ let $v = \overline{\operatorname{sgn}}\, u$. Then $\|v\|_\infty = 1$ and

$$\int_\Omega uv\,d\mu = \int_\Omega |u|\,d\mu = \|u\|_1.$$

Next, if $1 < p < \infty$ and $\|u\|_p < \infty$, note that if $\|u\|_p = 0$ there is nothing to prove; otherwise, let

$$v = \overline{\operatorname{sgn}}\, u\big(|u|/\|u\|_p\big)^{p/q}.$$

Then $\|v\|_q = 1$, $uv > 0$ and the conditions for equality in Hölder's inequality hold, so that

$$\int_\Omega uv\,d\mu = \int_\Omega |uv|\,d\mu = \|u\|_p\,\|v\|_q = \|u\|_p.$$

The proof of the second part is now done.

To prove the first half, note that the second half (together with Hölder's inequality) proves the first half whenever $1 \le p < \infty$ and $\|u\|_p < \infty$, and in fact the supremum is attained in these cases. We must show the remaining cases.

Recall that Ω is assumed to be σ-finite. Hence we can find measurable sets $E_1 \subset E_2 \subset \cdots \subset \Omega$, each with finite measure, so that $E_1 \cup E_2 \cup \cdots = \Omega$.

Assume $1 \le p < \infty$ and $\|u\|_p = \infty$. Write $D_k = \{t \in E_k\colon |u(t)| < k\}$. Then $D_1 \cup D_2 \cup \cdots = \Omega$ as well. Write $u_k = u\chi_{D_k}$.[4] Then $u_k \in L^p$. (In fact, $\|u_k\|_p^p \le k^p\mu(D_k)$ since $|u_k| \le k$.) Now there is some function v_k with $\|v_k\|_q = 1$ and $\int_\Omega u_k v_k\,d\mu = \|u_k\|_p$. This function must in fact be zero almost everywhere outside D_k. Thus $\int_\Omega uv_k\,d\mu = \int_\Omega u_k v_k\,d\mu = \|u_k\|_p$. But the monotone convergence theorem implies

$$\int_\Omega |u_k|^p\,d\mu \to \int_\Omega |u|^p\,d\mu = \infty \qquad (k \to \infty),$$

since $|u_k|$ increases pointwise to $|u|$. Thus $\|u_k\|_p \to \infty$, so we can find v with $\|v\|_q = 1$ and $\int_\Omega |uv|\,d\mu$ as large as we may wish.

Only the case $p = \infty$ remains. Whenever $M < \|u\|_\infty$ there is some measurable set E with $\mu(E) > 0$ and $|u| \ge M$ on E. Using the σ-finiteness of μ, we can ensure that $\mu(E) < \infty$ as well. Let $v = \chi_E/\mu(E)$. Then $v \in L^1$, $\|v\|_1 = 1$, and

$$\int_\Omega |uv|\,d\mu = \int_E \frac{|u|}{\mu(E)}\,d\mu \ge M.$$

In other words,

$$\sup_{\|v\|_1=1} \int_\Omega |uv|\,d\mu \ge M \qquad \text{whenever } M < \|u\|_\infty.$$

Letting $M \to \|u\|_\infty$ from below, we conclude that the supremum on the left is at least $\|u\|_\infty$. But by Hölder's inequality it can be no bigger, so we have equality. ∎

22 Proposition. (Minkowski's inequality) *Whenever* $1 \le p \le \infty$,

$$\|u + v\|_p \le \|u\|_p + \|v\|_p.$$

[4] χ_{D_k} is the *characteristic function* of D_k: It takes the value 1 on D_k and 0 outside D_k. (Statisticians often use the term *indicator function* because "characteristic function" has a different meaning in statistics.)

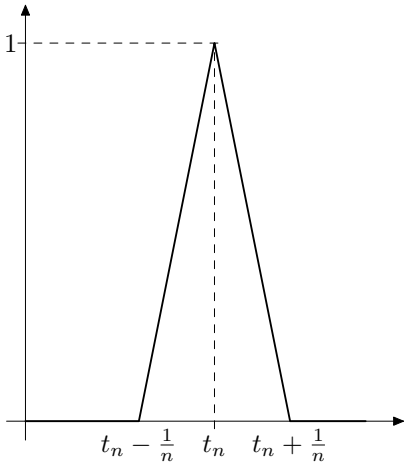
Proof: Let q be the conjugate exponent. From Corollary 21,

$$\begin{aligned}\|u+v\|_p &= \sup_{\|w\|_q=1}\int_\Omega |(u+v)w|\,d\mu \\ &\le \sup_{\|w\|_q=1}\Big(\int_\Omega |uw|\,d\mu + \int_\Omega |vw|\,d\mu\Big) \\ &\le \|u\|_p + \|v\|_p\end{aligned}$$

where we used the ordinary triangle inequality in the second line and Hölder's inequality in the final line. ∎

Now that we know that L^p is indeed a normed space when $p \ge 1$, it is time to tackle completeness. But first a word of caution.

Here is an example to demonstrate why showing completeness by considering general Cauchy sequences in L^p is difficult. Such sequences may converge in L^p, and yet diverge pointwise on a set of positive measure: Let u_n be the function shown on the right. Clearly, $\|u_n\|_p \to 0$ as $n \to \infty$ for any $p < \infty$. And yet, we can choose the center points t_n so that $u_n(t)$ does not converge to zero for any $t \in (0,1)$! Namely, let t_n be the fractional part of $\sum_{k=1}^n 1/n$. (I.e., so that the sum is an integer plus t_n, with $0 \le t_n < 1$.) Since the harmonic series diverges, we can see that $u_n(t) > \frac{1}{2}$ for infinitely many n (namely for $n = m$ or $n = m+1$ where $t_m \le t \le t_{m+1}$).



23 Lemma. *A normed space is complete if, and only if, every absolutely convergent series in the space is convergent.*

In other words, the criterion for completeness is

$$\text{if } \sum_{k=1}^\infty \|x_k\| < \infty \text{ then } \sum_{k=1}^\infty x_k \text{ converges.}$$

Proof: First, if the space is complete and $\sum_{k=1}^\infty \|x_k\| < \infty$, consider the partial sums $s_n = \sum_{k=1}^n x_k$. By the triangle inequality, for $m < n$ we get

$$\|s_n - s_m\| \le \sum_{k=m+1}^n \|x_k\| \le \sum_{k=m+1}^\infty \|x_k\| < \varepsilon$$

if m is big enough, which shows that the sequence $(s_n)_n$ is Cauchy and hence convergent.

Conversely, assume that every absolutely convergent series is convergent, and consider a Cauchy sequence $(u_n)_n$. Pick successively $k_1 < k_2 < \cdots$ so that $\|u_m - u_n\| < 2^{-j}$ whenever $m, n \geq k_j$. Put $x_1 = u_{k_1}$ and $x_j = u_{k_{j+1}} - u_{k_j}$. Then $\|x_j\| < 2^{-j}$ for $j \geq 2$, so $\sum_{j=1}^{\infty} x_j$ is absolutely convergent, and therefore convergent. Since $u_{k_{j+1}} = x_1 + x_2 + \cdots + x_j$, the sequence $(u_{k_j})_j$ is convergent. We have shown that any Cauchy sequence has a convergent subsequence, which is enough to prove completeness. ∎

24 Proposition. *L^p is complete (hence a Banach space) for $1 \leq p \leq \infty$.*

Proof: We first prove this for $1 \leq p < \infty$. Let $u_k \in L^p$, $\sum_{k=1}^{\infty} \|u_k\|_p = M < \infty$. By the Minkowski inequality,

$$\int_\Omega \Big(\sum_{k=1}^{n} |u_k|\Big)^p d\mu = \Big\|\sum_{k=1}^{n} |u_k|\Big\|_p^p \leq \Big(\sum_{k=1}^{n} \|u_k\|_p\Big)^p \leq \Big(\sum_{k=1}^{\infty} \|u_k\|_p\Big)^p = M^p.$$

By the monotone convergence theorem,

$$\int_\Omega \Big(\sum_{k=1}^{\infty} |u_k|\Big)^p d\mu \leq M^p$$

follows. Thus $\sum_{k=1}^{\infty} |u_k| < \infty$ almost everywhere, and so $\sum_{k=1}^{\infty} u_k(t)$ converges for almost every $t \in \Omega$. Let the sum be $s(t)$.

Now let $\varepsilon > 0$. Repeating the above computation with the sum starting at $k = m+1$, we find instead

$$\int_\Omega \Big(\sum_{k=m+1}^{\infty} |u_k|\Big)^p d\mu \leq \Big(\sum_{k=m+1}^{\infty} \|u_k\|_p\Big)^p < \varepsilon^p$$

if m is big enough. But

$$\Big|s(t) - \sum_{k=1}^{m} u_k(t)\Big| = \Big|\sum_{k=m+1}^{\infty} u_k(t)\Big| \leq \sum_{k=m+1}^{\infty} |u_k(t)|,$$

so

$$\Big\|s - \sum_{k=1}^{m} u_k\Big\|_p \leq \Big\|\sum_{k=m+1}^{\infty} |u_k|\Big\|_p < \varepsilon$$

for large enough m. Thus the sum converges in L^p to the limit s.

The case $p = \infty$ is similar, but simpler: Now

$$\sum_{k=1}^{\infty} |u_k(t)| \leq \sum_{k=1}^{\infty} \|u_k\|_\infty \leq \infty$$

for almost every $t \in \Omega$, so the sum $\sum_{k=1}^{\infty} u_k(t)$ is absolutely convergent, hence convergent, almost everywhere. Again let $s(t)$ be the sum. But now

$$\left| s(t) - \sum_{k=1}^{m} u_k(t) \right| = \left| \sum_{k=m+1}^{\infty} u_k(t) \right| \le \sum_{k=m+1}^{\infty} |u_k(t)| \le \sum_{k=m+1}^{\infty} \|u_k\|_\infty < \varepsilon$$

almost everywhere when m is big enough, which implies

$$\left\| s(t) - \sum_{k=1}^{m} u_k \right\|_\infty \le \varepsilon.$$

This finishes the proof. ∎

25 Lemma. *Assume that μ is a finite measure. Then, if $1 \le p < p' \le \infty$,*

$$\mu(\Omega)^{-1/p} \|u\|_p \le \mu(\Omega)^{-1/p'} \|u\|_{p'}.$$

When $p' = \infty$, we write $1/p' = 0$. Apart from the adjustment by powers of $\mu(\Omega)$, this roughly states that $\|\cdot\|_p$ is an increasing function of p. In particular, when $p < p'$ then $L^{p'} \subseteq L^p$: L^p *decreases* when p increases.

Proof: Assume $1 \le p < p' < \infty$. Write $p' = rp$, so that $r > 1$. Let s be the exponent conjugate to r. For simplicity, assume $u \ge 0$. Then

$$\begin{aligned} \|u\|_p^p &= \int_\Omega u^p \, d\mu = \int_\Omega u^p \cdot 1 \, d\mu \\ &\le \|u^p\|_r \, \|1\|_s = \left(\int_\Omega u^{rp} \, d\mu \right)^{1/r} \mu(\Omega)^{1/s} = \|u\|_{p'}^p \mu(\Omega)^{1/s} \end{aligned}$$

Raise this to the $1/p$th power and note that

$$\frac{1}{ps} = \frac{1}{p}\left(1 - \frac{1}{r}\right) = \frac{1}{p} - \frac{1}{p'},$$

and a simple rearrangement finishes the proof. ∎

When the measure space Ω is the set of natural numbers $\{1, 2, 3, \ldots\}$ and μ is the counting measure, the L^p spaces become *sequence spaces* ℓ^p, with norms

$$\|x\|_p = \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{1/p}.$$

The Hölder and Minkowski inequalities, and completeness, for these spaces are just special cases of the same properties for L^p spaces.

The inequalities between norms for different p get reversed however, compared to the case for finite measures.

26 Lemma. *Let $n \geq 2$ and $a_1, \dots, a_n$ be positive real numbers. Then*

$$(a_1^p + a_2^p + \cdots + a_n^p)^{1/p}$$

is a strictly decreasing function of p, for $0 < p < \infty$.

Proof: We prove this first for $n = 2$. We take the natural logarithm of $(a^p + b^p)^{1/p}$ and differentiate:

$$\frac{d}{dp} \ln\bigl((a^p + b^p)^{1/p}\bigr) = \frac{d}{dp} \frac{1}{p} \ln(a^p + b^p) = -\frac{1}{p^2} \ln(a^p + b^p) + \frac{a^p \ln a + b^p \ln b}{pa^p + pb^p}.$$

To show that this is negative is the same as showing that

$$pa^p \ln a + pb^p \ln b < a^p \ln(a^p + b^p) + b^p \ln(a^p + b^p).$$

But $a^p \ln(a^p + b^p) > a^p \ln a^p = pa^p \ln a$, so the first term on the righthand side is bigger than the first term on the lefthand side. The same thing happens to the second terms, so we are done with the case $n = 2$.

The general case can be proved in the same way, or one can proceed by induction. Consider the case $n = 3$: We can write

$$(a^p + b^p + c^p)^{1/p} = \bigl([(a^p + b^p)^{1/p}]^p + c^p\bigr)^{1/p}$$

Increasing p and pretending that the expression in square brackets is unchanged, we get a smaller value from the case $n = 2$. But then the square bracket also decreases, again from the case $n = 2$, and so the total expression decreases even more. The general induction step from n to $n+1$ terms is similar, using the 2-term case and the n-term case. ∎

27 Proposition. *For a sequence $x = (x_k)_k$, $\|x\|_p$ is a decreasing function of p, for $0 < p < \infty$. It is strictly decreasing provided at least two entries are nonzero, except where the norm is infinite.* ∎

Uniform convexity. A normed space is called *uniformly convex* if for every $\varepsilon > 0$ there is a $\delta > 0$ so that whenever x and y are vectors with $\|x\| = \|y\| = 1$, $\|x + y\| > 2 - \delta$ implies $\|x - y\| < \varepsilon$.

Perhaps a bit more intuitive is the following equivalent condition, which we might call *thin slices of the unit ball are small*: Given $\varphi \in X^*$ with $\|\varphi\| = 1$, define the *δ-slice* $S_{\varphi,\delta} = \{x \in X\colon \|x\| \leq 1 \text{ and } \operatorname{Re}\varphi(x) > 1 - \delta\}$. The "thin slices" condition states that for each $\varepsilon > 0$ there is some $\delta > 0$ so that, if $\varphi \in X^*$ with $\|\varphi\| = 1$, then $\|x - y\| < \varepsilon$ for all $x, y \in S_{\varphi,\delta}$.

This condition follows trivially from uniform convexity. The proof of the converse requires a minor trick: Given $x, y \in X$ with $\|x\| \le 1$, $\|y\| \le 1$ and $\|x+y\| > 2-\delta$, invoke the Hahn–Banach theorem to pick $\varphi \in X^*$ with $\|\varphi\| = 1$ and $\operatorname{Re}\varphi(x+y) > 2-\delta$. Then $\operatorname{Re}\varphi(x) = \operatorname{Re}\varphi(x+y) - \operatorname{Re}\varphi(y) > 2-\delta-1 = 1-\delta$, and similarly for y. If δ was chosen according to the "thin slices" condition, $\|x-y\| < \varepsilon$ follows.

Our interest in uniformly convexity stems from the following result. We shall prove it later – see Theorem 69 – as it requires some results we have not covered yet.

28 Theorem. (Milman–Pettis) *A uniformly convex Banach space is reflexive.* ∎

The following lemma is a very special case of the "thin slices" condition. In the lemma after it, we shall show that the special case is sufficient for a sufficiently general "thin slices" result to obtain uniform convexity of L^p.

29 Lemma. *Given $1 < p < \infty$ and $\varepsilon > 0$, there exists $\delta > 0$ so that, for every probability space (Ω, ν) and every measurable function z on Ω, $\|z\|_p \le 1$ and $\operatorname{Re}\int_\Omega z\,d\nu > 1-\delta$ imply $\|z-1\|_p < \varepsilon$.*

Proof: Consider the function

$$f(u) = |u|^p - 1 + p(1 - \operatorname{Re} u)$$

and note that $f(u) > 0$ everywhere except for the value $f(1) = 0$. (This is the case $a = u$, $b = 1$ in Young's inequality.) Further, note that $f(u)$ and $|u-1|^p$ are asymptotically equal as $|u| \to \infty$. Thus, given $\varepsilon > 0$, we can find some $\alpha > 1$ so that

$$|u-1|^p \le \alpha f(u) \text{ whenever } |u-1| \ge \varepsilon.$$

Assume that z satisfies the stated conditions, and let $E = \{\omega \in \Omega : |z(\omega) - 1| < \varepsilon\}$. Then

$$\begin{aligned}
\|z-1\|_p^p &= \int_E |z-1|^p\,d\nu + \int_{\Omega\setminus E} |z-1|^p\,d\nu \\
&\le \varepsilon^p + \alpha \int_\Omega f(z)\,d\nu \\
&\le \varepsilon^p + p\alpha\Big(1 - \int_\Omega \operatorname{Re} z\,d\nu\Big) \\
&< \varepsilon^p + p\alpha\delta.
\end{aligned}$$

Thus picking $\delta = \varepsilon^p/(p\alpha)$ is sufficient to guarantee $\|z-1\|_p < 2^{1/p}\varepsilon$. ∎

30 Lemma. *Given $1 < p < \infty$, $p^{-1} + q^{-1} = 1$ and $\varepsilon > 0$, there exists $\delta > 0$ so that the following holds: If u, w are measurable functions on a measure space Ω with $\|u\|_p \le 1$ and $\|w\|_q = 1$ and $\int_\Omega \operatorname{Re} uw\,d\mu > 1-\delta$, then $\|u - v\|_p < \varepsilon$, where v is the function satisfying $vw = |v|^p = |w|^q$ a.e.*

Proof: Let p and ε be given, and choose δ as in Lemma 29.

Let u, v and w be as stated above. Since nothing is changed by multiplying u, v by a complex function of absolute value 1, and dividing w by the same function, we may assume without loss of generality that $v \ge 0$ and $w \ge 0$.

Let $z = u/v$ where $v \ne 0$ and $z = 0$ where $v = 0$. Thus $zv = u$ where $v \ne 0$ and $zv = 0$ where $v = 0$. Since $(u - zv)zv = 0$ we find $\|u\|_p^p = \|u - zv\|_p^p + \|zv\|_p^p$. Also $\operatorname{Re}\int_\Omega zvw\,d\mu = \operatorname{Re}\int_\Omega uw\,d\mu > 1-\delta$, so $\|zv\|_p > 1-\delta$, and $\|u - zv\|_p^p < 1 - (1-\delta)^p$.

Let ν be the probability measure

$$d\nu = vw\,d\mu = v^p\,d\mu = w^q\,d\mu.$$

We find

$$\int_\Omega |z|^p\,d\nu = \int_{v\ne 0} |u|^p\,d\mu \le 1, \quad \operatorname{Re}\int_\Omega z\,d\nu = \operatorname{Re}\int_\Omega uw\,d\mu > 1-\delta.$$

By Lemma 29, we now get

$$\varepsilon^p > \int_\Omega |z-1|^p\,d\nu = \int_\Omega |z-1|^p v^p\,d\mu = \int_{v\ne 0} |u-v|^p\,d\mu.$$

On the other hand,

$$\int_{v=0} |u-v|^p\,d\mu = \int_\Omega |(u-zv|^p\,d\mu < 1 - (1-\delta)^p.$$

We therefore get $\|u - v\|_p^p < \varepsilon + 1 - (1-\delta)^p$, and the proof is complete. ∎

31 Theorem. (Clarkson) *L^p is uniformly convex when $1 < p < \infty$.*

Proof: Consider $x, y \in L^p$ with $\|x\|_p = \|y\|_p = 1$ and $\|x + y\|_p > 2 - \delta$. Let $v = (x+y)/\|x+y\|_p$, and choose $w \in L^q$ with $vw = |v|^p = |w|^q$. In particular $\|v\|_p = \|w\|_q = 1$. Then

$$\int_\Omega (x+y)w\,d\mu = \|x+y\|_p \int_\Omega vw\,d\mu = \|x+y\|_p > 2-\delta.$$

Since also $\operatorname{Re}\int_\Omega yw\,d\mu \le 1$, this implies $\operatorname{Re}\int_\Omega xw\,d\mu > 1-\delta$. If δ was chosen according to Lemma 30, we get $\|x - v\|_p < \varepsilon$. Similarly $\|y - v\|_p < \varepsilon$, and so $\|x - y\|_p < 2\varepsilon$. ∎

32 Corollary. *L^p is reflexive for $1 < p < \infty$.* ∎

It will often be useful to know that all L^p spaces share a common dense subspace. First, we prove a result showing that a L^p function cannot have very large values on sets of very large measure.

33 Lemma. (Chebyshev's inequality) *If $u \in L^p$ where $1 \le p < \infty$ then*

$$\mu\{x \in \Omega\colon |u(x)| \ge \varepsilon\} \le \varepsilon^{-p} \|u\|_p^p \qquad (\varepsilon > 0).$$

Proof: The proof is almost trivial, if we start at the righthand side:

$$\|u\|_p^p = \int_\Omega |u|^p \, d\mu \ge \int_E |u|^p \, d\mu \ge \varepsilon^p \mu(E),$$

where we used the fact that $|u|^p \ge \varepsilon^p$ on $E = \{x \in \Omega\colon |u(x)| \ge \varepsilon\}$. (In fact, this is one case where it seems easier to reconstruct the proof than to remember the exact inequality.) ∎

34 Proposition. *The space of functions $v \in L^\infty$ which vanish outside a set of finite measure, i.e., for which $\mu\{t \in \Omega\colon v(t) \ne 0\} < \infty$, is dense in L^p whenever $1 \le p < \infty$.*

We may further restrict the space to *simple functions*. This is an exercise in integration theory which we leave to the reader.

Proof: Let $u \in L^p$. For a given n, define v_n by

$$v_n(t) = \begin{cases} u(t), & 1/n \le |u(t)| \le n, \\ 0 & \text{otherwise.} \end{cases}$$

Then v_n belongs to the space in question, thanks to Chebyshev's inequality. Furthermore, $|u - v_n|^p \le |u|^p$ and $|u - v_n|^p \to 0$ pointwise as $n \to \infty$. It follows from the definition of the norm and the dominated convergence theorem that $\|u - v_n\|_p^p \to 0$. ∎

Whenever $Y \subseteq X^*$ is a subset of the dual space of a normed space X, we define its *pre-annihilator* as

$$Y_\perp = \{x \in X\colon f(x) = 0 \text{ for all } f \in Y\}.$$

35 Lemma. *Let X be a reflexive Banach space, and let $Y \subseteq X^*$ be a closed subspace with $Y_\perp = \{0\}$. Then $Y = X^*$.*

Proof: Assume $Y \neq X^*$. It is a simple consequence of the Hahn–Banach theorem that there is a *nonzero* bounded functional ξ on X^* which vanishes on Y; see Kreyszig Lemma 4.6-7 (page 243). (It is important for this that Y is closed.)

But X is reflexive, so there is some $x \in X$ so that $\xi(g) = g(x)$ for all $g \in X^*$. But then $g(x) = \xi(g) = 0$ for all $g \in Y$, so $x \in Y_\perp = \{0\}$. But this contradicts $\xi \neq 0$. ∎

36 Corollary. *If Z is a Banach space and Z^* is reflexive, then Z is reflexive.*

Proof: Apply the lemma with $X = Z^*$ and Y the image of Z under the canonical map $Z \to Z^{**} = X^*$: This maps $z \in Z$ to $\hat{z} \in Z^{**}$ defined by $\hat{z}(f) = f(z)$ where $f \in Z^*$. This mapping is isometric, so the image is closed. If $f \in Y_\perp$ then $f(z) = \hat{z}(f) = 0$ whenever $z \in Z$, so $f = 0$. Thus the conditions of the lemma are fulfilled. ∎

37 Theorem. *Let p and q be conjugate exponents with $1 \leq p < \infty$. Then every bounded linear functional on L^p has the form*

$$\tilde{v}(u) = \int_\Omega uv\,d\mu$$

where $v \in L^q$.

Moreover, $\|\tilde{v}\| = \|v\|_q$. Thus, L^q can be identified with the dual space of L^p.

This result is already known for $p = 2$, since L^2 is a Hilbert space, and this is just the Riesz representation theorem.

Proof: First, we note that as soon as the first part has been proved, the second part follows from Corollary 21.

We prove the first part for $1 < p < \infty$ first. Then L^p is reflexive. The space of all functionals $\tilde{v}$ on L^p, where $v \in L^q$, satifisies the conditions of Lemma 35. This completes the proof for this case.

It remains to prove the result for $p = 1$. We assume first that $\mu(\Omega) < \infty$. Let $f \in (L^1)^*$. Since then $\|u\|_1 \leq \mu(\Omega)^{1/2}\|u\|_2$ (Lemma 25), we find for $u \in L^2$ that $u \in L^1$ as well, and $|f(u)| \leq \|f\|\,\|u\|_1 \leq \mu(\Omega)^{1/2}\|f\|\,\|u\|_2$. So f defines a bounded linear functional on L^2 as well, and there is some $v \in L^2$ so that

$$f(u) = \int_\Omega uv\,d\mu \qquad (u \in L^2).$$

We shall prove that $v \in L^\infty$. Since L^2 is dense in L^1, the above equality will then extend to all $u \in L^1$ by continuity, and we are done.

Assume now that $\|v\|_\infty > M > \|f\|$. Then $|v| \geq M$ on a measurable set E with $\mu(E) > 0$. Let $u = \chi_E \overline{\operatorname{sgn}}\, v/\mu(E)$. Then $\|u\|_1 = 1$ and $u \in L^2$ as well, since it is bounded. Thus

$$\|f\| \geq f(u) = \int_\Omega uv\, d\mu = \frac{1}{\mu(E)} \int_E |v|\, d\mu \geq M,$$

which is a contradiction. This finishes the proof for $\mu(\Omega) < \infty$.

Otherwise, if $\mu(\Omega) = \infty$ but μ is σ-finite, write $\Omega = E_1 \cup E_2 \cup \cdots$ where each E_j has finite measure and all the sets E_j are pairwise disjoint. Use what we just proved to find $v_j \in L^\infty(E_j)$ with $f(u) = \int_{E_j} uv\, d\mu$ when $u \in L^1 E_j$, and $\|v_j\|_\infty \leq \|f\|$. Define v on Ω by setting $v(t) = v_j(t)$ when $t \in E_j$. Then, for $u \in L^1$,

$$u = \sum_{j=1}^{\infty} u\chi_{E_j}$$

(convergent in L^1), so

$$f(u) = \sum_{j=1}^{\infty} f(u\chi_{E_j}) = \sum_{j=1}^{\infty} \int_{E_j} uv\, d\mu = \int_\Omega uv\, d\mu.$$

This finishes the proof. ∎

Chapter 4

A tiny bit of topology

Basic definitions

The reader is supposed to be familiar with metric spaces. To motivate the definitions to come, consider a metric space (X,d). Note that many concepts from the theory of metric spaces can be formulated without reference to the metric d, so long as we know which sets are open. For example, a function $f\colon X \to Y$, where (Y,ρ) is another metric space, is continuous if and only if[1] $f^{-1}(V)$ is open in X for every open $V \subseteq Y$, and f is continuous at a point $x \in X$ if and only if $f^{-1}(V)$ is a neighbourhood of x whenever V is a neighbourhood of $f(x)$. (Here, a *neighbourhood* of a point is a set containing an open set containing the point.)

If you are unfamiliar with the above results, you are adviced to prove them to your satisfaction before proceeding.

Consider the open subsets of a metric space X. They have these properties:

T_1 $\varnothing$ and X are open subsets of X,
T_2 the intersection of two open sets is open,
T_3 an arbitrary union of open sets is open.

The basic notion of topology is to take these properties as axioms. Consider an arbitrary set X. A set $\mathscr{T}$ of subsets of X satisfying the conditions

T_1 $\varnothing \in \mathscr{T}$ and $X \in \mathscr{T}$,
T_2 $U \cap V \in \mathscr{T}$ whenever $U \in \mathscr{T}$ and $V \in \mathscr{T}$,
T_3 the union of the members of an arbitrary subset of $\mathscr{T}$ belongs to $\mathscr{T}$,

is called a *topology* on X. A *topological space* is a pair $(X,\mathscr{T})$ where $\mathscr{T}$ is a topology on X. The members of $\mathscr{T}$ are called *open sets*. It is worthwhile to restate the axioms in this language: We have seen that the open sets given by a metric form a topology. We shall call this the topology *induced* by the metric.

[1] The *inverse image* of V is $f^{-1}(V) = \{x \in X\colon f(x) \in V\}$. Despite the notation, f need not have an inverse for this definition to make sense.

Moreover, we shall call a topology (or a topological space) *metrizable* if there exists a metric which induces the given topology.

Examples. The *discrete topology* on a set X is the set of all subsets of X. I.e., every subset of X is open. This is induced by the discrete metric d, for which $d(x,y) = 1$ when $x \neq y$ (and $d(x,y) = 0$ when $x = y$ of course). It follows from T_3 that if $\{x\}$ is open for all $x \in X$, then the topology is discrete.

A *pseudometric* on a set X is a function $d\colon X \times X \to \mathbb{R}$ so that

PM_1 $d(x,y) \geq 0$ for all $x, y \in X$,
PM_2 $d(x,y) = d(y,x)$ for all $x, y \in X$,
PM_3 $d(x,z) \leq d(x,y) + d(y,z)$ for all $x, y, z \in X$.

In other words, it satisfies all the properties of a metric, except that we allow $d(x,y) = 0$ for some $x \neq y$. Now let $\mathscr{D}$ be an arbitrary set of pseudometrics on X. These induce a topology on X in a similar way that a single metric would: A subset $A \subseteq X$ is open in this topology if, whenever $a \in A$, there are a finite number of pseudometrics $d_1, \ldots, d_n \in \mathscr{D}$ and corresponding numbers $\varepsilon_1 > 0, \ldots, \varepsilon_n > 0$ so that

$$d_1(x,a) < \varepsilon_1, \ldots, d_n(x,a) < \varepsilon_n \Rightarrow x \in A.$$

If $\mathscr{D}$ is finite, this topology is induced by the single pseudometric $d_1 + \cdots + d_n$, where $\mathscr{D} = \{d_1, \ldots, d_n\}$. In fact, the same holds when $\mathscr{D}$ is *countably infinite*: If $\mathscr{D} = \{d_1, d_2, \ldots\}$, then

$$d(x,y) = \sum_{n=1}^{\infty} 2^{-n} \wedge d_n(x,y)$$

defines a pseudometric on X, where $\wedge$ selects the minimum of the numbers surrounding it. And this pseudometric induces the same topology as the entire collection $\mathscr{D}$.

The *trivial topology* on X consists of only the two sets $\varnothing$ and X itself. If X has at least two distinct points, this is not metrizable. In fact, it does not even satisfy the following *separation* property:

A topological space X is called *Hausdorff* if, whenever $x, y \in X$ with $x \neq y$, there are open sets U and V with $x \in U$, $y \in V$, and $U \cap V = \varnothing$.

Any metrizable space is Hausdorff: If d is a metric on X, we can let U and V be the open ε-balls around x and y, respectively, where $\varepsilon = \frac{1}{2} d(x,y)$.

A subset of a topological space is called *closed* if its complement is open. We might as well have stated the axioms for the closed sets:

T_1' $\varnothing$ and X are closed subsets of X,
T_2' the union of two closed sets is closed,

T_3' an arbitrary intersection of closed sets is closed.

The *interior* of a subset $A \subseteq X$ is the union of all open sets contained in A. It is the largest open subset of A. The *closure* of $A \subseteq X$ is the intersection of all closed sets containing A. Written $\overline{A}$, it is the smallest closed subset of X containing A. Finally, if we are given two topologies $\mathscr{T}_1$ and $\mathscr{T}_2$ on the same set X, and $\mathscr{T}_1 \supseteq \mathscr{T}_2$, we say that $\mathscr{T}_1$ is *stronger* than $\mathscr{T}_2$, or equivalently, that $\mathscr{T}_2$ is *weaker* than $\mathscr{T}_2$. Thus the trivial topology is the weakest of all topologies on a set, and the discrete topology is the strongest.

If X is a topological space and $Y \subseteq X$, then we can give Y a topology consisting of all sets of the form $Y \cap V$ where $V \subseteq X$ is open in X. The resulting topology, sometimes called the *relative topology*, is said to be *inherited* from X. Unless otherwise stated, this is always the topology we use on subsets of a topological space, when we wish to apply topological concepts to the subset itself. Beware though, of an ambiguity: When considering a subset $A \subset Y$, it may be open in Y, yet not open in X. Consider for example $X = \mathbb{R}$, $Y = [0,1]$, $A = (\frac{1}{2}, 1] = (\frac{1}{2}, \infty) \cap Y$. Similarly with closed subsets. So openness and closedness, and derived concepts like interior points, isolated points, etc., become *relative* terms in this situation.

We round off this section with some important notions from functional analysis.

Let X be a (real or complex) vector space. A *seminorm* on X is a function $p\colon X \to [0,\infty)$ which is homogeneous and subadditive. I.e., $p(\alpha x) = |\alpha| p(x)$ and $p(x+y) \leq p(x) + p(y)$ for $x, y \in X$ and every scalar α. Thus p is just like a norm, except it could happen that $p(x) = 0$ for some vectors $x \neq 0$.

From a seminorm p we can make a pseudometric $d(x,y) = p(x-y)$. Thus from a collection $\mathscr{P}$ of seminorms on X we get a collection of pseudometrics, which induces a topology on X. This topology is said to be *induced* by the given seminorms. The topology is Hausdorff if and only if for each $x \in X$ with $x \neq 0$ there is some $p \in \mathscr{P}$ with $p(x) \neq 0$. (In this case, we say $\mathscr{P}$ *separates points* in X.)

Of particular interest is the case where $\mathscr{P}$ is countable; then the induced topology is metrizable. Moreover the associated metric, of the form

$$d(x,y) = \sum_{k=1}^{\infty} 2^{-k} \wedge p_k(x-y),$$

is *translation invariant* in the sense that $d(x+z, y+z) = d(x,y)$.

An important class of topological vector spaces is the class *Fréchet spaces*, which are locally convex spaces[2] whose topology can be induced by a translation invariant metric which, moreover, makes the spaces complete.

[2] For a definition, see the chapter on topological vector spaces.

For example, let $X = C(\mathbb{R})$ be space of continuous functions defined on the real line, and consider the seminorms p_M given by

$$p_M(u) = \sup\{|u(t)| : |t| \le M\} \qquad (u \in C(\mathbb{R})).$$

This induces a topology on $C(\mathbb{R})$ which is called the topology of uniform convergence on compact sets, for reasons that may become clear later. This space is a Fréchet space.

Let X be a normed vector space, and X^* its dual. For every $f \in X^*$ we can create a seminorm $x \mapsto |f(x)|$ on X. The topology induced by all these seminorms is called the *weak topology* on X. Because each of the seminorms is continuous with respect to the norm, the weak topology is weaker than the norm topology on X. (The norm topology is the one induced by the metric given by the norm.) In fact, the weak topology is the weakest topology for which each $f \in X^*$ is continuous.

Similarly, each $x \in X$ defines a seminorm $f \mapsto |f(x)|$ on X^*. These together induce the *weak** topology on X^*. X^* also has a weak topology, which comes from pseudometrics defined by members of X^{**}. Since these include the former, the weak topology is stronger than the weak* topology, and both are weaker than the norm topology.

Neighbourhoods, filters, and convergence

Let X be a topological space. A *neighbourhood* of a point $x \in X$ is a subset of X containing an open subset which in it turn contains x. More precisely, N is a neighbourhood of x if there is an open set V with $x \in V \subseteq N \subseteq X$.

If $\mathscr{F}$ is the set of all neighbourhoods of x, then

F_1 $\varnothing \notin \mathscr{F}$ and $X \in \mathscr{F}$,
F_2 if $A, B \in \mathscr{F}$ then $A \cap B \in \mathscr{F}$,
F_3 if $A \in \mathscr{F}$ and $A \subseteq B \subseteq X$ then $B \in \mathscr{F}$.

Whenever $\mathscr{F}$ is a set of subsets of X satisfying F_1–F_3, we call $\mathscr{F}$ a *filter*. In particular, the set of all neighbourhoods of a point $x \in X$ is a filter, which we call the *neighbourhood filter* at x, and write $\mathscr{N}(x)$.

If $\{x\}$ is an open set, we call the point x *isolated*. If x is not isolated, then, any neighbourhood of x contains at least one point different from x. Let us define a *punctured neighbourhood* of x as a set U so that $U \cup \{x\}$ is a neighbourhood of x (whether or not $x \in U$). The set of all punctured neighbourhoods of a non-isolated point x forms a filter $\mathscr{N}'(x)$, called the *punctured neighbourhood filter* of x.

We may read F_3 as saying that if we know the *small* sets in a given filter, we know the whole filter. In other words, the only interesting aspect of a filter is its small members. More precisely, we shall call a subset $\mathscr{B} \subseteq \mathscr{F}$ a *base* for the filter $\mathscr{F}$ if every member of $\mathscr{F}$ contains some member of $\mathscr{B}$. In this case, whenever $F \subseteq X$,

$$F \in \mathscr{F} \Leftrightarrow B \subseteq F \text{ for some } B \in \mathscr{B}. \tag{4.1}$$

You may also verify these properties:

FB_1 $\varnothing \notin \mathscr{B}$ but $\mathscr{B} \neq \varnothing$,
FB_2 if $A, B \in \mathscr{B}$ then $C \subseteq A \cap B$ for some $C \in \mathscr{B}$.

Any set of subsets of X satisfying these two requirements is called a *filter base*. When $\mathscr{B}$ is a filter base, (4.1) defines a filter $\mathscr{F}$, which we shall call the filter *generated* by $\mathscr{B}$. Then $\mathscr{B}$ is a base for the filter generated by $\mathscr{B}$.

The above would seem like a lot of unnecessary abstraction, if neighbourhood filters were all we are interested in. However, filters generalize another concept, namely that of a sequence. Recall that the convergence, or not, of a sequence in a metric space only depends on what happens at the end of the sequence: You can throw away any initial segment of the sequence without changing its convergence properties. (But still, if you throw all initial segments away, nothing is left, so there is nothing that might converge.) Let $(x_k)_{k=1}^{\infty}$ be a sequence in X. We call any set of the form $\{x_k \colon k \geq n\}$ a *tail* of the sequence.[3] The set of tails of the sequence is a filter base. If X is a metric space then $x_n \to x$ if and only if every neighbourhood of x contains some tail of the sequence. But that means that every neighbourhood belongs to the filter generated by the tails of the sequence. This motivates the following definition.

A filter $\mathscr{F}_1$ is said to be *finer* than another filter $\mathscr{F}_2$ if $\mathscr{F}_1 \supseteq \mathscr{F}_2$. We also say that $\mathscr{F}_1$ is a *refinement* of $\mathscr{F}_2$. This corresponds to the notion of *subsequence*: The tails of a subsequence generate a finer filter than those of the original sequence.

A filter $\mathscr{F}$ on a topological space X is said to *converge* to $x \in X$, written $\mathscr{F} \to x$, if $\mathscr{F}$ is finer than the neighbourhood filter $\mathscr{N}(x)$. In other words, every neighbourhood of x belongs to $\mathscr{F}$.[4] We shall call x a *limit* of $\mathscr{F}$ (there may be several). A filter which converges to some point is called *convergent*. Otherwise, it is called *divergent*.

If X has the trivial topology, every filter on X converges to every point in X. This is clearly not very interesting. But if X is Hausdorff, no filter can converge

[3] Beware that in other contexts, one may prefer to use the word tail for the indexed family $(x_k)_{k=n}^{\infty}$ rather than just the set of values beyond the nth item in the list.

[4] This would look more familiar if it were phrased "every neighbourhood of x contains a member of $\mathscr{F}$". But thanks to F_3, we can express this more simply as in the text.

to more than one point. In this case, if the filter $\mathscr{F}$ converges to x, we shall call x *the* limit of $\mathscr{F}$, and write $x = \lim \mathscr{F}$.

We shall say that a sequence (x_k) converges to x if, for every neighbourhood U of x, there is some n so that $x_k \in U$ whenever $k \geq n$. This familiar sounding definition is equivalent to the statement that the filter generated by the tails of the sequence converges to x.

We shall present an example showing that sequence convergence is insufficient.[5] But first a definition and a lemma: A filter $\mathscr{F}$ on X is said to be *in* a subset $A \subseteq X$ if $A \in \mathscr{F}$. (Recall that filters only express what is in their smaller elements; thus, this says that all the interesting "action" of $\mathscr{F}$ happens within A.)

38 Lemma. *Let A be a subset of a topological space X. Then, for $x \in X$, we have $x \in \overline{A}$ if and only if there is a filter in A converging to x.*

Proof: Assume $x \in \overline{A}$. Then any neighbourhood of x meets A,[6] for otherwise x has an open neighbourhood U disjoint from A, and then $X \setminus U$ is a closed set containing A, so $\overline{A} \subseteq X \setminus U$. But $x \in U$, so this contradicts $x \in \overline{A}$. Now all the sets $A \cap U$ where U is a neighbourhood of x form a filter base, and the corresponding filter converges to x.

Conversely, assume $\mathscr{F}$ is a filter in A converging to x. Pick any neighbourhood U of x. Then $U \in \mathscr{F}$. Thus $U \cap A \neq \varnothing$, since $A \in \mathscr{F}$ as well. This proves that $x \in \overline{A}$. ∎

We are now ready for the example, which will show that we cannot replace filters by sequences in the above result. Let Ω be any uncountable set, and let X be the set of all families $(x_\omega)_{\omega \in \Omega}$, where each component x_ω is a real number. In fact, X is a real vector space. For each ω we consider the seminorm $x \mapsto |x_\omega|$. We now use the topology on X induced by all these seminorms.[7]

Let

$$Z = \{x \in X : |x_\omega| \neq 0 \text{ for at most a finite number of } \omega\},$$
$$e \in X, \quad e_\omega = 1 \text{ for all } \omega \in \Omega.$$

Then $e \in \overline{Z}$, but there is no sequence in Z converging to e.

In fact, for every neighbourhood U of e there are $\omega_1, \ldots, \omega_n \in \Omega$ and $\varepsilon > 0$ so that

$$|x_{\omega_k} - 1| < \varepsilon \text{ for } k = 1, \ldots, n \Rightarrow x \in U.$$

[5] Insufficient for *what*, you ask? In the theory of metric spaces, sequences and their limits are everywhere. But they cannot serve the corresponding role in general topology.

[6] We say that two sets *meet* if they have a nonempty intersection.

[7] This is called the *topology of pointwise convergence*.

But then we find $x \in Z \cap U$ if we put $x_{\omega_k} = 1$ for $k = 1, \dots, n$ and $x_\omega = 0$ for all other ω. Thus $e \in \overline{Z}$.

Furthermore, if $(x_n)_{n=1}^\infty$ is a sequence in Z then for each n (writing $x_{n\omega}$ for the ω-component of x_n) we have $x_{n\omega} \neq 0$ for only a finite number of ω. Thus $x_{n\omega} \neq 0$ for only a countable number of pairs (n, ω). Since Ω is not countable, there is at least one $\omega \in \Omega$ with $x_{n\omega} = 0$ for all n. Fix such an ω. Now $U = \{z \in X : |z_\omega| > 0\}$ is a neighbourhood of e, and $x_n \notin U$ for all n. Thus the sequence (x_n) does not converge to e.

A historical note. The theory of filters and the associated notion of convergence was introduced by Cartan and publicized by Bourbaki. An alternative notion is to generalize the sequence concept by replacing the index set $\{1, 2, 3, \dots\}$ by a partially ordered set A with the property that whenever $a, b \in A$ there is some $c \in A$ with $c \succ a$ and $c \succ b$. A family indexed by such a partially ordered set is called a *net*. The theory of convergence based on nets is called the *Moore–Smith* theory of convergence after its inventors. Its advantage is that nets seem very similar to sequences. However, the filter theory is much closer to being precisely what is needed for the job, particularly when we get to ultrafilters in a short while.

Continuity and filters

Assume now that X is a set and Y is a topological space, and let $f : X \to Y$ be a function. We say that $f(x) \to y$ as $x \to \mathscr{F}$ if, for each neighbourhood V of y, there is some $F \in \mathscr{F}$ so that $x \in F \Rightarrow f(x) \in V$.

The implication $x \in F \Rightarrow f(x) \in V$ is the same as $F \subseteq f^{-1}(V)$, so we find that $f(x) \to y$ as $x \to \mathscr{F}$ if and only if $f^{-1}(V) \in \mathscr{F}$ for every $V \in \mathscr{N}(y)$.

We can formalize this by defining a filter $f_*\mathscr{F}$ on Y by the prescription[8]

$$B \in f_*\mathscr{F} \Leftrightarrow f^{-1}(B) \in \mathscr{F},$$

so that $f(x) \to y$ as $x \to \mathscr{F}$ if and only if $f_*\mathscr{F} \to y$. We also use the notation

$$y = \lim_{x \to \mathscr{F}} f(x),$$

but only when the limit is unique, for example when Y is Hausdorff.

Think of the notation $x \to \mathscr{F}$ as saying that x moves into progressively smaller members of $\mathscr{F}$. The above definition says that, for every neighbourhood V of y, we can guarantee that $f(x) \in V$ merely by choosing a member $F \in \mathscr{F}$ and insisting that $x \in F$.

[8] It is customary to write f_*S for some structure S which is transported in the same direction as the mapping f, as opposed to f^*S for those that are mapped in the opposite direction. An example of the latter kind is the adjoint of a bounded operator.

Examples. Let $\mathscr{F}$ be the filter on the set $\mathbb{N}$ of natural numbers generated by the filter base consisting of the sets $\{n, n+1, \ldots\}$ where $n \in \mathbb{N}$. Let $(z_1, z_2, \ldots)$ be a sequence in Y. Then $z_n \to y$ as $n \to \infty$ if and only if $z_n \to y$ as $n \to \mathscr{F}$.

Next, consider a function φ on an interval $[a, b]$. Let X consist of all sequences

$$a = x_0 \le x_1^* \le x_1 \le x_2^* \le x_2 \le \cdots \le x_{n-1} \le x_n^* \le x_n = b$$

and consider the filter $\mathscr{F}$ on X generated by the filter base consisting of all the sets

$$\{x \in X : |x_k - x_{k-1}| < \varepsilon \text{ for } k = 1, \ldots, n\}$$

where $\varepsilon > 0$. Let $f : X \to \mathbb{R}$ be defined by the Riemann sum

$$f(x) = \sum_{k=1}^{n} \varphi(x_k^*)\,(x_k - x_{k-1}).$$

Then the existence and definition of the Riemann integral of φ over $[a, b]$ is stated by the equation

$$\int_a^b \varphi(t)\,dt = \lim_{x \to \mathscr{F}} f(x).$$

Let us return now to the notion of convergence. If X is a topological space and $w \in X$, we can replace $\mathscr{F}$ by the punctured neighbourhood filter $\mathscr{N}'(x)$. Thus we say $f(x) \to y$ as $x \to w$ if, for every neighbourhood V of y there is a punctured neighbourhood U of w so that $f(x) \in V$ whenever $x \in U$. This is the same as saying $f(x) \to y$ when $x \to \mathscr{N}'(x)$.

The function f is said to be *continuous* at $w \in X$ if either w is isolated, or else $f(x) \to f(w)$ as $x \to w$.

In this particular case, though, it is not necessary to avoid the point w in the definition of limit, so we might just as well define continuity as meaning: $f^{-1}(V)$ is a neighbourhood of x whenever V is a neighbourhood of $f(x)$.

We can also define continuity over all: $f : X \to Y$ is called *continuous* if $f^{-1}(V)$ is open in X for every open set $V \subseteq Y$. This is in fact equivalent to f being continuous at every point in X. (Exercise: Prove this!)

Compactness and ultrafilters

A topological space X is called *compact* if every open cover of X has a finite subcover.

What do these terms mean? A *cover* of X is a collection of subsets of X which covers X, in the sense that the union of its members is all of X. The cover if called *open* if it consists of open sets. A *finite subcover* is then a finite subset which still covers X.

By taking complements, we get this equivalent definition of compactness:

X is compact if and only if every set of closed subsets of X, with the finite intersection property, has nonempty intersection.

Here, the *finite intersection property* of a set of subsets of X means that every finite subset has nonempty intersection. So what the definition says, in more detail, is this: Assume $\mathscr{C}$ is a set of closed subsets of X. Assume that $C_1 \cap \ldots \cap C_n \neq \varnothing$ whenever $C_1, \ldots, C_n \in \mathscr{C}$. Then $\bigcap \mathscr{C} \neq \varnothing$ (where $\bigcap \mathscr{C}$ is the intersection of all the members of $\mathscr{C}$, perhaps more commonly (but reduntantly) written $\bigcap_{C \in \mathscr{C}} C$).

The translation between the two definitions goes as follows: Let $\mathscr{C}$ be a set of closed sets and $\mathscr{U} = \{X \setminus C \colon C \in \mathscr{C}\}$. Then $\mathscr{U}$ covers X if and only if $\mathscr{C}$ does *not* have nonempty intersection, and $\mathscr{U}$ has a finite subcover if and only if $\mathscr{C}$ does *not* have the finite intersection property.

One reason why compactness is so useful, is the following result.

39 Theorem. *A topological space X is compact if and only if every filter on X has a convergent refinement.*

In order to prove this, we shall detour through the notion of ultrafilters. Many arguments using compactness in metric spaces require some form of diagonal argument, selecting subsequences of subsequences and finally using the diagonal to get a sequence which is eventually a subsequence of any of the former subsequences. The theory of filters neatly sidesteps these arguments by passing to an ultrafilter, which is sort of the ultimate subsequence.

A filter $\mathscr{U}$ is called an *ultrafilter* if, for every $A \subseteq X$, either $A \in \mathscr{U}$ or $X \setminus A \in \mathscr{U}$.

The filter $\mathscr{U}$ is an ultrafilter if, and only if, there does not exist any strictly finer filter.

Indeed, if $\mathscr{F}$ is a filter but not an ultrafilter, and $A \subseteq X$ with $A \notin \mathscr{F}$ and $X \setminus A \notin \mathscr{F}$, then $A \cap B \neq \varnothing$ whenever $B \in \mathscr{F}$. For, if $A \cap B = \varnothing$ then $B \subseteq X \setminus A$, and so $X \setminus A \in \mathscr{F}$. It follows that all the sets $A \cap B$ where $B \in \mathscr{F}$ form a filter base for a filter that is strictly finer than $\mathscr{F}$. We leave the (easy) converse proof to the reader.

40 Lemma. (Ultrafilter lemma) *For every filter there is at least one finer ultrafilter.*

Proof: The proof is by appeal to Zorn's lemma: Given a set of filters totally ordered by inclusion, the union of all these filters is again a filter. This filter is, of course, finer than any of the original filters. Thus the set of filters finer than one given filter is inductively ordered. Then Zorn's lemma takes care of the rest. ∎

41 Theorem. *A topological space X is compact if and only if every ultrafilter on X is convergent.*

Proof: First, assume that every ultrafilter on X is convergent, let $\mathscr{C}$ be a set of closed subsets of X, and assume that $\mathscr{C}$ has the finite intersection property. Then the set $\mathscr{B}$ of all finite intersections of sets in $\mathscr{C}$, i.e.

$$\mathscr{B} = \{C_1 \cap \ldots \cap C_n \colon C_1, \ldots, C_n \in \mathscr{C}\},$$

is a filter base. It generates a filter $\mathscr{F}$, which has an ultrafilter refinement $\mathscr{U}$ according to the ultrafilter lemma. Let x be a limit of $\mathscr{U}$.

Now for any $C \in \mathscr{C}$, if $x \notin C$ then $X \setminus C$ is a neighbourhood of x. Thus $X \setminus C \in \mathscr{U}$, since $\mathscr{U}$ converges to x. But $C \in \mathscr{U}$ as well, and this is a contradiction, since $(X \setminus C) \cap C = \varnothing$. Thus $x \in C$ for every $C \in \mathscr{C}$. We have proved that $\mathscr{C}$ has nonempty intersection. Thus X is compact.

Conversely, assume that X is compact, and let $\mathscr{U}$ be an ultrafilter on X. Let $\mathscr{C} = \{\overline{U} \colon U \in \mathscr{U}\}$. By definition, $\mathscr{C}$ consists of closed sets, and it obviously inherits the finite intersection property from $\mathscr{U}$. Since X is compact, there is a point x in the intersection of $\mathscr{C}$.

To show that $\mathscr{U}$ converges to x, let V be an open neighbourhood of x. If $V \notin \mathscr{U}$ then $X \setminus V \in \mathscr{U}$, since $\mathscr{U}$ is an ultrafilter. But $X \setminus V$ is closed, so then $X \setminus V \in \mathscr{C}$, by construction. But since $x \notin X \setminus V$, this contradicts the choice of x. Thus every open neighbourhood of x belongs to $\mathscr{U}$, and so $\mathscr{U}$ converges to x. ∎

Proof of Theorem 39: First, assume that X is compact, and let $\mathscr{F}$ be a filter on X. Let $\mathscr{U}$ be an ultrafilter refining $\mathscr{F}$. By Theorem 41, $\mathscr{U}$ is convergent. Thus $\mathscr{F}$ has a convergent refinement.

Conversely, assume that every filter on X has a convergent refinement. Then any ultrafilter $\mathscr{U}$ on X must be convergent, for it has no proper refinement, so the convergent refinement of $\mathscr{U}$ which exists by assumption must be $\mathscr{U}$ itself. Thus every ultrafilter on X is convergent, so X is compact by Theorem 41. ∎

42 Proposition. *A continuous image of a compact space is compact.*

Proof: Let X and Y be topological spaces, with X compact. Assume $f \colon X \to Y$ is continuous and maps X onto Y. We shall show that then Y is compact.

Let $\mathscr{C}$ be an open cover of Y. Then $\{f^{-1}(C) \colon C \in \mathscr{C}\}$ is an open cover of X, since f is continuous. Because X is compact, there is a finite subcover $\{f^{-1}(C_1), \ldots, f^{-1}(C_n)\}$ of X. And then $\{C_1, \ldots, C_n\}$ covers Y since f is onto.

A similar proof using families of closed sets with the finite intersection property is just as easy to put together. ∎

Product spaces and Tychonov's theorem

Let X_j be a topological space for every $j \in J$, where J is some index set. The *direct product* $\prod_{j\in J} X_j$ consists of all families $(x_j)_{j\in J}$ with $x_j \in X_j$ for each $j \in J$. From now on, we will often simply write X for the product.

It is often useful to consider the *projection map* $\pi_j \colon X \to X_j$ given by $\pi_j(x) = x_j$, where $j \in J$ and $x \in X$.

We can define a topology on X as the weakest topology that makes every projection map π_j continuous: First, for each $j \in J$, let

$$\mathscr{T}_j = \{\pi_j^{-1}(U) \colon U \subseteq X_j \text{ is open}\}.$$

Then $\mathscr{T}_j$ is a topology on X. Next, let

$$\mathscr{S} = \{U_{j_1} \cap \ldots \cap U_{j_n} \colon j_k \in J, U_{j_k} \in \mathscr{T}_{j_k} \text{ for } k = 1, \ldots, n\}$$

(the set of all finite intersections from different $\mathscr{T}_j$). Finally, let $\mathscr{T}$ consist of all unions of sets from $\mathscr{S}$. It is clear that any topology on X which makes each π_j continuous must contain $\mathscr{T}$. But $\mathscr{T}$ is in fact a topology, so it is, as claimed, the weakest of all topologies on X making each π_j continuous. We call this the *product topology* on X.

43 Lemma. *Let $\mathscr{F}$ be a filter on $\prod_{j\in J} X_j$. Then $\mathscr{F}$ converges to $x \in \prod_{j\in J} X_j$ if and only if*

$$z_j \to x_j \text{ as } z \to \mathscr{F} \qquad \text{for all } j \in J.$$

Proof: First, assume that $\mathscr{F}$ converges to x. That $z_j \to x_j$ when $z \to \mathscr{F}$ is merely a consequence of the continuity of π_j.

Conversely, assume that $z_j \to x_j$ when $z \to \mathscr{F}$ for each j. Consider a neighbourhood U of x. By the definition of the product topology, there are indexes $j_1, \ldots, j_n$ and neighbourhoods U_{j_k} of x_{j_k} in X_{j_k} so that

$$z_{j_k} \in U_{j_k} \text{ for } k = 1, \ldots n \Rightarrow z \in U.$$

But by assumption, $\{z \colon z_{j_k} \in U_{j_k}\} \in \mathscr{F}$ for each k, so

$$\bigcap_{k=1}^{n} \{z \colon z_{j_k} \in U_{j_k}\} \in \mathscr{F}.$$

But the set on the lefthand side is a subset of U, so $U \in \mathscr{F}$. This proves that $\mathscr{F}$ converges to x. ∎

44 Theorem. (Tychonov) *Any product of compact spaces is compact.*

Proof: Keep the notation used earlier in this section. Let $\mathscr{U}$ be an ultrafilter on X. Then $(\pi_j)_*(\mathscr{U})$ is an ultrafilter on X_j for each j. (For if $A \subseteq X_j$, then either $\pi_j^{-1}(A)$ or its complement $\pi_j^{-1}(X_j \setminus A)$ belongs to $\mathscr{U}$, so either A or $X_j \setminus A$ belongs to $(\pi_j)_*(\mathscr{U})$.) But then, since X_j is compact, $(\pi_j)_*(\mathscr{U})$ has a limit[9] $x_j \in X_j$. In other words, $z_j \to x_j$ as $z \to \mathscr{F}$.

Put all these limits x_j together into the single element $x = (x_j)_{j\in J}$ of X. By the previous lemma, $\mathscr{U}$ converges to x, so our proof is complete. ∎

Normal spaces and the existence of real continuous functions

A topological space X is called *normal* if any two disjoint closed sets A, B can be separated by open sets, in the sense that there are open sets $U \supset A$ and $V \supset B$ with $U \cap V = \varnothing$.

We are interested in normal spaces because one can find a large number of continuous real functions on these spaces.

A *neighbourhood* of a set A is a set containing an open set which contains A. In other words, a set whose interior contains A.

An equivalent way to state that X is normal, is to say that if $A \subset X$ is closed, then any neighbourhood of A, contains a *closed* neighbourhood of A. (If W is an open neighbourhood of A, let $B = X \setminus W$, find disjoint open neighbourhoods $U \supset A$ and $V \supset B$, then $\overline{U}$ is a closed neighbourhood of A contained in W.)

45 Theorem. (Urysohn's lemma) *Let X be a normal topological space. Let $A_1 \subseteq X$ be a closed set, and A_0 a neighbourhood of A_1. Then there exists a continuous function $u\colon X \to [0,1]$ with $u(x) = 1$ for $x \in A_1$, and $u(x) = 0$ for $x \notin A_0$.*

Proof: Imagine that we have created sets A_t for all t in a dense subset T of $[0,1]$, so that A_0 and A_1 are the sets initially given, and A_s is a closed neighbourhood of A_t whenever $s < t$. Then define

$$u(x) = \begin{cases} 1, & x \in A_1, \\ 0, & x \notin A_0, \\ \sup\{t \in T\colon x \in A_t\} & \text{otherwise.} \end{cases}$$

To prove that u is continuous, pick a point $w \in X$ and let $v = u(w)$. If $0 < v < 1$, and $\varepsilon > 0$, pick $s, t \in T$ with $v - \varepsilon < s < v < t < v + \varepsilon$. Now $|f(x) - v| < \varepsilon$ whenever x

[9] If X_j is not Hausdorff, it may have several limits, so we need the axiom of choice to pick one of them. But this remark is not really important, as the use of ultrafilters relies heavily on the axiom of choice anyway.

is in the neighbourhood $A_s \setminus A_t$ of w. This argument works with trivial changes for $t = 0$ and $t = 1$. Thus continuity follows.

To create the sets A_t, let T be the set of dyadic rational numbers in $[0,1]$, i.e., the numbers that can be written $2^{-k}j$ where j and k are integers. We create sets $A_{2^{-k}j}$ by induction on k.

We make $A_{1/2}$ first: It shall be a closed neighbourhood of A_1 which is contained in the interior of A_0 (so A_0 is itself a neighbourhood of $A_{1/2}$).

Next, let $A_{1/4}$ be a closed neighbourhood of $A_{1/2}$ having A_0 as a neighbourhood. Then let $A_{3/4}$ be a closed neighbourhood of A_1 having $A_{1/2}$ as a neighbourhood.

Next, put $A_{1/8}$, $A_{3/8}$, $A_{5/8}$, $A_{7/8}$ between A_0, $A_{1/4}$, $A_{1/2}$, $A_{3/4}$, A_1, in the same way.

After each $A_{2^{-k}j}$ has been made for a given k, we create the sets $A_{2^{-k-1}j}$ (for odd j) in the same manner. ∎

Compact Hausdorff spaces are of special interest, because they occur so often.

46 Proposition. *Any compact Hausdorff space is normal.*

Proof: Let A and B be two closed subsets of a compact Hausdorff space X.

First, we show that we can separate A from any point outside A, in the sense that whenever $y \notin A$, we can find neighbourhoods U of A and V of y with $U \cap V = \varnothing$. Consider the set of all open U with $y \notin \overline{U}$. Because X is Hausdorff and $y \notin A$, these sets cover A. But A, being a closed subset of a compact space, is compact, so a finite number of these open sets can be found cover A. Say $A \subset U_1 \cup \ldots \cup U_n$ with U_k open and $y \notin \overline{U_k}$. Then $U_1 \cup \ldots \cup U_n$ is a neighbourhood of A whose closure $\overline{U_1} \cup \ldots \cup \overline{U_n}$ does not contain y, and from this our first claim follows.

Second, we repeat the above argument with a twist: Consider all of open sets U with $\overline{U} \cap A = \varnothing$. From what we proved in the previous paragraph, these sets cover B. So a finite number of them cover B, and we finish the proof in a way similar to the end of the previous paragraph. (Exercise: Fill in the details.) ∎

The Weierstrass density theorem

Bernshteĭn[10] polynomials are ordinary polynomials written on the particular form

$$b(t) = \sum_{k=0}^{n} \beta_k \binom{n}{k} t^k (1-t)^{n-k}, \tag{4.2}$$

[10] Named after Sergeĭ Natanovich Bernshteĭn (1880–1968). The name is often spelled "Bernstein".

where $\beta_0, \dots, \beta_n$ are given coefficients.[11] The special case where each $\beta_k = 1$ deserves mention: Then the binomial theorem yields

$$\sum_{k=0}^{n} \binom{n}{k} t^k (1-t)^{n-k} = \big(t + (1-t)\big)^n = 1. \tag{4.3}$$

We can show by induction on n that if $b = 0$ then all the coefficients β_n are zero. The base case $n = 0$ is obvious. When $n > 0$, a little bit of binomial coefficient gymnastics shows that the derivative of a Bernshteĭn polynomial can be written as another Bernshteĭn polynomial:

$$b'(t) = n \sum_{k=0}^{n-1} (\beta_{k+1} - \beta_k) \binom{n-1}{k} t^k (1-t)^{n-1-k}.$$

In particular, if $b = 0$ it follows by the induction hypothesis that all β_k are equal, and then they are all zero, by (4.3).

In other words, the polynomials $t^k(1-t)^{n-k}$, where $k = 0, \dots, n$, are linearly independent, and hence they span the $n+1$-dimensional space of polynomials of degree ≤ 1. Thus *all polynomials can be written as Bernshteĭn polynomials*, so there is nothing special about these – only about the way we write them.

To understand why Bernshteĭn polynomials are so useful, consider the individual polynomials

$$b_{k,n}(t) = \binom{n}{k} t^k (1-t)^{n-k}, \qquad k = 0, \dots, n. \tag{4.4}$$

If we fix n and t, we see that $b_{k,n}(t)$ is the probability of k heads in n tosses of a biased coin, where the probability of a head is t. The expected number of heads in such an experiment is nt, and indeed when n is large, the outcome is very likely to be near that value. In other words, most of the contributions to the sum in (4.2) come from k near nt. Rather than using statistical reasoning, however, we shall proceed by direct calculation – but the probability argument is still a useful guide.

47 Theorem. (Weierstrass) *The polynomials are dense in $C[0,1]$.*

This will follow immediately from the following lemma.

48 Lemma. *Let $f \in C[0,1]$. Let b_n be the Bernshteĭn polynomial*

$$b_n(t) = \sum_{k=0}^{n} f\Big(\frac{k}{n}\Big) \binom{n}{k} t^k (1-t)^{n-k}.$$

[11] When $n = 3$, we get a *cubic spline*. In this case, β_0, β_1, β_2 and β_3 are called the *control points* of the spline. In applications, they are usually 2- or 3-dimensional vectors.

Then $\|f - b_n\|_\infty \to 0$ *as* $n \to \infty$.

Proof: Let $t \in [0,1]$. With the help of (4.3) we can write

$$f(t) - b_n(t) = \sum_{k=0}^{n} \left(f(t) - f\Big(\frac{k}{n}\Big) \right) \binom{n}{k} t^k (1-t)^{n-k},$$

so that

$$|f(t) - b_n(t)| \le \sum_{k=0}^{n} \left| f(t) - f\Big(\frac{k}{n}\Big) \right| \binom{n}{k} t^k (1-t)^{n-k}, \tag{4.5}$$

We now use the fact that f is *uniformly continuous*: Let $\varepsilon > 0$ be given. There is then a $\delta > 0$ so that $|f(t) - f(s)| < \varepsilon$ whenever $|t - s| < \delta$. We now split the above sum into two parts, first noting that

$$\sum_{|k - nt| < n\delta} \left| f(t) - f\Big(\frac{k}{n}\Big) \right| \binom{n}{k} t^k (1-t)^{n-k} \le \varepsilon \tag{4.6}$$

(where we used $|f(t) - f(k/n)| < \varepsilon$, and then expanded the sum to all indexes from 0 to n and used (4.3)). To estimate the remainder, let $M = \|f\|_\infty$, so that

$$\sum_{|k - nt| \ge n\delta} \left| f(t) - f\Big(\frac{k}{n}\Big) \right| \binom{n}{k} t^k (1-t)^{n-k} \le 2M \sum_{|k - nt| \ge n\delta} \binom{n}{k} t^k (1-t)^{n-k}. \tag{4.7}$$

To finish the proof, we need to borrow from the Chebyshev inequality in order to show that the latter sum can be made small. First we find

$$\sum_{k=0}^{n} k \binom{n}{k} t^k (1-t)^{n-k} = nt \sum_{k=0}^{n-1} \binom{n-1}{k} t^k (1-t)^{n-1-k} = nt. \tag{4.8}$$

(Rewrite the binomial coefficient using factorials, perform the obvious cancellation using $k/k! = 1/(k-1)!$, put nt outside the sum, change the summation index, and use (4.3).) Next, using similar methods,

$$\sum_{k=0}^{n} k(k-1) \binom{n}{k} t^k (1-t)^{n-k} = n(n-1)\, t^2 \sum_{k=0}^{n-2} \binom{n-2}{k} t^k (1-t)^{n-2-k} = n(n-1)\, t^2.$$

Adding these two together, we get

$$\sum_{k=0}^{n} k^2 \binom{n}{k} t^k (1-t)^{n-k} = nt\big((n-1)\, t + 1\big). \tag{4.9}$$

Finally, using (4.3), (4.8) and (4.9), we find

$$\sum_{k=0}^{n}(nt-k)^2\binom{n}{k}t^k(1-t)^{n-k} = (nt)^2 - 2(nt)^2 + nt\big((n-1)t+1\big) = nt(1-t).$$

The most important feature here is that the n^2 terms cancel out. We now have

$$\begin{aligned} nt(1-t) &\geq \sum_{|k-nt|\geq n\delta}(nt-k)^2\binom{n}{k}t^k(1-t)^{n-k} \\ &\geq (n\delta)^2\sum_{|k-nt|\geq n\delta}\binom{n}{k}t^k(1-t)^{n-k}, \end{aligned}$$

so that

$$\sum_{|k-nt|\geq n\delta}\binom{n}{k}t^k(1-t)^{n-k} \leq \frac{t(1-t)}{n\delta^2} \leq \frac{1}{4n\delta^2}. \tag{4.10}$$

Combining (4.5), (4.6), (4.7) and (4.10), we end up with

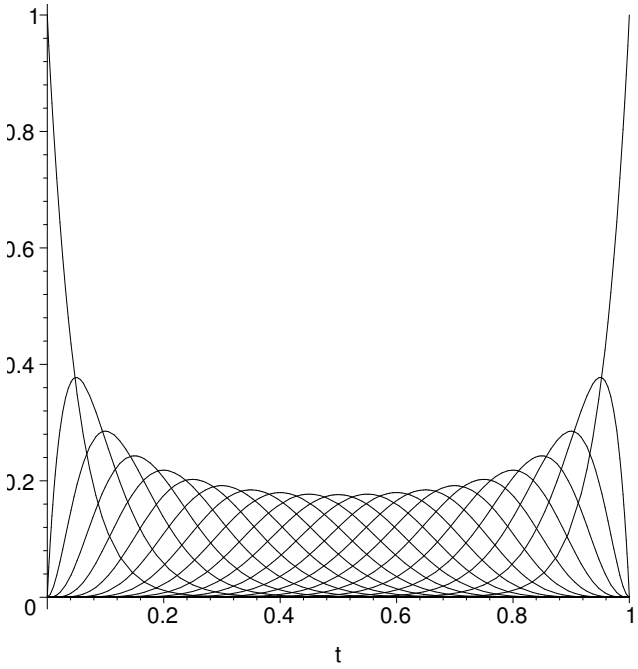
$$|f(t)-b_n(t)| < \varepsilon + \frac{M}{2n\delta^2}, \tag{4.11}$$

which can be made less than 2ε by choosing n large enough. More importantly, this estimate is independent of $t \in [0,1]$. ∎

One final remark: There is of course nothing magical about the interval $[0,1]$. Any closed and bounded interval will do. If $f \in C[a,b]$ then $t \mapsto f\big((1-t)a+tb\big)$ belongs to $C[0,1]$, and this operation maps polynomials to polynomials and preserves the norm. So the Weierstrass theorem works equally well on $C[a,b]$. In fact, it works on $C(K)$ where K is any compact subset of the real line: For any $f \in C(K)$ can be expanded to a continuous function on the smallest interval containing K by linearly interpolating the function in each bounded open interval in the complement of K.

The Stone–Weierstrass theorem is a bit more difficult: It replaces $[0,1]$ by any compact Hausdorff space X and the polynomials by any algebra of functions which separates points in X and has no common zero in X. (This theorem assumes real functions. If you work with complex functions, the algebra must also be closed under conjugation. But the complex version of the theorem is not much more than an obvious translation of the the real version into the complex domain.) One proof of the general Stone–Weierstrass theorem builds on the Weierstrass theorem. More precisely, the proof needs an approximation of the absolute value $|t|$ by polynomials in t, uniformly for t in a bounded interval.

An amusing (?) diversion. Any old textbook on elementary statistics shows pictures of the binomial distribution, i.e., $b_{k,n}(t)$ for a given n and t; see (4.4). But it can be interesting to look at this from a different angle, and consider each term as a function of t. Here is a picture of all these polynomials, for $n = 20$:



We may note that $b_{k,n}(t)$ has its maximum at $t = k/n$, and $\int_0^1 b_{k,n}(t)\,dt = 1/(n+1)$. In fact, $(n+1)b_{k,n}$ is the probability density of a beta-distributed random variable with parameters $(k+1, n-k+1)$. Such variables have standard deviation varying between approximately $1/(2\sqrt{n})$ (near the center, i.e., for $k \approx n/2$) and $1/n$ (near the edges). Compare this with the distance $1/n$ between the sample points.

It is tempting to conclude that polynomials of degree n can only do a good job of approximating a function which varies on a length scale of $1/\sqrt{n}$.

We can see this, for example, if we wish to estimate a Lipschitz continuous function f, say with $|f(t) - f(s)| \le L|t-s|$. Put $\varepsilon = L\delta$ in (4.11) and then determine the δ that gives the best estimate in (4.11), to arrive at $|f(t) - b_n(t)| < \frac{3}{2}M^{1/3}(L^2/n)^{2/3}$. So the n required for a given accuracy is proportional to L^2, in accordance with the analysis in the previous two paragraphs.

Reference: S. N. Bernshteĭn: A demonstration of the Weierstrass theorem based on the theory of probability. *The Mathematical Scientist* **29**, 127–128 (2004).

By an amazing coincidence, this translation of Bernshteĭn's original paper from 1912 appeared recently. I discovered it after writing the current note.

Chapter 5

Topological vector spaces

Definitions and basic properties

A *topological vector space* is a (real or complex) vector space equipped with a Hausdorff topology, so that the vector space operations are continuous. That is, if the space is called X then vector addition is a continuous map $X \times X \to X$, and multiplication by scalars is a continuous map $\mathbb{C} \times X \to X$ (or $\mathbb{R} \times X \to X$).

One simple and common way to define a topological vector space, is by starting with a vector space X and a family $\mathscr{P}$ of *seminorms* on X, separating the points of X. (For details see the preceding chapter.) The topology generated by $\mathscr{P}$ does in fact make X a topological vector space. (*Exercise*: Prove this.)

It is useful to note that the topology of a topological vector space is completely specified by the neighbourhoods of 0. For addition by any constant x is continuous with a continuous inverse (addition by $-x$), so the neighbourhoods of x are the sets of the form $x+U$ where U is a neighbourhood of 0.

A topological vector space is called *locally convex* if every neighourhood of 0 contains a convex neighbourhood of 0. If the topology is generated by a family of seminorms, the space is locally convex, since every neighbourhood of 0 contains a neighbourhood of the form

$$\{x \in X\colon\ p_j(x) < \varepsilon,\ j = 1, \dots, n\} \qquad (p_1, \dots, p_n \in \mathscr{P}),$$

which is clearly convex.

Examples include the *weak* topology on a normed space X, and the weak* topology on its dual.

In fact, the topology of any locally convex space is generated by seminorms. For any neighbourhood of 0 contains a neighbourhood U that is not only convex but *balanced*, in the sense that $\alpha x \in U$ whenever $x \in U$ and α is a scalar with $|\alpha| = 1$. For such a neighbourhood U we can create a seminorm p by

$$p(x) = \inf\{t > 0\colon\ x/t \in U\},$$

and the set of all such seminorms generate the topology. (p is a *semi*norm because the supremum may in fact be infinite, so that $p(x) = 0$ for some x.) Since we shall not need this result, we omit the proof. (But proving it could be a useful exercise.)

We already know that bounded linear functionals on a Banach space are continuous. Below is the corresponding result for topological vector spaces.

49 Proposition. *For a linear functional f on a topological vector space X, the following are equivalent:*

1. *f is continuous,*
2. *$\operatorname{Re} f$ is continuous,*
3. *$\operatorname{Re} f$ is bounded below or above on some nonempty open set,*

Proof: That $1 \Rightarrow 2$ is obvious.

$2 \Rightarrow 1$: In the case of real scalars, there is nothing to prove. In the case of complex scalars, use the identity $f(x) = \operatorname{Re} f(x) - i \operatorname{Re} f(ix)$.

$2 \Rightarrow 3$: If $\operatorname{Re} f$ is continuous then $\operatorname{Re} f$ is bounded on the nonempty open set $\{x \in X : |\operatorname{Re} f(x)| < 1\}$, which implies the third condition.

$3 \Longrightarrow 2$: Finally, if $\operatorname{Re} f$ is bounded below on some nonempty open set, say $\operatorname{Re} f > a$ on U, let $u \in U$. Then $\operatorname{Re} f$ is bounded below on the neighbourhood $U - u$ of 0, and above on the neighbourhood $u - U$ of 0. Hence $\operatorname{Re} f$ is bounded on a neighbourhood $V = (U-u) \cap (u-U)$ of 0.

For simplicity, say $|\operatorname{Re} f(x)| \le M$ for $x \in V$. Given $\varepsilon > 0$, then $|\operatorname{Re} f(y) - \operatorname{Re} f(x)| < \varepsilon$ on the neighbourhood $x + (\varepsilon/M)V$ of x. ∎

For a linear functional f on X, we define the *kernel* of f to be

$$\ker f = \{x \in X : f(x) = 0\}.$$

50 Proposition. *A linear functional on a locally convex topological vector space is continuous if and only if its kernel is closed.*

Proof: If f is continuous, then $\ker f = f^{-1}(\{0\})$ is closed, since $\{0\}$ is closed.

Next, assume that $\ker f$ is closed. If $f = 0$, then f is continuous. Assume therefore $f \neq 0$. Pick $w \in X$ with $f(w) = 1$. Since $\ker f$ is closed, f is nonzero on a neighbourhood W of w. Since X is locally convex, W can be taken to be convex. If X is a real space, we must have $f|_W > 0$, so f is bounded below on W, and f must be continuous.

If X is a complex space, we note that $\{f(x) : x \in W\}$ is a convex subset of $\mathbb{C}$ which does not contain 0. Then that set lies in some halfplane, meaning that there is some real number θ with $\operatorname{Re}(e^{i\theta} f(x)) \ge 0$ for all $x \in W$. But then $\operatorname{Re}(e^{i\theta} f)$ is bounded below on W, so $e^{i\theta} f$ is continuous. But then so is f. ∎

51 Proposition. *Let X be a normed vector space. A linear functional on X^* is of the form $f \mapsto f(x)$ where $x \in X$ if and only if it is weakly* continuous.*

Proof: The functionals $f \mapsto f(x)$ are weakly* continuous by the construction of the weak* topology on X^*.

So we only need to consider a weakly* continuous functional ξ on X^*, and must show it can be written $f \mapsto f(x)$.

By the construction of the weak* topology, there are $x_1, \dots, x_n \in X$ and $\delta > 0$ with $|\xi(f)| < 1$ whenever $|f(x_k)| < \delta$ for $k = 1, \dots, n$. In particular,

$$\xi(f) = 0 \text{ whenever } \big(f(x_1), \dots, f(x_n)\big) = (0, \dots, 0)$$

(for then $|\xi(tf)| < 1$ for all t).

It follows that we can define a linear functional ζ on $\mathbb{R}^n$ by setting

$$\zeta\big(f(x_1),\dots,f(x_n)\big) = \xi(f), \qquad f \in X^*.$$

This is well defined, for if $\big(f(x_1),\dots,f(x_n)\big) = \big(g(x_1),\dots,g(x_n)\big)$ then $\xi(f-g) = 0$ by the above. Strictly speaking, this may define ζ only on a subspace of $\mathbb{R}^n$, but the functional can be extended to all of $\mathbb{R}^n$. Write $\zeta(y) = c_1 y_1 + \cdots + c_n y_n$. Then

$$\xi(f) = c_1 f(x_1) + \cdots + c_n f(x_n) = f(x), \qquad \text{where } x = c_1 x_1 + \cdots + c_n x_n.$$

This completes the proof. ∎

The weak topology on a Banach space truly is weaker than the norm topology, at least if the space is infinite-dimensional. For any weak neighbourhood of 0 will contain a set of the form $\{x\colon |f_k(x)| < \varepsilon,\ k = 1,\dots,n\}$ where $f_k \in X^*$. In particular, it contains $\{x\colon |f_k(x)| = 0,\ k = 1,\dots,n\}$, which is an infinite-dimensional space.

A word on notation: It is common to write $\rightharpoonup$ for weak convergence and $\overset{*}{\rightharpoonup}$ for weak* convergence. In the case of *reflexive* Banach spaces, such as Hilbert spaces and L^p spaces for $1 < p < \infty$, there is of course no essential difference, and the notation $\rightharpoonup$ is generally preferred.

In a weaker topology, convergence is easier to achieve because there are fewer neighbourhoods. Here is a simple example. Let $1 < p \le \infty$ and let $e_k \in \ell^p$ be the sequence which has 1 at position k and 0 at all other positions. Then, because $\|e_j - e_k\|_p = 2^{1/p}$ when $j \ne k$, the sequence $(e_k)_{k=1}^\infty$ is not Cauchy, and therefore not convergent in norm. But still, $e_k \rightharpoonup 0$ (weakly) as $k \to \infty$, because whenever $x \in \ell^q$ with $1 \le q < \infty$ then $x_k \to 0$.

In some important cases, however, weak convergence with an added condition does imply norm convergence. Recall that L^p spaces are uniformly convex for $1 < p < \infty$. In the existence theory of partial differential equations, one commonly proves weak convergence of approximate solutions in some L^p space, and one then needs norm convergence to complete the existence proof. The following result is useful in such situations.

52 Proposition. *Let X be a uniformly convex normed space, and let $(x_k)_{k=1}^\infty$ be a weakly convergent sequence in X with weak limit x. Then $x_k \to x$ (in norm) if and only if $\|x_k\| \to \|x\|$.*

Proof: If $x_k \to x$ in norm then $\|x_k\| \to \|x\|$. We need to prove the converse. If $x = 0$, the converse is nothing but the definition of norm convergence. So we may assume that $x \ne 0$, and (dividing everything by $\|x\|$) we may as well assume that $\|x\| = 1$ and $\|x_k\| \to 1$. We may even replace x_k by $x_k/\|x_k\|$.

So now our assumptions are $\|x_k\| = \|x\| = 1$ and $x_k \rightharpoonup x$, and we need to prove that $\|x - x_k\| \to 0$.

By (a corollary to) the Hahn–Banach theorem there is a linear functional $f \in X^*$ with $\|f\| = f(x) = 1$. So $f(x_k) \to 1$. Let $\varepsilon > 0$, and pick $\delta > 0$, thanks to the uniform convexity of X, so that $\|u - v\| < \varepsilon$ whenever $\|u\| = \|v\| = 1$ and $\|u + v\| < 2 - \delta$. Pick a number N so that

$\operatorname{Re} f(x_k) > 1-\delta$ whenever $k \geq N$. For such a k, then, $\operatorname{Re} f(x+x_k) = f(x) + \operatorname{Re} f(x_k) > 2-\delta$, so $\|x+x_k\| > 2-\delta$. Thus $\|x-x_k\| < \varepsilon$ for such k. ∎

The Banach–Alaoglu theorem

The following theorem is also known by Alaouglu's name alone.[1]

53 Theorem. (Banach–Alaoglu) *Let X be a normed space. The closed unit ball in the dual space X^* is compact in the weak* topology.*

Proof: Basically, we can identify the closed unit ball $\overline{B}$ of X^* with a closed subset of the space

$$\Xi = \prod_{x\in X} \{z \in \mathbb{C} : |z| \leq \|x\|\},$$

with the product topology. More precisely, if $f \in X^*$ and $\|f\| \leq 1$, we write f_x in place of $f(x)$, and so $f = (f_x)_{x\in X}$ *is* the wanted element of Ξ. A general element $f \in \Xi$ belongs to $\overline{B}$ if and only if it is linear, i.e.,

$$f(x+y) = f(x) + f(y), \quad f(\alpha x) = \alpha f(x), \qquad x, y \in X, \alpha \in \mathbb{C}.$$

(The bound $\|f\| \leq 1$ is already built into Ξ.) But for given $x, y \in X$ and $\alpha \in \mathbb{C}$, the quantities $f(x+y)$, $f(x)$, $f(y)$, $f(\alpha x)$ are continuous functions of f with respect to the product topology (in an earlier notation, they are the projections $\pi_{x+y}(f)$, etc.), which proves that X^* is indeed closed in Ξ. Also, the weak* topology on X^* is just the topology which X^* inherits as a subspace of Ξ. Thus X^* is a closed subspace of a compact space, and therefore itself compact. ∎

This is well and good, but compactness only guarantees the existence of convergent *filters*. We wish to have convergent *sequences*. The following lemma will help.

54 Lemma. *Assume that X is a separable normed space. Then the weak* topology on the closed unit ball of X^* is metrizable.*

Proof: Let $\{x_k : k = 1, 2, \ldots\}$ be a dense subset of X. We define a metric d on X^* by

$$d(f,g) = \sum_{k=1}^{\infty} 2^{-k} \wedge |f(x_k) - g(x_k)|.$$

[1] This theorem was also featured in the Norwegian edition of Donald Duck magazine quite a long time ago (seventies?). In the story, Gyro Gearloose (Petter Smart in Norwegian) has created a machine that makes its users very smart. He tries it on Donald's nephews, and the first thing they say when they emerge from the machine is the statement of Alaouglu's theorem. Well, almost. The exact text (as far as I remember) was: "Enhetsballen er kompakt i den svarte stjernetypologien." Close enough. Apparently, in the American edition, they said something of a rather more trivial nature.

(Here $a \wedge b$ is short for the minimum of a and b.) That d satisfies the axioms for a metric is obvious, with the possible exception of the requirement that $d(f,g) > 0$ when $f \neq g$. But if $f \neq g$ then there is some $x \in X$ with $f(x) \neq g(x)$. If k is chosen so that $\|x_k - x\|$ is small enough, we should get $f(x_k) \neq g(x_k)$, so that one of the terms in the sum defining $d(f,g)$ is nonzero. In fact

$$\begin{aligned} |f(x_k) - g(x_k) - \big(f(x) - g(x)\big)| &= |(f-g)(x_k - x)| \\ &\leq \|f-g\|\,\|x - x_k\| < |\big(f(x) - g(x)\big)| \end{aligned}$$

if $\|x - x_k\| \leq |\big(f(x) - g(x)\big)| / \|f - g\|$, and then $f(x_k) - g(x_k)$ must be nonzero.

We might hope that d generates the weak* topology on X^*, but that turns out not to be so. But it does generate the relative topology inherited to the closed unit ball B of X^*, as we now show.

First consider a d-neighbourhood of $g \in B$. It will contain an ε-ball, that is, $\{f \in B\colon d(f,g) < \varepsilon\}$. Pick n so that $\sum_{k=n+1}^{\infty} 2^{-k} < \frac{1}{2}\varepsilon$. Then the given d-neighbourhood contains the weak*-neighbourhood

$$\Big\{f \in B\colon |f(x_k) - g(x_k)| < \frac{\varepsilon}{2n} \text{ for } k = 1,\dots,n\Big\}.$$

On the other hand, consider a weak*-neighbourhood of $g \in B$. It contains a set of the form

$$V = \big\{f \in B\colon |f(z_j) - g(z_j)| < \varepsilon \text{ for } j = 1,\dots,m\big\},$$

where $z_1,\dots,z_m \in X$. Now pick some n so that, for each $j = 1,\dots,n$, there is some $k \leq n$ with $\|x_k - z_j\| < \frac{1}{4}\varepsilon$. Let $\delta = 2^{-n} \wedge \frac{1}{2}\varepsilon$. We claim that V contains the d-neighbourhood $\{f \in B\colon d(f,g) < \delta\}$.

To see this, note that if $d(f,g) < \delta$ then $|f(x_k) - g(x_k| < \frac{1}{2}\varepsilon$ for $k = 1,\dots,n$. For a given j, pick some k with $\|x_k - z_j\| < \frac{1}{4}\varepsilon$. Then, since $\|f - g\| \leq \|f\| + \|g\| \leq 2$,

$$|f(z_j) - g(z_j)| \leq |(f-g)(z_j - x_k)| + |f(x_k) - g(x_k)| < \tfrac{1}{2}\varepsilon + \tfrac{1}{2}\varepsilon = \varepsilon,$$

where have used that $\|f - g\| \leq 2$. (This is where we need to restrict our attention to a *bounded* subset of X^*.)

This proves the second half. ∎

55 Proposition. *Let X be a separable normed space. Then any bounded sequence in X^* has a weakly* convergent subsequence.*

Proof: Let $(f_k)_{k=1}^{\infty}$ be a bounded sequence in X^*. We may assume, without loss of generality, that $\|f_k\| \leq 1$ for all k. The unit ball of X^* with the weak* topology is compact and metrizable, but it is well known that any sequence in a compact metric space has a convergent subsequence.

We could pretend that we do not know this fact, and proceed directly: This is, if nothing else, a handy exercise to learn to translate back and forth between filters and sequences. The set of tails $\{f_k\colon k \geq n\}$ of the sequence generates a filter $\mathscr{F}$ on B. By compactness,

there is a refinement $\mathscr{G}$ which converges to some $g \in B$. Then for each $\varepsilon > 0$, the ball $\{f \in B\colon d(f,g) < \varepsilon\}$ belongs to $\mathscr{G}$, and so does every tail of the original sequence. Hence the intersection of the ball and the tail belongs to $\mathscr{G}$, and this intersection is therefore not empty. In plain language, this means that for each $\varepsilon > 0$ and each n there is some $k > n$ with $d(f_k, g) < \varepsilon$. From this knowledge, building a convergent subsequence is easy: Pick k_1 with $d(f_{k_1}, g) < 1$, then pick $k_2 > k_1$ with $d(f_{k_2}, g) < \frac{1}{2}$, then pick $k_3 > k_2$ with $d(f_{k_3}, g) < \frac{1}{3}$, and so forth. ∎

It is useful to know a sort of converse to the above result: To have any reasonable hope of getting a weakly* convergent subsequence, we had better start with a bounded sequence.

56 Proposition. *Any weakly convergent sequence on a normed space, or weakly* convergent sequence on the dual of a Banach space, is necessarily bounded.*

Proof: Let X be a Banach space, and $(f_k)_{k=1}^{\infty}$ a weakly* convergent sequence. For each $x \in X$, the sequence $\big(f_k(x)\big)_{k=1}^{\infty}$ is convergent, and hence bounded. By the Banach–Steinhaus theorem (uniform boundedness principle), the sequence $(f_k)_{k=1}^{\infty}$ is bounded.

The corresponding result for weakly convergent sequences on a normed space is proved the same way, but now using the Banach–Steinhaus theorem for functionals on the dual space, which as we know is complete, so the theorem is applicable. ∎

It is important in the second part of the previous result that the space be complete. To see why, let $\ell^1_{\mathrm{c}} \subset \ell^1$ be the set of sequences all of whose entries except a finite number are zero. We use the ℓ^1 norm on this space. Let f_k be the functional $f_k(x) = kx_k$. Then $f_k \xrightarrow{*} 0$ on ℓ^1_{c}, but $\|f_k\| = k \to \infty$.

The geometric Hahn–Banach theorem and its immediate consequences

Let X be a real vector space. Recall that a *sublinear functional* on X is a function $p\colon X \to \mathbb{R}$ so that $p(\alpha x) = \alpha p(x)$ for $\alpha \in \mathbb{R}^+$, $x \in X$ and $p(x+y) \le p(x) + p(y)$ for $x, y \in X$. Let us say that p *dominates* a linear functional f if $f(x) \le p(x)$ for all x. Finally, recall that the *Hahn–Banach theorem* states that, if p is a sublinear functional on X and f is a linear functional on a subspace of X, dominated by p, then f has an extension to all of X which again is dominated by p.

We shall be interested in the geometric implications of the Hahn–Banach theorem. We shall call x an *interior* point of a set $C \subseteq X$ for every $y \in X$ there is some $\varepsilon > 0$ so that $x + ty \in C$ whenever $|t| < \varepsilon$.[2] It is easy to show that, if p is a sublinear functional on X, then $\{x \in X\colon p(x) < 1\}$ and $\{x \in X\colon p(x) \le 1\}$ are convex sets with 0 an interior point. The converse is also true:

[2] There is potential for confusion here, as *interior* is usually a *topological* concept. If X is a topological vector space then an interior point of C (in the topological sense, i.e., C is a neighbourhood of x) is also an interior point in the sense defined here. The converse is not true.

57 Lemma. *Let C be a convex subset of a real vector space X, and assume that 0 is an interior point of C. Then the function p defined by*

$$p(x) = \inf\{t > 0 \colon x/t \in C\}$$

is a sublinear functional on X. Moreover, x is an interior point of C if and only if $p(x) < 1$, and $x \notin C$ if $p(x) > 1$.

The functional p is sometimes called the *gauge* of C.

Proof: The homogeneity condition $p(\alpha x) = \alpha p(x)$ when $\alpha > 0$ is fairly obvious.

For subadditivity, if $x/s \in C$ and $y/t \in C$ then we can form the convex combination

$$\frac{s}{s+t}\frac{x}{s} + \frac{t}{s+t}\frac{y}{t} = \frac{x+y}{s+t}$$

so that $(x+y)/(s+t) \in C$, and $p(x+y) \le s+t$. Taking the infimum over all s and t satisfying the conditions, we find $p(x+y) \le p(x) + p(y)$.

If $p(x) > 1$ then $x/t \notin C$ for some $t > 1$. Equivalently, $sx \notin C$ for some $s < 1$. But then $x \notin C$, since $sx = sx + (1-s)0$ is a convex combination of x and 0 and C is convex.

If $p(x) = 1$ then x is not an interior point of C. For then $x + tx = (1+t)x \notin C$ for any $t > 0$.

If $p(x) < 1$, then $x \in C$ first of all, for then there is some $t < 1$ so that $x/t \in C$. And so $x = t(x/t) + (1-t)0 \in C$ because C is convex. Next, if $y \in X$ and $t \in \mathbb{R}$ then $p(x+ty) \le p(x) + |t|\big(p(y) \vee p(-y)\big) < 1$ when $|t|$ is small enough, so $x + ty \in C$ when $|t|$ is small enough. Thus x is an interior point of C. (Here $a \vee b$ is the maximum of a and b.) ∎

58 Theorem. (Hahn–Banach separation theorem I)
Let X be a real vector space, and C a convex subset of X with at least one interior point. Let $x \in X \setminus C$. Then there exists a linear functional f on X so that $f(z) \le f(x)$ for every $z \in C$, and $f(z) < f(x)$ when z is an interior point of C.

Proof: We may, without loss of generality, assume that 0 is an interior point of C. (Otherwise, replace C by $C - w$ and x by $x - w$ where w is an interior point of C.)

Let p be the gauge of C. Define $f(tx) = tp(x)$ for $t \in \mathbb{R}$; then f is a linear functional on the one-dimensional space spanned by x which is dominated by p. By the Hahn–Banach theorem, we can extend f to all of X, with the extension still dominated by p. For $z \in C$ we find $f(z) \le p(z) \le 1 \le p(x) = f(x)$, with the middle inequality being strict when z is interior in C. ∎

59 Theorem. (Hahn–Banach separation theorem II)
Let X be a real vector space, and U and V two nonempty disjoint convex subsets of X, at least one of which contains an interior point. Then there is a nonzero linear functional f on X and a constant c so that $f(u) \le c \le f(v)$ for all $u \in U$, $v \in V$. If u is an interior point of U then $f(u) < c$. Similarly, if v is an interior point of V then $f(v) > c$.

Proof: Let $C = U - V = \{u - v\colon u \in U, v \in V\}$. Then $0 \notin C$ because $U \cap V = \varnothing$. Moreover, C is convex and has an interior point. Thus there is a linear functional f so that $f(x) \le f(0) = 0$ for $x \in C$, and $f(x) < 0$ for any interior point x of C. Thus, if $u \in U$ and $v \in V$, we get $f(u - v) \le 0$, and so $f(u) \le f(v)$. It follows that $\sup_{u\in U} f(u) \le f(v)$ for all $v \in V$, and then also $\sup_{u\in U} f(u) \le \inf_{v\in V} f(v)$. We can pick c between these two numbers to finish the proof. If u is an interior point of U and $v \in V$, let $z \in X$ with $f(z) > 0$. Then $u + tz \in U$ when t is small and positive, so $f(u) < f(u+tz) \le c$. The proof of the corresponding strict inequality for an interior point v of V is proved the same way. ∎

We now investigate the consequences of the Hahn–Banach theorem in locally convex spaces.

60 Corollary. *Let X be a locally convex topological vector space, C a closed, convex subset of X, and $w \in X \setminus C$. Then there is a continuous linear functional f on X and a constant c so that* $\operatorname{Re} f(x) \le c < \operatorname{Re} f(w)$ *for all $x \in C$.*

Proof: Let V be a convex neighbourhood of w with $V \cap C = \varnothing$. Apply the previous theorem to C and V, noting that w is an interior point of V. If X is a complex space, apply this result to X as a real space, then use the fact that any real linear functional on X is the real part of a complex linear functional on X. The continuity of the functional follows from Proposition 49. ∎

61 Corollary. *Let X be a locally convex topological vector space, and $A \subset X$. Then $w \in \overline{\operatorname{co}}\, A$ if and only if* $\operatorname{Re} f(w) \le c$ *for every continuous linear functional f on X and every scalar c satisfying* $\operatorname{Re} f(x) \le c$ *for all $x \in A$.*

Proof: Let C be the set of points satisfying the stated condition. Clearly, C is closed and convex, and $A \subseteq C$. Thus $\overline{\operatorname{co}}\, A \subseteq C$.

If $w \notin \overline{\operatorname{co}}\, A$, then by Corollary 60 (with $\overline{\operatorname{co}}\, A$ in the place of C) there is a continuous linear functional f and a constant c so that $\operatorname{Re} f(x) \le c < \operatorname{Re} f(w)$ for all $x \in \overline{\operatorname{co}}\, A$. In particular this holds for all $x \in A$, and so $w \notin C$. Thus $C \subseteq \overline{\operatorname{co}}\, A$. ∎

62 Corollary. *Let X be a locally convex topological vector space, and let $w \in X$ be a nonzero vector. Then there is a continuous linear functional f on X with $f(w) \neq 0$.*

Proof: Apply the previous corollary with $C = \{0\}$. ∎

63 Corollary. *Let C be a (norm) closed convex subset of a normed space X. Then C is weakly closed.*

Proof: Let $w \in X \setminus C$, and pick a linear functional f as in Corollary 60. Then the weak neighbourhood $\{x\colon f(x) > c\}$ of w is disjoint from C. ∎

This has interesting consequences for weakly convergent sequences. From a weakly convergent sequence one can form a new sequence by taking convex combinations of members of the original sequence to obtain a norm convergent sequence. The details follow these definitions.

Given a set of points $A \subset X$, a *convex combination* of points in A is a point which can be written

$$\sum_{k=1}^{n} t_k a_k, \qquad a_k \in A, \quad t_k \geq 0, \quad \sum_{k=1}^{n} t_k = 1.$$

The set of all convex combinations of points in A forms a convex set, called the *convex hull* of A, and written $\operatorname{co} A$. If X is a topological vector space, the closure of $\operatorname{co} A$ is also convex. It is called the *closed convex hull* of A, and written $\overline{\operatorname{co}}\, A$.

64 Corollary. *Let X be a normed space, assume $x_k \in X$ and $x_k \rightharpoonup x$ (weakly). Then, given $\varepsilon > 0$ and a number N, there exists some convex combination z of $x_N, x_{N+1}, \ldots$ with $\|x - z\| < \varepsilon$.*

Proof: Let $A = \{x_N, x_{N+1}, \ldots\}$, and consider $\overline{\operatorname{co}}\, A$ (with closure in the norm topology). By Corollary 63, $\overline{\operatorname{co}}\, A$ is weakly closed as well. Since $x_k \rightharpoonup x$ and $x_k \in A$ for $k \geq N$, $x \in \overline{\operatorname{co}}\, A$. The conclusion is immediate. ∎

The *Banach–Saks theorem* is a special case of the above corollary for L^p spaces. In this case, the convex combinations can be chosen to be of the special form $\left(x_{n_1} + \cdots + x_{n_m}\right) / m$.

We round off this section with some other useful results. The first lemma is of interest in its own right.

65 Lemma. *The closed unit ball in the dual space of a normed space is weakly* closed.*

Proof: Let X be a normed space, and B the closed unit ball of X^*, Then B is defined by inequalities $|f(x)| \leq 1$ where $x \in X$ and $\|x\| \leq 1$, and the maps $f \mapsto f(x)$ are weakly* continuous. ∎

66 Theorem. (Goldstine) *Let X be a normed space. Identify X with its canonical image in X^{**}. Then the closed unit ball of X^{**} is the weak*-closure of the unit ball of X.*

Proof: By the above lemma applied to X^* instead of X, the closed unit ball of X^{**} is weakly* closed.

Let B be the closed unit ball of X, and $\overline{B}$ its weak*-closure. If $\overline{B}$ is not the entire unit ball of X^{**} there is some $\xi \in X^{**}$ with $\|\xi\| \leq 1$ and $\xi \notin \overline{B}$. We can find a weakly* continuous linear functional F on X^{**}, and a constant c, so that $\operatorname{Re} F(x) \leq c < \operatorname{Re} F(\xi)$ whenever $x \in \overline{B}$. But according to Proposition 51, this functional must be of the form $F(\zeta) = \zeta(f)$ where $f \in X^*$. Bearing in mind that we have identified X with its canonical image in X^{**}, we then have $\operatorname{Re} f(x) \leq c < \operatorname{Re} \xi(f)$ for every $x \in B$.

The first inequality implies $\|f\| \le c$. The other implies $c < \|\xi\| \cdot \|f\|$, and combining these two inequalities we obtain $\|\xi\| > 1$. But this contradicts the assumption $\|\xi\| \le 1$, so we are done. ∎

67 Theorem. (Kakutani) *A Banach space is reflexive if and only if its closed unit ball is weakly compact.*

Proof: Recall that the weak topology on X is just the subspace topology inherited from the weak* topology on X^{**}. Thus, if X is reflexive, the compactness of its unit ball follows from the Banach–Alaoglu theorem. Conversely, if X has a weakly compact unit ball B, then B is also a weak*-compact subset of X^{**}, and therefore weak*-closed. By Goldstine's theorem, it must be the entire unit ball of X^{**}. ∎

Uniform convexity and reflexivity

We are now ready to prove the theorem that will finish our proof on the dual space of L^p for $1 < p < \infty$. The proof of Theorem 69 presented here is essentially due to J. R. Ringrose (J. London Math. Soc. **34** (1959), 92.)

The following lemma will perhaps make Ringrose's extremely short proof clearer. Recall that the *diameter* of a subset A of a normed space is defined to be $\operatorname{diam} A = \sup\{\|x - y\| : x, y \in A\}$.

68 Lemma. *Let $A \subseteq X^*$ where X is a normed space. Then the diameter of the weak*-closure of A equals the diameter of A.*

Proof: Write $\overline{A}$ for the weak*-closure of A. Since $A \subseteq \overline{A}$, it is clear that $\operatorname{diam} A \le \operatorname{diam} \overline{A}$.

The opposite inequality is an immediate consequence of Lemma 65. Indeed, write $d = \operatorname{diam} A$, and note that by assumption, $f - g \in dB$ for all $f, g \in A$, where B is the closed unit ball of X^*. But B, and hence dB, is weakly* closed, and since subtraction is weakly* continuous, we also get $f - g \in dB$ for all $f, g \in \overline{A}$. ∎

Note that the norm on X^* is not weakly* continuous. It is, however, *upper weakly* semicontinuous*, as an easy adaption of the above proof shows. The lemma is an immediate consequence of the upper semicontinuity of the norm.

69 Theorem. (Milman–Pettis) *A uniformly convex Banach space is reflexive.*

Proof: Let X be a uniformly convex Banach space. Let $\xi \in X^{**}$. We must show that $\xi \in X$ (where we have again identified X with its canonical image in X^{**}). We may assume that $\|\xi\| = 1$.

Let $\varepsilon > 0$, and pick $\delta > 0$ so that $\|x - y\| < \varepsilon$ whenever $x, y \in X$, $\|x\| \le 1$, $\|y\| \le 1$ and $\|x + y\| > 2 - \delta$.

Let B be the closed unit ball of X (as a subset of X^{**}), and let $\overline{B}$ be its weak*-closure, which as we know is the unit ball in X^{**}. In particular, $\xi \in \overline{B}$.

From the definition of norm in X^{**}, we can find $f \in X^*$ with $\|f\| = 1$ and $\xi(f) > 1 - \frac{1}{2}\delta$. Let $V = \{\zeta \in X^{**} : \zeta(f) > 1 - \frac{1}{2}\delta\}$. Then V is weakly* open.

Now $\xi \in V \cap \overline{B} \subseteq \overline{V \cap B}$, where the bar denotes weak*-closure. (See below.)

If $x, y \in V \cap B$ then $\|x+y\| > 2-\delta$ because $f(x+y) = f(x) + f(y) > 2-\delta$. Thus $\|x-y\| < \varepsilon$, and so $\operatorname{diam} V \cap B \leq \varepsilon$. By the previous lemma, $\operatorname{diam} \overline{V \cap B} \leq \varepsilon$ as well. In particular, $\|\xi - x\| \leq \varepsilon$, where $x \in V \cap B$.

We have shown that ξ lies in the norm closure of X. But since X is a Banach space, it is complete and therefore norm closed in X^{**}. Thus $\xi \in X$, and the proof is complete.

Note: The inclusion $V \cap \overline{B} \subseteq \overline{V \cap B}$ used above holds in any topological space, where V is open. For then the complement W of $\overline{V \cap B}$ is open, and $W \cap V \cap B = \varnothing$. And since $W \cap V$ is open, this implies $W \cap V \cap \overline{B} = \varnothing$, which is another way to state the desired inclusion. ∎

The Krein–Milman theorem

Any point in a cube, or an octahedron, or any other convex polygon, can be written as a convex combination of the corners of the polygon. The Krein–Milman theorem is an infinite-dimensional generalization of this fact.

Let K be a convex set in some vector space. A *face* of K is a nonempty convex subset $F \subseteq K$ so that, whenever $u, v \in K$, $0 < t < 1$, and $tu + (1-t)v \in F$, it follows that $u, v \in F$. An *extreme point* of K is a point x so that $\{x\}$ is a face of K. In other words, whenever $u, v \in K$, $0 < t < 1$, and $tu + (1-t)v = x$, it follows that $u = v = x$. The set of extreme points of K is also called the *extreme boundary* of K, and written $\partial_e K$. (Make a drawing illustrating these concepts!)

As an example, assume real scalars (for simplicity) and let f be a linear functional so that $f(x) \leq c$ for all $x \in K$. Then $\{x \in K : f(x) = c\}$ is a face of K, if it is nonempty.

As another example, consider an ordinary closed cube in $\mathbb{R}^3$. The faces of this cube are: The cube itself, its sides (what we think of as its faces in everyday language), its edges, and its corners (or rather the singleton sets made up of its corners). In particular, the corners are the extreme points of the cube.

70 Theorem. (Krein–Milman) *Any compact convex subset of a locally convex vector space is the closed convex hull of its extreme boundary.*

More briefly, if K is such a set then $K = \overline{\operatorname{co}}\, \partial_e K$.

Proof: We prove this for the case of real scalars. The case for complex scalars follows in the usual way.

We shall begin by proving a much weaker statement, namely that *any compact convex set* (in a locally convex space) *contains at least one extreme point.*

To this end, let K be a compact convex set, and consider the collection Φ of all its closed faces. We shall prove that there exists a *minimal* closed face, and then we shall prove that a minimal closed face contains only one point.

For the first part, we shall use Zorn's lemma. We need to show that Φ is inductively ordered by *inverse* inclusion. If $\mathscr{C}$ is a chain in Φ, then $\bigcap\mathscr{C} \neq \varnothing$ because $\mathscr{C}$ consists of closed sets, it has the finite intersection property (for it is totally ordered, so any intersection of a finite number of members of $\mathscr{C}$ is simply the smallest of them), and K is compact. It is also a face (exercise: show this). Thus Zorn's lemma proves the existence of a *minimal* member of Φ.

Now let F be a minimal closed face of K. Assume that F contains two distinct points x and y. Let f be a continuous linear functional with $f(x) \neq f(y)$. Then f attains its maximum on F, because F is compact (it is a closed subset of the compact set K). Writing c for this maximum, then $\{z \in F\colon f(z) = c\}$ is a face of F, and therefore a face of K as well. Since F was a *minimal* face, this new face must be all of F, so that $f(z) = c$ for all $z \in F$. But this contradicts $f(x) \neq f(y)$. Thus F cannot contain two distinct points, so it consists of a single point. Thus K contains at least one extreme point.

Now, assume that $K \neq \overline{\operatorname{co}}\,\partial_e K$. Since obviosly $\overline{\operatorname{co}}\,\partial_e K \subseteq K$, we can now use Corollary 61 to find a continuous linear functional f and a constant c so that $f(x) \leq c$ for all $x \in \overline{\operatorname{co}}\,\partial_e K$, but $f(x) > c$ for at least some points in K. Let m be the maximum value of f over K. It exists because K is compact. Then $\{x \in X\colon f(x) = m\}$ is a closed face of K. In particular, this set is compact and convex, so it contains at least one extreme point (of itself, and therefore of K). But now if this extreme point is called x then $f(x) = m$, but also $f(x) \leq c < m$ because $x \in \partial_e K \subseteq \overline{\operatorname{co}}\,\partial_e K$. This contradiction completes the proof. ∎

If we throw away a corner of a polyhedron such as a cube or octahedron in three dimensional space, the remaining corners will have a smaller convex hull. On the other hand, we can trow away single points from the extreme boundary of a disk, and the closed convex hull of the remaining points will still be the entire disk. The next theorem, known as Milman's converse to the Krein–Milman theorem, tells us the precise story of what can be thrown away and what must be kept.

71 Theorem. (Milman) *Let K be a compact convex subset of a locally convex topological vector space X. Let $A \subseteq K$ so that $K = \overline{\operatorname{co}}\,A$. Then $\partial_e K \subseteq \overline{A}$.*

Proof: Assuming the conclusion is wrong, we let $w \in \partial_e K$ while $w \notin \overline{A}$.

There is an open convex neighbourhood U of 0 with $(w + U - U) \cap A = \varnothing$. Then $(w + U) \cap (\overline{A} + U) = \varnothing$ as well. Since $\overline{A}$ is a closed subset of the compact set K, it is compact, so there are $x_1, \dots, x_n \in \overline{A}$ so that the n sets $x_k + U$ cover $\overline{A}$.

Let $K_k = \overline{\operatorname{co}}\big((x_k + U) \cap \overline{A}\big)$. Each K_k is a compact convex subset of K, and $K = \overline{\operatorname{co}}(K_1 \cup \dots \cup K_n)$ because $A \subseteq K_1 \cup \dots \cup K_n$.

Even better: Any point $x \in K$ can be written as a convex combination

$$x = \sum_{k=1}^{n} t_k z_k, \qquad t_k \geq 0, \quad z_k \in K_k,\ k = 1, \dots, n, \quad \sum_{k=1}^{n} t_k = 1.$$

For the set of all sums as given above is a compact set (it is a continuous image of the compact set $T \times K_1 \times \cdots \times K_n$, where $T = \{t \in \mathbb{R}^n\colon t_k \geq 0,\ \sum_{k=1}^{n} t_k = 1\}$) and also convex, so that set *is* the closed convex hull of $K_1, \dots, K_n$.

In particular, w can be written that way. But $w \notin K_k$ for any k, for $K_k \subseteq \overline{x_k + U}$ which is disjoint from $w + U$. Thus, when we write $w = \sum_{k=1}^{n} t_k z_k$ at least two of the coefficients, say t_k and $t_{k'}$, are nonzero. But then, varying t_k and $t_{k'}$ while keeping their sum constant, we get a line segment lying in K with w in its interior. This is impossible, since w is assumed to be an extreme point of K.

There is a tiny hole in the above argument, for we could have $z_k = z_{k'}$ (the sets K_k will overlap), in which case the line segment degenerates to a point. But this hole is easy to plug: We cannot have *all* those z_k for which $t_k > 0$ being identical, for then we would have $w \in K_k$, and that is not the case. ∎

Chapter 6

Spectral theory

Chapter abstract. This chapter is intended to be a supplement to Kreyszig's chapter 7 and the opening parts of chapter 9. We are trying to exhibit some of the more important points and leaving some details aside. Kreyszig is great for details, and I will not try to compete with him on that front. However, I will correct him on one point, namely the proof of his theorem 7.4-2 (Spectral mapping theorem for polynomials), which he has made to be much more involved than it needed to be. The rest of the way, this chapter is intended to keep an almost conversational tone.

Operators and Banach algebras

We are of course primarily interested in the spectral theory for bounded operators on a Banach space X. However, much of what we need to do is almost purely algebraic, and so the ideas are actually more clearly exposed if we forget that the operators are in fact operators. There are cases where abstraction goes too far, obscuring rather than revealing, and making the theory more difficult than necessary. And there are cases where abstraction helps by getting rid of irrelevant detail and letting us concentrate on what is essential to the problem at hand. I think the introduction of Banach algebras into spectral theory is an example of the latter sort.

The space $B(X) = B(X, X)$ of bounded linear operators from the Banach space X to itself is the typical example of a *Banach algebra*. Composition of operators becomes simply multiplication in this algebra, and the question of the existence of an inverse becomes divorced from any concrete questions about the kernel and range of an operator.

The algebraic theory of the spectrum

In this section, A will be a complex algebra with a unit e. (See Kreyszig section 7.6.) Recall that $a \in A$ is called *invertible* with inverse b if $ab = ba = e$ where $b \in A$. The usual proof of the uniqueness of the inverse does in fact show something much more interesting and useful:

72 Lemma. *Assume that $a \in A$ has a left inverse x and a right inverse y. In other words, assume that $xa = e$ and $ay = e$. Then $x = y$, and a is invertible.*

Proof: Use the associative law: $x = xe = x(ay) = (xa)y = ey = y$. ∎

The beginner may be confused by the fact that sometimes ab can be invertible even though a and b are not. For example, the right shift operator S on ℓ^2 is not invertible, and neither is its adjoint, the left shift operator S^*. However, S^*S is the identity operator, which is clearly invertible. But SS^* is not invertible: It is the projection on those elements of ℓ^2 whose first component is zero. This situation is typical:

73 Lemma. *Assume that ab and ba are both invertible. Then a and b are both invertible.*

Proof: $ab(ab)^{-1} = e$, so $b(ab)^{-1}$ is a right inverse for a. And $(ba)^{-1}ba = e$, so $(ba)^{-1}b$ is a left inverse for a. Thus a is invertible. The invertibility of b can be proved similarly. ∎

Two elements a, b of A are said to *commute* if $ab = ba$. An immediate consequence of the above lemma is that if ab is invertible and a and b commute, then a and b are both invertible. The converse is of course obvious. This is easily extended by induction to more factors:

74 Lemma. *If $a_1, \dots, a_n$ are mutually commuting elements of A then $a_1 \cdots a_n$ is invertible if and only if $a_1, \dots, a_n$ are all invertible.* ∎

The notion of *spectrum* works just fine for arbitrary complex algebras.

$$\sigma(a) = \{\lambda \in \mathbb{C} \colon a - \lambda e \text{ is not invertible}\}.$$

We can now both generalize Kreyszig's 7.4-2 and simplify its proof:

75 Proposition. (Spectral mapping theorem) *Let A be a complex algebra with unit, let $a \in A$, and let $p(\lambda)$ be a polynomial in one variable λ. Then*

$$\sigma\big(p(a)\big) = p\big(\sigma(a)\big).$$

Proof: We rely on the fundamental theorem of algebra, which says that every nonconstant complex polynomial has at least one root, and thus can be factored into a product of first-degree polynomials. So we fix a constant κ, and factor the polynomial $p(\lambda) - \kappa$:

$$p(\lambda) - \kappa = \alpha(\lambda - \mu_1) \cdots (\lambda - \mu_n), \tag{6.1}$$

where $\alpha, \mu_1, \dots, \mu_n \in \mathbb{C}$. In this identity we can now substitute a for λ, and get

$$p(a) - \kappa e = \alpha(a - \mu_1 e) \cdots (a - \mu_n e).$$

Notice that the lefthand side is non-invertible if and only if $\kappa \in \sigma\big(p(a)\big)$, while the factor $a - \mu_j e$ on the righthand side is non-invertible if and only if $\mu_j \in \sigma(a)$. Since all the factors on the righthand side commute with each other, it follows that

$$\kappa \in \sigma\big(p(a)\big) \text{ if and only if } \mu_j \in \sigma(a) \text{ for some } j.$$

But now we go back to equation (6.1). $\kappa \in p\big(\sigma(a)\big)$ means that the lefthand side is zero for some $\lambda \in \sigma(a)$. But a look at the righthand side of (6.1) reveals that this means that some μ_j equals some $\lambda \in \sigma(a)$, which we have just seen is equivalent to $\kappa \in \sigma\big(p(a)\big)$. ∎

Geometric series in Banach algebras

The whole elementary theory of Banach algebras hinges on the following, almost trivial, observation.

76 Proposition. *Given an element x in a Banach algebra A with unit, $e-x$ is invertible with inverse*

$$(e-x)^{-1} = \sum_{k=0}^{\infty} x^k$$

provided the series on the righthand side converges – and it will, if $\|x\| < 1$. More generally, the series converges if and only if $\|x^k\| < 1$ for some k.

Proof: Clearly, if the series converges then $\|x^k\| \to 0$ as $k \to \infty$. The conclusion then follows immediately by considering the identity (telescoping series)

$$(e-x)\sum_{k=0}^{n} x^k = \Big(\sum_{k=0}^{n} x^k\Big)(e-x) = e - x^{n+1}$$

and letting $n \to \infty$.

If $\|x\| < 1$ then of course $\|x^k\| \le \|x\|^k$, so $\sum_{k=0}^{\infty}\|x^k\| \le \sum_{k=0}^{\infty}\|x\|^k = 1/(1-\|x\|)$, and the sum is absolutely convergent, and hence convergent, since A is complete. More generally, if $\|x^k\| < 1$ for some k, then any sum obtained by including only every kth summand, i.e. $\sum_{q=0}^{\infty} x^{qk+r} = x^r \sum_{q=0}^{\infty}(x^k)^q$, converges. Adding these together for $r = 0, 1, \ldots, k-1$ we get the original sum, which therefore converges as well. ∎

So any member of A sufficiently close to the identity is invertible. This handy result is easily extended to a neigbourhood of *any* invertible element. Assume that w is invertible and write $w - x = w(e - w^{-1}x)$. This factorization shows that $w-x$ is invertible if $\|w^{-1}x\| < 1$. It is useful to have an expression for the inverse as well:

$$(w-x)^{-1} = (e-w^{-1}x)^{-1}w^{-1} = \sum_{k=0}^{\infty}(w^{-1}x)^k w^{-1} \qquad (\|x\| < \|w^{-1}\|^{-1}). \tag{6.2}$$

If you feel disturbed by the asymmetry of the summand $(w^{-1}x)^k w^{-1}$, don't. If you expand the expression, the asymmetry disappears. Also, $(w^{-1}x)^k w^{-1} = w^{-1}(xw^{-1})^k$.

We have proved

77 Proposition. *The set of invertible elements of a Banach algebra is open.* ∎

The resolvent

The *resolvent set* of a member x in a Banach algebra A is the complement of the spectrum: $\rho(x) = \mathbb{C}\setminus\sigma(x)$. It is an immediate consequence of Proposition 77 that the resolvent set is open. And it is an immediate consequence of Proposition 76 that $\lambda \in \rho(x)$ whenever $|\lambda| > \|x\|$ (for if $\lambda \neq 0$ then $\lambda \in \rho(x)$ if and only if $e - x/\lambda$ is invertible).

In other words, the spectrum of x is closed and bounded ($|\lambda| \leq \|x\|$ whenever $\lambda \in \sigma(x)$):

78 Proposition. *The spectrum of any member of a Banach algebra is compact.* ∎

The *resolvent* of x is $R_\lambda(x) = (x - \lambda e)^{-1}$, which is defined for $\lambda \in \rho(x)$. When the context is clear, we just write R_λ.

We can use equation (6.2) to find a useful geometric series for the resolvent in the neighbourhood of any given point: Let $\lambda \in \rho(x)$ and let $\zeta \in \mathbb{C}$ be small enough. Then (6.2) with the obvious variable substitutions yields

$$R_{\lambda+\zeta} = (x - \lambda e - \zeta e)^{-1} = \sum_{k=0}^{\infty} (R_\lambda \zeta)^k R_\lambda = \sum_{k=0}^{\infty} R_\lambda^{k+1} \zeta^k$$

The important fact to note is that this is a power series in ζ with coefficients in A. This proves

79 Proposition. *R_λ is a holomorphic function of λ on $\rho(x)$, for any $x \in A$.* ∎

There are several possible definitions of holomorphic for a vector valued function. The representation in the neighbourhood of every point as a norm convergent power series is the strongest of them, from which all the other forms follow. For example, that $f(R_\lambda)$ be a holomorphic function of λ for each $f \in A^*$. If $A = B(X, X)$, some (apparently) even weaker definitions can be found in Kreyszig's section 7.5 (p. 388).

80 Proposition. *The spectrum of any member of a Banach algebra is nonempty.*

Proof: Assume that $\sigma(x) = \varnothing$. Then R_λ is a holomorphic function defined on the whole complex plane. Moreover, it is bounded at infinity, since Proposition 76 implies

$$-R_\lambda = \lambda^{-1}(e - x/\lambda)^{-1} = \sum_{k=0}^{\infty} \lambda^{-k-1} x^k \qquad (|\lambda| > \|x\|), \tag{6.3}$$

so

$$|R_\lambda| \leq \sum_{k=0}^{\infty} \|\lambda^{-k-1} x^k\| \leq \sum_{k=0}^{\infty} |\lambda|^{-k-1} \|x\|^k = \frac{1}{|\lambda|} \cdot \frac{1}{1 - \|x\|/|\lambda|} \to 0 \quad \text{when } \lambda \to \infty.$$

Thus R_λ is a bounded holomorphic function on $\mathbb{C}$ and by Liouville's theorem, it is constant.

But wait – Liouville's theorem is about complex functions, not vector valued ones. However, if $f \in A^*$ then Liouville's theorem does show that $f(R_\lambda)$ is independent of λ.

And then the Hahn–Banach theorem shows that R_λ is constant: For if $R_\lambda \neq R_\mu$ then we can find some $f \in A^*$ with $f(R_\lambda) \neq f(R_\mu)$, and this is what we just saw is impossible.

Having R_λ constant is absurd, however. For then its inverse $R_\lambda^{-1} = x - \lambda e$ is constant, which it clearly is not. This contradiction finishes the proof. ∎

The occasionally useful *resolvent identity* appears simpy by multiplying the rather trivial relationship

$$(x - \lambda e) - (x - \mu e) = (\mu - \lambda)e$$

from the left by R_μ and from the right by R_λ to get

$$R_\mu - R_\lambda = (\mu - \lambda) R_\mu R_\lambda, \qquad \lambda, \mu \in \rho(x).$$

Holomorphic functional calculus[1]

Multiply equation (6.3) by λ^n and integrate anticlockwise around a large circle Γ. Recall that for an integer m

$$\int_\Gamma \lambda^m \, d\lambda = \begin{cases} 2\pi i, & m = -1, \\ 0 & \text{otherwise,} \end{cases}$$

so we get

$$-\int_\Gamma \lambda^n R_\lambda \, d\lambda = 2\pi i x^n,$$

which we write in the much more suggestive form

$$\frac{1}{2\pi i} \int_\Gamma \frac{\lambda^n}{\lambda e - x} \, d\lambda = x^n, \tag{6.4}$$

from which we conclude that

$$\frac{1}{2\pi i} \int_\Gamma \frac{p(\lambda)}{\lambda e - x} \, d\lambda = p(x)$$

for any polynomial $p(\lambda)$. (Here, of course, dividing by some element of A means multiplying by its inverse. Normally, this operation is not well defined, because we could require multiplication on the left or on the right. But in this case, the numerator is a scalar, so this ambiguity does not happen.)

Note the similarity with Cauchy's integral formula

$$f(z) = \frac{1}{2\pi i} \int_\Gamma \frac{f(\zeta)}{\zeta - z} \, d\zeta$$

valid for any holomorphic function f, where Γ is a closed path surrounding z once in the positive direction, and where f has no singularities inside Γ.

Motivated by this, one can *define* $f(x)$ to be

$$f(x) = \int_\Gamma \frac{f(\lambda)}{\lambda e - x} \, d\lambda$$

[1] This part is not covered in Kreyszig's book, but I cannot resist.

when f is a holomorphic function defined on a neighbourhood of $\sigma(x)$, and Γ is a path inside the domain of f, surrounding each point of $\sigma(x)$ once in the positive direction, but not surrounding any point outside the domain of f.[2]

One can then prove such results as the spectral mapping theorem $\sigma\big(f(x)\big) = f\big(\sigma(x)\big)$, and relationships such as $(f+g)(x) = f(x)+g(x)$ (nearly obvious) and $(fg)(x) = f(x)g(x)$ (far less obvious).

But this is taking us too far afield. Just one more digression before we leave this subject: The *spectral radius* of x is the maximum value of λ where $\lambda \in \sigma(x)$ (in other words, the radius of the spectrum relative to the origin).

81 Proposition. (Spectral radius formula) *The spectral radius of any element x in a complex Banach algebra is*

$$r(x) = \overline{\lim_{n\to\infty}} \|x^n\|^{1/n}.$$

The notation $\overline{\lim}$ stands for *upper limit* or *limit superior*, defined by

$$\overline{\lim_{n\to\infty}} a_n = \lim_{n\to\infty} \sup_{k>n} a_k.$$

Note that the supremum is a decreasing function of n, so the limit is guaranteed to exist in $[-\infty,\infty)$.

Proof: We have proved (6.4) for large circles Γ, where the corresponding geometric series converges. However, the integral can be defined over any circle Γ centered at the origin and with radius $R > r(x)$, and the integral is independent of R by standard complex function theory generalised to the Banach space setting. Hence the formula remains true for any such R. From it, we get the estimate

$$\|x^n\| \le MR^{n+1}, \qquad M = \max_{\lambda\in\Gamma} \|R_\lambda\|.$$

Taking n-th roots and letting $n \to \infty$ we obtain

$$\overline{\lim_{n\to\infty}} \|x^n\|^{1/n} \le R.$$

As this holds for any $R > r(x)$, we conclude

$$\overline{\lim_{n\to\infty}} \|x^n\|^{1/n} \le r(x).$$

The opposite inequality follows from Proposition 76 applied to x/λ whenever $\lambda > \overline{\lim}_{n\to\infty} \|x^n\|^{1/n}$: For then there is some k with $\lambda > \|x^k\|^{1/k}$ and hence $\|x/\lambda\|^k < 1$. ∎

[2] It is possible that $\sigma(x)$ has several components, each in its own component of the domain of f. In this case, Γ must consist of several closed paths.

Spectral properties of self-adjoint operators on a Hilbert space

For the remainder of this note, H will be a Hilbert space. We write $B(H) = B(H,H)$ for the space of bounded linear operators $H \to H$, and $B(H)_{\mathsf{sa}}$ for the self-adjoint members of $B(H)$.

An operator $T \in B(H)$ is called *normal* if it commutes with its adjoint: $TT^* = T^*T$. It is easily proved that if T is normal then $\|T^*x\| = \|Tx\|$ for every $x \in H$.

82 Lemma. *Assume $T \in B(H)$ is normal. Then T is invertible if and only if there is some constant $c > 0$ so that $\|Tx\| \geq c\|x\|$ for every $x \in H$.*

Proof: If T is invertible, then $\|T^{-1}y\| \leq \|T^{-1}\|\|y\|$. With $y = Tx$ this becomes $\|x\| \leq \|T^{-1}\|\|Tx\|$. Put $c = \|T^{-1}\|^{-1}$ to get $\|Tx\| \geq c\|x\|$.

Conversely, if $\|Tx\| \geq c\|x\|$ for each $x \in H$, then T is an isomorphism from H onto $\operatorname{im} T$. In particular, since H is complete, then so is $\operatorname{im} T$, and hence $\operatorname{im} T$ is closed. It only remains to show that $\operatorname{im} T = H$, and for this we only need to show that $\operatorname{im} T$ is dense in H. And this is equivalent to $(\operatorname{im} T)^\perp = \{0\}$. But

$$\begin{aligned} y \in (\operatorname{im} T)^\perp &\Leftrightarrow \langle Tx, y\rangle = 0 \quad \text{for every } x \in H \\ &\Leftrightarrow \langle x, T^*y\rangle = 0 \quad \text{for every } x \in H \\ &\Leftrightarrow T^*y = 0 \Leftrightarrow Ty = 0 \Leftrightarrow y = 0, \end{aligned}$$

where the next-to-last equivalence is due to T being normal (so $\|T^*y\| = \|Ty\|$), and the final one follows directly from the assumptions. ∎

83 Proposition. *If $T \in B(H)_{\mathsf{sa}}$ then $\sigma(T) \subseteq \mathbb{R}$.*

Proof: We first show that $T + iI$ is invertible. It is certainly normal, since it commutes with its adjoint $T - iI$. And

$$\begin{aligned} \|(T+iI)x\|^2 &= \langle Tx + ix, Tx + ix\rangle = \langle Tx, Tx\rangle - i\langle Tx, x\rangle + i\langle x, Tx\rangle + \langle x, x\rangle \\ &= \|Tx\|^2 + \|x\|^2 \geq \|x\|^2, \end{aligned}$$

so $T + iI$ is invertible by the previous Lemma.

In general, if $\lambda = \alpha + i\beta \notin \mathbb{R}$ with $\alpha, \beta \in \mathbb{R}$ then $\beta \neq 0$, so $\beta^{-1}(T + \alpha I) + iI$ is invertible by what we just proved. Multiplying this by β, we conclude that $T + \lambda I$ is invertible. ∎

A *sesquilinear form*[3] on H is a mapping $B\colon H \times H \to \mathbb{C}$ which is linear in its first variable and conjugate linear in its second variable. The corresponding *quadratic form* is given by $Q(x) = B(x,x)$. Note, in particular, that $Q(tx) = B(tx,tx) = tB(x,tx) = t\bar{t}B(x,x) = |t|^2Q(x)$.

84 Lemma. (Polarization identity) *If B is a sesquilinear form on H and Q the corresponding quadratic form, then*

$$B(x,y) = \frac{1}{4}\sum_{k=0}^{3} i^k Q(x + i^k y).$$

[3]The prefix sesqui- is supposed to come from the Latin term for *one and a half.*

Proof: Expand $Q(x+i^k y) = B(x+i^k y, x+i^k y) = Q(x) + i^k B(y,x) + i^{-k} B(x,y) + Q(y)$ and use the fact that $\sum_{k=0}^3 i^k = \sum_{k=0}^3 i^{2k} = 0$. ∎

The sesquilinear form B is called *Hermitean* if $B(y,x) = \overline{B(x,y)}$ for all $x, y \in H$. If B is Hermitean, then clearly Q takes only real values. Conversely, using the polarization identity, it is not hard to show that if Q is real-valued then B is Hermitean. B is called *non-negative* if $Q(x) \geq 0$ for all x.

85 Lemma. (Cauchy–Schwarz) *If B is a non-negative Hermitean form and Q is the associated quadratic form then*

$$|B(x,y)|^2 \leq Q(x)Q(y)$$

Proof: You have seen this before; the proof is precisely the same as the proof of the Cauchy–Schwarz inequality $|\langle x, y\rangle| \leq \|x\|\|y\|$ in an inner product space. Here is a quick version: Whenever $t \in \mathbb{R}$ then

$$0 \leq Q(x+ty) = Q(y)t^2 + 2\operatorname{Re} B(x,y)\, t + Q(x).$$

So the discriminant of this second degree polynomial is non-positive, which leads to $|\operatorname{Re} B(x,y)|^2 \leq Q(x)Q(y)$. Now adjust the phase of $B(x,y)$, making it real by replacing x by $e^{i\theta}x$ for a suitable real number θ. ∎

The following result is absolutely essential for the whole spectral theory of self-adjoint operators.

86 Proposition. (Kreyszig 9.2-2) *If $T \in B(H)_{\mathsf{sa}}$ then*

$$\|T\| = \sup_{\|x\|=1} |\langle Tx, x\rangle|.$$

Proof: Let K be the supremum on the righthand side. Clearly, if $\|x\| = 1$ then $|\langle Tx, x\rangle| \leq \|Tx\|\|x\| \leq \|T\|$, so $K \leq \|T\|$.

For the opposite inequality, let B be the sesquilinear form $B(x,y) = \langle Tx, y\rangle$, and let Q be the corresponding quadratic form.

Note that $|Q(x)| = |\langle Tx, x\rangle| \leq K\|x\|^2$ for all x.

If $x, y \in H$ then the polarization identity yields

$$\langle Tx, y\rangle = \frac{1}{4}\sum_{k=0}^{3} i^k Q(x+i^k y).$$

Taking real parts, we throw away the imaginary ($k=1$ and $k=3$) terms and get

$$\begin{aligned}\operatorname{Re}\langle Tx, y\rangle &= \frac{1}{4}\big(Q(x+y) - Q(x-y)\big) \leq \frac{K}{4}\big(\|x+y\|^2 + \|x-y\|^2\big) \\ &= \frac{K}{2}\big(\|x\|^2 + \|y\|^2\big) = K, \qquad \text{if } \|x\| = \|y\| = 1.\end{aligned}$$

In general, given $x, y \in H$ we can always adjust phases so that $\langle T(e^{i\theta}x), y\rangle \geq 0$, and applying the above inequality we end up with $|\langle Tx, y\rangle| \leq K$. Since this is so whenever $\|y\| = 1$ we get $\|Tx\| \leq K$, and since that is true whenever $\|x\| = 1$ we get $\|T\| \leq K$. ∎

Our next task is to relate the spectrum of T to the extreme values of $\langle Tx, x\rangle$. $T \in B(H)_{\mathsf{sa}}$ is called *positive* if $\langle Tx, x\rangle \geq 0$ for all $x \in H$. We write $T \geq 0$ if this is the case.

87 Lemma. *If $T \geq 0$ then $\sigma(T) \subseteq [0, \infty)$.*

Proof: Whenever $\lambda < 0$ then

$$\langle (T - \lambda I)x, x\rangle = \langle Tx, x\rangle - \lambda\langle x, x\rangle \geq -\lambda\|x\|^2$$

since $T \geq 0$. Thus

$$-\lambda\|x\|^2 \leq \langle (T - \lambda I)x, x\rangle \leq \|(T - \lambda I)x\|\,\|x\|$$

by the Cauchy–Schwarz inequality. Division by $\|x\|$ yields $-\lambda\|x\| \leq \|(T - \lambda I)x\|$, which shows that $T - \lambda I$ is invertible by Lemma 82. We have shown that $\lambda < 0$ implies $\lambda \notin \sigma(T)$. Since $\sigma(T) \subseteq \mathbb{R}$, the proof is complete. ∎

88 Lemma. *If $T \geq 0$ and $\inf_{\|x\|=1} \langle Tx, x\rangle = 0$ then T is not invertible.*

Proof: We use the Cauchy–Schwarz inequality on the sesquilinear form $B(x, y) = \langle Tx, y\rangle$ and its associated quadratic form Q to obtain

$$|\langle Tx, y\rangle|^2 \leq \langle Tx, x\rangle\langle Ty, y\rangle \leq \langle Tx, x\rangle\|T\|\,\|y\|^2.$$

Apply this with $y = Tx$ to get

$$\|Tx\|^4 \leq \langle Tx, x\rangle\|T\|^3\|x\|^2.$$

Since by assumption we can make $\langle Tx, x\rangle$ as small as we wish while $\|x\| = 1$, we can make $\|Tx\|$ as small as we wish with $\|x\| = 1$. Thus T is not invertible. ∎

89 Proposition. *If $T \in B(H)_{\mathsf{sa}}$, let*

$$m = \inf_{\|x\|=1} \langle Tx, x\rangle, \quad M = \sup_{\|x\|=1} \langle Tx, x\rangle.$$

Then $\sigma(T) \subseteq [m, M]$ and $m, M \in \sigma(T)$.

Proof: We find $T - mI \geq 0$, so $\sigma(T - mI) \subseteq [0, \infty)$. This implies $\sigma(T) \subseteq [m, \infty)$. Similarly, we find $MI - T \geq 0$, so $\sigma(MI - T) \subseteq [0, \infty)$. This implies $\sigma(T) \subseteq (-\infty, M]$. Together, we get $\sigma(T) \subseteq [m, M]$.

Next, it is an immediate result of the previous lemma that $T - mI$ is not invertible, so $m \in \sigma(T)$. For the same reason, $MI - T$ is not invertible, so $M \in \sigma(T)$. ∎

The *spectral radius* of an operator T is the maximum value of $|\lambda|$ where $\lambda \in \sigma(T)$. When T is self-adjoint, the spectral radius will be the maximum of $|m|$ and $|M|$, which is the supremum of $|\langle Tx, x\rangle|$ for $\|x\| = 1$, which is the norm of T.

90 Theorem. *The norm of a self-adjoint, bounded linear operator on a Hilbert space is the same as its spectral radius.* ∎

Functional calculus

Recall the spectral mapping theorem, Proposition 75. Apply this to a polynomial p with real coefficients and a self-adjoint operator T. Then $p(T)$ is also self-adjoint, so its norm is its spectral radius, which is the maximum of $|\mu|$ where $\mu \in \sigma(p(T)) = p(\sigma(T))$, and so

$$\|p(T)\| = \max\{|p(\lambda)| \colon \lambda \in \sigma(T)\}.$$

We write this more suggestively as

$$\|p(T)\| = \|p|_{\sigma(T)}\|_\infty.$$

In other words, $p \mapsto p(T)$ is an isometric mapping from those functions on $\sigma(T)$ which are (restrictions of) real polynomials. But the Weierstrass theorem assures us that the real polynomials are dense in $C(\sigma(T))$. Therefore, the mapping considered has a unique continuous extension to a map from $C(\sigma(T))$ (where we consider only *real* continuous functions on the spectrum) to $B(H)_{\mathsf{sa}}$.

We use the notation $f(T)$ for the value of this extended map when applied to a function f.

91 Lemma. *If $f, g \in C(\sigma(T))$ then $f(T)g(T) = (fg)(T)$.*

Proof: This is true for polynomials, by simple algebra: So $p(T)q(T) = (pq)(T)$ for all polynomials p and q. Now let $q \to g$ uniformly. Then $pq \to pg$ uniformly, so in the limit we get $p(T)g(T) = (pg)(T)$. Next, let $p \to f$ and conclude in the same way that $f(T)g(T) = (fg)(T)$. ∎

The spectral mapping theorem also extends from polynomials to continuous functions:

92 Proposition. *If $T \in B(H)_{\mathsf{sa}}$ and $f \in C(\sigma(T))$ then $\sigma(f(T)) = f(\sigma(T))$.*

Proof: First, if $\mu \notin f(\sigma(T))$ then $f(\lambda) - \mu \neq 0$ for $\lambda \in \sigma(T)$, so that $g(\lambda) = 1/\big(f(\lambda) - \mu\big)$ defines a continuous function g on the spectrum. Since $(f-\mu)g = 1$ we get $(f(T) - \mu I)g(T) = I$, so $f(T) - \mu I$ is invertible, i.e., $\mu \notin \sigma(f(T))$. We have proved that $\sigma(f(T)) \subseteq f(\sigma(T))$.

To prove the opposite inclusion, let $\mu \in f(\sigma(T))$. Say $\mu = f(\lambda)$ with $\lambda \in \sigma(T)$.

Let $\varepsilon > 0$. By continuity, there is a $\delta > 0$ so that $|f(t) - \mu| < \varepsilon$ whenever $|t - \lambda| < \delta$. Let g be the restriction to $\sigma(T)$ of the continuous function which is 1 at λ, 0 outside $[\lambda - \delta, \lambda + \delta]$, and is linear in each of the two halves of that interval. Then $\|g\|_\infty = 1$, while $\|(f-\mu)g\|_\infty < \varepsilon$. So $\|g(T)\| = 1$ and $\|(f(T) - \mu I)g(T)\| < \varepsilon$. We can find $x \in H$ with $\|g(T)x\| > \frac{1}{2}$. At the same time, $\|(f(T) - \mu I)g(T)x\| < \varepsilon$. Or put differently, with $y = g(T)x$, $\|y\| > \frac{1}{2}$ while $\|(f(T) - \mu I)y\| < \varepsilon$. This proves that $f(T) - \mu I$ is not invertible, so $\mu \in \sigma(f(T))$. ∎

The above proof also produces this interesting result:

93 Corollary. *An isolated point in the spectrum of a self-adjoint operator is an eigenvalue.*

Proof: Let $\lambda \in \sigma(T)$ be isolated, i.e., it λ is the only point in some interval $(\lambda-\delta, \lambda+\delta)$ belonging to the spectrum. Let g be the function constructed in the above proof, and pick x with $g(T)x \neq 0$. By construction, $tg(t) = \lambda g(t)$ for all $t \in \sigma(T)$. Thus $Tg(T) = \lambda g(T)$, and therefore $Tg(T)x = \lambda g(T)x$. This shows that $g(T)x$ is an eigenvector of T with eigenvalue λ. ∎

94 Lemma. *If $T \in B(H)_{\mathsf{sa}}$, $g \in C(\sigma(T))$ and $f \in C(g(\sigma(T)))$ then $f(g(T)) = (f \circ g)(T)$.*

Proof: Lemma 91 and induction on n yields $g^n(T) = g(T)^n$ for $n = 1, 2, \ldots$. This is also trivially true for $n = 0$. Thus, $p(g(T)) = (p \circ g)(T)$ for every polynomial p. Let $p \to f$ to get the final conclusion. ∎

All the operators $f(T)$ commute with T. Even better:

95 Proposition. *When $f \in C(\sigma(T))$ then $f(T)$ commutes with every bounded operator on H which commutes with T.*

Proof: Assume $S \in B(H)$ and $ST = TS$. Then $ST^2 = STT = TST = TTS = T^2S$, and in general it is an easy induction to show that $ST^n = T^nS$. Thus $Sp(T) = p(T)S$ for every polynomial p. This equality extends by continuity:

$$Sf(T) = \lim_{p \to f} Sp(T) = \lim_{p \to f} p(T)S = f(T)S$$

where p in the limit is a polynomial, and $p \to f$ means $\|p - f\|_\infty \to 0$. ∎

Square roots, absolute values etc. If $T \geq 0$ then $\sigma(T) \subseteq [0, \infty)$. So the square root is well defined and continuous on the spectrum. Applying this to T we obtain a positive *square root* of T. We write $T^{1/2}$ for this square root. There is only one positive square root of T, for if S is any positive square root of T we can apply Lemma 94 to get $S = (S^2)^{1/2} = T^{1/2}$.

Similarly we can define the *absolute value* $|T|$, and the *positive and negative parts* T^+ and T^- by applying the corresponding functions. (The positive part of a real number is $t^+ = t$ if $t \geq 0$, $t^+ = 0$ if $t < 0$. And the negative part is $t^- = 0$ if $t > 0$, $t^- = -t$ if $t \leq 0$.) Then $|T|$, T^+ and T^- are all positive, $|T| = T^+ + T^-$, $T = T^+ - T^-$, and $T^+T^- = 0$.

We can also use the functional calculus to prove the following converse to Lemma 87:

96 Lemma. *If $T \in B(H)_{\mathsf{sa}}$ and $\sigma(T) \subseteq [0, \infty)$ then $T \geq 0$.*

Proof: The given assumption is enough to show the existence of the square root $S = T^{1/2}$. Thus $T = S^2$, and so $\langle Tx, x\rangle = \langle S^2x, x\rangle = \langle Sx, Sx\rangle \geq 0$. ∎

The spectral theorem

We shall now see how the functional calculus can be used to arrive at the spectral theorem for bounded, self-adjoint operators. But first, some words about projections.

Projections. First of all, any mapping P of a space into itself is called *idempotent* if $P^2 = P$. In other words, for each x in the space, $P(P(x)) = P(x)$. An obvious rephrasing of this is to say that $P(y) = y$ for every y in the range of the map.

Next, if our idempotent P is actually a linear map on a vector space X, a simple calculation shows that $I - P$ is also idempotent. For every $x \in X$, the identity $x = Px + (I-P)x$ shows that the two subspaces $\operatorname{im} P = \ker(I-P)$ and and $\operatorname{im}(I-P) = \ker P$ span X. Moreover, the intersection of the two subspaces is clearly trivial, so X is a *direct sum* of the two subspaces. Conversely, if X is the direct sum of subspaces U and V, we can create an idempotent linear map P by insisting that $Pu = u$ for $u \in U$, and $Pv = 0$ for $v \in V$. Thus $\operatorname{im} P = U$ and $\ker P = V$.

Assuming next that X is a Banach space, it is clear that if P is a *bounded* idempotent linear map, then the two subspaces $\operatorname{im} P$ and $\ker P$ are closed. As an exercise, use the open mapping theorem to show the converse: If $X = U \oplus V$ with U and V closed, then the corresponding idempotent P is bounded. (Hint: Use the norm $\|(u,v)\| = \|u\| + \|v\|$ on $U \times V$. Then $(u,v) \mapsto u + v$ is a bounded bijection from $U \times V$ to X. Consider its inverse.)

Finally, consider an idempotent bounded linear map P on a Hilbert space H: We call P a *projection* if the subspaces $\operatorname{im} P$ and $\ker P = \operatorname{im}(I-P)$ are orthogonal.[4] Clearly, this is equivalent to having $\langle Px, (I-P)y\rangle = 0$ for all $x, y \in H$. But from this we get $\langle (P - P^*P)x, y\rangle = 0$, so that $(P - P^*P)x = 0$ since this holds for all y, and therefore $P - P^*P = 0$ since x was arbitrary. Therefore $P = P^*P$, and since P^*P is self-adjoint, then so is P. This argument works equally well backwards, so we have proved

97 Lemma. *A bounded idempotent on a Hilbert space is a projection if and only if it is self-adjoint.* ∎

Obviously, there is a one-to-one correspondence between closed subspaces of H and projections: Given a closed subspace U of H, let P be the idempotent whose image is U and whose kernel is $U^\perp$. We simply call P the projection on U. Then of course $I - P$ is the projection on $U^\perp$. It is common to write $P^\perp = I - P$.

Next, we consider what it means for an operator and a projection to commute: A subspace U is called *invariant* for a map A if A maps U into itself.

98 Lemma. *Let P be a projection on H and $A \in B(H)$. Then $PA = AP$ if and only if both $\operatorname{im} P$ and $\ker P$ are invariant subspaces for A.*

Proof: Assume $PA = AP$. Then for any $x \in \operatorname{im} P$, $PAx = APx = Ax$, so $Ax \in \operatorname{im} P$. Thus $\operatorname{im} P$ is invariant. Since A also commutes with $I - P$, this shows that $\ker P = \operatorname{im}(I-P)$ is invariant as well.

[4] Sometimes, people call any idempotent a projection, in which case what we call a projection must be termed an *orthogonal projection*. But since we are only interested in the latter kind, we use the shorter term.

Conversely, if $\operatorname{im} P$ is invariant then for all x we find $APx \in \operatorname{im} P$, so that $PAPx = APx$. Thus $AP = PAP$. Similarly, if $\operatorname{im}(I-P)$ is also invariant then $A(I-P) = (I-P)A(I-P)$, which simplifies into $PA = PAP$. Thus $AP = PAP = PA$. ∎

A projection splits the Hilbert space into two orthogonal subspaces. We have just seen that an operator commutes with this projection if and only if the operator also splits into two parts, each mapping one of the two orthogonal subspaces into itself.

Let us apply this to another projection Q: If the projections P and Q commute, then certainly PQ is a projection: For $(PQ)^2 = PQPQ = PPQQ = PQ$, and $(PQ)^* = Q^*P^* = QP = PQ$. Similarly, $PQ^\perp$, $P^\perp Q$ and $P^\perp Q^\perp$ are all projections, and we find $I = (P+P^\perp)(Q+Q^\perp) = PQ + PQ^\perp + P^\perp Q + P^\perp Q^\perp$. These four projections are projections onto four mutually orthogonal spaces whose sum is the whole space.

We consider the ordering of projections next:

99 Lemma. *For any projection P on a Hilbert space, $0 \le P \le I$. Also, $\sigma(P) = \{0,1\}$ except for the two special cases $\sigma(0) = \{0\}$ and $\sigma(I) = \{1\}$.*

Further, let P and Q be projections. Then $P \le Q$ if and only if $\operatorname{im} P \subseteq \operatorname{im} Q$. In this case, P and Q commute, and $PQ = P$.

Proof: First, for any projection P, $\langle Px, x\rangle = \langle P^2x, x\rangle = \langle Px, Px\rangle = \|Px\|^2$, so $P \ge 0$. But $P^\perp = I - P$ is also a projection, so $I - P \ge 0$ as well. Thus $P \le I$.

Since $P^2 - P = 0$ the spectral mapping theorem shows that the function $f(t) = t^2 - t$ vanishes on $\sigma(P)$. Thus $\sigma(P) \subseteq \{0,1\}$. If $\sigma(P) = \{0\}$ then $P = 0$ (this is true for any self-adjoint operator). If $\sigma(P) = \{1\}$ then $\sigma(I-P) = \{0\}$, so $I - P = 0$.

If $P \le Q$ and $Qx = 0$ then $\|Px\|^2 = \langle Px, x\rangle \le \langle Qx, x\rangle = 0$, so $Px = 0$. Thus $\ker Q \subseteq \ker P$, and taking orthogonal complements, we find $\operatorname{im} P \subseteq \operatorname{im} Q$.

Conversely, if $\operatorname{im} P \subseteq \operatorname{im} Q$ then $QP = P$ (consider what happens to QPx). Taking adjoints, we also get $PQ = P$. Finally, we get $P = PQ = QPQ$, so that $\langle Px, x\rangle = \langle PQx, Qx\rangle \le \|Qx\|^2 = \langle Qx, x\rangle$, and thus $P \le Q$. ∎

The spectral family. Given an operator $T \in B(H)_{\mathsf{sa}}$, we are now ready to define the *spectral family* of this operator: Simply let E_λ be the projection onto $\ker(A - \lambda I)^+$, for any $\lambda \in \mathbb{R}$.

Before listing the elementary properties of the spectral family, we need a lemma or two.

100 Lemma. *Assume $A, B \in B(H)_{\mathsf{sa}}$ and $0 \le A \le B$. Then $\ker B \subseteq \ker A$.*

Proof: If $Bx = 0$ then $\langle Ax, x\rangle \le \langle Bx, x\rangle \le 0$ since $A \le B$. Since $A \ge 0$, we not only conclude $\langle Ax, x\rangle = 0$, but also from the Cauchy–Schwarz inequality applied to $L(x,y) = \langle Ax, y\rangle$ we get $|\langle Ax, y\rangle|^2 \le \langle Ax, x\rangle\langle Ay, y\rangle = 0$ for all y, so that $Ax = 0$. ∎

101 Lemma. *If $A \ge 0$, $B \ge 0$ and $AB = BA$ then $AB \ge 0$.*

Proof: A commutes with $B^{1/2}$ as well, so $AB = B^{1/2}AB^{1/2} \geq 0$, since $\langle ABx, x\rangle = \langle B^{1/2}AB^{1/2}x, x\rangle = \langle AB^{1/2}x, B^{1/2}x\rangle \geq 0$. ∎

102 Proposition. (Elementary properties of the spectral family)

(*i*) *If $\lambda < \mu$ then $E_\lambda \leq E_\mu$.*
(*ii*) *If λ lies to the left of $\sigma(T)$ then $E_\lambda = 0$.*
(*iii*) *If λ lies to the right of $\sigma(T)$ then $E_\lambda = I$.*
(*iv*) *If $\lambda < \mu$ and $[\lambda, \mu] \cap \sigma(T) = \varnothing$ then $E_\lambda = E_\mu$.*
(*v*) *Each E_λ commutes with every operator which commutes with T.*
(*vi*) *$E_\lambda T \leq \lambda E_\lambda$, and $E_\lambda^\perp T \geq \lambda E_\lambda^\perp$.*

Proof: When $\lambda < \mu$ then $(t-\lambda)^+ \geq (t-\mu)^+$ for all t. Thus $(T-\lambda I)^+ \geq (T-\mu I)^+$. Lemma 100 then completes the proof of (*i*).

If λ lies to the left of $\sigma(T)$ then $(t-\lambda)^+ > 0$ for all $t \in \sigma(T)$, so $(T-\lambda I)^+$ is invertible, and (*ii*) follows.

If λ lies to the right of $\sigma(T)$ then $(T-\lambda I)^+ = 0$, and (*iii*) follows.

Next, under the conditions in (*iv*) we can find a constant $c > 0$ so that $(t-\mu)^+ \geq c(t-\lambda)^+$ for all $t \in \sigma(T)$, so that $(T-\mu I)^+ \geq c(T-\lambda I)^+$. By Lemma 100 $E_\mu \leq E_\lambda$ follows, and so $E_\mu = E_\lambda$ according to (*i*).

To prove (*v*), let $S \in B(H)$ commute with T. Then S commutes with $(A-\lambda I)^+$, and therefore $\ker(A-\lambda I)^+$ and $\operatorname{im}(A-\lambda I)^+$ are both invariant for S. The orthogonal complement of $\ker(A-\lambda I)^+$ is the closure of $\operatorname{im}(A-\lambda I)^+$, and therefore also invariant, by Lemma 98.

To prove (*vi*), note that $T-\lambda I = (T-\lambda I)^+ - (T-\lambda I)^-$ and multiply by E_λ. This results in $E_\lambda(T-\lambda I) = -E_\lambda(T-\lambda I)^- \leq 0$ by Lemma 101 and (*v*). This shows the first inequality.

The second inequality in (*vi*) is proved the same way: We only need to know that $E_\lambda^\perp(T-\lambda I)^- = 0$. As noted above, $E_\lambda^\perp$ is the projection onto the closure of $\operatorname{im}(T-\lambda)^+$. But $(T-\lambda)^-(T-\lambda)^+ = 0$, so $(T-\lambda)^-$ vanishes on $\operatorname{im}(T-\lambda)^+$, and therefore $(T-\lambda)^- E_\lambda^\perp = 0$. Since these operators commute by (*v*), the proof is complete. ∎

We combine Lemma 102 (*v*) and (*vi*) with Lemma 101 to obtain the inequalities

$$\lambda E_\mu E_\lambda^\perp \leq E_\mu E_\lambda^\perp T \leq \mu E_\mu E_\lambda^\perp, \qquad \lambda < \mu,$$

which somehow corresponds to the obvious statement $\lambda\chi_{[\lambda,\mu]} \leq t\chi_{[\lambda,\mu]} \leq \mu\chi_{[\lambda,\mu]}$.

We also note that when $\lambda < \mu$ then $E_\mu E_\lambda^\perp = E_\mu(I - E_\lambda) = E_\mu - E_\mu E_\lambda = E_\mu - E_\lambda$, so we can write the above as

$$\lambda(E_\mu - E_\lambda) \leq (E_\mu - E_\lambda)T \leq \mu(E_\mu - E_\lambda), \qquad \lambda < \mu.$$

To make use of this, consider a partition $\lambda_0 < \lambda_1 < \cdots < \lambda_n$ with λ_0 to the left of $\sigma(T)$ and λ_n to the right of $\sigma(T)$, and add:

$$\sum_{i=1}^n \lambda_{i-1}(E_{\lambda_i} - E_{\lambda_{i-1}}) \leq \sum_{i=1}^n (E_{\lambda_i} - E_{\lambda_{i-1}})T \leq \sum_{i=1}^n \lambda_i(E_{\lambda_i} - E_{\lambda_{i-1}})$$

We notice that the sum in the middle telescopes, so we are in fact left with

$$\sum_{i=1}^{n} \lambda_{i-1}(E_{\lambda_i} - E_{\lambda_{i-1}}) \leq T \leq \sum_{i=1}^{n} \lambda_i(E_{\lambda_i} - E_{\lambda_{i-1}}). \tag{6.5}$$

Finally, note that the difference between the upper and lower estimates for T is

$$\sum_{i=1}^{n} \lambda_i(E_{\lambda_i} - E_{\lambda_{i-1}}) - \sum_{i=1}^{n} \lambda_{i-1}(E_{\lambda_i} - E_{\lambda_{i-1}})$$
$$= \sum_{i=1}^{n} (\lambda_i - \lambda_{i-1})(E_{\lambda_i} - E_{\lambda_{i-1}}) \leq \max_i(\lambda_i - \lambda_{i-1})I$$

Thus, as the partition becomes finer and $\max_i(\lambda_i - \lambda_{i-1}) \to 0$, the two sums in (6.5) converge in norm towards T.

These sums look very much like Riemann sums, and in fact we use them to define the *spectral integral*:

$$\int_{\mathbb{R}} \lambda \, dE_\lambda = \lim \sum_{i=1}^{n} \lambda_i^*(E_{\lambda_i} - E_{\lambda_{i-1}})$$

where λ_i^* can be any value in the interval $[\lambda_{i-1}, \lambda_i]$ and the limit is taken as $\max_i(\lambda_i - \lambda_{i-1}) \to 0$. We have arrived at

103 Theorem. (Spectral theorem) *Any bounded, self-adjoint operator T on a Hilbert space can be represented as an integral*

$$T = \int_{\mathbb{R}} \lambda \, dE_\lambda$$

where (E_λ) is the associated spectral family of T.

We can easily recover the functional calculus from the spectral theorem: In fact we find

$$f(T) = \int_{\mathbb{R}} f(\lambda) \, dE_\lambda.$$

Our next task is to make this integral meaningful.

Spectral families and integrals

We now study spectral families in general, without any implied connection to a given operator. A *spectral family* is a family $(E_\lambda)_{\lambda \in \mathbb{R}}$ of projections on a Hilbert space H, so that $E_\lambda \leq E_\mu$ whenever $\lambda \leq \mu$, and so that $\|E_\lambda x\| \to 0$ as $\lambda \to -\infty$ and $\|E_\lambda x - x\| \to 0$ as $\lambda \to \infty$, for all $x \in H$.[5]

[5] One commonly adds a requirement of one-sided strong operator continuity, but we shall dispense with this technicality.

Though we shall not give meaning to the symbol dE (appearing as dE_λ in the spectral integral), we shall still make the following definition: We say that dE *vanishes* on an open interval I if E_λ is the same projection for all $\lambda \in I$. Further, we say that dE is *supported* on a closed subset $F \subseteq \mathbb{R}$ if it vanishes on every open interval which does not meet[6] F. Finally, the *support* of dE, written $\operatorname{supp} dE$, is the complement of the union of all open intervals on which dE vanishes. Our next lemma shows that the support deserves its name.

104 Lemma. *The support of dE is closed, and* $\operatorname{supp} dE$ *supports dE. It is, in fact, the smallest closed set that supports dE.*

Proof: Any union of open intervals is an open set, so clearly $\operatorname{supp} dE$ (which is the complement of such a union) is closed.

Now let I be an open interval that does not meet $\operatorname{supp} dE$. We wish to show that dE vanishes on I; this will prove that $\operatorname{supp} dE$ supports dE. Let $s < t$ be two points of I. We can cover $[s,t]$ by open intervals on which dE vanishes, and by compactness we can find a finite number of them covering $[s,t]$. Thus we can find[7] points $t_0 = s < t_1 < t_2 < \cdots < t_n = t$ so that $[t_{i-1}, t_i]$ is contained in one of these intervals for $i = 1, 2, \ldots, n$. Thus $E_{t_{i-1}} = E_{t_i}$ for $i = 1, 2, \ldots, n$, and so $E_s = E_t$.

If $F \subseteq \mathbb{R}$ supports dE and $x \notin F$ then there is an open interval $I \ni x$ which does not meet F. Then by assumption dE vanishes on I, so I does not meet $\operatorname{supp} dE$ either, by the definition of that set. In particular $x \notin \operatorname{supp} dE$. It follows that $\operatorname{supp} dE \subseteq F$, so the final statement is proved. ∎

Since $\operatorname{supp} dE$ is closed, this set is compact if and only if it is bounded. From now on, we shall only investigate spectral families for which this is true.

We shall consider partitions of the form $\lambda_0 < \lambda_1 < \cdots < \lambda_n$ with λ_0 to the left of $\operatorname{supp} dE$ and λ_n to the right of $\operatorname{supp} dE$.

For any such partition, we pick some $\lambda_i^* \in [\lambda_{i-1}, \lambda_i]$ and write up the Riemann sum

$$\sum_{i=1}^{n} f(\lambda_i^*)(E_{\lambda_i} - E_{\lambda_{i-1}}).$$

We then let the mesh size $\max|\lambda_i - \lambda_{i-1}|$ go to zero. If the limit exists, no matter how this is done, that limit is the integral $\int_{\mathbb{R}} f(\lambda)\, dE_\lambda$.

105 Proposition. *If dE has bounded support and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then the spectral integral $\int_{\mathbb{R}} f(\lambda)\, dE_\lambda$ exists.*

Proof sketch: Given $\varepsilon > 0$, use the uniform continuity of f on any bounded interval: There exists $\delta > 0$ so that $|f(s) - f(t)| < \varepsilon$ whenever $|s - t| < \delta$ and $\min(\operatorname{supp} dE) - \delta < s, t < \max(\operatorname{supp} dE)$.

[6] Two sets *meet* if they have nonempty intersection

[7] Not hard to believe, but it takes a bit of work to prove it rigorously.

Pick any partition with mesh finer than δ. Given choices $\lambda_i^*, \lambda_i^{**} \in [\lambda_{i-1}, \lambda_i]$, the difference between the corresponing Riemann sums is

$$\sum_{i=1}^{n} \big(f(\lambda_i^*) - f(\lambda_i^{**})\big)(E_{\lambda_i} - E_{\lambda_{i-1}})$$

where the coefficient of each $E_{\lambda_i} - E_{\lambda_{i-1}}$ has absolute value less than ε. Since these are mutually orthogonal projections, it is not hard to see that the whole expression has norm less than ε.

Even more interestingly, we are permitted to choose λ_i^{**} *outside* $[\lambda_{i-1}, \lambda_i]$, so long as $|\lambda_i^{**} - \lambda_i^*| < \delta$ still holds.

Apply this special case to *two* partitions, one of which is a refinement of the other. This means that each point in one of the partitions is also a point of the other (the refinement of the first one). If the first partition has mesh size less than δ, so has the refinement, of course. And any Riemann sum associated with the first partition can be rewritten as an "almost Riemann sum" associated with the refinement, simply by replacing each difference $E_{\lambda_i} - E_{\lambda_{i-1}}$ from the coarser partition by the sum of projections corresponding to those subintervals in the refinement whose union is $[\lambda_{i-1}, \lambda_i]$.

The estimate obtained above still works, which shows that the two Riemann sums are closer together that ε.

Now if we have any two partitions with mesh size smaller than δ, we can produce a common refinement of the two, simply by taking the union of their points. Then a quick use of the triangle inequality will show that any two Riemann sums associated with these two partitions are closer than 2ε.

Thus the family of Riemann sums, indexed by partitions, have a sort of Cauchy property, which will guarantee the existence of the limit, since the space $B(H)_{\mathsf{sa}}$ is complete. (Take any sequence of partitions whose mesh sizes go to zero. This leads to a Cauchy sequence of operators, which has a limit. Any two such sequence get arbitrary close when you go far out in them, so the limits are the same.) ∎

106 Lemma. *The map $f \mapsto \int_{\mathbb{R}} f(\lambda)\, dE_\lambda$ is an algebra homomorphism.*

In other words, it is linear and multiplicative.

Proof sketch: Linearity is fairly obvious. For multiplicativity, what we need is the fact that if $P_1, \dots, P_n$ are mutually orthogonal projections then

$$\Big(\sum_{i=1}^{n} \mu_i P_i\Big)\Big(\sum_{i=1}^{n} \nu_i P_i\Big) = \sum_{i=1}^{n} \mu_i \nu_i P_i,$$

so Riemann sums have the multiplicative property, and going to the limit, so does the integral. ∎

107 Lemma. *If f is a continuous function and $f(\lambda) \geq 0$ for all $\lambda \in \operatorname{supp} dE$ then $\int_{\mathbb{R}} f(\lambda)\, dE_\lambda \geq 0$.*

Proof: For each index i in a Riemann sum, if $E_{\lambda_{i-1}} \neq E_{\lambda_i}$ then we can pick $\lambda_i^* \in [\lambda_{i-1}, \lambda_i]$ so that in fact $\lambda_i^* \in \operatorname{supp} dE$. But then $f(\lambda_i^*) \geq 0$. Adding terms, the whole Riemann sum is nonnegative, and hence so is the limit. ∎

As a corollary to this lemma, if $f = 0$ on $\operatorname{supp} dE$ then the spectral integral is zero. Thus we can in fact define the integral for any real continuous function on $\operatorname{supp} dE$, since all we have to do is to extend f in a continuous fashion to all of $\mathbb{R}$, and what we have just proved shows that the integral does not depend on how we perform the extension.

We now have a new version of the spectral mapping theorem.

108 Proposition. *If $f \in C(\operatorname{supp} dE, \mathbb{R})$ then*

$$\sigma\Big(\int_{\mathbb{R}} f(\lambda)\, dE_\lambda\Big) = f(\operatorname{supp} dE).$$

Proof: If $\mu \notin f(\operatorname{supp} dE)$ then $f - \mu$ has an inverse $1/(f-\mu)$ which is continuous on $\operatorname{supp} dE$. Then

$$\int_{\mathbb{R}} f(\lambda)\, dE_\lambda \cdot -\mu I = \int_{\mathbb{R}} \big(f(\lambda) - \mu\big)\, dE_\lambda$$

has the inverse $\int_{\mathbb{R}} 1/\big(f(\lambda)-\mu\big)\, dE_\lambda$, so $\mu \notin \sigma\Big(\int_{\mathbb{R}} f(\lambda)\, dE_\lambda\Big)$.

Conversely, if $\mu \in f(\operatorname{supp} dE)$ we mimic the proof of Proposition 92 and find, for given $\varepsilon > 0$, a function g so that $\|\int_{\mathbb{R}} \big(f(\lambda)-\mu\big) g(\lambda)\, dE_\lambda\| < \varepsilon$ while $\|\int_{\mathbb{R}} g(\lambda)\, dE_\lambda\| = 1$. Only the latter equality pose a problem. The function g needs to be chosen with $0 \leq g \leq 1$ and so that $g(t) = 1$ for t in a neighbourhood of some $\lambda \in \operatorname{supp} dE$. This means there are $s < \lambda < t$ with $g = 1$ in $[s,t]$. Picking always partitions containing s and t, it is not hard to show that any corresponding Riemann sum is $\geq E_t - E_s$. Thus the same is true of the limit: $\int_{\mathbb{R}} g(\lambda)\, dE_\lambda \geq E_t - E_s$ as well. Since the righthand side is a *nonzero* projection, the integral must have norm at least 1. ∎

We now return to the case of a given self-adjoint operator T and its spectral family (E_λ). One half of the following result is already contained in Lemma 102 (*iv*).

109 Corollary. $\operatorname{supp} dE = \sigma(T)$.

Proof: Apply the above proposition to the identity function $f(\lambda) = \lambda$. ∎

110 Theorem. (Spectral theorem, part 2) *For any bounded self-adjoint operator T and its associated spectral family (E_λ) we have*

$$f(T) = \int_{\mathbb{R}} f(\lambda)\, dE_\lambda, \qquad f \in C(\sigma(T), \mathbb{R}).$$

Proof: From part 1 of the spectral theorem this is true when f is the identity function. Both sides depend linearly and multiplicatively on f; hence they are equal for any polynomial f. They are also continuous functions of f, and polynomials are dense by the Weierstrass theorem. Thus the general result follows by continuity. ∎

Chapter 7

Compact operators

Compact operators

Let X and Y be Banach spaces. Write X_1 for the closed unit ball $\{x \in X : \|x\| \leq 1\}$ of X. We call a linear operator $T : X \to Y$ *compact* if the image TX_1 is precompact in Y (in the norm topology). A *precompact* set is one whose closure is compact.

It is useful to know the following equivalent formulation of compactness in metric spaces. First, a metric space is called *totally bounded* if it contains a finite ε-net for each $\varepsilon > 0$. A subset $S \subseteq X$ is called an ε-net if each $x \in X$ is closer than ε to some member of S.

111 Proposition. *A metric space (X, d) is compact if, and only if, it is complete and totally bounded.*

Proof sketch: First, assume X is compact. To show it is complete, pick any Cauchy sequence in X. By compactness, some subsequence converges. But since the original sequence is Cauchy, the original sequence must converge to the limit of the subsequence. To show total boundedness, cover X by a finite number of open balls $B_\varepsilon(x)$.

Second, assume X is complete and totally bounded, and let $\mathscr{F}$ be a set of closed subsets of X, with the finite intersection property. Pick numbers $\varepsilon_1, \varepsilon_2, \ldots$ with $\varepsilon_k \to 0$. Let S_1 be an ε_1-net, so that the balls $B_{\varepsilon_1}(s)$, where $s \in S_1$, cover X. In particular, there must be at least one $s_1 \in S_1$ so that $\{F \cap B_{\varepsilon_1}(s_1) : F \in \mathscr{F}\}$ has the finite intersection property.

Next, let S_2 be an ε_2-net, and pick $s_2 \in S_2$ so that $\{F \cap B_{\varepsilon_1}(s_1) \cap B_{\varepsilon_2}(s_2) : F \in \mathscr{F}\}$ has the finite intersection property. Continuing in this way, we end up with a sequence (s_k) so that $\{F \cap B_{\varepsilon_1}(s_1) \cap \cdots \cap B_{\varepsilon_n}(s_n) : F \in \mathscr{F}, n = 1, 2, \ldots\}$ has the finite intersection property.

In particular, $B_{\varepsilon_1}(s_1) \cap \cdots \cap B_{\varepsilon_n}(s_n) \neq \varnothing$, so we can pick $x_n \in B_{\varepsilon_1}(s_1) \cap \cdots \cap B_{\varepsilon_n}(s_n)$. By construction (x_n) is a Cauchy sequence, so it is convergent. Let x be its limit.

We claim $x \in \bigcap \mathscr{F}$. Indeed let $F \in \mathscr{F}$. But since $F \cap B_{\varepsilon_k}(s_k) \neq \varnothing$, we have $\operatorname{dist}(x_k, F) < 2\varepsilon_k$ for all k. But then $\operatorname{dist}(x, F) < d(x, x_k) + 2\varepsilon_k \to 0$ as $k \to \infty$. Since F is closed, $x \in F$. ∎

112 Corollary. *A subset of a complete metric space (in particular, a Banach space) is precompact if and only if it is totally bounded.* ∎

An easy way to remember our next result is this: *Precompact sets are almost finite-dimensional.*

113 Lemma. *A subset A of a Banach space X is precompact if and only if it is bounded and, for each $\varepsilon > 0$, there is a finite-dimensional subspace $N \subseteq X$ so that* $\operatorname{dist}(x, N) < \varepsilon$ *for all $x \in A$.*

Proof: If A is totally bounded then A is clearly bounded, and a finite ε-net in A spans a subspace N with the given property.

Conversely, assume A is bounded, let $\varepsilon > 0$, and assume $N \subseteq X$ is a subspace with the stated properties. Let $B = \{y \in N\colon \operatorname{dist}(y, A) < \varepsilon\}$. Since A is bounded, then so is B. And since N is finite-dimensional, B is totally bounded. Let $\{y_1, \dots, y_n\}$ be an ε-net in B. For $k = 1, \dots, n$ let $x_k \in A$ with $\|x_k - y_k\| < \varepsilon$.

If $x \in A$ then since $\operatorname{dist}(x, B) < \varepsilon$ there is some $y \in B$ with $\|x - y\| < \varepsilon$. Next, there is some k with $\|y - y_k\| < \varepsilon$. Finally, $\|x - x_k\| < \|x - y\| + \|y - y_k\| + \|y_k - x_k\| < 3\varepsilon$, so that $\{x_1, \dots, x_n\}$ is a 3ε-net in A. If we can do this for all $\varepsilon > 0$, then this shows that A is totally bounded. ∎

114 Corollary. *The set of compact operators on a Banach space X is closed in $B(X)$.*

Proof: Let T belong to the closure of the set of compact operators. If $\varepsilon > 0$, there is some compact operator S with $\|T - S\| < \varepsilon$. And there is a finite-dimensional space N so that $\operatorname{dist}(Sx, N) < \varepsilon$ for all $x \in X_1$. Since also $\|Tx - Sx\| < \varepsilon$, then $\operatorname{dist}(Tx, N) < 2\varepsilon$ for all $x \in X_1$. This proves the compactness of T. ∎

115 Proposition. *The set of compact operators on a Hilbert space H is precisely the closure of the set of bounded finite rank operators on H.*

Proof: By the previous Corollary, and the obvious fact that all bounded finite rank operators are compact, it only remains to prove that the finite rank operators are dense in the set of compact operators. So let $T \in B(H)$ be compact, and let $\varepsilon > 0$. Let N be a finite-dimensional subspace so that $\operatorname{dist}(Tx, N) < \varepsilon$ for each $x \in X_1$. Let E be the orthogonal projection onto N. Then ET has finite rank, and if $x \in X_1$, $\|Tx - ETx\| = \operatorname{dist}(Tx, N) < \varepsilon$. Thus $\|T - ET\| < \varepsilon$. ∎

To see that the unit ball of an infinite dimensional Banach space does not satisfy the condition of Lemma 113, we need the following lemma, whose proof is found in Kreyszig p. 78:

116 Lemma. (F. Riesz) *Let Y be a closed, proper subspace of a normed space X. Then, for each $\varepsilon > 0$, there is some vector $x \in X$ with $\|x\| = 1$ and* $\operatorname{dist}(x, Y) > 1 - \varepsilon$. ∎

Recall that we are only dealing with the norm topology presently, so the next result does not contradict the Banach–Alaoglu theorem in any way:

117 Proposition. *No infinite-dimensional Banach space has a compact unit ball.*

Proof: If N is any finite dimensional subspace of the given space X, apply Lemma 116 with $Y = N$. It follows from Lemma 113 that X_1 is not (pre)compact. ∎

118 Corollary. *Any eigenspace, for a nonzero eigenvalue, of a compact operator is finite-dimensional.*

Proof: Let T be a compact operator with an eigenvalue $\lambda \neq 0$. Let $Y = \ker(T - \lambda I)$ be the corresponding eigenspace. Then $TX_1 \supseteq TY_1 = \lambda Y_1$. Since TX_1 is precompact, then so is λY_1, and therefore so is Y_1 (just multiply by λ^{-1}). Thus Y is finite-dimensional. ∎

119 Corollary. *No compact operator on an infinite-dimensional Banach space can be invertible.*

Proof: Assume $T \in B(X)$ is compact and invertible. Then $\overline{TX_1}$ is compact, and hence so is $T^{-1}\overline{TX_1} \supseteq X_1$. This is impossible. ∎

It is time to apply what we have learned to the spectral theory of compact operators. Let H be a Hilbert space and let $T \in B(H)_{\mathsf{sa}}$ be compact. Let (E_λ) be the associated spectral family.

Recall that $TE_\lambda^\perp \geq \lambda E_\lambda^\perp$. When $\lambda > 0$, this implies that the restriction of T to the image of $E_\lambda^\perp$ is an invertible map on that image. Since this restriction, like T itself, is compact, Corollary 119 implies that $E_\lambda^\perp$ has finite rank. This projection is also a decreasing function of λ for $\lambda > 0$, which means it must have at most a finite number of discontinuities in any interval $[\varepsilon, \infty]$ where $\varepsilon > 0$. Each discontinuity corresponds to an eigenspace of T.

We can apply the same argument to E_λ for $\lambda < 0$, and so we finally arrive at

120 Theorem. *Any compact, self-adjoint operator on a Hilbert space H can be written*

$$T = \sum_{k=1}^{\infty} \lambda_k E_k$$

where $|\lambda_k| \to 0$, and the E_k are mutually orthogonal projections of finite rank. ∎

Hilbert–Schmidt operators

In this section, H will be a *separable*, infinite-dimensional Hilbert space. In particular, H can be equipped with an orthonormal basis (e_n). For any bounded operator $T \in B(H)$, we define the *Hilbert–Schmidt norm* of T to be

$$\|T\|_{\mathsf{HS}} = \Big(\sum_{k=1}^{\infty} \|Te_k\|^2\Big)^{1/2}.$$

Furthermore, we call T a *Hilbert–Schmidt operator* if $\|T\|_{\mathsf{HS}} < \infty$.

Though these definitions seem to depend on the chosen basis, in fact they do not. To see this, first pick any $x \in H$ and use $x = \sum_k \langle x, e_k\rangle e_k$ to compute:

$$\|Tx\|^2 = \langle Tx, Tx\rangle = \sum_k \langle x, e_k\rangle\langle Te_k, Tx\rangle \tag{7.1}$$

We now let (e'_l) be a different orthonormal basis, put $x = e'_l$, and sum over l to get

$$\sum_l \|Te'_l\|^2 = \sum_l \sum_k \langle e'_l, e_k\rangle\langle Te_k, Te'_l\rangle = \sum_k \sum_l \langle e'_l, e_k\rangle\langle Te_k, Te'_l\rangle = \sum_k \langle Te_k, Te_k\rangle = \sum_k \|Te_k\|^2$$

where we have used $e_k = \sum_l \langle e_k, e'_l\rangle e'_l$ on the right.

From (7.1) we also get $\|Tx\|^2 \le \sum_k |\langle x, e_k\rangle| \|Te_k\| \|Tx\|$, and therefore

$$\|Tx\| \le \sum_k |\langle x, e_k\rangle| \|Te_k\| \le \Big(\sum_k |\langle x, e_k\rangle|^2\Big)^{1/2} \Big(\sum_k \|Te_k\|^2\Big)^{1/2} = \|x\| \|T\|_{\mathsf{HS}}$$

(with a little help from the Cauchy–Shwarz inequality for sequences). Thus we get the estimate

$$\|T\| \le \|T\|_{\mathsf{HS}}.$$

It is now easy to prove the following result:

121 Proposition. *Any Hilbert–Schmidt operator is compact.*

Proof: Let E_n be the projection onto $\operatorname{lin}\{e_1, \dots, e_n\}$: $E_n x = \sum_{k=1}^n \langle x, e_k\rangle e_k$. An easy calculation yields

$$\|T - TE_n\|_{\mathsf{HS}}^2 = \sum_{k=n+1}^{\infty} \|Te_k\|^2 \to 0 \qquad (n \to \infty),$$

In other words $TE_n \to T$ in Hilbert–Schmidt norm, and therefore also in operator norm, as $n \to \infty$. By Proposition 115, this finishes the proof. ∎

As an example of a Hilbert–Schmidt operator, consider the Hilbert space $L^2[a,b]$ over a real interval, with the Lebesgue measure. Consider an integral kernel $k \in L^2([a,b]\times[a,b])$, and define K on $L^2[a,b]$ by

$$(Kf)(s) = \int_a^b k(s,t) f(t)\, dt.$$

A quick appeal to the Cauchy–Shwarz inequality gives $|(Kf)(s)|^2 \le \int_a^b |k(s,t)|^2\, dt \cdot \|f\|_2^2$, which we integrate to get $\|Kf\|_2 \le \|k\|_2 \cdot \|f\|_2$. Thus K is in fact bounded.

Next, let (e_n) be an orthonormal basis for $L^2[a,b]$. Write $e_m \otimes \bar{e}_n(s,t) = e_m(s)\overline{e_n(t)}$. Then the set of all the functions $e_m \otimes \bar{e}_n$ form an orthonormal basis for $L^2([a,b]\times[a,b])$, so we can write

$$k = \sum_{m,n} \alpha_{m,n} e_m \otimes \bar{e}_n, \quad \|k\|_2^2 = \sum_{m,n} |\alpha_{m,n}|^2.$$

Then we find $Ke_n = \sum_m \alpha_{m,n} e_m$, and therefore $\|Ke_n\|^2 = \sum_m |\alpha_{m,n}|^2$, and finally $\sum_n \|Ke_n\|^2 = \sum_{m,n} |\alpha_{m,n}|^2 = \|k\|_2^2$. In other words, $\|K\|_{\mathsf{HS}} = \|k\|_2$. In particular, K is a Hilbert–Schmidt operator.

Sturm–Liouville theory

Sturm–Liouville theory is the study of the differential operator

$$Lf = (pf')' + qf \qquad \text{on } [a,b]$$

together with boundary conditions of the form

$$\alpha f(a) + \alpha' f'(a) = 0, \quad \beta f(b) + \beta' f'(b) = 0.$$

These problems arise while solving linear partial differential equations by separation of variables.

Here p and q are given continuous functions. We shall assume that $p > 0$ on $[a,b]$.

In order to show that the solutions discovered this way span all solutions, it is necessary to show that L possesses a complete system of eigenvectors, i.e., functions solving $Lf + \lambda f = 0$, where f is supposed to satisfy the boundary conditions above.

There is a difficulty here: That the operator L is unbounded. However, it turns out to have an inverse that is not only bounded, but a (self-adjoint) Hilbert–Schmidt operator. Thus the spectral theorem for self-adjoint compact operators will take care of the rest.

Our purpose here is just to briefly outline the procedure.

The general theory of ordinary differential equations tells us that the homogeneous equation

$$Lf = (pf')' + qf = 0$$

has a two-dimensional solution space: The solution becomes unique if we specify both f and f' at a given point. We shall instead fix two nonzero solutions u and v, each satisfying one of our two boundary conditions:

$$\alpha u(a) + \alpha' u'(a) = 0, \quad \beta v(b) + \beta' v'(b) = 0.$$

We shall assume that these are linearly independent, so that u, v form a basis for the solution space of the homogeneous equation.[1]

The Wronskian $uv' - u'v$ has no zeroes in $[a,b]$. For if it zero at some point x, then we could find a nontrival solution (ξ, η) of the equations

$$\xi u(x) + \eta v(x) = 0, \quad \xi u'(x) + \eta v'(x) = 0,$$

and then $y = \xi u + \eta v$ would be a non-trivial solution with $y(x) = y'(x) = 0$, and that is impossible. We can define

$$w = p \cdot (uv' - u'v)$$

and notice by differentiation (exercise!) that w is constant.

If we now wish to solve the non-homogeneous equation

$$Lf = (pf')' + qf = g,$$

[1] If not, then 0 is an eigenvalue of L, and it turns out that we can avoid the problem by adding a suitable constant to q.

one common soluton strategy is by *variation of parameters*: Since the general solution of the homogeneous equation is given by $\varphi u + \psi v$ with constants φ and ψ, we try to solve the non-homogeneous problem by using the same formula with *functions* φ and ψ instead. Since we are thus using two functions to represent one unknown function, we have a bit of freedom to write an extra equation:

$$f = \varphi u + \psi v, \quad f' = \underbrace{\varphi' u + \psi' v}_{=0} + \varphi u' + \psi v'$$

where we use our new freedom to throw away the terms involving the first order derivatives of the unknown functions. Now writing $pf' = \varphi \cdot pu' + \psi \cdot pv'$, we differentiate once more and get

$$(pf')' + qf = \varphi \cdot \underbrace{((pu')' + qu)}_{=0} + \psi \cdot \underbrace{((pv')' + qv)}_{=0} + \varphi' pu' + \psi' pv'$$

where the indicated terms drop out because u, v solve the homogeneous equation. We are thus led to the system

$$\begin{bmatrix} u & v \\ pu' & pv' \end{bmatrix} \begin{bmatrix} \varphi' \\ \psi' \end{bmatrix} = \begin{bmatrix} 0 \\ g \end{bmatrix}$$

which has the solution

$$\begin{bmatrix} \varphi' \\ \psi' \end{bmatrix} = \frac{1}{w} \begin{bmatrix} pv' & -v \\ -pu' & u \end{bmatrix} \begin{bmatrix} 0 \\ g \end{bmatrix} = \frac{1}{w} \begin{bmatrix} -vg \\ ug \end{bmatrix}.$$

The boundary condition $\alpha f(a) + \alpha' f'(a) = 0$ yields

$$\varphi(a) \cdot \underbrace{(\alpha u(a) + \alpha' u'(a))}_{=0} + \psi(a) \cdot \underbrace{(\alpha v(a) + \alpha' v'(a))}_{\neq 0},$$

so that $\psi(a) = 0$. Similarly we find $\varphi(b) = 0$. So we can write

$$\varphi(x) = \int_b^x \frac{-v(\tau)g(\tau)}{w}\, d\tau = \int_x^b \frac{v(\tau)g(\tau)}{w}\, d\tau, \quad \psi(x) = \int_a^x \frac{u(\tau)g(\tau)}{w}\, d\tau$$

leading finally to the solution

$$f(x) = \int_x^b \frac{v(\tau)u(x)}{w} g(\tau)\, d\tau + \int_a^x \frac{u(\tau)v(x)}{w} g(\tau)\, d\tau = \int_a^b k(x,\tau) g(\tau)\, d\tau$$

where

$$k(x,\tau) = \begin{cases} \dfrac{u(x)v(\tau)}{w} & a \le x \le \tau \le b, \\ \dfrac{v(x)u(\tau)}{w} & a \le \tau \le x \le b. \end{cases}$$

In other words, the inverse of the operator L – with the given boundary conditions – is the Hilbert–Schmidt operator K given by the integral kernel k, which is called the *Green's function* for the problem.

Index