LÁSZLÓ MÁTÉ

HILBERT SPACE METHODS IN SCIENCE AND ENGINEERING

1828

Hilbert Space Methods in Science and Engineering

Except for size (i.e. dimension) one Hilbert space is very like another. To make a Hilbert space more interesting than its neighbours, it is necessary to enrich it by the addition of some external structure

P R Halmos

Hilbert Space Methods in Science and Engineering

László Máté

Technical University of Budapest, Hungary

Akadémiai Kiadó · Budapest 1989

English text revised by D. Petz

ISBN 963 05 4925 5

Joint edition published by Akadémiai Kiadó, Budapest, Hungary
and Adam Hilger, Bristol, England and New York, USA

Printed in Hungary by Szegedi Nyomda, Szeged.

Contents

Preface

The object of this book is to present Hilbert space theory as a useful language for applied mathematics and to present the basic facts and methods in a form suitable for engineers and scientists who apply mathematics.

For this purpose the text also contains many applications of Hilbert space theory, and we have emphasised the methods that are based on Hilbert space theory rather than giving a lot of material. The bulk of the applications revolve around reproducing kernel Hilbert spaces and causal operators. Several applications are treated here for the first time at an introductory level.

We have made an effort to make the book self-contained; however, an important problem remains concerning integration theory. To avoid this problem, the theorems are also formulated for non-complete spaces if it is possible and natural to do so, and the non-complete L_0^2-spaces, consisting of continuous functions with an L^2-norm, are introduced. In spite of this, the concepts of measurable function and Lebesgue integral cannot be completely avoided since the most fundamental theorems are valid only for complete spaces. The reader can find a short and satisfactory integral theory in Gohberg and Goldberg (1981), Appendix 2 (see also Naylor and Sell (1982), Appendix D).

The content of the book can be summarised as follows. Chapter 1 gives the fundamental concepts that are indispensable for understanding modern technical–mathematical literature dealing with normed spaces. The major part of this chapter is the Contractive Mapping Principle, which shows, with a few devices, the power of abstract space methods.

Chapter 2 gives a detailed and rather elementary account on Hilbert space geometry centred around the Projection Principle.

Chapter 3 comprises a reproducing kernel Hilbert space (RKHS) theory. The emphasis is not on the usual application in analytic function theory, but rather on the various RKHS models in differential equations, interpolation and control where the Hilbert space structure is enriched by the addition of some external structure. However, the rich applications in stochastic processes have been almost completely omitted because of the complicated background required to understand them.

Chapter 4 contains standard material on spectral theory, presented in as simple a form as possible.

Chapter 5 gives a mathematical theory of causal operators. Causal operators can be established in L^2-spaces by Hilbert space methods using truncation operators; however, in some other cases this does not work. For example, in most RKHS the truncations are not operators of the space. A unified Hilbert space approach is developed in the monograph by Feintuch and Saeks (1982) and, following the classical L^2-theory, the latter approach is also presented. This section is supplemented by recent results not included in previous textbooks.

Among the exercises at the end of chapters, the *easier* ones are marked with a circle (○) and the same mark is used to indicate the easier texts in the further reading section at the end of the book. On the other hand, we have used an asterisk ($*$) to indicate the sections that are *more difficult*, and the beginner is advised to skip over them at first reading. However, §§ 1.5, 1.6 and 2.5 contain material that is necessary for the subsequent chapters.

Many parts of this book have evolved during my one-semester courses in functional analysis, held for engineering students since the early 1970s, and the book is especially appropriate for similar one- or two-semester courses. We would be very grateful to be kept informed of reactions to this proposal.

Our thanks are due to the editors for the careful technical preparation, to C Kocak and O Gulyás for useful remarks concerning Chapter 3, and to D Petz, who read through the entire manuscript, for many helpful criticisms.

L Máté
Technical University of Budapest

1

Fundamentals

1.1 Linear spaces

The natural one-to-one correspondence between triplets (x, y, z) of real numbers and space vectors provides a geometrical model for many problems in physics, economics and biology.

If the data of the objects are described by more than three numbers then the geometrical model with space vectors is not applicable, and the visualisation of the problem in terms of a model of 'geometric nature' led to the concept of n-dimensional linear space.

Hilbert space theory can be considered as a further development in this direction. If infinite-dimensional spaces are also invoked, 'classical' mathematical analysis and geometry can be connected and a more effective device for solving mathematical problems than the 'pure classical analysis' is obtained.

As a first step towards Hilbert space theory we begin with a concise review of the fundamental concepts of linear algebra. A more detailed account can be found in Gelfand (1961).

1.1.1. The set of 'directed straight lines' with the usual rules of linear operations — addition and multiplication of scalars and the scalar product of two vectors — will be called *geometric vector space.* The main properties of the linear operations in the geometric vector space X are as follows. If $x, y, z \in X$, then

(i) $(x+y)+z=x+(y+z)$;
(ii) $x+y=y+x$;
(iii) there exists a zero element θ such that for every $x \in X$,

$$x+\theta = x.$$

Remark 1. A directed straight line with the properties of θ does not exist; however, without the zero element θ we could not answer the question: what is $x+(-1)x$? Hence for the main algebraic rules to be satisfied we add the symbol θ to the set of directed line segments as the 'zero vector'.

If λ and μ are scalars (i.e. real numbers) then

(iv) $\lambda(x+y)=\lambda x+\lambda y$;

(v) $(\lambda+\mu)x=\lambda x+\mu x$;

(vi) $(\lambda\mu)x=\lambda(\mu x)$;

(vii) $1.x=x$.

The properties (i)–(vii) alone ensure that the usual algebraic rules can be applied in the vector calculus.

A *linear space* is an abstraction of the geometric vector space. If Φ is a field, then a set X is called a linear space over Φ if for any pair $x, y\in X$ there is a unique element $x+y\in X$ called the *sum* of x and y and for any pair $\lambda\in\Phi$, $x\in X$ there is a unique $\lambda x\in X$, called the *product* of x with the scalar λ, such that the properties (i)–(vii) are satisfied. The elements of a linear space are also called *vectors*.

Remark 2. In the case of the geometric vector space the vectors are the directed straight lines and Φ is the field of real numbers. For any linear space X it is usual that Φ consists of either the real or the complex numbers. In the first case X is called *real linear space*. Unless otherwise stated, Φ is taken as *the field of complex numbers* in what follows.

A sum of the form

$$\lambda x+\mu y \qquad x, y\in X, \quad \lambda, \mu\in\Phi$$

is called *a linear combination* of x and y, and a linear combination of any *finite* number of vectors is defined in a similar way. A subset $\mathscr{H}\subset X$ is called a *linear subspace* or *subspace* for short if it follows from $h_1, h_2, \ldots, h_r\in\mathscr{H}$ that $\lambda_1 h_1+\lambda_2 h_2+\ldots+\lambda_r h_r\in\mathscr{H}$, where λ_k; $k=1, 2, \ldots, n$ are scalars. Obviously, a linear subspace can be considered as a linear space in itself.

A set $\mathscr{H}$ of vectors is called *linearly independent* if

$$\lambda_1 x_1+\lambda_2 x_2+\ldots+\lambda_m x_m = \theta$$

only in the case

$$\lambda_1 = \lambda_2 = \ldots = \lambda_m = 0$$

for any finite numbers of vectors $x_i\in\mathscr{H}$, $i=1, 2, \ldots, m$.

If a linear space X contains n linearly independent vectors a_k; $k=1, 2, \ldots, n$ such that for every $x\in X$ there exist $\{\xi_k\in\Phi,\ k=1, 2, \ldots, n\}$ such that

$$x = \xi_1 a_1+\xi_2 a_2+\ldots+\xi_n a_n$$

(i.e. every $x\in X$ is a linear combination of the vectors a_k; $k=1, 2, \ldots, n$)

then X is called n *dimensional* and $\{a_k\}$, $k=1, 2, \dots, n$ is called a *finite basis* or *basis* for short. One can show that n is independent of the choice of basis.

We can talk in similar terms about the dimension and the basis of any linear subspace $\mathscr{H} \subset X$.

Remark 3. An n-dimensional space or subspace is also called *finite dimensional* without mention of the dimension of the space or subspace.

If X or $\mathscr{H}$ does not contain a finite basis, then X or $\mathscr{H}$ is called *infinite dimensional.*

The main subjects of functional analysis are infinite-dimensional linear spaces. The absence of a finite basis creates major difficulties in the investigations of functional analysis and this has led to the concept of separability and infinite basis (§§ 1.5, 2.2).

1.1.2. A mapping T: $X \to Y$, i.e. a mapping T from a linear space X into a linear space Y, is called a *linear operator* if

$$T(\lambda_1 x_1 + \lambda_2 x_2) = \lambda_1 T x_1 + \lambda_2 T x_2$$

$$x_1, x_2 \in X \qquad \lambda_1, \lambda_2 \in \Phi.$$

The set of $x \in X$ for which Tx has a meaning is called the *domain* $\mathscr{D}(T)$ of T and $\{Tx;\ x \in \mathscr{D}(T)\}$ is called the *range* of T.

If T_1: $X \to Y$ and T_2: $X \to Y$ *have the same domain,* then the linear combination $\lambda_1 T_1 + \lambda_2 T_2$ is defined as

$$[\lambda_1 T_1 + \lambda_2 T_2] x := \lambda_1 T_1 x + \lambda_2 T_2 x \qquad x \in X;\ \lambda_1, \lambda_2 \in \Phi$$

and the product of two operators T_1 and T_2 as

$$T_2 T_1 x := T_2 [T_1 x] \qquad x \in X$$

if X, Y and Z are linear spaces, T_1: $X \to Y$, T_2: $Y \to Z$ and the range of T_1 is included in the domain of T_2. Both the linear combination and the product of linear operators are linear (see figure 1.1).

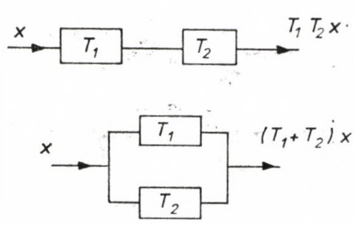

fig. 1.1

Remark 4. The product and linear combination are also defined for non-linear mappings in the same manner. If the mapping (operator) F is not linear then the value is denoted by $F(x)$ instead of Fx. With a few exceptions, we shall deal with linear operators only.

If the operator T is everywhere defined and $Y=X$, then we shall say that T is *an operator on* X, and if $Y=\Phi$ then the operator is called *functional.*

1.2 Normed spaces

The fundamental concepts in mathematical analysis are the various types of convergence and the limit. Derivatives, integrals, series expansions etc are based on these notions. It is therefore inevitable to define convergence in a mathematical structure to be applied in the problems of mathematical analysis. Introducing a norm in a linear space is one of the methods for introducing convergence.

1.2.1. The absolute value (modulus) of a vector in the geometric vector space has the following properties. Denoting the absolute value of the vectors x, y in an unusual manner as $\|x\|$, $\|y\|$ we have

(i) $\|x\| \geqslant 0$ and $\|x\|=0$ if and only if $x=\theta$;

(ii) $\|x+y\| \leqslant \|x\|+\|y\| \qquad x, y \in X$

(iii) $\|\lambda x\| = |\lambda| \, \|x\| \qquad x \in X, \ \lambda \in \Phi$

where Φ denotes scalars. The fact that a sequence $\{x_n\}$ of vectors converges to the vector x can be expressed as $\|x-x_n\| \to 0$.

A norm of a vector in a linear space X is an abstraction of the absolute value of a vector in the geometric vector space.

1.2.1.1 Definition. A linear space X is called a *normed space* if there exists a mapping $x \to \|x\|$ from X into the set of non-negative numbers defined for every $x \in X$ such that the properties (i)–(iii) are satisfied. The non-negative number $\|x\|$ is called the *norm* of x.

1.2.1.2 Definition. A sequence $\{x_n\}$ of vectors in X is called *convergent* if there exists $x \in X$ such that the sequence $\|x-x_n\|$ of non-negative numbers tends to zero. In this case, x is called the *limit* of the convergent sequence $\{x_n\}$.

Notation: $\lim\limits_{n\to\infty} x_n = x$ or $x_n \to x$

Example 1. An immediate generalisation of the absolute value of geometrical vectors for the linear space of n-tuples of real numbers is the following: if $x := \{x_1, x_2, \ldots, x_n\}$ then

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}.$$

Notice that for $n \leqslant 3$ this is the absolute value of the corresponding vector in the geometric vector space.

It is obvious that the properties (i) and (iii) of the norm are satisfied. However, to verify (ii) we need non-trivial considerations (see, for example, Gelfand (1961), Lusternik and Sobolev (1961)).

Example 2. Other useful norms in the linear space of n-tuples of real numbers are

(*a*) $$\|x\|_1 := \sum_{i=1}^{n} |x_i|$$

(*b*) $$\|x\|_\infty := \max_i |x_i|.$$

For the connections between various norms in an n-dimensional linear space we refer to § 1.7.

Remark. In the previous examples we can also take n-tuples of *complex* numbers.

Example 3. The infinite sequences $x := \{x_i;\ i = 1, 2, \ldots\}$ of complex (or real) numbers satisfying

$$\sum_{i=1}^{\infty} |x_i|^2 < \infty$$

form a normed space called *l^2-space* with the norm

$$\|x\|_2 := \left(\sum_{i=1}^{\infty} |x_i|^2\right)^{1/2}.$$

Indeed, properties (i) and (iii) of the norm are satisfied trivially and on the basis of Example 1 we have

$$\left(\sum_{i=1}^{m} |x_i + y_i|^2\right)^{1/2} \leqslant \left(\sum_{i=1}^{m} |x_i|^2\right)^{1/2} + \left(\sum_{i=1}^{m} |y_i|^2\right)^{1/2}$$

for any finite sum and hence, passing to the limit, if

$$\sum_{i=1}^{\infty} |x_i|^2 < \infty$$

and

$$\sum_{i=1}^{\infty} |y_i|^2 < \infty$$

then

$$\sum_{i=1}^{\infty} |x_i + y_i|^2 < \infty$$

and hence property (ii) of the norm is also satisfied.

Example 4. It is easy to see that the bounded infinite sequences with the norm

$$\|x\|_\infty := \sup_i |x_i|$$

and the absolute summable sequences with the norm

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

form normed spaces; they are called l^∞-*space* and l^1-*space*, respectively (the superscript 1 is sometimes omitted).

Example 5. The linear space of functions continuous on a closed interval $[a, b]$ is a normed space with the norm

$$\|f\|_\infty := \sup \{|f(t)|;\ t \in [a, b]\}.$$

It is called C$[a, b]$-*space* (or C-*space* for short if the domain $[a, b]$ is clear from the context).

That the sequence $\{f_n\}$ of continuous functions tends to $f \in \mathrm{C}$ in this normed space means exactly that $f_n \to f$ *uniformly on* $[a, b]$ and hence this is the most important normed space of continuous functions.

Remark. We can see clearly in this example, why it is necessary to adopt a new name: norm, and a new notation: $\|\,.\,\|$ for this generalisation of the absolute value of geometrical vectors. We can speak about the absolute value $|f(t)|$ of a continuous function f and also about the norm of f, i.e. the maximum value of $|f(t)|$ in $[a, b]$.

Example 6. Other useful norms in the linear space of continuous functions on a closed finite $[a, b]$ are the following:

(*a*) $$\|f\|_2 := \Big(\int_a^b |f(t)|^2\, \mathrm{d}t\Big)^{1/2} \quad (\mathrm{L}_0^2\text{-space})$$

(*b*) $$\|f\|_1 := \int_a^b |f(t)|\, \mathrm{d}t \quad (\mathrm{L}_0^1\text{-space})$$

(the superscript 1 is sometimes omitted). The only non-trivial part in proving that the properties (i), (ii) and (iii) of the norm are satisfied is that (ii) is satisfied for $\|\,.\,\|_2$; this is postponed until § 2.1.

1.2.2. The following important theorems on convergent sequences of real numbers are valid for every normed space with the same proof (but substituting 'norm' for 'absolute value' of course).

1.2.2.1 Theorem. A sequence $\{x_n\}$ has at most one limit.

1.2.2.2 Theorem. If $x_n \to x$ then also $x_{n_i} \to x$ for every subsequence $\{x_{n_i}\}$.

1.2.2.3 Theorem. If $x_n \to x$ then $\|x_n\| \to \|x\|$.

1.2.2.4 Definition. $\{x_n\}$ is *bounded* if there exists $K \geqslant 0$ (common for every n!) such that $\|x_n\| \leqslant K$.

1.2.2.5 Theorem. If $\{x_n\}$ is convergent, then $\{x_n\}$ is bounded.

1.2.2.6 Theorem. The linear operations are continuous in the following sense:

(*a*) If $x_n \to x$ and $y_n \to y$ then $x_n + y_n \to x + y$;

(*b*) If $\lambda_n \to \lambda$ and $x_n \to x$ then $\lambda_n x_n \to \lambda x$.

1.2.3. There are important theorems, however, which are not valid for every normed space.

1.2.3.1 Definition. A sequence $\{x_n\}$ in a normed space X is called *convergent in itself* or *Cauchy-sequence* if for every $\varepsilon > 0$ there exists $N = N(\varepsilon)$ such that

$$\|x_n - x_m\| < \varepsilon \qquad \text{for } n, m > N(\varepsilon).$$

1.2.3.2 Theorem. Every convergent sequence $\{x_n\}$ is also convergent in itself.
Proof. Since

$$\|x_n - x_m\| = \|(x_n - x) + (x - x_m)\| \leqslant \|x_n - x\| + \|x - x_m\| \qquad x, x_n, x_m \in X$$

it follows that $x_n \to x$ implies that $\{x_n\}$ is a Cauchy sequence.

The Cauchy Convergence Theorem says that for a sequence $\{x_n\}$ of real (or complex) numbers the converse of the above theorem is also valid: convergent and Cauchy sequences are the same and this is also true in the geometric vector space. However, the following example shows that the Cauchy Convergence Theorem is not valid in every normed space.

Example 1. Let us consider in $L_0[-1, +1]$ the functions

$$x_n(t) = \begin{cases} 0 & \text{if } t < 0 \\ nt & \text{if } 0 \leq t \leq 1/n \\ 1 & \text{if } t > 1/n. \end{cases}$$

(See figure 1.2.) The sequence x_n is convergent in itself since

$$\|x_n - x_m\|_1 = \int_{-1}^{+1} |x_n(t) - x_m(t)| \, dt$$

$$= \int_0^{1/m} (mt - nt) \, dt + \int_{1/m}^{1/n} (1 - nt) \, dt$$

$$= \int_0^{1/m} mt \, dt + \int_{1/m}^{1/n} dt - \int_0^{1/n} nt \, dt$$

$$= \frac{1}{2m} + \left(\frac{1}{n} - \frac{1}{m}\right) - \frac{1}{2n} \qquad n < m.$$

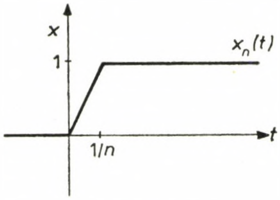

fig. 1.2

We shall show that $\{x_n\}$ is not convergent, i.e. there is no *continuous* function $x = x(t)$ such that $\|x - x_n\|_1 \to 0$.

For this purpose, let us consider a 'larger' normed space containing both the continuous functions and step functions. For a *step function* $y = y(t)$, the interval $[-1, +1]$ can be divided into a finite number of subintervals $[t_i, t_{i+1}]$ such that the value of $y = y(t)$ is constant in every (t_i, t_{i+1}) and the union of the intervals $[t_i, t_{i+1}]$ is $[-1, +1]$ (see figure 1.3).

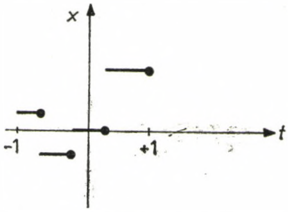

fig. 1.3

We now consider the linear space generated by the continuous functions and step functions on $[-1, +1]$ with norm

$$\|x\| = \int_{-1}^{+1} |x(t)|\,dt.$$

Then for

$$l_+(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}$$

we have

$$\int_{-1}^{+1} |l_+(t) - x_n(t)|\,dt = \int_{0}^{1/n} (1-nt)\,dt < \frac{1}{n}$$

and hence $\|l_+ - x_n\|_1 \to 0$.

Since, by 1.2.2.1, there is at most one limit for $\{x_n\}$ in a normed space and $l_+ \notin L_0(-1, +1)$, we have proved that *there is no limit in* L_0 for $\{x_n\}$.

1.2.3.3 Definition. A normed space X is called *complete* if every Cauchy sequence has a limit in X. A complete normed space is also called a *Banach space*.

The above example in $L_0[-1, +1]$ shows that $L_0[-1, +1]$ is not complete and it can be shown by a similar example that neither is $L_0^2[a, b]$.

The space $C[a, b]$ is complete. This follows from the following classical result in mathematical analysis: 'If a sequence $\{f_n\}$ of continuous functions converges uniformly on $[a, b]$ to a function f, then f is continuous on $[a, b]$.' Indeed, from Example 5 in § 1.2.1, this is the same as saying: 'The normed space $C[a, b]$ is complete.'

The remaining examples in § 1.2.1 of normed spaces are also complete, i.e. they are Banach spaces. (Although we shall not prove this, it is by no means trivial.)

1.2.3.4 Definition. The normed space X (or a subset $\mathscr{H} \subset X$) is called *compact* if every sequence $\{x_n\}$ belonging to $\mathscr{H}$ contains a convergent subsequence tending to a vector $x \in \mathscr{H}$.

The Bolzano–Weierstrass Theorem says that every bounded closed subset of the real (or complex) numbers is compact, and this is true also for the bounded closed subsets of the geometrical vector space. However, the following example shows that the Bolzano–Weierstrass Theorem is not valid in every normed space.

Example 2. In l^2 the sequence $\{e_k\}$; $k = 1, 2, \ldots,$ where

$$e_k = \{0, 0, \ldots, 0, \overset{k}{1}, 0, \ldots\}$$

belongs to the closed unit sphere $\{x: \|x\|_2 \leq 1\}$. However, there is no convergent subsequence of $\{e_k\}$; $k=1, 2, \ldots$ since for any pair e_n, e_m,

$$\|e_n - e_m\| = \sqrt{2}.$$

Examples for compact subsets $\mathscr{H}$ and further theorems concerning complete space and compact subsets will be found in § 1.6.

1.2.4. In a linear space, as we have seen, many different norms can be defined and hence different normed spaces are obtained. For example, from the linear space of continuous functions on the closed finite interval $[a, b]$, the Banach space $C[a, b]$ is obtained if it is supplied with the norm $\|\,.\,\|_\infty$; however, the linear space of continuous functions with the norm $\|\,.\,\|_1$ is the (non-complete) normed space $L_0[a, b]$ and with the norm $\|\,.\,\|_2$, the (non-complete) normed space $L_0^2[a, b]$. It is therefore useful to compare the different norms in the following sense.

1.2.4.1 Definition. If the linear space X is supplied with two different norms, say $\|\,.\,\|$ and $\|\,.\,\|^*$, then the norm $\|\,.\,\|$ is termed *stronger* than $\|\,.\,\|^*$ (and hence $\|\,.\,\|^*$ is weaker than $\|\,.\,\|$) if

$$\|x\|^* \leq K\|x\| \qquad x \in X$$

for $K>0$ (common for every $x \in X$).

1.2.4.2 Definition. The norms $\|\,.\,\|$ and $\|\,.\,\|^*$ are called *equivalent* if there exist positive K_1 and K_2 for which

$$K_1\|x\|^* \leq \|x\| \leq K_2\|x\|^* \qquad x \in X.$$

Example 1. In the linear space of continuous functions on the finite interval $[a, b]$, the norm $\|\,.\,\|_\infty$ is stronger than the norm $\|\,.\,\|_1$ since for every continuous function f,

$$\int_a^b |f(t)|\,dt \leq (b-a) \sup \{|f(t)|;\ t \in [a, b]\}.$$

Example 2. In the linear space of n-tuples of complex numbers, the norms $\|\,.\,\|_1$, $\|\,.\,\|_2$ and $\|\,.\,\|_\infty$ are equivalent. Indeed for $x=\{\xi_k;\ k=1, \ldots, n\}$,

$$\max_k |\xi_k| \leq \sum_{k=1}^{n} |\xi_k| \leq n \max_k |\xi_k|$$

and hence

$$\|x\|_\infty \leq \|x\| \leq n\|x\|_\infty \qquad x \in X.$$

The equivalence of $\|\,.\,\|_\infty$ and $\|\,.\,\|_2$ is obtained in a similar manner.

Example 3. Let us consider the linear space of functions on the real line which are continuous on a finite interval I_f and are zero outside I_f, i.e. $f(t)=0$ if $t \notin I_f$ and I_f *depends on f.* The norms $\| \cdot \|_\infty$ and $\| \cdot \|_1$ are also defined in this linear space; however, they cannot be compared in the sense of definition 1.2.4.1. Indeed, considering

$$f_n(t) = \begin{cases} 1/(1+|t|) & \text{if} \quad |t| < n \\ 0 & \text{otherwise} \end{cases}$$

we have $\|f_n\|_\infty = 1$ for every n but $\|f_n\|_1 \to \infty$; considering

$$h_n(t) = \begin{cases} n & \text{if} \quad |t| < 1/n \\ 0 & \text{otherwise} \end{cases}$$

we have $\|h_n\|_\infty \to \infty$ but $\|h_n\|_1 = 1$ for every n.

1.2.4.3 Theorem. If the two norms $\| \cdot \|$ and $\| \cdot \|^*$ are equivalent in X then the same sequences are convergent in both spaces $\{X; \| \cdot \|\}$ and $\{X; \| \cdot \|^*\}$. Consequently $\{X; \| \cdot \|\}$ is complete if and only if $\{X; \| \cdot \|^*\}$ is complete and $\mathscr{H} \subseteq \{X: \| \cdot \|\}$ is compact if and only if $\mathscr{H}$ is compact as a subset of $\{X; \| \cdot \|^*\}$.

While this is a very important theorem, its proof is obvious.

1.3 Contractive mappings

1.3.1. If $y = F(x)$ is a function with a continuous derivative on the real line, then the solution of the equation

$$x = F(x)$$

is the limit of the recursive sequence

$$x_0 = a \qquad x_{n+1} = F(x_n) \qquad n = 0, 1, 2, \ldots$$

if the following conditions are satisfied:

(a)
$$\left| \frac{\mathrm{d}}{\mathrm{d}x} F(x) \right| < 1 \qquad x \in \mathscr{D}$$

where $\mathscr{D}$ is a closed domain of the real line;

(b) $\quad a \in \mathscr{D}$ and if $x \in \mathscr{D}$ then $F(x) \in \mathscr{D}$.

A very important example is $F(x)=\frac{1}{2}(x+A/x)$; in this case the sequence

$$x_0 = A > 0 \qquad x_{n+1} = \frac{1}{2}(x_n+A/x_n) \qquad n = 0, 1, 2,\ldots$$

tends to $A^{1/2}$, i.e. to the solution of the equation

$$x = \frac{1}{2}(x+A/x).$$

The process of solving an equation by a recursive sequence like this is visualised in figures 1.4(a), (b).

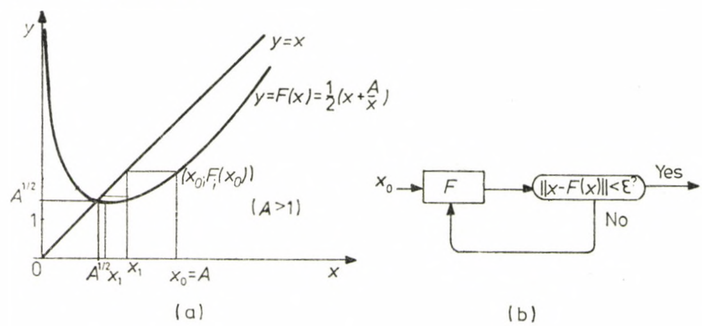

fig. 1.4

1.3.2. The subject of this section is the generalisation of the above process to the case when F is a mapping of a Banach space.

1.3.2.1 Definition. A mapping (operator) F of a Banach space B is *contractive* in $\mathscr{D}\subseteq B$ if

(*a*) from $x\in\mathscr{D}$ it follows that $F(x)\in\mathscr{D}$;
(*b*) $\|F(x)-F(y)\|\leq q\|x-y\|$, $0<q<1$, $x, y\in\mathscr{D}$.

1.3.2.2 Theorem. If F is a contractive mapping in $\mathscr{D}$ then the recursive sequence

$$x_0\in\mathscr{D} \qquad x_{n+1} = F(x_n) \qquad n = 0, 1, 2, \ldots \tag{$*$}$$

is convergent. Moreover, if $x=\lim_n x_n$, then

$$x = F(x) \tag{$**$}$$

and the solution of this equation $(**)$ is unique in $\mathscr{D}$.

Proof. First, it will be shown that $\{x_n\}$ is a Cauchy sequence and hence the limit $x=\lim_n x_n$ exists since B is a Banach space.

$$\|x_{n+1}-x_n\| = \|F(x_n)-F(x_{n-1})\| \leq q\,\|x_n-x_{n-1}\| \qquad n = 1, 2, \ldots$$

and in full,

$$\|x_2-x_1\| = \|F(x_1)-F(x_0)\| \leqslant q\|x_1-x_0\|$$

$$\|x_3-x_2\| = \|F(x_2)-F(x_1)\| \leqslant q\|x_2-x_1\| \leqslant q^2\|x_1-x_0\|$$

$$\|x_4-x_3\| = \|F(x_3)-F(x_2)\| \leqslant q\|x_3-x_2\| \leqslant q^3\|x_1-x_0\|$$

and hence, by induction,

$$\|x_{n+1}-x_n\| \leqslant q^n\|x_1-x_0\| \qquad n = 1, 2, \ldots$$

which means that $\{\|x_{n+1}-x_n\|\}$ is less than a geometric sequence with quotient $0<q<1$. Consider the Cauchy Convergence Theorem and, for example, $n>m$:

$$\begin{aligned}\|x_n-x_m\| &\leqslant \|x_{m+1}-x_m\|+\|x_{m+2}-x_{m+1}\|+\ldots+\|x_n-x_{n-1}\| \\ &\leqslant (q^m+q^{m+1}+\ldots+q^{n-1})\|x_1-x_0\| \\ &\leqslant q^m(1+q+q^2+\ldots)\|x_1-x_0\| \\ &= q^m\frac{1}{1-q}\|x_1-x_0\|\end{aligned}$$

and hence $\{x_n\}$ is a Cauchy sequence. For the limit x of the sequence $\{x_n\}$,

$$\|x-F(x)\| \leqslant \|x-x_n\|+\|x_n-F(x_n)\|+\|F(x_n)-F(x)\|$$

and hence if $x \in \mathscr{D}$ and $\|x-x_n\|<\varepsilon/4$ then

$$\|x-F(x)\| \leqslant \varepsilon/4+\varepsilon/2+q\varepsilon/4 < \varepsilon$$

which means that $x=F(x)$. If $x \notin \mathscr{D}$, then $F(x)$ is defined as $\lim_n F(x_n)$, and the same conclusion is reached.

If we suppose that $z=F(z)$ and $z \neq x$, i.e. there are two solutions of $(**)$, x and z, then

$$\|x-z\| = \|F(x)-F(z)\| < q\|x-z\| \qquad 0 < q < 1$$

which is impossible if $x \neq z$.

Example 1. The usual iterative method for solving first-order differential equations,

$$x'(t) = f(t, x(t)) \qquad x(t_0) = x_0$$

when a Lipschitz condition

$$|f(x_1, t)-f(x_2, t)| \leqslant M|x_1-x_2| \qquad \text{if} \quad |t-t_0| < d$$

is satisfied is a special case of $(*)$ if

$$F(x(t)) := x_0 + \int_{t_0}^{t} f(\tau, x(\tau))\,d\tau$$

in $C[a, b]$. Indeed, if $\alpha < \min(d, 1/M)$ and $a = t_0 - \alpha$, $b = t_0 + \alpha$, then for $x, y \in C[a, b]$,

$$\begin{aligned} |F(x)-F(y)| &= \left|\int_{t_0}^{t} [f(\tau, x(\tau)) - f(\tau, y(\tau))]\,d\tau\right| \\ &\leq \int_{t_0}^{t} \left|f(\tau, x(\tau)) - f(\tau, y(\tau))\right|\,d\tau \\ &\leq M \int_{t_0}^{t} |x(\tau)-y(\tau)|\,d\tau \\ &\leq M|t-t_0|\;\|x-y\|_\infty. \end{aligned}$$

Hence for $|t-t_0| < \alpha$,

$$\|F(x)-F(y)\|_\infty \leq M\alpha\|x-y\|_\infty$$

and $M\alpha < 1$.

Remark. The iterative method for solving first-order differential equations also works in more general conditions.

Example 2. The equation

$$x(t) - \lambda \int_0^t K(t-\tau)x(\tau)\,d\tau = f(t)$$

where $f, K \in C[0, T]$ for any $T > 0$ is called a convolution equation of the second order. What is the condition for F being a contractive mapping in $C[0, T]$ if

$$F(x(t)) := f(t) + \lambda \int_0^t K(t-\tau)x(\tau)\,d\tau?$$

To answer this question, consider

$$\begin{aligned} \|F(x)-F(y)\|_\infty &= \sup\left(\left|\lambda \int_0^t K(t-\tau)(x(\tau)-y(\tau))\,d\tau\right|;\; t\in[0, T]\right) \\ &\leq |\lambda| \int_0^T |K(t-\tau)|\,d\tau \sup(|x(t)-y(t)|;\; t\in[0, T]) \\ &\leq \|x-y\|_\infty |\lambda| \sup\left(\int_0^T |K(t-\tau)|\,d\tau;\; t\in[0, T]\right). \end{aligned}$$

So, if $|\lambda|$ is small enough, or more precisely if

$$|\lambda| < \left[\sup\left(\int_0^T |K(t-\tau)|\,d\tau;\ t\in[0,T]\right)\right]^{-1}$$

then F is contractive in $C[0, T]$ and hence the sequence

$$x_0(t) = f(t) \qquad x_{n+1}(t) = f(t)+\lambda \int_0^t K(t-\tau)x_n(\tau)\,d\tau \qquad n = 0, 1, 2, \ldots$$

converges uniformly in $[0, T]$ to the solution of the convolution equation.

This example has several generalisations. A more thorough investigation will show that the recursive sequence $\{x_n(t)\}$ tends to the solution of the convolution equation *also for every* λ. By a similar argument one can demonstrate the convergence of the iterative method for the more general integral operator

$$F(x(t)) := f(t)+\lambda \int_a^b K(t, \tau)x(\tau)\,d\tau.$$

More about contractive mapping, including the problems just mentioned, can be found in §§ 1.8.22–1.8.28.

1.4 Continuous linear operators

A continuous operator T: $X\to Y$ sends a convergent sequence $\{x_n\}$ into a convergent sequence $\{Tx_n\}$. A bounded operator T sends a bounded subset of X into a bounded subset of Y. For a *linear* operator T these two properties coincide.

1.4.1. In the following, unless otherwise stated, an operator T: $X\to Y$ is assumed to be defined everywhere.

1.4.1.1 Definition. The operator T: $X\to Y$ is called *continuous in* $x_0\in X$ if for every $\{x_n\}$ tending to x_0 it follows that $Tx_n\to Tx_0$.

As in classical analysis the following theorem can be proved.

1.4.1.2 Theorem. The operator T: $X\to Y$ is continuous at $x_0\in X$ if and only if for every $\varepsilon>0$ we have $\delta=\delta(\varepsilon)>0$ such that $\|x-x_0\|<\delta$ implies $\|Tx-Tx_0\|<\varepsilon$.

1.4.1.3 Theorem. If a linear operator T is continuous at $\theta\in X$, then T is continuous at every $x\in X$.

Proof. Since T is linear,

$$Tx = T(x+\theta) = Tx+T\theta \qquad x \in X$$

and hence $T\theta = \theta$. Now, if $x_n \to x_0$ then $x_n - x_0 \to \theta$ and hence $T(x_n - x_0) \to \theta$ since T is linear and continuous in θ. But $T(x_n - x_0) = Tx_n - Tx_0$ and hence from $x_n \to x_0$ it follows that $Tx_n \to Tx_0$ for any $x_0 \in X$.

1.4.1.4 Definition. A linear operator T is *bounded* if there exists $(M > 0)$

$$\|Tx\| \leqslant M\|x\| \qquad x \in X.$$

In this case M is called an *upper bound* or a *bound* of T.

We now turn to the fundamental property of linear operators which tells us that for a linear operator continuity is the same as boundedness.

1.4.1.5 Theorem. A linear operator T is continuous if and only if T is bounded.
Proof. If T is bounded with a bound M, then for every $\varepsilon > 0$, $\|Tx\| < \varepsilon$ if $\|x\| < \varepsilon/M$ and hence T is continuous at θ. It follows from the foregoing theorem that T is continuous at every $x \in X$, i.e. T is a continuous linear operator.

Conversely, if T is continuous at θ, then $\|Tx\| < 1$ if $\|x\| < \delta(1)$, and hence for every non-zero $x \in X$,

$$\frac{\delta(1)}{\|x\|}\|Tx\| = \left\|T\left(\delta(1)x/\|x\|\right)\right\| < 1$$

and so

$$\|Tx\| < \frac{1}{\delta(1)}\|x\|$$

i.e. T is a bounded operator with a bound $M = 1/\delta(1)$.

1.4.2. Examples of bounded linear operators in the various normed spaces defined in § 1.2.1.

Example 1. Let us consider the linear space of n-tuples of real numbers with the norm $\|x\|_\infty = \max_i |x_i|$ and let the operator T be the multiplication with an $n \times n$ matrix $\mathbf{A}$ with elements $\{a_{ik}\}$, i.e. for $x = \{x_i;\ i = 1, 2, \ldots, n\}$,

$$Tx := \left\{\sum_{k=1}^{n} a_{ik}x_k;\ i = 1, 2, \ldots, n\right\}.$$

This is a bounded linear operator with a bound

$$M = \max_i \sum_{k=1}^{n} |a_{ik}|$$

since

$$\Big|\sum_{k=1}^{n} a_{ik}x_k\Big| \leqslant \sum_{k=1}^{n} |a_{ik}||x_k| \leqslant \|x\|_\infty \sum_{k=1}^{n} |a_{ik}| \qquad i = 1, 2, \ldots, n$$

and hence

$$\|Tx\|_\infty := \max_i \Big|\sum_{k=1}^{n} a_{ik}x_k\Big| \leqslant \|x\|_\infty \max_i \sum_{k=1}^{n} |a_{ik}|.$$

Example 2. If $\{a_{ik};\ i=1, 2, \ldots, k=1, 2, \ldots\}$ is a 'double' infinite sequence called an *infinite matrix,* with the condition

$$\sup_i \sum_{k=1}^{\infty} |a_{ik}| < \infty$$

then

$$Tx := \Big\{\sum_{k=1}^{\infty} a_{ik}x_k;\ i = 1, 2, \ldots\Big\}$$

is a bounded linear operator in l^∞ with bound

$$M = \sup_i \sum_{k=1}^{\infty} |a_{ik}|.$$

Indeed, for every n and $x=\{x_i;\ i=1, 2, \ldots\}$,

$$\Big|\sum_{k=1}^{n} a_{ik}x_k\Big| \leqslant \sum_{k=1}^{n} |a_{ik}||x_k| \leqslant \|x\|_\infty \sum_{k=1}^{n} |a_{ik}| \leqslant \|x\|_\infty \sup_i \sum_{k=1}^{n} |a_{ik}|$$

and hence

$$\Big\{\sum_{k=1}^{\infty} a_{ik}x_k;\ i = 1, 2, \ldots\Big\}$$

is a bounded infinite sequence and

$$\Big\|\Big\{\sum_{k=1}^{\infty} a_{ik}x_k\Big\}\Big\|_\infty \leqslant \|x\|_\infty \sup_i \sum_{k=1}^{\infty} |a_{ik}|.$$

Example 3. If $K=K(t, \tau)$ is a continuous function on the finite closed (two-dimensional) interval $[a, b]\times[a, b]$ then

$$Tx := \int_a^b K(t, \tau)x(\tau)\,\mathrm{d}\tau \tag{$*$}$$

is a bounded linear operator in $\mathrm{C}[a, b]$. Indeed, it is obvious that T is linear and

$$\Big|\int_a^b K(t, \tau)x(\tau)\,\mathrm{d}\tau\Big| \leqslant \int_a^b |K(t, \tau)||x(\tau)|\,\mathrm{d}\tau \leqslant \|x\|_\infty \int_a^b |K(t, \tau)|\,\mathrm{d}\tau$$

and hence

$$\|Tx\|_\infty \leq \|x\|_\infty \sup\left\{\int_a^b |K(t,\tau)|\,d\tau;\ t\in[a,b]\right\}.$$

Remark. If $K(t, \tau)$ is continuous on $[a, b]\times[a, b]$, then

$$\int_a^b |K(t,\tau)|\,d\tau$$

is continuous on $[a, b]$ and hence it is bounded.

Example 4. The integration operator on $C[a, b]$,

$$Tx := \int_a^t x(\tau)\,d\tau \qquad t\in[a,b]$$

is a bounded operator with a bound $M=b-a$ since

$$\left|\int_a^t x(\tau)\,d\tau\right| \leq \|x\|_\infty(b-a).$$

Remark. The integration operator is in the form $(*)$ if

$$K(t,\tau) = \begin{cases} 1 & \text{if } \tau < t \\ 0 & \text{otherwise.} \end{cases}$$

Example 5. The most important example of a linear operator which is not continuous is the differentiation operator in $C[a, b]$,

$$Tx := \frac{d}{dt}(x(t)).$$

To see that the differentiation operator is unbounded in $C[0, 2\pi]$, let us consider the bounded sequence $\{\sin nt;\ n=1, 2, \ldots\}$:

$$T\sin nt := \frac{d}{dt}(\sin nt) = n\cos nt$$

and hence

$$\|T\sin nt\|_\infty = n\|\sin nt\|_\infty$$

since

$$\sup\{\cos nt;\ t\in[0,2\pi]\} = \sup\{\sin nt;\ t\in[0,2\pi]\}.$$

We conclude that T sends the bounded sequence $\{\sin nt;\ n=1, 2, \ldots\}$ into the unbounded $\{n\cos nt;\ n=1, 2, \ldots\}$ and hence T is not a bounded operator.

Remark. We can consider $T = \mathrm{d}/\mathrm{d}t$ as an operator of $C^1[a, b]$, the subspace of $C[a, b]$ consisting of functions with a continuous derivative. In this case, the differentiation operator is an everywhere-defined $T\colon C^1 \to C$ operator.

Example 6. In $C^1[a, b]$, the linear space of functions defined in $[a, b]$ with a continuous derivative, we can define

$$\|f\| := \|f\|_\infty + \left\|\frac{\mathrm{d}}{\mathrm{d}t} f\right\|_\infty .$$

It is easy to show that the properties (i)–(iii) of the norm in § 1.2.1 are satisfied for this $\|\,.\,\|$. If $C^1[a, b]$ is the normed space supplied with this norm (instead of $\|\,.\,\|_\infty$, as we had in the previous example), the differentiation operator is a bounded (and hence continuous) operator from C^1 into C.

Example 7. It is obvious that the operator (i.e. functional) which sends an infinite sequence $x = \{x_k\}$; $k = 1, 2, \ldots$ from l^∞ or l^1 into its nth element, x_n, is bounded and linear. Similarly, the mapping

$$f \to f(t_0) \qquad t_0 \in [a, b],\ f \in C[a, b]$$

called the *evaluation functional* is also bounded and linear.

Example 8. The evaluation functional is unbounded on $L_0[-1, +1]$. Indeed, if

$$x_n(t) = \begin{cases} n^2 t + n & \text{if } -1/n \leqslant t < 0 \\ -n^2 t + n & \text{if } 0 \leqslant t \leqslant 1/n \\ 0 & \text{elsewhere} \end{cases}$$

(see figure 1.5) then $\|x_n\|_1 = 1$, $x_n(0) = n$ and hence the evaluation functional $f \to f(0)$ sends the bounded sequence $\{x_n\}$ into the unbounded $\{x_n(0)\}$.

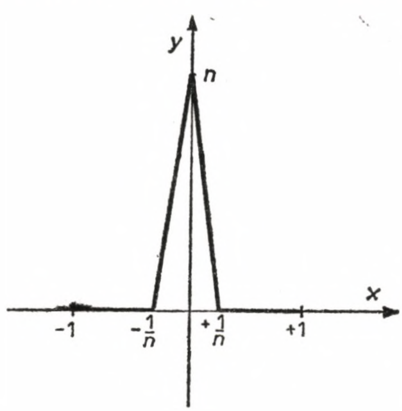

fig. 1.5

Similarly, the evaluation functional is unbounded also on $L_0^2[a, b]$ for any interval $[a, b]$.

Example 9. Let us consider the Fourier series

$$\frac{a_0}{2} + \sum_{k=1}^{\infty} a_k \cos kt + b_k \sin kt$$

of a real-valued continuous function $f = f(t)$, where

$$a_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \cos kt \, dt$$

$$b_k = \frac{1}{\pi} \int_0^{2\pi} f(t) \sin kt \, dt.$$

It is known that

$$\frac{a_0^2}{4} + \sum_{k=1}^{\infty} a_k^2 + b_k^2 = \frac{1}{\pi} \int_0^{2\pi} f(t)^2 \, dt$$

which means that the operator F from the real $L_0^2[0, 2\pi]$ into the real l^2 defined by

$$Ff = \{a_0/2, b_1, a_1, b_2, a_2, \ldots\}$$

is a bounded linear operator with bound $\pi^{1/2}$.

Remark. Modifying slightly the norm in the real $L_0^2[0, 2\pi]$-space,

$$\|f\|_2 := \frac{1}{\pi^{1/2}} \Big(\int_0^{2\pi} f(t)^2 \, dt \Big)^{1/2}$$

we have the stronger property $\|Ff\| = \|f\|$ for every $f \in L_0^2$. This kind of operator is called an *isometry*.

1.4.3. If T_1 and T_2 are bounded linear operators with bounds M_1 and M_2, then $\lambda T_1 + \mu T_2$ is also a bounded operator with bound $M = \max\{|\lambda| M_1, |\mu| M_2\}$ for any scalars λ, μ, and hence the bounded linear operators $T: X \to Y$ form a linear space.

Remark. $|\lambda| M_1 + |\mu| M_2$ is also a bound for $\lambda T_1 + \mu T_2$ since

$$\|(\lambda T_1 + \mu T_2)x\| \leqslant |s| \|T_1 x\| + |\mu| \|T_2 x\| < |\lambda| M_1 + |\mu| M_2.$$

1.4.3.1 Definition. The least upper bound of a bounded linear operator T is called the *norm* of T:

$$\|T\| := \inf\{M: \|Tx\| \leqslant M\|x\|;\ x \in X\}.$$

Remark. Since $x \in X$ and $Tx \in Y$, $\|x\|$ and $\|Tx\|$ should have very different meanings if $X \neq Y$, as can be seen in Examples 6–9 in § 1.4.2.

1.4.3.2 Theorem.

$$\|T\| = \sup \{\|Tx\| : \|x\| = 1\}.$$

Proof. If $M_0 = \sup \{\|Tx\|; \|x\| = 1\}$ then for every x: $x \neq \theta$,

$$\left\| T \frac{x}{\|x\|} \right\| \leq M_0$$

i.e.

$$\|Tx\| \leq M_0 \|x\|$$

and hence M_0 is an upper bound of T.

Conversely, if M is any other bound of T, then $\|Tx\| \leq M\|x\|$ and, for x: $x \neq \theta$,

$$\left\| T \frac{x}{\|x\|} \right\| = \frac{1}{\|x\|} \|Tx\| \leq M.$$

Hence

$$M_0 = \sup \{\|Tx\|; \|x\| = 1\} = \sup_{x \neq \theta} \left\| T \frac{x}{\|x\|} \right\| \leq M$$

which means precisely that M_0 is the *least* upper bound.

An immediate consequence of the foregoing theorem is the following.

1.4.3.3 Theorem. If T, T_1, T_2 are bounded linear operators from X into Y then

(i) $\|T\| \geq 0$ and $\|T\| = 0$ if and only if $Tx = \theta$ for every $x \in X$;

(ii) $\|T_1 + T_2\| \leq \|T_1\| + \|T_2\|$;

(iii) $\|\lambda T\| = |\lambda| \, \|T\|$; $\lambda \in \Phi$

and hence the linear space of bounded linear operators T: $X \to Y$ is a normed space.

In addition, the following connection holds between the norm and the product of linear operators.

1.4.3.4 Theorem. If the product $T_1 T_2$ is defined, then $T_1 T_2$ is also a *bounded* linear operator and

$$\|T_1 T_2\| \leq \|T_1\| \, \|T_2\|.$$

Proof. For every $x \in X$,

$$\|T_1T_2x\| \leqslant \|T_1\| \, \|T_2x\| \leqslant \|T_1\| \, \|T_2\| \, \|x\|$$

and hence

$$\|T_1T_2\| := \sup\{\|T_1T_2x\| : \|x\| = 1\} \leqslant \|T_1\| \, \|T_2\|.$$

$*$1.5 The geometry of normed spaces; separability

In linear algebra, in an n-dimensional space, the computations are fulfilled with the aid of a finite basis. Hence the lack of basis in an infinite-dimensional space creates difficulties in applications. Fortunately, many infinite-dimensional normed spaces X contain an *infinite* sequence $\{a_k;\ k=1, 2, \ldots\}$ of vectors called a fundamental sequence, such that every $x \in X$ can be approximated by linear combinations of vectors belonging to $\{a_k;\ k=1, 2, \ldots\}$. Such spaces will be called separable, and in a separable normed space a certain fundamental sequence plays the role of a basis in most cases.

1.5.1. We begin with some ideas of a geometrical nature. The set

$$\{x : \|x - x_0\| < r\}$$

is called an *open sphere* with centre x_0 and radius r.

1.5.1.1 Definition. A subset $\mathscr{H}$ of a normed space X is called *open* if from $x_0 \in \mathscr{H}$ it follows that there exists $r > 0$ such that $\{x:\ \|x - x_0\| < r\} \subset \mathscr{H}$. A subset $\mathscr{H}$ of X is *closed* if $\{x:\ x \notin \mathscr{H}\}$ is open.

There is also a more direct definition of a closed set via closure points using the concept of closure.

1.5.1.2 Definition. $x_0 \in X$ belongs to the *closure* of the subset $\mathscr{H} \subset X$ if any open sphere with centre x_0 contains a vector $x \in \mathscr{H}$.

1.5.1.3 Theorem. $x_0 \in X$ belongs to the closure of $\mathscr{H} \subset X$ if and only if there exists a sequence $\{x_n;\ x_n \in \mathscr{H}\}$ such that $x_n \to x_0$.

Proof. If x_0 belongs to the closure of $\mathscr{H}$, then every open sphere $\{x:\ \|x - x_0\| < 1/n\};\ n=1, 2, \ldots$ contains $x_n \in \mathscr{H}$ and hence $\|x_n - x_0\| < 1/n;\ n=1, 2, \ldots$, i.e. $x_n \to x_0$.

Conversely, if $x_n \in \mathscr{H}$ and $x_n \to x_0$, then by definition, all but a finite number of $\{x_n\}$ are contained in every open sphere with centre x_0.

1.5.1.4 Definition. Let $\overline{\mathscr{H}}$ be the closure of $\mathscr{H}$; then $\mathscr{H}$ is *closed* if $\overline{\mathscr{H}}=\mathscr{H}$.

In other words, by Theorem 1.5.1.3, *a subset $\mathscr{H}$ is closed if (and only if) it follows from $x_n \in \mathscr{H}$ and $x_n \to x$ that $x \in \mathscr{H}$ also.*

1.5.2. A subset $\mathscr{H}$ is called *dense* in X if $\overline{\mathscr{H}}=X$; i.e., by Theorem 1.5.1.3, $\mathscr{H}$ is dense in X if for every $x \in X$ there is a sequence $\{x_n\}$; $x_n \in \mathscr{H}$ such that $x_n \to x$.

1.5.2.1 Definition. The normed space X is *separable* if there exists a countable dense subset in X.

1.5.2.2 Definition. The linear space X (or subspace $\mathscr{H}$) is generated by the subset $\mathscr{S}$ if every $x \in X$ (or $x \in \mathscr{H}$) is a linear combination of elements of $\mathscr{S}$.

1.5.2.3 Theorem (without proof). If the linear space generated by a countable subset $\{a_n;\ n=1, 2, \ldots\}$ is dense in X, then X is separable.

The foregoing theorem says, practically, that if there is a sequence $\{a_n;\ n=1, 2, \ldots\}$ such that every $x \in X$ can be approximated to any required accuracy by sums

$$\sum_{i=1}^{m} \lambda_i a_{n_i}$$

then X is separable. So, the condition that a countable set $\{a_n;\ n=1, 2, \ldots\}$ be dense in X is *only apparently stronger* than the condition that the set of linear combinations from $\{a_n;\ n=1, 2, \ldots\}$ be dense in X. This is very important, since in most spaces only the latter can easily be verified, as will be seen in the following examples.

Example 1. Consider the countable set $\{t^k;\ n=1, 2, \ldots\}$ in $\mathrm{C}[a, b]$. By the well-known Weierstrass Theorem, every continuous function on the closed finite interval $[a, b]$ can be approximated by polynomials with respect to the uniform convergence. Since uniform convergence on $[a, b]$ is the same as convergence in the Banach space $\mathrm{C}[a, b]$, it follows that $\mathrm{C}[a, b]$ is separable.

Example 2. Consider the countable set $\{e_n;\ n=1, 2, \ldots\}$ in the l^2-space, where e_n is the sequence whose nth element is 1 and any other element is 0. It is obvious that every finite sequence (i.e. a sequence with all but a finite number of elements equal to 0) is a linear combination of elements from $\{e_n;\ n=1, 2, \ldots\}$. Moreover, the finite sequences form a dense subset in l^2. Indeed, if

$$x = \{\xi_k;\ k = 1, 2, \ldots\} \in l^2$$

and

$$x_n = \{\xi_{nk};\ k = 1, 2, \ldots\}$$

where

$$\xi_{nk} = \begin{cases} \xi_k & \text{if } k \leqslant n \\ 0 & \text{otherwise} \end{cases}$$

then obviously $\|x - x_n\|_2 \to 0$.

We conclude that the l^2-space is separable.

Example 3. The finite sequences do not form a dense subset in the space l^∞. Indeed, if

$$v = \{1, 1, \ldots\} \in l^\infty$$

then for *any* finite sequence $x_n = \{\xi_1, \xi_2, \ldots, \xi_n, 0, \ldots\}$, $\|v - x_n\| \geqslant 1$.

It can be proved (it is by no means trivial!) that there is no countable dense subset in l^∞, i.e. *l^∞ is not a separable space.*

Example 4. Every real-valued continuous function in $[0, 2\pi]$ has a Fourier expansion, convergent in $L_0^2[0, 2\pi]$ by the well-known Riesz–Fischer Theorem, and hence the linear space generated by $\{1, \sin nt, \cos nt;\ n = 1, 2, \ldots\}$ is dense in $L_0^2[0, 2\pi]$. We conclude that L_0^2 is separable.

Remark 1. The approximation by polynomials in $C[a, b]$ and the approximation by trigonometric polynomials in $L_0^2(0, 2\pi)$ via the Fourier expansion are of a very different nature. For every $f \in L_0^2$ we can construct an infinite series, the Fourier series, and *f is approximated by the partial sum of the series.* However, in the form of a power series, called the Taylor expansion, only certain infinitely differentiable functions can be approximated in $C[a, b]$.

So, we have to make a distinction between two different cases. If there is a countable subset $\{a_n;\ n = 1, 2, \ldots\}$ such that the linear space generated by $\{a_n;\ n = 1, 2, \ldots\}$ is dense in X, i.e. every $x \in X$ can be approximated to any required accuracy by (finite) sums

$$\sum_{i=1}^{m} \lambda_i a_{n_i}$$

then $\{a_n;\ n = 1, 2, \ldots\}$ is called a *fundamental sequence.* This is the case of $\{t^n;\ n = 1, 2, \ldots\}$ in $C[a, b]$.

If every $x \in X$ can be given in the form of an infinite series

$$x = \sum_{k=1}^{\infty} \xi_k a_k$$

then a_n; $n=1, 2, \ldots$ is called an *infinite basis* (see § 2.2). This is the case of $\{e^{int}; n=0, \pm 1, \pm 2, \ldots\}$ in $L_0^2[0, 2\pi]$.

In both cases X is a separable space.

Remark 2. We may also speak of a dense subset $\mathscr{H}$ of $\mathscr{M} \subset B$. A subset $\mathscr{M}$ of B is called separable if there exists a countable dense subset in $\mathscr{H}$.

To summarise, in a separable subset of a normed space there is a sequence $\{a_n; n=1, 2, \ldots\}$ such that every element is either the linear combination or the limit of the linear combinations of $\{a_n; n=1, 2, \ldots\}$.

∗1.6 More about complete spaces and compact sets

1.6.1. Some of the important normed spaces of continuous functions are not complete, e.g. the L_0- and L_0^2-spaces. However, the important contractive mapping theorem, as we have seen in § 1.3, is valid only for complete normed spaces — i.e. Banach spaces — and later on we will encounter more and more such theorems.

If a certain normed space is not complete, then we can restrict ourselves to an appropriate subspace which is a complete normed space, or we can complete the normed space into a Banach space by a completion process. In the latter case a new problem of identifying the new elements arises.

1.6.1.1 Definition. The linear subspace $\mathscr{M}$ of a normed space X is called *complete* if every Cauchy sequence $\{x_n\}$; $x_n \in \mathscr{M}$ has a limit $x \in \mathscr{M}$.

Example 1. Every finite-dimensional subspace of a normed space is complete, as will be shown in § 1.7.

Example 2. The space C_{00} of continuous functions with bounded support (see § 1.8.2) on the line with norm

$$\|x\|_\infty = \sup |x(t)|$$

is not complete. Indeed, if $x=x(t)$ is a *positive* continuous function on the real line such that $\lim_{t \to \pm\infty} x(t)=0$ and $x_n=x_n(t)$ is a continuous function with bounded support such that $|x_n(t)| \leqq x(t)$ and $x_n(t)=x(t)$ for $|t|<n$ (see figure 1.6), then $x_n \in C_{00}$ and the sequence $\{x_n(t)\}$ tends to $x(t)$ uniformly on the real line. So, by reasoning similar to that in Example 1 of § 1.2.3, the Cauchy sequence $\{x_n\}$; $x_n \in C_{00}$ has no limit in C_{00}.

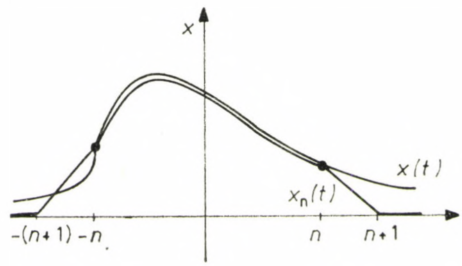

fig. 1.6

Example 3. In contrast to the previous example the closed subspace of C_{00} consisting of functions with support in $[-1, +1]$ is complete. Indeed, if the support of $x_n=x_n(t)$ is contained in $[-1, +1]$ for every n and $x_n(t)\to x(t)$ uniformly, then the support of $x=x(t)$ is also contained in $[-1, +1]$.

The next theorem tells us that every normed space has a complete extension.

1.6.1.2 Theorem. For every normed space X there exists a complete normed space (i.e. a Banach space) B such that for a dense subspace B_0 of B there is a 1–1 linear mapping (operator) L from X onto B_0 with the property

$$\|Lx\| = \|x\| \qquad x\in X.$$

B is called a *completion* of X.

Note: all completions of X are *isomorphic*, i.e. they can be identified with each other by an isometric mapping.

Instead of a proof this theorem will be elucidated with the aid of some examples.

Example 4. As was shown in the previous example, C_{00} is not complete; however, C_{00} is a dense subspace of C_0, the linear space of continuous functions on the real line with $\lim_{t\to\pm\infty} x(t)=0$ and with the norm $\|\,.\,\|_\infty$.

Since the uniform limit of $\{x_n;\ x_n\in C_0\}$ also belongs in C_0, i.e. the limit x is also a continuous function with $\lim_{t\to\pm\infty} x(t)=0$, C_0 is a complete normed space. Moreover, C_0 is the completion of C_{00} by the description in Theorem 1.6.1.2.

Example 5. If we consider the rational numbers as a normed space over the field Φ of rational numbers where the norm is the absolute value of the rational number, then the normed space thus obtained is not complete. For example, the sequence of rational numbers

$$1,\ 1.4,\ 1.41,\ 1.414,\ 1.4144,\ \ldots$$

(the digit-by-digit approximation of $\sqrt{2}$) is a Cauchy sequence and has no limit among the rational numbers.

The completion of the normed space of rational numbers is the space of real numbers, also with absolute value as a norm. To show this it is enough to remember that every real number is the limit of a sequence of rational numbers and the Cauchy convergence theorem is valid for real numbers.

More particularly, consider the Cauchy sequences of rational numbers and identify two sequences if their difference tends to zero. Then every Cauchy sequence tending to a rational number is identified with the rational number and the remaining Cauchy sequences are the new elements, the non-rational numbers.

Example 6. Consider the Cauchy sequences in $L_0^2[a, b]$ and identify two sequences $\{x_n\}$ and $\{y_n\}$ if their difference tends to zero, i.e.

$$\int_a^b |x_n(t)-y_n(t)|^2 \, dt \to 0.$$

Then every Cauchy sequence tending to a continuous function $x \in L_0^2[a, b]$ is identified with x. The remaining sequences (i.e. the remaining equivalent 'classes' of sequences) are the new elements which, together with the continuous functions, form the completion of $L_0^2[a, b]$ called $L^2[a, b]$.

One of the achievements of the Riesz–Fischer Theorem is the representation of this completion, $L^2[a, b]$, by measurable functions f, the Lebesgue integral of which obeys

$$\int_a^b |f(t)|^2 \, dt < \infty.$$

Remark. Fortunately, every bounded function occurring in applications is contained in the linear subspace of L^2 generated by step functions and continuous functions. So, if a normed space consisting of such functions is not complete then the new elements obtained by completion are needed only for 'theoretical' purposes, i.e. the results of Banach Space Theory should be applied.

Now, it is worth describing the completion process of a normed space X which we have seen in particular cases in the previous examples.

The completion consists of the equivalence classes of Cauchy sequences $\{x_n\}$ $(x_n \in X)$ in the following sense.

Two Cauchy sequences $\{x_n\}$ and $\{y_n\}$ are equivalent, or belong to the same class, if $x_n - y_n \to \theta$.

If the sequence $\{x_n\}$ converges to an element x of X, then the class of Cauchy sequences which are equivalent to $\{x_n\}$ is identified with $x \in X$.

The classes of the remaining Cauchy sequences are the new elements which, together with the elements of X, form the completion B of X. If $x \in B$, then

$$\|x\| := \lim_n \|x_n\|$$

where $\{x_n\}$ is a Cauchy sequence of the class of x.

Although obvious, the following theorem is important.

1.6.1.3 Theorem. The completion of a separable normed space X is also separable.

1.6.2. Remember that a normed space X (or a subset $\mathscr{H} \subset X$) is called compact if the Bolzano–Weierstrass Theorem is valid in X (or in $\mathscr{H}$). (See Definition 1.2.3.4.)

Further examples for non-compact bounded closed sets in infinite-dimensional normed spaces are the following.

Example 1. In $L_0^2[-\pi, +\pi]$ the sequence $\{(1/\pi^{1/2}) \sin nt;\ n=1, 2, \ldots\}$ belongs to the closed unit sphere $\{x\colon \|x\|_2 \leq 1\}$ since

$$\frac{1}{\pi} \int_{-\pi}^{+\pi} \sin^2 nt \,\mathrm{d}t = \frac{1}{2\pi} \int_{-\pi}^{+\pi} (1-\cos 2nt)\,\mathrm{d}t = 1.$$

However, there is no convergent subsequence of $\{(1/\pi^{1/2}) \sin nt;\ n=1, 2, \ldots\}$ since

$$\frac{1}{\pi} \int_{-\pi}^{+\pi} (\sin kt - \sin mt)^2 \,\mathrm{d}t = 2$$

for any pair k, m.

Example 2. Consider the following 'double' sequence in C[0, 1]:

$$x_{nk}(t) = \begin{cases} 2^{n+1}\left(t-\dfrac{k}{2^n}\right) & \text{if } \dfrac{k}{2^n} \leq t < \dfrac{2k+1}{2^{n+1}} \\ -2^{n+1}\left(t-\dfrac{k+1}{2^n}\right) & \text{if } \dfrac{2k+1}{2^{n+1}} \leq t < \dfrac{k+1}{2^n} \\ 0 & \text{elsewhere.} \end{cases}$$

Then $\|x_{nk}\|_\infty = 1$ and for any two different functions $x_{n,k}$ and $x_{n',k'}$,

$$\sup\{|x_{n,k}(t) - x_{n',k'}(t)|;\ t \in [0, 1]\} = 1$$

and hence there is no convergent subsequence of $\{x_{nk}\}$.

To find compact subsets in an infinite-dimensional normed space is not an easy task. Among the most important examples are the following.

A subset $\mathscr{H}$ of continuous functions is called *equicontinuous* if for every $\varepsilon > 0$ there is a common $\delta = \delta(\varepsilon)$ *for every* $x \in \mathscr{H}$ such that $|t_2 - t_1| < \delta(\varepsilon)$ implies that $|x(t_2) - x(t_1)| < \varepsilon$.

1.6.2.1 Theorem. A closed set $\mathscr{H} \subset C[a, b]$ is compact if and only if $\mathscr{H}$ is an equicontinuous subset of functions.

If for $f \in L^2[a, b]$ we define

$$f_h(t) := \frac{1}{2h} \int_{t-h}^{t+h} f(\tau)\, d\tau$$

then we have the following theorem.

1.6.2.2 Theorem. A closed set $\mathscr{H} \subset L^2[a, b]$ is compact if and only if for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that $|h| < \delta(\varepsilon)$ implies for every $x \in \mathscr{H}$ that $\|x - x_h\|_2 < \varepsilon$.

Remark. Although we omit the proofs of Theorems 1.6.2.1 and 1.6.2.2, they are by no means trivial (see, for example, Lusternik and Sobolev, 1961).

There are very fine properties for compact subsets:

1.6.2.3 Theorem. If $\mathscr{H} \subset X$ is compact then $\mathscr{H}$ is complete, bounded and separable.

Proof. If $x_n \in \mathscr{H}$ and $\{x_n\}$ is a Cauchy sequence, then there exists a subsequence $\{x_{n_i}\}$ tending to an $x \in \mathscr{H}$ since $\mathscr{H}$ is compact. It follows from the inequality

$$\|x_n - x\| \leq \|x_n - x_{n_i}\| + \|x_{n_i} - x\|$$

that $x_n \to x$ in this case and hence $\mathscr{H}$ is complete.

If $\mathscr{H}$ is unbounded then for every positive integer n one can find $x_n \in \mathscr{H}$ such that $\|x_n\| > n$, and hence there is no convergent subsequence $\{x_{n_i}\}$ of $\{x_n\}$, i.e. $\mathscr{H}$ is not compact.

1.6.2.4 Definition. A subset $\mathscr{H} \subset X$ is called *pre-compact* if the closure of $\mathscr{H}$ is compact.

It follows that for a closed subset $\mathscr{H}$, compactness and pre-compactness are the same properties.

1.7 Finite-dimensional normed spaces

Every n-dimensional normed space is equivalent, in the sense of 1.2.4.2, to the n-dimensional Euclidean space. An easy but important consequence of this equivalence is that every linear operator is continuous in a finite-dimensional space.

1.7.1. First we shall show that the three basic norms $\|\,.\,\|_1$, $\|\,.\,\|_2$ and $\|\,.\,\|_\infty$ are equivalent.

1.7.1.1 Theorem. In the linear space X of n-tuples $\{\xi_k\}$; $k=1, 2, \ldots, n$ of complex numbers the norms

$$\|x\|_1 := \sum_{k=1}^{n} |\xi_k|$$

$$\|x\|_2 := \left(\sum_{k=1}^{n} |\xi_k|^2\right)^{1/2}$$

$$\|x\|_\infty = \sup_k |\xi_k|$$

are equivalent.

Proof. If $x=\{\xi_k\}$; $k=1, 2, \ldots, n$ then

$$\sup_k |\xi_k| \leqslant \sum_{k=1}^{n} |\xi_k| \leqslant n \sup_k |\xi_k|$$

and hence $\|\,.\,\|_1$ and $\|\,.\,\|_\infty$ are equivalent norms; moreover

$$\sup_k |\xi_k|^2 \leqslant \sum_{k=1}^{n} |\xi_k|^2 \leqslant n \sup_k |\xi_k|^2.$$

Hence

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

for every $x \in X$, i.e. the norms $\|\,.\,\|_\infty$ and $\|\,.\,\|_2$ are equivalent.

1.7.1.2 Theorem. The linear space X of n-tuples of complex numbers with the norm $\|\,.\,\|_\infty$ is a complete normed space.

Proof. If

$$x_i := \{\xi_k^{(i)};\ k = 1, 2, \ldots, n\}$$

and $\{x_i\}$ is a Cauchy sequence, then

$$\sup_k |\xi_k^{(i)} - \xi_k^{(j)}| < \varepsilon \qquad \text{if} \quad i, j > N(\varepsilon).$$

In particular, $\{\xi_k^{(i)}\}$ is a Cauchy sequence of complex numbers for $k=1, 2, \ldots, n$

and hence there exists $x:=\{\xi_k\}$; $k=1, 2, \ldots, n$ such that

$$\xi_k^{(i)} \to \xi_k \qquad k = 1, 2, \ldots, n.$$

Moreover,

$$\sup \{|\xi_k^{(i)} - \xi_k|;\ k = 1, 2, \ldots, n\} \to 0$$

or, in other words,

$$\|x_i - x\|_\infty \to 0.$$

1.7.1.3 Theorem. If X is the linear space of n-tuples of complex numbers with the norm $\|\,.\,\|_\infty$, then every bounded closed subset of X is compact.
Proof. For a bounded subset $\mathscr{M}$ there exists $K>0$ such that $\|x\|_\infty < K$ for every $x \in \mathscr{M}$ and hence

$$|\xi_k^{(i)}| < K \qquad i = 1, 2, \ldots, \quad k = 1, 2, \ldots, n$$

for any infinite sequence $\{x_i\}$; $i=1, 2, \ldots$ belonging to $\mathscr{M}$. In particular, $|\xi_1^{(i)}| < K$ and hence it follows from the Bolzano–Weierstrass Theorem that there is a convergent subsequence $\{\xi_1^{(i_k)}$; $k=1, 2, \ldots\}$; let

$$\xi_1 = \lim_{k\to\infty} \xi_1^{(i_k)}.$$

There is also a convergent subsequence of $\{\xi_2^{(i_k)}\}$; $k=1, 2, \ldots$ since we also have $|\xi_2^{(i_k)}| < K$. Using the same notation for the subsequence, let $\xi_2 = \lim_{k\to\infty} \xi_2^{(i_k)}$.

Repeating this procedure, after n steps a subsequence $\{x_{i_k}\}$; $k=1, 2, \ldots$ is obtained, each coordinate of which is convergent. If

$$x = \{\xi_1, \xi_2, \ldots, \xi_n\}$$

then we conclude, as in the proof of the previous theorem, that

$$\|x_{i_k} - x\|_\infty \to 0.$$

Remark 1. On the basis of 1.2.4.3 and 1.7.1.1, Theorems 1.7.1.2 and 1.7.1.3 are also valid if the norm $\|\,.\,\|_\infty$ is replaced by $\|\,.\,\|_1$ or $\|\,.\,\|_2$ or any norm equivalent to $\|\,.\,\|_\infty$.

Remark 2. It is easy to see that a sequence is convergent in $\|\,.\,\|_\infty$ (and hence in any norm which is equivalent to $\|\,.\,\|_\infty$) if and only if it is convergent 'coordinate-wise'.

1.7.2. We shall show that any norm of a finite-dimensional normed space is equivalent to the Euclidean norm $\|\,.\,\|_2$.

Let us consider the mapping

$$\tau: \{\xi_k;\ k = 1, 2, \ldots, n\} \to x = \sum_{k=1}^{n} \xi_k a_k$$

from the linear space of n-tuples of complex numbers supplied with the norm $\| \,.\, \|_1$ into an n-dimensional normed space X, where

$$a_k;\quad k = 1, 2, \ldots, n$$

is a basis in X. It is obvious that τ is a linear 1–1 map onto X and hence there exists the linear inverse mapping τ^{-1}.

1.7.2.1 Theorem. The mappings τ and τ^{-1} are bounded linear operators.
Proof. The operator τ is bounded since

$$\|x\| \leqslant \sum_{k=1}^{n} |\xi_k| \, \|a_k\| \leqslant M \sum_{k=1}^{n} |\xi_k|; \quad x \in X$$

where $M = \sup\{\|a_k\|;\ k=1, 2, \ldots, n\}$.

For the inverse operator τ^{-1} let us suppose that there exists a sequence $\{x_n\}$ tending to zero and $\|\tau^{-1}(x_n)\| = 1$ for $n = 1, 2, \ldots$, i.e. it is supposed that τ^{-1} is not continuous at θ. It follows from Theorem 1.7.1.3 (see also Remark 1) that there exists a convergent subsequence $\{\tau^{-1}(x_{n_i})\}$ and it is obvious that $\lim_i \tau^{-1}(x_{n_i}) = a \neq \theta$; moreover,

$$x_{n_i} = \tau[\tau^{-1}(x_{n_i})] \to \tau(a)$$

since τ is continuous. Thus we have a contradiction, since $a \neq \theta$, $\tau(a) = \theta$ and τ is 1–1 and linear.

It now follows from 1.4.1.5 that τ^{-1} is a bounded operator.

1.7.2.2 Theorem. Any two norms in a finite-dimensional normed space X are equivalent.
Proof. If $\{a_k;\ k=1, 2, \ldots, n\}$ is a basis in X and

$$x = \sum_{k=1}^{n} \xi_k a_k$$

then it is easy to show that

$$\|x\|' := \sum_{k=1}^{n} |\xi_k|$$

is a norm (i.e. the axioms (i)–(iii) in § 1.2.1 are satisfied). It follows from the previous theorem that

$$\|x\| \leqslant M \sum_{k=1}^{n} |\xi_k| \qquad x \in X$$

and

$$\sum_{k=1}^{n} |\xi_k| \leqslant \|\tau^{-1}\| \, \|x\|.$$

Hence

$$\frac{1}{\|\tau^{-1}\|} \sum_{k=1}^{n} |\xi_k| \leqslant \|x\| \leqslant M \sum_{k=1}^{n} |\xi_k|$$

which means that any norm is equivalent to $\| \, . \, \|'$. Finally it is obvious that if we have two norms in a linear space, both of which are equivalent to $\| \, . \, \|'$, then the two norms are equivalent.

It also follows from the foregoing theorem that *every finite-dimensional normed space is complete and every bounded closed subset in a finite-dimensional normed space is compact* (see Remark 1 after Theorem 1.7.1.3).

1.7.3. As is well known, a linear operator in a finite-dimensional linear space X with a given basis has the form of a matrix multiplication (see § 4.1.1). Moreover, it was shown in Example 1.4.2.1 that the matrix multiplication is a bounded operator if X is supplied with the norm $\| \, . \, \|_\infty$.

It is easy to show that if a linear operator T is bounded in $\| \, . \, \|_\infty$ then it is bounded with respect to every norm that is equivalent to $\| \, . \, \|_\infty$. Indeed, if $\| \, . \, \|$ and $\| \, . \, \|_\infty$ are equivalent, i.e.

$$K_1 \|x\|_\infty < \|x\| < K_2 \|x\|_\infty \qquad x \in X$$

for suitable K_1 and K_2, then

$$K_1 \|Tx\|_\infty < \|Tx\| < K_2 \|Tx\|_\infty \qquad x \in X$$

and hence

$$\|Tx\| < K_2 \|Tx\|_\infty < K_2 \|T\| \, \|x\|_\infty < \frac{K_2}{K_1} \|T\| \, \|x\|$$

which means that T is a bounded operator in the equivalent norm also with bound $\dfrac{K_2}{K_1} \|T\|$. Thus we have proved the following theorem.

1.7.3.1 Theorem. In a finite-dimensional normed space every linear operator is continuous.

1.8 Problems and notes

1.8.1. Complete the proofs considered to be easy or obvious in the text.

1.8.2. The *support* of a function f, defined in the n-dimensional Euclidean space E_n (particularly in the real line), is called the closure of the set

$$\{t: f(t) \neq 0; \quad t \in E_n\}.$$

A function φ with bounded support on the real line is called a *step function* if its domain can be divided into a finite number of subintervals so that on the inside of each subinterval, φ has a constant value (see figure 1.3). A function is *piecewise continuous* if it is the sum of a continuous function and a step function.

∘**1.8.3.** For which $\alpha > 0$ is the sequence

$$\left\{\frac{1}{n^\alpha}\right\}; \quad n = 1, 2, \ldots$$

(*a*) in l^2, (*b*) in l, (*c*) in l^∞?

∘**1.8.4.** The interval $[-1, +1]$ is divided into 2000 parts and those step functions that have a constant value in each of these subintervals are considered. Show that a finite-dimensional linear space of step functions is thereby obtained, and give a basis and hence the dimension of this space.

1.8.5. The set of functions

$$\{A \sin(t+\varphi); \; A \geqslant 0, \; 0 \leqslant \varphi < 2\pi\}$$

is a two-dimensional linear subspace of $L^2[0, 2\pi]$. Indeed,

$$A \sin(t+\varphi) = A \cos\varphi \sin t + A \sin\varphi \cos t.$$

Hence $\{\sin t, \sin(t-\pi/2)\}$ is a basis and

$$\tau: \; A \sin(t+\varphi) \to (A \cos\varphi, A \sin\varphi)$$

is a linear 1–1 mapping (operator) from this linear space onto the linear space of pairs of real numbers (see figure 1.7).

Representing each pair of real numbers as a geometrical vector of the plane in the usual way, the norm of $y = A \sin(t+\varphi)$ is equal to the absolute value of the corresponding vector, since

$$\frac{1}{\pi} \int_0^{2\pi} A^2 \sin^2(t+\varphi) \, dt = A^2.$$

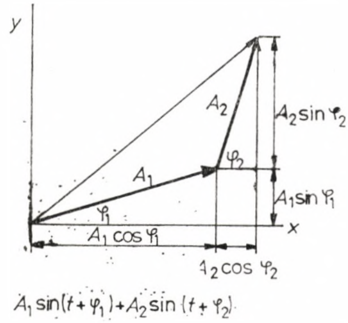

fig. 1.7

∘**1.8.6.** Using axioms (i)–(iii) of the norm in § 1.2.1, prove that

$$\|x-y\| \geqslant \|x\|-\|y\|.$$

1.8.7. Show that for a continuous function f in [0, 1],

$$\|f\|_1 \leqslant \|f\|_2 \leqslant \|f\|_\infty.$$

1.8.8. A sequence is called *finite* if all but a finite number of elements are zero. Show that for a finite sequence x,

$$\|x\|_1 \geqslant \|x\|_2 \geqslant \|x\|_\infty.$$

1.8.9. A generalisation of the usual norms for continuous functions of several variables is as follows.

Let $\mathscr{D}$ be a bounded closed domain of the n-dimensional Euclidean space; then

$$\|f\|_\infty = \sup\{|f(t)|;\ t\in\mathscr{D}\}$$

$$\|f\|_2 = \Big(\int_{\mathscr{D}} |f(t)|^2\, \mathrm{d}t\Big)^{1/2}$$

$$\|f\|_1 = \int_{\mathscr{D}} |f(t)|\, \mathrm{d}t.$$

Here, $\mathrm{d}t=\mathrm{d}t_1\ldots\mathrm{d}t_n$ and $\int_{\mathscr{D}}$ is the multiple integral.

1.8.10. The common generalisation of the norms $\|\,.\,\|_1$ and $\|\,.\,\|_2$ is the following.

(*a*) For functions,

$$\|x\|_p := \Big(\int_a^b |x(t)|^p\, \mathrm{d}t\Big)^{1/p} \qquad 1 \leqslant p < \infty.$$

(*b*) For sequences,

$$\|x\|_p := \Big(\sum_{k=1}^{\infty} |x_k|^p\Big)^{1/p} \qquad 1 \leqslant p < \infty.$$

Prove that for continuous functions in the finite closed interval $[a, b]$ and for finite sequences,

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

1.8.11. Prove that the continuous functions in $[0, 1]$ form a *Banach space* with the norm

$$\|f\| := \|f\|_2 + \|f\|_\infty.$$

○**1.8.12.** Which of the following functionals p are norms?

(*a*) On the space of functions with continuous derivatives in $[0, 1]$,

$$p(x) := \sup\left\{\left|\frac{\mathrm{d}}{\mathrm{d}t}\, x(t)\right|;\ t\in[0, 1]\right\}.$$

(*b*) On the linear space of $n\times n$ (quadratic) matrices,

$$p(\boldsymbol{A}) := \sup\Big\{\sum_{k=1}^{n} |a_{ik}|;\ i = 1, 2, \ldots, n\Big\}.$$

(*c*) On the linear space of continuous functions of two variables in $[0, 1]\times[0, 1]$,

$$p(\boldsymbol{K}) := \sup\Big\{\int_0^1 |K(t, \tau)|\,\mathrm{d}\tau;\ t\in[0, 1]\Big\}.$$

(*d*) On the space of complex functions, analytic in the open unit disc $\{z\colon |z|<1\}$ and continuous on the circle $|z|=1$,

$$p(f) := \left(\sum_{k=0}^{\infty}\left|\frac{f^{(k)}(0)}{k!}\right|^2\right)^{1/2}$$

where $f^{(k)}$ is the kth derivative.

1.8.13. If $\|\cdot\|_1, \|\cdot\|_2, \ldots, \|\cdot\|_n$ are norms on a certain linear space X, then

$$p(x) = \max\{\|x\|_k;\ \ k = 1, 2, \ldots, n\}$$

and

$$q(x) = \sum_{k=1}^{n} \|x\|_k$$

are also norms on X.

1.8.14. A geometrical meaning of the boundedness (continuity) of a linear operator $T\colon X\to Y$ is the following. There is a sphere with centre θ in Y such that the image of the unit sphere in X is contained in this sphere. The radius of the smallest such sphere in Y is just $\|T\|$.

1.8.15. It was proved in Theorem 1.4.3.3 that the linear operators $\{T\colon X\to Y\}$ form a normed space. What is the condition for $\{T\colon X\to Y\}$ to be a Banach space? Answer: if Y is a Banach space then $\{T\colon X\to Y\}$ is also a Banach space.

Proof. If $\{T_n\}$ is a Cauchy sequence, then $\{T_n x\}$ is even more a Cauchy sequence in Y for every $x\in X$; hence $y=\lim_n T_n x$ exists since Y is supposed to be complete. Define

$$Tx := \lim_n T_n x \qquad x\in X.$$

It will be proved that T is bounded and $\|T_n-T\|\to 0$.

If $\|x\|\leq 1$, then for every $\varepsilon>0$,

$$\|T_n x-Tx\| = \|(T_n x-T_m x)+(T_m x-Tx)\| \leq \|T_n x-T_m x\| + \|T_m x-Tx\| \leq \|T_n-T_m\|+\|T_m x-Tx\| < \varepsilon$$

if $n, m>N(\varepsilon/2)$ and $m>N(\varepsilon/2;\, x)$.

It follows that for $n>N(\varepsilon/2)$

$$\sup\{\|T_n x-Tx\|;\ \|x\| = 1\} < \varepsilon$$

which means that T_n-T is a bounded operator for $n>N(\varepsilon/2)$ and $T_n-T\to 0$.

Since $T=T_n-(T_n-T)$, T is also bounded.

1.8.16. Consider the linear space X of *finite sequences* (see § 1.8.8) and the linear operator

$$Tx := \{\sum_k a_{ik}x_k;\ i = 1, 2, \dots\} \qquad x\in X$$

X being supplied with the norm $\|\,.\,\|_1$; what is the condition on the 'double' infinite sequence

$$\{a_{ik}\} \qquad i = 1, 2, \dots,\ k = 1, 2, \dots$$

called the infinite matrix, that T be a bounded linear operator on X? What is the norm of the operator T?

1.8.17. Is the operator

$$Tf := \int_0^1 K(t, \tau)f(\tau)\,\mathrm{d}\tau$$

continuous in $L_0[0, 1]$ if K is a continuous function in $[0, 1]\times[0, 1]$? What is the norm of this operator?

∘**1.8.18.** Is the operator

$$Tf := \int_0^1 K(t, \tau) f(\tau)\, d\tau$$

bounded in $L_0^2[0, 1]$ if K is a continuous function in $[0, 1]\times[0, 1]$?

1.8.19. What is the condition for the corresponding infinite matrix that the operator

$$Tx := \{\sum_{k=1}^{\infty} a_{ik} x_k;\ i = 1, 2, \ldots\}$$

be bounded in the l^2-space?

∘**1.8.20.** The operator of the form

$$Tf := \int_0^t K(t-\tau) f(\tau)\, d\tau$$

is called the *convolution* operator and $K=K(t)$ is called the *kernel.* Prove that the sum and the product of convolution operators are also convolution operators.

1.8.21. Recall that if $T\colon X\to Y$ is a mapping then $T^{-1}\colon Y\to X$ defined by

$$T^{-1}\colon Tx \to x$$

is called an inverse mapping. A linear 1–1 operator T is called *invertible* if the inverse mapping T^{-1} is a bounded linear operator defined for every $y\in Y$.

By the same considerations as in the proof of Theorem 1.7.2.1 it can be shown that if T is a bounded 1–1 linear operator from the n-dimensional Euclidean space *onto* a Banach space Y then T is invertible, i.e. *the inverse is a bounded operator.*

The well-known *Banach Inverse Mapping Theorem* says that this remains true if the domain of T is any Banach space X and not only an n-dimensional Euclidean space.

1.8.22. If F is a contractive mapping and G is another mapping of the Banach space X with the commutation property

$$GF = FG$$

then $x_0=F(x_0)$ implies $x_0=G(x_0)$.

The proof of this important observation is easy. If $x_0=F(x_0)$ then $G(x_0)=GF(x_0)=FG(x_0)$ and hence $G(x_0)$ is also the solution of the equation

$$x = F(x).$$

But the solution of this equation is unique, since F is contractive; hence $x_0=G(x_0)$.

1.8.23. Prove that if F^n, the nth iterate of F, is contractive in $\mathscr{D}$ for $n>n_0$ and the condition 1.3.2.1 (*a*) is satisfied then the recursive sequence

$$x_0\in\mathscr{D} \qquad x_{n+1} = F(x_n)$$

converges to the solution of $x=F(x)$ although the mapping F is not contractive. (Is the solution unique?)

1.8.24. It was shown in Example 2 of § 1.3.2, that the operator

$$F(x(t)) := f(t)+\lambda \int_0^t K(t-\tau)x(\tau)\,d\tau$$

is a contractive mapping if $|\lambda|$ is small enough.

Applying the considerations in 1.8.22 and 1.8.23, it turns out that the recursive sequence

$$x_0(t) = f(t) \qquad x_{n+1}(t) = f(t)+\lambda \int_0^t K(t-\tau)x_n(\tau)\,d\tau \qquad n = 0, 1, 2, \ldots$$

tends to the solution of the corresponding convolution equation for *any* λ.

By applying 1.8.22 and 1.8.23 to the iterative solution of the differential equation in Example 1 of § 1.3.2, a more general condition for the existence of the solution is obtained.

∘**1.8.25.** Consider the operator equation

$$\lambda x-Tx = f$$

where T is a bounded linear operator of a Banach space B and $f\in B$. Now

$$F(x) = \frac{1}{\lambda}(f+Tx)$$

is a contractive operator if $|\lambda|$ is large enough. More particularly,

$$\|F(x)-F(z)\| = \left\|\frac{1}{\lambda}T(x-z)\right\| \leqslant \left|\frac{1}{\lambda}\right| \|T\|\,\|x-z\|$$

and hence F is contractive if $\|T\|<|\lambda|$.

In this case the recursive sequence corresponding to F, given by Theorem 1.3.2.2, is the following:

$$x_1 = \frac{1}{\lambda}(f+Tf) \qquad \text{if} \quad x_0 = f$$

$$x_2 = \frac{1}{\lambda}(f+Tx_1) = \frac{1}{\lambda}f+\frac{1}{\lambda^2}Tf+\frac{1}{\lambda^2}T^2f$$

$$x_3 = \frac{1}{\lambda}(f+Tx_2) = \frac{1}{\lambda}f+\frac{1}{\lambda^2}Tf+\frac{1}{\lambda^3}T^2f^2+\frac{1}{\lambda^3}T^3f^3$$

$$\vdots$$

$$x_{n+1} = \sum_{k=0}^{n}\frac{1}{\lambda^{k+1}}T^kf+\frac{1}{\lambda^{n+1}}T^{n+1}f$$

and hence the solution of the operator equation is given in the form of the infinite series

$$x = \sum_{k=0}^{\infty}\frac{1}{\lambda^{k+1}}T^kf$$

called the *Neumann series.*

Following the proof of 1.3.2.2, show that the Neumann series is convergent for $|\lambda|>\|T\|$.

1.8.26. An easy application of the above is the following iterative solution of a system of linear algebraic equations, called the *Gauss–Seidel method.* The method is applicable if the number of equations is equal to the number of unknowns. Consider the matrix form

$$\mathbf{Ax} = \mathbf{b}. \qquad (*)$$

It can be supposed that the diagonal elements $a_{kk}\neq 0$; $k=1, 2, \ldots, n$; divide the kth equation by a_{kk}; a system is then obtained which is equivalent to the original one and is in the form

$$(\mathbf{E}+\mathbf{B})\mathbf{x} = \mathbf{b}$$

where the diagonal elements of $\mathbf{B}$ are zeros. Now *the recursive sequence*

$$\mathbf{x}_0 = \mathbf{b} \qquad \mathbf{x}_{n+1} = \mathbf{b}-\mathbf{Bx}_n$$

converges to the solution of the system of linear algebraic equations $(*)$ *if*

$$F(\mathbf{x})\colon \mathbf{b}-\mathbf{Bx}$$

is a contractive mapping.

Since on the space of n-tuples of complex numbers, every norm is equivalent (from § 1.7.2) we choose the most convenient one and state that

$$\|F(\mathbf{x})-F(\mathbf{z})\|_\infty = \|\mathbf{B}(\mathbf{x}-\mathbf{z})\|_\infty \leqslant q\|x-z\|_\infty$$

where $0<q<1$. In fact, we have seen in Example 1 of § 1.4.2 that

$$\|\mathbf{B}(\mathbf{x}-\mathbf{z})\|_\infty \leqslant \sup_i \sum_{k=1}^{n} |b_{ik}| \, \|x-z\|_\infty$$

and hence if

$$\sum_{k=1}^{n} |b_{ik}| < 1 \qquad i = 1, 2, \ldots, n$$

then F is contractive and the Gauss–Seidel method is convergent.

For the matrix $\mathbf{A}$ of the original system of equations this condition can be formulated as follows:

$$|a_{ii}| > \sum_{k\neq i} a_{ik} \quad i = 1, 2, \ldots, n \qquad (**)$$

since $b_{ik}=a_{ik}/a_{ii}$ for $i\neq k$.

1.8.27. Based on the considerations in 1.8.26, the Gauss–Seidel method is as follows.

(i) Divide the kth equation by $a_{kk}\neq 0$ and substitute the diagonal elements by zeros to obtain the matrix $\mathbf{B}$.

(ii) Form the recursive sequence of n-tuples of numbers

$$\mathbf{x}_{n+1} = \mathbf{b}-\mathbf{B}\mathbf{x}_n \qquad (\mathbf{x}_0 = \mathbf{b})$$

up to $\|\mathbf{x}_{n+1}-\mathbf{x}_n\|<\varepsilon$.

Think about the following modifications.

(*a*) Using the norms $\|\,.\,\|_1$ or $\|\,.\,\|_2$ do you obtain a more convenient or weaker condition than $(**)$?

(*b*) Do you find cases when $(**)$ is not satisfied for $\mathbf{A}$ but it is satisfied for $\mathbf{A}^n$, $n>n_0$?

○**1.8.28.** On the basis of the Contractive Mapping Theorem and 1.8.25 prove that for any bounded linear operator T the operator

$$\lambda E-T$$

(E is the identity operator) is invertible if $|\lambda|$ is large enough.

Give the smallest number C such that $\lambda E-T$ is invertible for $|\lambda|>C$.

1.8.29. Show that in the Banach space $C[-\pi, +\pi]$, the bounded sequence

$$x_n(t) = \sin nt$$

has no convergent subsequence.

∘**1.8.30.** Let C be the linear space of bounded continuous functions on the real line with the usual norm $\|\,.\,\|_\infty$. Consider C_{00}, the continuous functions with bounded support, as a subspace of $L_0^2(-\infty, +\infty)$ or C. Prove or disprove the following statements:

(i) C_{00} is dense in $L_0^2(-\infty, +\infty)$;
(ii) C_{00} is dense in C.

1.8.31. If $\mathscr{S}$, $\mathscr{M}$ and X all have the same norm, and

$$\mathscr{S} \subset \mathscr{M} \subset X$$

then $\mathscr{S}$ *is dense in* $\mathscr{M}$ if $\overline{\mathscr{S}} \supseteq \mathscr{M}$. Here $\overline{\mathscr{S}}$ is the closure of $\mathscr{S}$ in X. In particular, if $\mathscr{M}$ is a closed subset and $\mathscr{S}$ is dense in $\mathscr{M}$ then $\overline{\mathscr{S}} = \mathscr{M}$.

1.8.32. If f is a real-valued function on a normed space X and f is continuous on a compact subset $\mathscr{M}$, then f takes its minimum and maximum on $\mathscr{M}$, i.e. there exists $m_0 \in \mathscr{M}$ such that

$$f(m_0) = \sup \{f(m);\; m \in \mathscr{M}\}.$$

To prove this, we first suppose that f is unbounded on $\mathscr{M}$; there then exists a sequence $\{x_n\}$; $x_n \in \mathscr{M}$ such that $f(x_n) > n$ and, since $\mathscr{M}$ is compact, a convergent subsequence $\{x_{n_i}\}$ can be chosen and $x_{n_i} \to x \in \mathscr{M}$. Thus we find that $f(x_{n_i}) \to f(x)$ and $f(x_{n_i}) > n_i$, which is a contradiction.

Now let M be the least upper bound for $\{f(x);\; x \in \mathscr{M}\}$; there then exists $\{x_k\}$ such that

$$f(x_k) > M - 1/k \qquad k = 1, 2, \ldots.$$

Again there is a convergent subsequence $\{x_{k_i}\}$ since $\mathscr{M}$ is compact and if $x_{k_i} \to x_0$, then $x_0 \in \mathscr{M}$ since a compact subset is closed and

$$M \geqslant f(x_0) > M - 1/k \qquad k = 1, 2, \ldots.$$

This means that

$$f(x_0) = M = \sup \{f(m);\; m \in \mathscr{M}\}.$$

1.8.33. For a *non-compact* bounded, closed subset $\mathscr{M}$ we can find a continuous function f on X such that it does not take its minimum on $\mathscr{M}$. For example,

$$\mathscr{M} = \{x\colon x(0) = 0;\; x(1) = 1;\; \|x\| \leqslant 1\}$$

is a closed subset of C[0, 1] and

$$f(x) := \int_0^1 |x(t)|\, dt$$

is a continuous function on C[0, 1]. f does not take its minimum on $\mathcal{M}$. Indeed,

$$t^n \in \mathcal{M} \qquad \text{and} \qquad f(t^n) := \int_0^1 t^n\, dt = \frac{1}{n+1} \qquad n = 1, 2, \ldots$$

and hence

$$\inf \{f(x):\ x \in \mathcal{M}\} = 0.$$

On the other hand, there is obviously no continuous function $x = x(t)$ satisfying the following conditions:

$$x(0) = 0 \qquad x(1) = 1 \qquad \int_0^1 |x(t)|\, dt = 0.$$

1.8.34 There is a generalisation of 1.8.32. If $F\colon X \to Y$ is a continuous mapping, where X and Y are normed spaces and $\mathcal{M} \subset X$ is compact, then there exists $x_0 \in \mathcal{M}$ such that

$$\|F(x_0)\| = \sup \{\|F(x)\|;\ x \in \mathcal{M}\}.$$

To prove this we consider $f(x) := \|F(x)\|$ and apply 1.8.32.

2

The Geometry of Hilbert Spaces

2.1 Scalar product

In the geometric vector space the norm is derived from the scalar product; however, until now the norm has been defined directly. Moreover, in the geometric vector space geometric concepts of a different character, such as orthogonality and projection, can be expressed by a scalar product.

Our next subject for discussion is the 'geometry' of linear spaces, in which a scalar product is defined in a certain axiomatic way and the norm is derived from this scalar product as in the geometric vector space. We shall see that the normed spaces thus obtained have richer structure and are 'more similar' to the geometric vector space that those not having this property.

2.1.1. The scalar product of vectors a and b in the geometric vector space is defined by

$$(a|b) := \|a\|, \|b\| \cos \gamma$$

where $\| . \|$ is the absolute value of the vector and γ is the angle between a and b. It is easy to check that the following properties are satisfied for any vectors x, y, z and scalar λ.

(i) $(x|y) = (y|x)$

(ii) $(x+y|z) = (x|z)+(y|z)$

(iii) $(\lambda x|y) = \lambda(x|y)$

(iv) $(x|x) \geqslant 0$ and $(x|x) = 0$ if and only if $x = \theta$.

Notice that the properties (i)–(iv), called the *axioms of the scalar product*, are more important than its geometrical meaning and this is the motivation for what follows.

2.1.1.1 Definition. A mapping

$$(x, y) \mapsto (x|y)$$

from the ordered pairs (x, y) of elements of a linear space X into the field of scalars (complex numbers) satisfying axioms (ii)–(iv) and

(i)′ $(x|y)=\overline{(y|x)}$

where $\overline{(y|x)}$, the complex conjugate of $(y|x)$, is called a *scalar product* of x and y.

Remark. In the case of geometric vector space and in any real linear space, axioms (i) and (i)′ are the same, since in these cases the scalars are real numbers. However, in the case of a complex linear space, i.e. when the scalars are complex numbers, it is necessary to alter axiom (i) since there is a contradiction between axioms (i) and (iv) in this case.

Indeed, using axioms (i) and (iii),

$$(\mathrm{i}x|\mathrm{i}x) = \mathrm{i}(x|\mathrm{i}x) = \mathrm{i}(\mathrm{i}x|x) = \mathrm{i}^2(x|x) = -(x|x)$$

for any $x \neq \theta$ and hence at least one of the values $(\mathrm{i}x|\mathrm{i}x)$ and $(x|x)$ is not positive, in contradiction of axiom (iv). However, using axiom (i)′ instead of (i),

$$(\mathrm{i}x|\mathrm{i}x) = \mathrm{i}(x|\mathrm{i}x) = \mathrm{i}\overline{(\mathrm{i}x|x)} = \mathrm{i}(-\mathrm{i})(x|x) = (x|x)$$

and hence axiom (iv) is not violated.

2.1.1.2 Definition. If a scalar product is defined in a linear space X then X is called a *scalar product space* or *pre-Hilbert space*.

In the geometric vector space the norm (i.e. the absolute value) of a vector x is expressed by the scalar product

$$\|x\| = (x|x)^{1/2}.$$

In the next section it will be shown that the axioms (i)–(iii) of the norm in § 1.2.1 are satisfied by $\|x\| := (x|x)^{1/2}$ also in any scalar product space; thus scalar product spaces are (special cases of) normed spaces.

2.1.2. In the geometric vector space,

$$|(x|y)| \leqslant \|x\|\,\|y\|$$

for any pair of vectors x, y. We shall prove this inequality, which is known as the Cauchy–Schwarz inequality, purely from the axioms (i)′–(iv) of the scalar product without reference to any geometry.

2.1.2.1 Theorem. In a pre-Hilbert space $\mathscr{H}$,

$$|(x|y)| \leqslant \|x\|\,\|y\| \qquad x, y \in \mathscr{H}$$

where $\|x\| := (x|x)^{1/2}$ and $\|y\| := (y|y)^{1/2}$.

Proof. If $x=\theta$ or $y=\theta$, then obviously the equality holds. If $x\neq\theta$ and $y\neq\theta$ then for any scalar λ we have

$$\begin{aligned}0 \leqslant (x-\lambda y|x-\lambda y) &= (x|x)+|\lambda|^2(y|y)-(\lambda y|x)-(x|\lambda y)\\ &= (x|x)+|\lambda|^2(y|y)-(\lambda y|x)-\overline{(\lambda y|x)}\\ &= (x|x)+|\lambda|^2(y|y)-2\,\mathrm{Re}\,(\lambda(y|x)).\end{aligned}$$

For $\lambda=(x|y)/(y|y)$ we have

$$0 \leqslant (x|x)-\frac{|(y|x)|^2}{(y|y)}$$

and the Cauchy–Schwarz inequality is obtained by suitable rearrangement.

Remark. The motivation for the proof is the solution of the following optimisation problem.

Choose λ for fixed $x, y\in\mathscr{H}$ in the real scalar product space $\mathscr{H}$ so that the 'distance' $\|x-\lambda y\|$ is minimal (see figure 2.1). Solution:

$$\begin{aligned}\|x-\lambda y\|^2 := (x-\lambda y|x-\lambda y)\\ = (x|x)-2\lambda(x|y)+\lambda^2(y|y)\end{aligned}$$

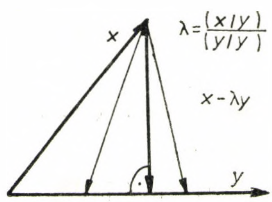

fig. 2.1

since $\mathscr{H}$ is a ***real*** scalar product space. Hence

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(x-\lambda y|x-\lambda y) = 2\lambda(y|y)-2(x|y)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}(x-\lambda y|x-\lambda y) = 0 \qquad \text{if } \lambda = \frac{(x|y)}{(y|y)}.$$

It follows that the 'distance' $\|x-\lambda y\|$ is minimal if

$$\lambda = \frac{(x|y)}{(y|y)}.$$

2.1.2.2 Theorem. $\|x\|:=(x|x)^{1/2}$ is a norm.

Proof. It is obvious that axioms (i) and (iii) of the norm are satisfied. To check axiom (ii), known as the triangle inequality, we have

$$\|x+y\|^2 := (x+y|x+y) = (x|x)+(y|y)+(x|y)+(y|x)$$
$$\leqslant \|x\|^2+\|y\|^2+2\|x\|\,\|y\| = (\|x\|+\|y\|)^2$$

by the Cauchy–Schwarz inequality, and hence

$$\|x+y\| \leqslant \|x\|+\|y\|.$$

As we have seen in 1.2.2.6, multiplication with a scalar and addition are continuous operations in a normed space. In pre-Hilbert spaces the scalar product is also continuous.

2.1.2.3 Theorem. If $x_n \to x$ and $y_n \to y$, then

$$(x_n|y_n) \to (x|y).$$

Proof.

$$|(x_n|y_n)-(x|y)| \leqslant |(x_n|y_n)-(x_n|y)|+|(x_n|y)-(x|y)|$$
$$\leqslant \|x_n\|\,\|y_n-y\|+\|x_n-x\|\,\|y\| \leqslant K(\|y_n-y\|+\|x_n-y\|)$$

where K is the common upper bound for $\|y\|$ and the convergent sequences $\{\|x_n\|\}$ and $\{\|y_n\|\}$.

It follows that if $\|y_n-y\|\to 0$ and $\|x_n-x\|\to 0$ then

$$|(x_n|y_n)-(x|y)| \to 0.$$

By Theorem 1.6.1.2, every normed space has a completion. Is the completion of a scalar product space again a scalar product space? The answer is yes.

In fact, if $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences of elements of a pre-Hilbert space $\mathscr{H}$, then

$$|(x_n|y_n)-(x_k|y_k)| = |(x_n|y_n-y_k)+(x_n-x_k|y_k)|$$
$$\leqslant |(x_n|y_n-y_k)|+|(x_n-x_k|y_k)| \leqslant \|x_n\|\,\|y_n-x_k\|+\|x_n-x_k\|\,\|y_k\|$$
$$\leqslant K(\|y_n-y_k\|+\|x_n-x_k\|)$$

where K is a common upper bound of the Cauchy sequences $\{\|x_n\|\}$ and $\{\|y_n\|\}$; hence $\{(x_n|y_n)\}$ is a Cauchy sequence of complex numbers. Now, we define

$$(x|y) := \lim\,(x_n|y_n) \qquad (*)$$

for the elements x and y of the completion of $\mathscr{H}$ defined by the Cauchy sequences $\{x_n\}$ and $\{y_n\}$ according to § 1.6.1. The definition $(*)$ is consistent with the axioms of the scalar product and the completion process described in

§ 1.6.1. For example, if $\{x'_n\}$ and $\{y'_n\}$ are Cauchy sequences and

$$x'_n - x_n \to \theta \qquad y'_n - y_n \to \theta$$

(i.e. the same element is defined by $\{x'_n\}$ and $\{x_n\}$ (respectively $\{y'_n\}$ and $\{y_n\}$)) then $\lim (x'_n | y'_n) = \lim (x_n | y_n)$. We can summarise this as follows.

2.1.2.4 Theorem. The completion of a scalar product space is a scalar product space with the scalar product (*).

A complete scalar product space is called a *Hilbert space.*

Example 1. If $\{e_k;\ k=1, 2, 3\}$ is an orthonormed basis in the geometric vector space; i.e.

$$(e_i|e_j) = \begin{cases} 1 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$$

and

$$x = \xi_1 e_1 + \xi_2 e_2 + \xi_3 e_3$$

$$y = \eta_1 e_1 + \eta_2 e_2 + \eta_3 e_3$$

then

$$(x|y) = \xi_1 \eta_1 + \xi_2 \eta_2 + \xi_3 \eta_3$$

which is sometimes called the 'coordinate form' of the scalar product of the vectors. The next generalisation for n-tuples of complex numbers is motivated by this form.

If $x = \{x_k;\ k=1, 2, \ldots, n\}$ and $y = \{y_k;\ k=1, 2, \ldots, n\}$ then

$$(x|y) := \sum_{k=1}^{n} x_k \bar{y}_k$$

(where $\bar{y}_k$ is the complex conjugate of y_k) is a scalar product in the linear space of n-tuples of complex numbers. We emphasise the Cauchy–Schwarz inequality in this case,

$$\sum_{k=1}^{n} |x_k \bar{y}_k|^2 \leqslant \sum_{k=1}^{n} |x_k|^2 \sum_{k=1}^{n} |y_k|^2$$

and the scalar product space thus obtained is complete, and hence a Hilbert space, since the linear space in question is finite-(n-) dimensional. If we restrict ourselves to real numbers, then the Hilbert space thus obtained will also be called *n-dimensional Euclidean space.*

Example 2. In the l^2-space,

$$(x|y) := \sum_{k=1}^{\infty} x_k \bar{y}_k$$

is a scalar product and l^2 is a scalar product space since

$$\|x\|_2 := \sum_{k=1}^{\infty} |x_k|^2 = (x|x)^{1/2}.$$

We have only to prove that the infinite numerical series defining the scalar product is convergent. It follows from the Cauchy–Schwarz inequality of the previous example that

$$\Big|\sum_{k=m}^{n} x_k \bar{y}_k\Big|^2 \leqslant \sum_{k=n}^{m} |x_k|^2 \sum_{k=n}^{m} |y_k|^2$$

and hence, if

$$\sum_{k=1}^{\infty} |x_k|^2 < \infty$$

and

$$\sum_{k=1}^{\infty} |y_k|^2 < \infty$$

then the series

$$\sum_{k=1}^{\infty} x_k \bar{y}_k$$

must be convergent.

l^2 is complete, i.e. a Hilbert space (the proof is not trivial and this is equivalent to the Riesz–Fischer Theorem).

Example 3. In the space $L_0^2(a, b)$,

$$(f|g) := \int_a^b f(t)\overline{g(t)}\, \mathrm{d}t$$

is a scalar product and L_0^2 is a scalar product space since

$$\|f\|_2 := \int_a^b |f(t)|^2\, \mathrm{d}t = (f|f)^{1/2}.$$

According to § 1.2.3 and, in particular, the considerations following Definition 1.2.3.3, L_0^2 is not complete. The completion is the L^2-space introduced in Example 6 of § 1.6.1.

Example 4. The linear space H_0^2 of complex functions analytic in the disc $\{z\colon |z|<1\}$ and continuous on the boundary $\{z\colon |z|=1\}$ is a scalar product space with

$$(f|g) = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} f(z)\overline{g(z)}\frac{1}{z}\, \mathrm{d}z.$$

There is a close connection between H_0^2 and $L_0^2(0, 2\pi)$; substituting $z=e^{it}$, we obtain

$$\frac{1}{2\pi i}\oint_{|z|=1} f(z)\overline{g(z)}\frac{1}{z}\,dz = \frac{1}{2\pi}\int_0^{2\pi} f(e^{it})\overline{g(e^{it})}\,dt \qquad (*)$$

and hence it is not surprising that the completion of H_0^2, called H^2 (Hardy) space, consists of those analytic functions in $\{z: |z|<1\}$ for which $f(e^{it})\in L^2[0, 2\pi]$.

A very important property of the H_0^2-space is the following.

If $f_n\to f$ *in the scalar product space* H_0^2 then $f_n(z)\to f(z)$ *uniformly in every closed disc* $\{z: |z|\leqslant r<1\}$. To prove this we use the estimate

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|e^{it}-z_0|^2}\,dt < \frac{1}{(1-r)^2} \qquad \text{if } |z_0|\leqslant r$$

(see figure 2.2). Now by the Cauchy integral formula and the Cauchy–Schwarz inequality, in $L_0^2[0, 2\pi]$,

$$|f(z_0)-f_n(z_0)| = \frac{1}{2\pi}\left|\oint_{|z|=1}\frac{f(z)-f_n(z)}{z-z_0}\,dz\right|$$

$$= \frac{1}{2\pi}\left|\int_0^{2\pi}(f(e^{it})-f_n(e^{it}))\frac{e^{it}}{e^{it}-z_0}\,dt\right|$$

$$\leqslant \frac{1}{(1-r)^2}\,\|f-f_n\|_2.$$

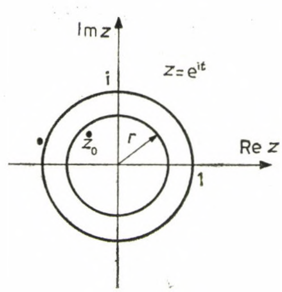

fig. 2.2

Hence the assertion on the uniform convergence is clear, considering the connection $(*)$ between H_0^2 and $L_0^2[0, 2\pi]$.

Example 5. If

$$Dy := y''+b(t)y'+c(t)y$$

where $b=b(t)$ and $c=c(t)$ are continuous functions on a closed finite interval $[a, b]$ and y' and y'' refer to the first and second derivatives, respectively, of $y=y(t)$, then $\mathscr{H}_D$ is the linear space of functions $y=y(t)$ for which Dy is a continuous function on $[a, b]$. $\mathscr{H}_D$ is a pre-Hilbert space with the scalar product

$$(y|z) := y(a)\overline{z(a)}+y'(a)\overline{z'(a)}+\int_a^b (Dy)(\tau)(Dz)(\tau)\,d\tau$$

and the completion of $\mathscr{H}_D$ is a special type of Sobolev space.

Example 6. If (Ω, A, P) is a probability space, then the random variables $\xi=\xi(\omega)$ with finite variance form a Hilbert space $L^2(\Omega, A, P)$ with scalar product

$$(\xi|\eta) := M(\xi, \eta) \qquad \xi, \eta\in L^2(\Omega, A, P)$$

where M is the mean value of the random variable. The important closed linear subspace of $L^2(\Omega, A, P)$ consists of random variables ξ with zero mean, i.e. $M(\xi)=0$. In the language of Hilbert space geometry, this is the closed linear subspace, which is orthogonal to the constant random variables (see Definition 2.2.1.1).

For the applications of Hilbert space theory to probability problems see Lamperti (1977).

2.1.3. There are normed spaces whose norm cannot be generated as $\|x\| :=(x|x)^{1/2}$ from a scalar product. An important example of such a space is $C[a, b]$. To see this, let us consider the *parallelogram law*

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

which is valid for any scalar product space, as can easily be verified. The parallelogram law is not satisfied for $x(t)=t^2$ and $y(t)=1$ in $C[0, 1]$. Indeed, $\|y\|_\infty=1$ and $\|x\|_\infty=\sup\{t^2;\ t\in[0, 1]\}=1$; moreover,

$$\|x+y\|_\infty = \sup\{t^2+1;\ t\in[0, 1]\} = 2$$

and

$$\|x-y\|_\infty = \sup\{1-t^2;\ t\in[0, 1]\} = 1.$$

2.2 Orthogonal systems (sequences)

In the geometric vector space, every vector is the linear combination of fixed sets of three orthogonal vectors and any finite-dimensional linear space possesses a finite basis, i.e. n fixed vectors such that every vector of the linear space is a linear combination of these fixed vectors. In certain infinite-dimensional spaces a fixed infinite sequence of elements can be found with similar properties (see, for example, Remark 1 in § 1.5.2).

The main subject of this section is the construction of an infinite basis in a separable Hilbert space or even in a separable scalar product space, which is orthogonal in the sense of Definition 2.2.1.2.

2.2.1. The orthogonality is defined in a pre-Hilbert space in the following natural way.

2.2.1.1 Definition. The vectors x and y in the pre-Hilbert space $\mathscr{H}$ are called *orthogonal* if $(x|y)=0$.

Remark. In a real pre-Hilbert space the angle α between x and y can be defined by the formula

$$\frac{(x|y)}{\|x\|\,\|y\|} = \cos\alpha$$

but we need only consider the case when $\cos\alpha=0$, i.e. when x and y are orthogonal.

2.2.1.2 Definition. A sequence $\{e_k\}$ in a scalar product space $\mathscr{H}$ is called *orthogonal* if

$$(e_i|e_j) = 0 \qquad \text{if} \quad i \neq j.$$

If $\|e_k\|=1$ for $i=1, 2, \dots$ is also satisfied, then $\{e_k;\ k=1, 2, \dots\}$ is called *orthonormal.* An orthogonal or orthonormed sequence is also called an orthogonal or orthonormal *system.*

We begin with the following *minimum problem.* Let $\{e_k\}$ be an orthonormal system, n a fixed integer, and x an element of $\mathscr{H}$; determine the scalars γ_k; $k=1, 2, \dots, n$ in such a way that the 'distance'

$$\left\|x-\sum_{k=1}^{n}\gamma_k e_k\right\|$$

is minimal. First we give the solution for *real* scalar product space. In this case,

$$\left\|x-\sum_{k=1}^{n}\gamma_n e_n\right\|^2 = \left(x-\sum_{k=1}^{n}\gamma_k e_k \,\middle|\, x-\sum_{k=1}^{n}\gamma_k e_k\right)$$

$$= (x|x)-2\sum_{k=1}^{n}\gamma_k(x|e_k)+\sum_{k=1}^{n}\gamma_k^2.$$

For the minimum of this quadratic form

$$\frac{\partial}{\partial\gamma_k}\left\|x-\sum_{k=1}^{n}\gamma_k e_k\right\|^2 = 2\gamma_k-2(x|e_k) = 0 \qquad k = 1, 2, \ldots, n;$$

it follows that the desired minimum is obtained if and only if

$$\gamma_k = (x|e_k) \qquad k = 1, 2, \ldots .$$

In the case of a complex space $\mathscr{H}$ the solution is the same but a more lengthy calculation is required, since in this case we have to seek the minimum of a quadratic form of *2n real variables.*

We conclude that the minimum of the 'distance'

$$\left\|x-\sum_{k=1}^{n}\gamma_k e_k\right\|$$

is obtained if and only if $\gamma_k=(x|e_k)$; $k=1, 2, \ldots,$ and then

$$\left\|x-\sum_{k=1}^{n}(x|e_k)e_k\right\|^2 = \|x\|^2-\sum_{k=1}^{n}|(x|e_k)|^2. \tag{*}$$

Remark. The consideration connected with the Cauchy–Schwarz inequality after Theorem 2.1.2.1 is the special case of this problem for $n=1$.

Example 1. If e_k is the infinite sequence whose kth element is 1 and all other elements are 0, then $e_k \in l^2$ and $\{e_k;\ k=1, 2, \ldots\}$ is an orthonormal system in l^2.

Notice that this $\{e_k;\ k=1, 2, \ldots\}$ is an immediate generalisation of the 'fundamental basis'

$$\begin{aligned} e_1 &= \{1, 0, \ldots, 0, \ldots, 0\} \\ e_2 &= \{0, 1, \ldots, 0, \ldots, 0\} \\ &\vdots \\ e_k &= \{0, 0, \ldots, 1, \ldots, 0\} \\ &\vdots \\ e_n &= \{0, 0, \ldots, 0, \ldots, 1\} \end{aligned}$$

in the linear space of n-tuples of complex numbers.

For $x=\{x_k;\ k=1, 2, \ldots\}$, $(x|e_k)=x_k$ and hence $\sum_{k=1}^{n}(x|e_k)e_k=\{x_1, x_2, \ldots$ $\ldots, x_n, 0, 0, \ldots\}$. Moreover, if $n\to\infty$ then

$$\|x\|-\sum_{k=1}^{n}|(x|e_k)|^2 \to 0 \tag{**}$$

in this case. This is therefore a trivial example of the results in this section.

Example 2. The sequence

$$\left\{\frac{1}{(2\pi)^{1/2}} e^{ikt};\ k = 0, \pm 1, \pm 2, \ldots\right\}$$

is an orthonormal system in $L_0^2[0, 2\pi]$. For $x \in L^2[0, 2\pi]$ and $e_k = e^{ikt}$,

$$(x|e_k) := \frac{1}{(2\pi)^{1/2}} \int_0^{2\pi} x(t) e^{-ikt} \, dt,$$

and

$$\sum_{k=-n}^{n} (x|e_k) e_k$$

is the nth partial sum of the (complex) Fourier series of $x \in L^2[0, 2\pi]$. So this is a well-known example but is by no means trivial.

It is also known that if $n \to \infty$ then $(**)$ also holds in this case.

Remark. On the basis of this example, the coefficients $(x|e_k)$; $k=1, 2, \ldots$ can be considered as the generalisation of the Fourier coefficients. These coefficients are therefore called *the Fourier coefficients of x with respect to the orthogonal system* $\{e_k;\ k=1, 2, \ldots\}$.

Example 3. The sequence

$$\left\{\frac{1}{\pi^{1/2}} \sin kt;\ k = 1, 2, \ldots\right\}$$

is an orthonormal system also in the *real* $L^2[-\pi, +\pi]$-space. Now

$$(x|e_k) := \frac{1}{\pi^{1/2}} \int_{-\pi}^{+\pi} x(t) \sin kt \, dt$$

but if $n \to \infty$ then $(**)$ is not fulfilled. More particularly, $(**)$ is fulfilled only if $x = x(t)$ is an odd function and in any other case

$$\left\{\|x\|^2 - \sum_{k=1}^{n} (x|e_k);\ n = 1, 2, \ldots\right\}$$

is just a decreasing sequence of positive numbers.

Example 4. The sequence $\{z^n;\ k=0, 1, 2, \ldots\}$ is an orthonormal system in H_0^2. Indeed,

$$(z^n|z^m) = \frac{1}{2\pi i} \oint_{|z|=1} z^n \bar{z}^m \frac{1}{z} \, dz = \frac{1}{2\pi i} \oint_{|z|=1} z^{n-m-1} \, dz$$

$$= \begin{cases} 1 & \text{if } n = m \\ 0 & \text{if } n \neq m \end{cases}$$

since $\bar{z}=1/z$ on the circle $|z|=1$. In this case,

$$(f|z^k) = \frac{1}{2\pi i}\oint_{|z|=1} f(z)\bar{z}^k \frac{1}{z}\,dz = \frac{1}{2\pi i}\oint_{|z|=1} \frac{f(z)}{z^{k+1}}\,dz = \frac{f^{(k)}(0)}{k!} \qquad k=0,1,2,\ldots$$

by the generalised Cauchy integral formula. It follows that $(f|z^k)$ is the kth coefficient of the Taylor expansion (at $z_0=0$) of the analytic function f.

Notice that this Taylor expansion is connected to the (complex) Fourier series of $f(e^{it})$ in the following way:

$$f(e^{ikt}) = \sum_{k=0}^{\infty} (f|z^k)\, e^{ikt}$$

and hence if $f\in H_0^2$ then the kth Fourier coefficient of $f(e^{ikt})$ is zero if $k<0$.

It follows, that $(**)$ holds in this case too if $n\to\infty$.

2.2.2. We now modify the problem posed in the previous subsection as follows.

For a given orthonormal system $\{e_k\}$, find those elements x of $\mathscr{H}$ that can be written in the form of an orthogonal series

$$x = \sum_{k=1}^{\infty} \xi_k e_k.$$

What are the coefficients ξ_k in this case?

If we suppose that x is in this form, then

$$(x|e_i) = \sum_{k=1}^{\infty} \xi_k (e_k|e_i) \qquad i=1,2,\ldots$$

by axioms (ii) and (iii) and the continuity of the scalar product; hence

$$\xi_k = (x|e_k) \qquad k=1,2,\ldots.$$

If we now apply also equation $(*)$ from § 2.2.1 the following result is obtained.

2.2.2.1 Theorem. (*a*) If $x\in\mathscr{H}$ can be given in the form of an orthogonal series, then

$$x = \sum_{k=1}^{\infty} (x|e_k) e_k. \tag{$*$}$$

(*b*) An element x of the pre-Hilbert space $\mathscr{H}$ can be given in the form $(*)$ if and only if

$$\|x\|^2 = \sum_{k=1}^{\infty} |(x|e_k)|^2.$$

Remark 1. It follows also from equation $(*)$ in § 2.2.1 that for every $x \in \mathscr{H}$ and orthonormal system $\{e_k\}$,

$$\sum_{k=1}^{n} |(x|e_k)|^2 \leqslant \|x\|^2$$

since the norm is a non-negative number.

Remark 2. The orthogonal series $(*)$ is sometimes called *the orthogonal expansion of* $x \in \mathscr{H}$ by the orthogonal system $\{e_k;\ k=1, 2, \ldots\}$.

2.2.3. What is the condition for an orthonormal system $\{e_k\}$ that § 2.2.1. $(*)$ holds *for every* $x \in \mathscr{H}$? This is a natural question following on from Theorem 2.2.2.1.

2.2.3.1 Definition. A sequence $\{a_k;\ k=1, 2, \ldots\}$ is called complete if $(x|a_k)=0$ for $k=1, 2, \ldots$ implies $x=\theta$.

2.2.3.2 Theorem. For every $x \in \mathscr{H}$,

$$x = \sum_{k=1}^{\infty} (x|e_k)e_k$$

if and only if the orthonormal system $\{e_k\}$ is complete.
Proof. The sequence $\{s_n\}$,

$$s_n = \sum_{k=1}^{n} (x|e_k)e_k$$

is a Cauchy sequence since

$$\|s_n - s_m\|^2 = \Big(\sum_{k=m}^{n} (x|e_k)e_k \Big| \sum_{k=m}^{n} (x|e_k)e_k\Big) = \sum_{k=m}^{n} |(x|e_k)|^2$$

and the series

$$\sum_{k=1}^{\infty} |(x|e_k)|^2$$

is convergent by Remark 1 following Theorem 2.2.2.1. Hence $\{s_n\}$ *has a limit in the completion of* $\mathscr{H}$ and

$$\Big(x - \sum_{k=1}^{\infty} (x|e_k)e_k \Big| e_i\Big) = (x|e_i) - (x|e_i) = 0 \qquad i = 1, 2, \ldots.$$

Now if we suppose that $\{e_k\}$ is complete then it follows from the foregoing equality that

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

and hence § 2.2.2. ($*$) holds.

Conversely, if $\{e_k\}$ is not complete, then there exists $x \neq \theta$ such that $(x|e_k)=0$ for $k=1, 2, \ldots$ and hence § 2.2.2. ($*$) does not hold for this $x \in \mathscr{H}$.

It is not easy to verify whether a given orthonormal system is complete or not. In §§ 2.14.11–14 we shall give theorems and examples related to this problem. Furthermore, we shall show in § 2.2.5 that any separable Hilbert space contains a *complete* orthonormal system.

2.2.4. A standard method for the construction of orthonormal systems, called the *Gram–Schmidt process*, proceeds as follows. Let $\{a_k;\ k=1, 2, \ldots\}$ be linearly independent (§ 1.1.1). The first member of the orthonormal system is

$$e_1 = \frac{a_1}{\|a_1\|}$$

and for the second member e_2,

$$z_2 = a_2 - \lambda_{21} e_1$$

where the scalar λ_{21} is determined by the condition

$$(z_2|e_1) = (a_2|e_1) - \lambda_{21} = 0$$

and hence $\lambda_{21} = (a_2|e_1)$. So, if $e_2 = z_2/\|z_2\|$ then $\{e_1, e_2\}$ is an orthonormal system with two elements (see figure 2.3 (*a*)).

For the third member, e_3,

$$z_3 = a_3 - \lambda_{31} e_1 - \lambda_{32} e_2$$

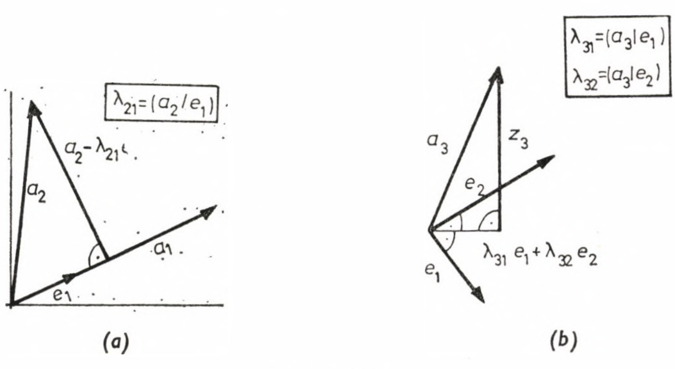

fig. 2.3

where the scalars λ_{31} and λ_{32} are determined by the conditions

$$(z_3|e_1) = (a_3|e_1) - \lambda_{31} = 0$$

$$(z_3|e_2) = (a_3|e_2) - \lambda_{32} = 0.$$

Hence $\lambda_{31} = (a_3|e_1)$ and $\lambda_{32} = (a_3|e_2)$. So if $e_3 = z_3/\|z_3\|$ then $\{e_1, e_2, e_3\}$ is an orthonormal system with three members (figure 2.3(*b*)).

Now if $e_1, e_2, \ldots, e_n$ have already been obtained, then for e_{n+1},

$$z_{n+1} = a_{n+1} - \sum_{k=1}^{n} \lambda_{n+1,k} e_k$$

where the scalars $\lambda_{n+1,k}$; $k = 1, 2, \ldots, n$ are determined by the condition

$$(z_{n+1}|e_j) = (a_{n+1} - \sum_{k=1}^{n} \lambda_{n+1,k} e_k | e_j) = (a_{n+1}|e_j) - \lambda_{n+1,j} = 0$$

and hence $\lambda_{n+1,k} = (a_{n+1}|e_k)$; $k = 1, 2, \ldots, n$. So if $e_{n+1} = z_{n+1}/\|z_{n+1}\|$, then $\{e_k;\ k = 1, 2, \ldots, n+1\}$ is an orthonormal system obtained from the linear space generated by the $n+1$ vectors a_k; $k = 1, 2, \ldots, n+1$.

Remark. In the Gram–Schmidt process, n linearly independent vectors a_k; $k = 1, 2, \ldots, n$ are converted into the n elements of an orthonormal system $\{e_k;\ k = 1, 2, \ldots\}$. In this process e_1 is the scalar multiple of a_1, e_2 is a linear combination of a_2 and e_1, e_3 is a linear combination of a_3, e_1 and e_2 and so on. Notice that the computation is organised in terms of the minimal number of vectors and operations.

Example 1. Using the Gram–Schmidt process for

$$1, t, t^2, \ldots, t^n, \ldots$$

in $L^2[-1, +1]$, a sequence of orthogonal polynomials known as Legendre polynomials is obtained, the nth element of which is of exactly $(n-1)$th degree; the first four members of this sequence are

$$1, \quad t, \quad \frac{1}{2}(3t^2 - 1), \quad \frac{1}{2}(5t^3 - 3t).$$

A general formula for the Legendre polynomials (the so-called Rodriguez formula) is

$$L_n(t) = \frac{1}{2^n n!} \frac{\mathrm{d}^n}{\mathrm{d}t^n} (t^2 - 1)^n \qquad n = 1, 2, \ldots$$

and a recursive formula is

$$L_0(t) = 1 \qquad L_1(t) = t$$

$$(n+1)L_{n+1}(t) = (2n+1)\,tL_n(t) - nL_{n-1}(t) \qquad n = 1, 2, \ldots.$$

Example 2. If the Gram–Schmidt process in l^2 is applied to the vectors

$$a_k = \{\alpha_1, \alpha_2, \ldots, \alpha_k, 0, 0, \ldots\}$$

where α_i; $i=1, 2, \ldots, k$ are arbitrary numbers, the orthonormal system described in Example 2.2.1.1 is obtained.

Example 3. Consider the analytic functions in a suitable open domain $\mathscr{D}$ for which

$$\iint_{\mathscr{D}} |f(z)|^2 \,\mathrm{d}x\,\mathrm{d}y < \infty.$$

Then

$$(f|g) := \iint_{\mathscr{D}} f(z)\overline{g(z)}\,\mathrm{d}x\,\mathrm{d}y$$

is a scalar product and a sequence of (complex) orthogonal polynomials can be obtained in this scalar product space if the Gram–Schmidt process is applied to the sequence $1, z, z^2, \ldots, z^n, \ldots$.

Further examples for orthogonal systems can be found in the next section.

2.2.5. A complete orthonormal system is also called an *orthonormal (orthogonal) basis* since it is a basis for the scalar product space $\mathscr{H}$ in the sense of § 1.1.1. or § 1.5.2 Remark 1. An important theoretical conclusion of the above is as follows.

2.2.5.1 Theorem. Every separable scalar product space contains a (finite or infinite) basis.

Proof. If $\{a_n;\ n=1, 2, \ldots\}$ is a countable dense subset of $\mathscr{H}$, then, applying the Gram–Schmidt process to $\{a_n;\ n=1, 2, \ldots\}$, a complete orthonormal system $\{e_k\}$; $k=1, 2, \ldots$ is obtained. In fact, if

$$(x|e_k) = 0 \qquad k = 1, 2, \ldots$$

for $x \in \mathscr{H}$, then we also have $(x|a_n)=0$; $n=1, 2, \ldots$ (see § 2.2.4 Remark) and hence $x=\theta$ by § 2.14.42.

Remark. It is not necessary for $\{a_n;\ n=1, 2, \ldots\}$ to be linearly independent. (What happens if $\{a_n;\ n=1, 2, \ldots\}$ is not linearly independent when the Gram–Schmidt process is applied?)

2.3 Some important orthonormal systems in the L^2-spaces

As in the case of the geometric vector space, there are many complete orthonormal systems in a given separable scalar product space. In this section, complete orthonormal systems in L^2-spaces that have proved to be useful in practice will be discussed.

2.3.1. Our first example is closely connected with the (complex) Fourier series expansion. In many cases we have only a sample of N elements

$$f\left(k\frac{2\pi}{N}\right); \quad k = 0, 1, 2, \ldots, N-1 \tag{$*$}$$

for $f \in L_0[0, 2\pi]$ and we shall show how we obtain an analogy of the Fourier series from this sample.

In this case we consider

$$c_{Nn} = \frac{1}{N} \sum_{j=0}^{N-1} f\left(j\,\frac{2\pi}{N}\right) e^{-inj(2\pi/N)}$$

which is an approximating sum of the integral

$$c_n = \int_0^{2\pi} f(t) e^{int}\, dt$$

i.e. the nth Fourier coefficient of $f=f(t)$. Now the analogy of the Fourier series is the sum

$$T_N(t) = \sum_{n=0}^{N-1} c_{Nn} e^{int} \tag{$**$}$$

and $\{c_{Nn}\}$; $n=0, 1, 2, \ldots, N-1$ is called the discrete Fourier transform of f based on the sample ($*$).

The important feature of the *discrete Fourier transform* is the following.

2.3.1.1 Theorem.

$$T_N\left(k\frac{2\pi}{N}\right) = f\left(k\frac{2\pi}{N}\right) \qquad k = 0, 1, \ldots, N-1.$$

Proof. Consider the mapping

$$f \to \left\{f\left(k\frac{2\pi}{N}\right)\right\}; \; k = 0, 1, \ldots, N-1 \tag{$*$}$$

from $L_0^2[0, 2\pi]$ into the N-dimensional linear space of N-tuples of complex numbers with scalar product

$$(f|g) := \frac{1}{N}\sum_{k=0}^{N-1} f\left(k\frac{2\pi}{N}\right)\overline{g\left(k\frac{2\pi}{N}\right)}.$$

Then

$$e^{int} \to \{e^{ink(2\pi/N)}\}; \quad k = 0, 1, \dots, N-1$$

and

$$\frac{1}{N}\sum_{j=0}^{N-1} e^{inj(2\pi/N)}e^{imj(2\pi/N)} = \begin{cases} 0 & \text{if } n \neq m \\ 1 & \text{if } n = m \end{cases}$$

and hence *the mapping* $(*)$ *carries the orthonormal system* $\{e^{int}\}$; $n=0, \pm 1, \dots$ *into a complete orthonormal system of the N-dimensional Hilbert space thus obtained.*

If we expand the sample $\{f(k2\pi/N)\}$; $k=0, 1, \dots, N-1$ in this orthonormal system (the copy of $\{e^{int}\}$; $n=0, \pm 1, \pm 2, \dots$ in the mapping $(*)$) then

$$f\left(k\frac{2\pi}{N}\right) = \sum_{n=0}^{N-1} c_{Nn}e^{ink(2\pi/N)} = T_N\left(k\frac{2\pi}{N}\right).$$

Remark 1. The copy of $\{e^{int}\}$; $n=0, \pm 1, \pm 2, \dots$ is a *complete* system since the range of the mapping $(*)$ is N dimensional.

Remark 2. Obviously, for most $f, g \in L_0^2[0, 2\pi]$,

$$\int_0^{2\pi} f(t)\overline{g(t)}\,dt \neq \frac{1}{N}\sum_{j=0}^{N-1} f\left(j\frac{2\pi}{N}\right)\overline{g\left(j\frac{2\pi}{N}\right)}.$$

2.3.2. If

$$e_n(t) = \begin{cases} N^{1/2} & \text{if } t \in \left(\frac{n-1}{N}, \frac{n}{N}\right) \\ 0 & \text{otherwise} \end{cases}$$

then we obtain the simplest orthonormal system in $L^2[0, 1]$ consisting of N elements. It is called a *system of square impulses*.

If we consider the double sequence

$$e_{nN}(t) = \begin{cases} N^{1/2} & \text{if } t \in \left(\frac{n-1}{N}, \frac{n}{N}\right); \quad N = 1, 2, \dots \quad n = 1, 2, \dots, N \\ 0 & \text{otherwise} \end{cases}$$

we obtain a *complete* orthonormal sequence. In fact, if

$$\int_0^1 f(t)e_{nN}(t)\,dt = N^{1/2}\int_{(n-1)/N}^{n/N} f(t)\,dt = 0$$

for $N=1, 2, \ldots$ and $n=1, 2, \ldots, N$ then the value of the integral function

$$F(t) := \int_0^t f(\tau)\, d\tau$$

is zero for every rational t and hence it is identically zero since $F=F(t)$ is a continuous function.

The oldest discrete orthonormal system is the Haar system in $L^2[0, 1]$ (figure 2.4):

$$h_0(0), h_0^{(1)}, \ldots, h_m^{(1)}, h_m^{(2)}, \ldots, h_m^{(2^m)}, \ldots \qquad m = 1, 2, \ldots$$

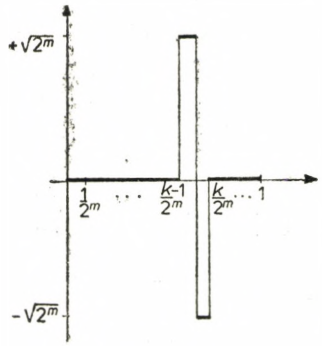

fig. 2.4

where

$$h_0^{(0)}(t) = 1; \quad h_0^{(1)}(t) = \begin{cases} 1 & \text{if } 0 \leqslant t < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} < t \leqslant 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$h_m^{(k)}(t) = \begin{cases} 2^{m/2} & \text{if } \frac{k-1}{2^m} \leqslant t < \frac{k-\frac{1}{2}}{2^m} \\ -2^{m/2} & \text{if } \frac{k-\frac{1}{2}}{2^m} \leqslant t < \frac{k}{2^m} \\ 0 & \text{elsewhere.} \end{cases}$$

If $m=n$ and $k \neq l$, then $h_m^{(k)}(t)h_n^{(l)}(t)=0$ in all but a finite number of points in [0, 1] and hence the orthogonality is clear. If $m \neq n$, say $m > n$, then

$$\int_0^1 h_m^{(k)}(t)\, h_n^{(l)}(t)\, \mathrm{d}t = \int_{(l-1)/2^m}^{l/2^m} h_m^{(k)}(t)\, h_k^{(l)}(t)\, \mathrm{d}t$$

$$= (2^n)^{1/2} \int_{(l-1)/2^n}^{(l-1/2)/2^n} h_m^{(k)}(t)\, \mathrm{d}t - (2^n)^{1/2} \int_{(l-1/2)/2^n}^{l/2^n} h_m^{(k)}(t)\, \mathrm{d}t = 0$$

since each of the last integrals is zero. Moreover, it can immediately be seen that

$$\int_0^1 h_m^{(k)}(t)^2\, \mathrm{d}t = 1.$$

The Haar system is complete, i.e. if the function $f \in L^2[0, 1]$ is orthogonal to every Haar function then $f(t) = 0$. In fact, if

$$F(t) = \int_0^t f(\tau)\, \mathrm{d}\tau$$

then F is a continuous function and $F(0) = 0$. Moreover,

$$0 = \int_0^1 f(\tau)\, h_0^{(0)}(\tau)\, \mathrm{d}\tau = F(1) - F(0)$$

and hence $F(1) = 0$ also. Taking the second Haar function,

$$0 = \int_0^1 f(\tau)\, h_0^{(1)}(\tau)\, \mathrm{d}\tau = \left[F\left(\frac{1}{2}\right) - F(0)\right] - \left[F(1) - F\left(\frac{1}{2}\right)\right] = 2F\left(\frac{1}{2}\right).$$

For the next Haar function,

$$0 = \int_0^1 f(\tau)\, h_1^{(1)}(\tau)\, \mathrm{d}\tau = \sqrt{2} \int_0^{1/4} f(\tau)\, \mathrm{d}\tau - \sqrt{2} \int_{1/4}^{1/2} f(\tau)\, \mathrm{d}\tau$$

$$= \sqrt{2}\left[F\left(\frac{1}{4}\right) - F(0) + F(1/u) - F\left(\frac{1}{2}\right)\right] = 2\sqrt{2}\, F\left(\frac{1}{4}\right) \quad \text{etc.}$$

It turns out that $F(k/2^m) = 0$ for $k = 0, 1, \ldots, 2^m$ and $m = 0, 1, 2, \ldots$, i.e. the value of the continuous function F is zero on a dense subset of [0, 1] and hence $F(t) = 0$ for every $t \in [0, 1]$. It follows that $f(t) = 0$ (almost everywhere).

2.3.3. The *Rademacher system* $\{r_n\}$ is another discrete orthonormal system in $L^2[0, 1]$. Its formation is even simpler than that of the Haar system: $r_0(t) = 1$ and for $n = 1, 2, \ldots$ the nth element is obtained if the interval [0, 1] is divided into 2^n parts; in the kth interval the value of $r_n(t)$ is $+1$ or -1 according to whether k is odd or even (see figure 2.5).

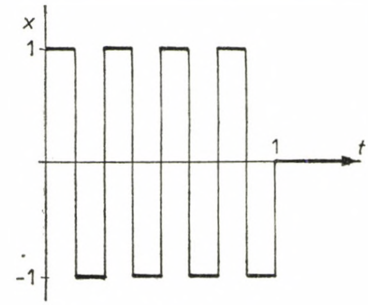

fig. 2.5

It is easily verified that the following connection exists between Rademacher functions and Haar functions:

$$r_n = (2^{-n})^{1/2} \sum_{k=1}^{2^n} h_n^{(k)}.$$

The disadvantage of the Rademacher system is that it is not complete. Indeed, any continuous function in [0, 1] with the properties

$$f(1-t) = f(t) \qquad \int_0^1 f(t)\,\mathrm{d}t = 0$$

is orthogonal to every Rademacher function. A simple example for a function of this type is $f(t)=\cos 2\pi t$.

A completion of the Rademacher system is the *Walsh system* $\{w_n\}$.

$w_0=r_0$ and the consecutive members are constructed according to the following rule:

$$\text{if} \quad n = \sum_{k=1}^{p} 2^{\nu_k} \quad \text{then} \quad w_n = r_{\nu_1+1} r_{\nu_2+1} \cdots r_{\nu_p+1}.$$

Applying this rule, it turns out that

$$w_{2^k} = r_{k+1} \qquad k = 0, 1, \ldots$$

and some of the first Walsh functions are

$$\begin{aligned} w_3 &= r_1 r_2 & \text{since } 3 &= 2+2^0 = 11_2 \\ w_4 &= r_3 & 4 &= 2^2 = 100_2 \\ w_5 &= r_1 r_3 & 5 &= 2^2+2^0 = 101_2 \\ w_6 &= r_2 r_3 & 6 &= 2^2+2 = 110_2 \\ w_7 &= r_1 r_2 r_3 & 7 &= 2^2+2+2^0 = 111_2 \\ w_8 &= r_4 & 8 &= 2^4 = 1000_2 \end{aligned}$$

(see figure 2.6).

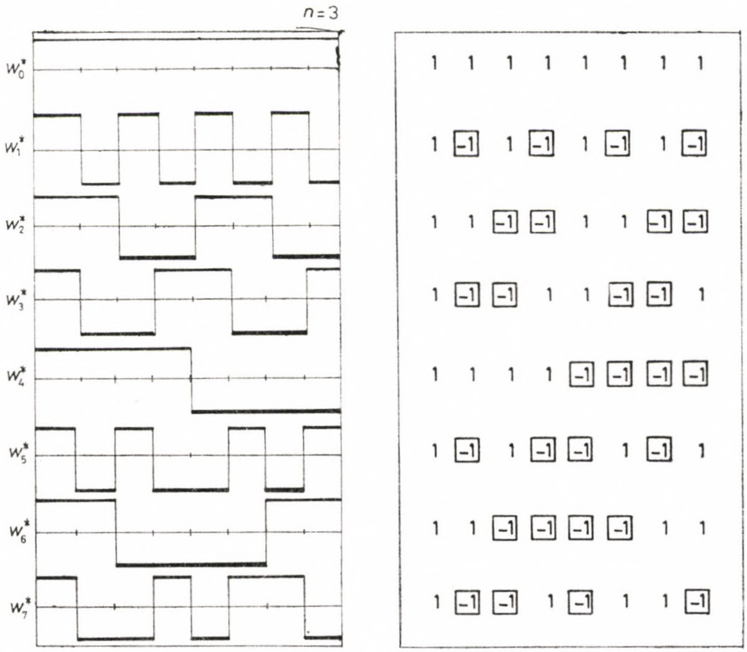

fig. 2.6

2.3.3.1 Theorem. The Walsh functions form an orthonormal system in $L^2[0, 1]$.
Proof. It is obvious that

$$\int_0^1 w_n(t)^2 \, dt = 1 \qquad n = 0, 1, 2, \ldots.$$

If $n \neq m$, then

$$w_n(t) w_m(t) = r_{n_1}(t) r_{n_2}(t) \ldots r_{n_k}(t)$$

where $0 < n_1 < n_2 < \ldots < n_k$, since $r_n^2(t) = 1$ for every Rademacher function r_n. Moreover, it follows from the definition or construction of Rademacher functions that

$$\int_0^1 w_n(t) w_m(t) \, dt = \int_0^1 r_{n_1}(t) r_{n_2}(t) \ldots r_{n_k}(t) \, dt = \pm \int_I r_{n_k}(t) \, dt = 0$$

where I is a subinterval of $[0, 1]$ with length 2^{1-n_k}. Another construction of Walsh functions is the following. We define the quadratic matrices

$$\mathbf{H}_1 = 1 \qquad \mathbf{H}_{2N} = \begin{pmatrix} H_N & H_N \\ H_N & -H_N \end{pmatrix} \qquad N = 1, 2, \ldots, 2^k, \ldots$$

which are called the Hadamard matrices of $2N$th degree. For example,

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad \mathbf{H}_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

Now, dividing the interval [0, 1] into $2N$ parts, the function $w_n^*(t)$ for $n<2N$ is defined as the (step) function whose value in the ith subinterval is $h_{n+1,i}$, the value of the ith element of the $(n+1)$th row of $\mathbf{H}_{2N}$.

It can be proved that the functions w_n^*; $n=0, 1, 2, \ldots$ thus obtained are Walsh functions; more precisely, $\{w_k^*\}$ is a certain rearrangement of the Walsh system $\{w_k\}$. Furthermore, the matrices $\mathbf{H}_{2N}$; $N=1, 2, \ldots$ are invertible (see §§ 2.14.16–17).

With the Hadamard matrices $\mathbf{H}_{2N}$ the computation of the Fourier coefficients with respect to the Walsh system is easy:

$$(f|w_k^*) := \int_0^1 f(t) w_k^*(t)\,\mathrm{d}t = \sum_{i=1}^{2N} h_{k+1,i} \int_{(i-1)/2N}^{i/2N} f(t)\,\mathrm{d}t$$

for $2N>k$ and hence, if the corresponding integral function is $F=F(t)$ as in the case of the Haar system, then

$$(f|w_k^*) = \sum_{i=1}^{2N} h_{k+1,i}\left(F\left(\frac{i}{2N}\right) - F\left(\frac{i-1}{2N}\right)\right) \tag{$*$}$$

where $h_{k+1,i}$ is the ith element of the $(k+1)$th row of the matrix $\mathbf{H}_{2N}$.

Using this construction of Walsh functions and Walsh–Fourier coefficients, we show that the following theorem is true.

2.3.3.2 Theorem. Let $g \in L^2[0, 1]$ be a function with the property $(g|w_k^*)=0$ for $k=0, 1, 2, \ldots$; then $g(t)=0$ (almost everywhere).

In fact, in this case

$$\sum_{i=1}^{2N} h_{k+1,i}\left(G\left(\frac{i}{2N}\right) - G\left(\frac{i-1}{2N}\right)\right) = 0 \qquad k = 0, 1, 2, \ldots, 2N-1$$

by $(*)$, where $2N>k$ and

$$G(t) := \int_0^t g(\tau)\,\mathrm{d}\tau.$$

Since every $\mathbf{H}_{2N}$ is invertible, it follows that

$$G\left(\frac{i}{2N}\right) - G\left(\frac{i-1}{2N}\right) = 0 \qquad i = 1, 2, \ldots, 2N$$

and $G(0)=0$; hence $G(i/2N)=0$ for $i=1, 2, \ldots, 2N$.

Thus we have proved that the values of the continuous functions $G=G(t)$ are zero on a dense subset of [0, 1]. It follows that $G(t)=0$ for every $t\in[0, 1]$ and $g(t)=0$.

2.3.4. The orthonormal system obtained in $L^2[0, \infty)$ from the Gram–Schmidt process from the sequence $e^{-\alpha t} t^n$ $(\alpha>0)$; $n=0, 1, 2, \ldots$ is called the *Laguerre system*. It can be proved (e.g. by induction) that for the nth Laguerre function L_n $(n=1, 2, \ldots)$,

$$L_n(t) = \frac{(2\alpha)^{1/2}}{n!} e^{\alpha t} \frac{d^n}{dt^n} (t^n e^{-2\alpha t})$$

(the Rodriguez formula) and the Laguerre system is complete since the sequence

$$a_k(t) = e^{-\alpha t} t^k \qquad k = 0, 1, 2, \ldots$$

is complete in $L^2[0, \infty)$. (Although we do not give the proof here, this is not an easy theorem.)

The importance of the Laguerre system is that the Fourier and Laplace transforms of L_n,

$$\mathscr{F}[L_n] = \frac{\sqrt{(2\alpha)^{1/2}}}{i\omega+\alpha} \left(\frac{i\omega-\alpha}{i\omega+\alpha}\right)^n$$

and

$$\mathscr{L}[L_n] = \frac{\sqrt{(2\alpha)^{1/2}}}{s+\alpha} \left(\frac{s-\alpha}{s+\alpha}\right)^n$$

respectively, i.e. $\mathscr{F}[L_n]$ and $\mathscr{L}[L_n]$ are rational functions of $i\omega$ and s, respectively, and this is the only orthonormal system with this property.

2.4 The projection principle for finite-dimensional subspace

In the previous section we obtained results concerning approximation in L^2-spaces by the partial sums of orthogonal series. However, in many cases it is more convenient to approximate instead by the sum of functions that have other advantageous properties than orthogonality.

2.4.1. Let $\{a_k;\ k=1, 2, \ldots, n\}$ be linearly independent elements of a scalar product space $\mathscr{H}$ and $x\in\mathscr{H}$. Find $\{\gamma_k;\ k=1, 2, \ldots, n\}$ such that

$$\left\|x - \sum_{k=1}^{n} \gamma_k a_k\right\|$$

is minimal. This problem is the immediate generalisation of the problem posed in § 2.2.1 for an orthogonal system of n elements.

Again, as in § 2.2, we turn to a geometrical analogy of this problem. If a_1, a_2 are linearly independent vectors of the geometrical vector space, then the linear subspace

$$\{\gamma_1 a_1+\gamma_2 a_2\} \qquad \gamma_1, \gamma_2 \in \Phi$$

can be visualised as a plane passing through the origin θ, and the problem is to find the point on this plane that is at a minimal distance from x.

It is well known that the solution of this geometrical problem is unique, and that it is the orthogonal projection of x onto the plane (see figure 2.7). This geometrical picture will be followed in the subsequent analysis.

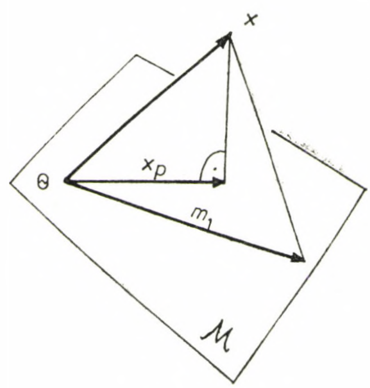

fig. 2.7

2.4.1.1 Definition. Let $\mathscr{M}$ be a linear subspace of the pre-Hilbert space $\mathscr{H}$; then $x_{\mathscr{M}} \in \mathscr{M}$ is called the *nearest vector* or the *best approximation* of $x \in \mathscr{H}$ if

$$\|x-x_{\mathscr{M}}\| \leqslant \|x-m\| \qquad m \in \mathscr{M}.$$

2.4.1.2 Definition. Let $\mathscr{M}$ be a linear subspace of the pre-Hilbert space $\mathscr{H}$; then $x_p \in \mathscr{M}$ is called the *(orthogonal) projection* of $x \in \mathscr{H}$ if, for every $m \in \mathscr{M}$,

$$(x-x_p|m) = 0.$$

Through these abstract formulations we have a connection between projection and best approximation similar to that in the geometric vector space.

2.4.1.3 Theorem. $x_p \in \mathscr{M}$ is the best approximation of $x \in \mathscr{H}$ if and only if x_p is the orthogonal projection of x in $\mathscr{M}$.

Proof. If x_p is the projection of x, i.e. for every $m \in \mathscr{M}$,

$$(x-x_p|m) = 0 \qquad (*)$$

and $m_1 \in \mathcal{M}$, different from x_p, then

$$\|x-m_1\|^2 = \|(x-x_p)+(x_p-m_1)\|^2 = \|x-x_p\|^2+\|x_p-m_1\|^2$$

and hence

$$\|x-m_1\| > \|x-x_p\|.$$

This means that x_p is the best approximation of x.

If $m_0 \in \mathcal{M}$ and there exists $m_1 \in \mathcal{M}$ such that

$$(x-m_0|m_1) \neq 0$$

i.e. m_0 is not the (orthogonal) projection of x, then there exists $m_0' \in \mathcal{M}$ nearer to x than m_0. In fact, if

$$m_0' = m_0+(x-m_0|m_1')m_1'$$

where $m_1'=m_1/\|m_1\|$, then $m_0' \in \mathcal{M}$ and

$$\|x-m_0'\|^2 = \|x-m_0\|^2-|(x-m_0|m_1)|^2.$$

Hence

$$\|x-m_0'\| < \|x-m_0\|.$$

Remark. The first part of the proof is visualised in figure 2.7 and the second part in figure 2.8.

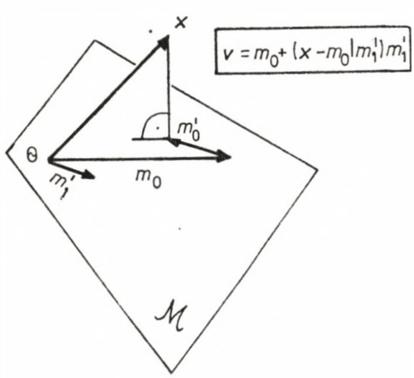

fig. 2.8

Now, let $\mathcal{M}$ be a *finite-dimensional* subspace of $\mathcal{H}$ generated by $\{a_k;\ k=1, 2, \ldots, n\}$, i.e. the set of linear combinations

$$\sum_{k=1}^{n} \gamma_k a_k \qquad \gamma_k \in \Phi.$$

2.4.1.4 Theorem. If $\mathscr{M}$ is the linear subspace generated by $\{a_1, a_2, \ldots, a_n\}$ then the orthogonal projection of $x \in \mathscr{H}$ onto $\mathscr{M}$ is

$$x_p = \sum_{k=1}^{n} \gamma_k a_k$$

where $\{\gamma_k;\ k=1, 2, \ldots, n\}$ is the solution of the following system of linear equations:

$$\sum_{k=1}^{n} (a_k|a_i)\gamma_k = (x|a_i) \qquad i = 1,2,\ldots,n. \qquad (*)$$

Proof. From 2.4.1.2 we have

$$(x - \sum_{k=1}^{n} \gamma_k a_k | a_i) = 0 \qquad i = 1, 2, \ldots, n \qquad (**)$$

since $\{a_k\}$; $k=1, 2, \ldots, n$ is a basis of $\mathscr{M}$ and $(*)$ is obtained from $(**)$ by obvious computations using the properties (i)–(iv) of the scalar product in § 2.1.1.

For the *existence* of the projection x_p in the case of finite-dimensional $\mathscr{M}$ we have the following. The matrix with elements $a_{ik}=(a_k|a_i)$ is called the *Gram matrix* of a_k; $k=1, 2, \ldots, n$ and the latter are linearly independent vectors if and only if the determinant of their Gram matrix is not zero. We shall prove this in the following more general formulation.

2.4.1.5 Theorem.

$$\sum_{k=1}^{n} \gamma_k a_k = 0$$

if and only if $\{\gamma_k;\ k=1, 2, \ldots, n\}$ is the solution of the system of linear equations

$$\sum_{k=1}^{n} (a_k|a_i)\gamma_k = 0 \qquad i = 1, 2, \ldots, n.$$

Proof. If

$$\sum_{k=1}^{n} \gamma_k a_k = 0$$

then the system of linear equations is obtained by multiplying both sides by a_i; $i=1, 2, \ldots, n$.

Conversely, if $\{\gamma_k;\ k=1, 2, \ldots, n\}$ satisfy the system of linear equations, then it follows from

$$\Big(\sum_{k=1}^{n} \gamma_k a_k \Big| \sum_{k=1}^{n} \gamma_k a_k\Big) = \sum_{i=1}^{n} \sum_{j=1}^{n} \gamma_i \bar{\gamma}_j (a_i|a_j) = \sum_{j=1}^{n} \bar{\gamma}_j \Big(\sum_{i=1}^{n} \gamma_i (a_i|a_j)\Big)$$

that

$$\left\|\sum_{k=1}^{n} \gamma_k a_k\right\|^2 = 0.$$

Example 1. If $\mathscr{H} = L_0^2[0, 1]$ and $a_k = t^{k-1}$; $k=1, 2, \dots, n$ then for the Gram matrix,

$$a_{ik} = (a_k|a_i) = \int_0^1 t^{k-1} t^{i-1}\,\mathrm{d}t = \frac{1}{k+i-1} \qquad k = 1, 2, \dots, n; \quad i = 1, 2, \dots, n$$

and by applying Theorem 2.4.1.4 we can obtain the best approximation of a function $f \in L_0^2[0, 1]$ by polynomials of nth degree in L^2-norm.

Example 2. Let the interval [0, 1] be divided into N parts by the points

$$0 = t_0 < t_1 < \dots < t_{N-1} < t_N = 1$$

and considering the continuous functions

$$a_i = a_i(t) = \begin{cases} 1 & \text{if} \quad t = t_i \\ 0 & \text{if} \quad t \notin (t_{i-1}, t_{i+1}) \\ \text{polynomial of 1st degree in the intervals} \\ \qquad [t_{i-1}, t_i] \quad \text{and} \quad [t_i, t_{i+1}] \end{cases}$$

it is easy to show that $\{a_i;\ i=0, 1, \dots, N\}$ form a basis for the linear subspace $\mathscr{M}$ of functions which may be plotted as a broken line with nodes only at t_i; $i=0, 1, \dots, N$. By applying 2.4.1.4 for $\mathscr{H} = L_0^2[0, 1]$ and this subspace $\mathscr{M}$, the best approximation of function f in $L_0^2[0, 1]$ by a 'broken line with nodes only at t_i; $i=0, 1, \dots, N$' will be obtained. A remarkable property of the Gram matrix in this case is that

$$a_{kj} = 0 \qquad \text{if} \quad |k-j| \geqslant 2.$$

This is called a *three-banded matrix.*

Example 3. Let $\mathscr{M}_N$ be the $(N+1)$-dimensional subspace of $L^2[-\pi, +\pi]$ generated by $\{\cos kt;\ k=0, 1, \dots, N\}$ and $f \in L^2[-\pi, +\pi]$; then

$$f_N(t) = \sum_{k=0}^{N} \gamma_k \cos kt$$

is the best approximation of f in $\mathscr{M}_N$ if γ_k; $k=0, 1, \dots, N$ are the Fourier coefficients

$$\gamma_k = \frac{1}{\pi} \int_{-\pi}^{+\pi} f(t) \cos kt\,\mathrm{d}t$$

and hence f_N is the projection of f onto $\mathcal{M}$. In fact, for any orthonormal system $\{a_k;\ k=1, 2, ..., n\}$,

$$(a_k|a_i) = \begin{cases} 0 & \text{if } k \neq i \\ 1 & \text{if } k = i \end{cases}$$

and hence the Gram matrix is the unit matrix.

∗2.5 The projection principle (general case)

In this section we shall drop the restriction of $\mathcal{M}$ being finite dimensional and the existence of the projection onto an infinite-dimensional subspace $\mathcal{M}$ will be investigated. Two other forms of the projection principle, which will become important later on, will also be given.

2.5.1. In studying the proof of Theorem 2.4.1.3 we notice that it is valid for any linear subspace $\mathcal{M}$ and hence *the best approximation and projection are also the same thing for an infinite-dimensional subspace* $\mathcal{M}$. This is not the case for the *existence* of x_p.

Recall that a linear subspace $\mathcal{M}$ of $\mathcal{H}$ is called *complete* if the Cauchy Convergence Theorem is valid in $\mathcal{M}$. If $\mathcal{H}$ is a Hilbert space and not only a scalar product space, then every closed subspace $\mathcal{M}$ is complete; moreover, every finite-dimensional $\mathcal{M}$ is complete.

2.5.1.1 Theorem. If $\mathcal{M}$ is a complete subspace of a scalar product space $\mathcal{H}$, then there exists a projection $x_p \in \mathcal{M}$ for every $x \in \mathcal{H}$.

Proof. If $d = \inf\{\|x-m\|;\ m \in \mathcal{M}\}$, then there exists a sequence $\{m_k\}$ such that $\|x-m_k\| \to d$. What we have to show is that $\{m_k\}$ is a Cauchy sequence. In this case $x_p = \lim m_k$.

To estimate $\|m_i - m_j\|$, we apply the parallelogram law for $x-m_i$ and $x-m_j$:

$$2\|x-m_i\|^2 + 2\|x-m_j\|^2 = \|m_i-m_j\|^2 + 4\|x-(m_i+m_j)/2\|^2$$

and hence for every $\varepsilon > 0$,

$$\begin{aligned} \|m_i-m_j\|^2 &= 2\|x-m_i\|^2 + 2\|x-m_j\|^2 - 4\|x-(m_i+m_j)/2\|^2 \\ &\leq 2\|x-m_i\|^2 + 2\|x-m_j\|^2 - 4d^2 < \varepsilon \end{aligned}$$

if i, j are large enough since $(m_i+m_j)/2 \in \mathcal{M}$ and $\|x-m_k\| \to d$.

The projection is *unique*. In fact, if x_p and x'_p are projections of $x \in \mathscr{H}$ onto $\mathscr{M}$, then

$$(x-x_p|m) = 0 \quad \text{and} \quad (x-x'_p|m) = 0 \qquad m \in \mathscr{M}.$$

Subtracting these two equalities, we obtain

$$(x'_p-x_p|m) = 0 \qquad m \in \mathscr{M}$$

and hence

$$\|x'_p-x_p\|^2 = (x'_p-x_p|x'_p-x_p) = 0$$

since $x'_p, x_p \in \mathscr{M}$.

Example 1. Consider now the infinite-dimensional linear subspace $\mathscr{M}$ of $L^2[-\pi, +\pi]$ consisting of the linear combinations of the functions

$$\cos kt \qquad k = 0, 1, 2, \ldots$$

(i.e. there is no restriction on the frequency k). In this linear subspace there is no nearest element to

$$f(t) = \sin t + \sum_{k=1}^{\infty} \frac{1}{k} \cos kt.$$

In fact, if

$$f_N(t) = \sum_{k=1}^{N} \frac{1}{k} \cos kt$$

then by 2.2.1 (*),

$$\|f-f_N\|_2^2 = \|f\|_2^2 - \pi \sum_{k=1}^{N} \frac{1}{k^2}$$

and hence

$$\|f-f_M\|_2 < \|f-f_N\|_2 \qquad \text{if} \quad M > N.$$

Thus we have shown that there is no projection of $f \in L^2[-\pi, +\pi]$ onto $\mathscr{M}$.

Example 2. Let us consider the *closed* linear subspace $\mathscr{M}$ of $L^2[-\pi, +\pi]$ generated by $\{\cos kt;\ k=0, 1, 2, \ldots\}$. Then the nearest element to f is

$$f_M(t) = \sum_{k=1}^{\infty} \frac{1}{k} \cos kt$$

and hence the projection exists in this closed linear subspace $\mathscr{M}$. In fact,

$$f(t)-f_M(t) = \sin t \quad \text{and} \quad \int_{-\pi}^{+\pi} \sin t \cos kt \, dt = 0 \qquad k = 0, 1, 2, \ldots.$$

2.5.2. In the geometric vector space, the set of vectors

$$\{x: (x-x_0|m) = 0\}$$

can be visualised as a plane passing through the point x_0 and orthogonal to the vector m (see figure 2.9).

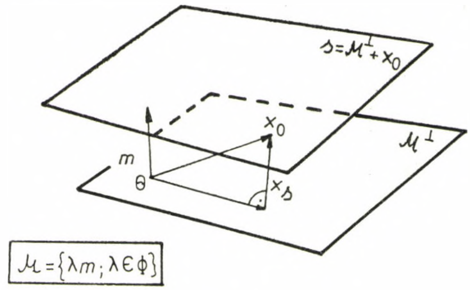

fig. 2.9

2.5.2.1 Definition. If $\mathscr{M}$ is a linear subspace of a Hilbert space $\mathscr{H}$ and $x_0 \in \mathscr{H}$, then

$$\mathscr{S} = \{x\colon (x-x_0|m) = 0; \quad m\in\mathscr{M}\}$$

is called the *hyperplane* $\mathscr{S}$ passing through the point x_0 and orthogonal to $\mathscr{M}$.

$\mathscr{M}^{\perp}=\{z\colon (z|m)=0;\ m\in\mathscr{M}\}$ is called the *orthogonal complement* of $\mathscr{M}$ in $\mathscr{H}$.

It is obvious that

$$\mathscr{M}^{\perp} = \{x-x_0;\ x\in\mathscr{S}\}$$

and $\mathscr{M}^{\perp}$ is a closed linear subspace of $\mathscr{H}$; hence the projection $x_{0p}\in\mathscr{M}^{\perp}$ of x_0 exists. If $\mathscr{M}$ is closed then $\mathscr{M}^{\perp\perp}=\mathscr{M}$. This will follow from Theorem 2.5.4.2.

Notice that $x_{0p}=x_0-x_{\mathscr{S}}$, where $x_{\mathscr{S}}\in\mathscr{S}$ and, by Theorem 2.4.1.3,

$$\|x_{\mathscr{S}}\| = \|x_0-x_{0p}\| \leqslant \|x_0-(x_0-x)\| \qquad x\in\mathscr{S}$$

which means that $x_{\mathscr{S}}$ is the element of the hyperplane $\mathscr{S}$ of minimum norm. Thus we have the following theorem.

2.5.2.2 Theorem. If $\mathscr{M}$ is a linear subspace of a Hilbert space $\mathscr{H}$ and $\mathscr{S}$ is the hyperplane passing through $x_0\in\mathscr{H}$ and orthogonal to $\mathscr{M}$, then

$$\mathscr{M}^{\perp} = \{x_0-x;\ x\in\mathscr{S}\}$$

is a closed linear subspace of $\mathscr{H}$ and, if x_{0p} is the projection of x_0 onto $\mathscr{M}$, then

$$x_{\mathscr{S}} = x_0-x_{0p}$$

is the element of $\mathscr{S}$ of minimum norm.

The theorem is illustrated in figure 2.9 for the case where $\mathscr{H}$ is the geometric vector space and $\mathscr{M}$ is the one-dimensional subspace generated by m.

2.5.3. We now apply the results of § 2.5.2 to the solution of the following problem.

'For a_k; $k=1, 2, \ldots, n$ and given numbers η_k; $k=1, 2, \ldots, n$, find x such that

$$(x|a_k) = \eta_k \qquad k = 1, 2, \ldots, n \tag{$*$}$$

and the norm is minimal.'

If $\mathscr{M}$ is a finite-dimensional subspace of $\mathscr{H}$ with basis $\{a_k\}$; $k=1, 2, \ldots, n$ then the points x of the hyperplane $\mathscr{J}$ passing through $x_0 \in \mathscr{H}$ and orthogonal to $\mathscr{M}$ are characterised by

$$\{x: (x-x_0|a_k) = 0;\ k = 1, 2, \ldots, n\}$$

and hence, introducing the notation $\eta_k = (x_0|a_k)$, the 'geometrical formulation' of the posed problem is to find $x \in \mathscr{J}$ with minimum norm.

Moreover, it follows from the considerations that led to Theorem 2.5.2.2. that $x_{\mathscr{J}} \in \mathscr{J}$ with minimum norm belongs to $\mathscr{M}^{\perp\perp} = \mathscr{M}$ and hence in our case it is in the form

$$x_{\mathscr{J}} = \sum_{k=1}^{n} \xi_k a_k .$$

Substituting this form into ($*$) we obtain

$$\sum_{k=1}^{n} \xi_k (a_k|a_i) = \eta_i \qquad i = 1, 2, \ldots, n$$

i.e. this problem leads to a system of linear equations similar to 2.4.1.4 ($*$).

2.5.4. A third formulation of the projection principle says that a Hilbert space can be decomposed into the orthogonal direct sum of closed subspaces.

2.5.4.1 Definition. Let N_i; $i=1, 2, \ldots, m$ be closed subspaces of the Hilbert space $\mathscr{H}$ such that

$$(z_i|z_j) = 0 \qquad \text{for} \qquad z_i \in N_i,\ z_j \in N_j \qquad i \neq j. \tag{$*$}$$

Then the *orthogonal direct sum* $\bigoplus_{i=1}^{n} N_i$ is the set of the sums

$$\sum_{i=1}^{n} z_i \qquad z_i \in N_i .$$

It is obvious that $\bigoplus_{i=1}^{n} N_i$ is a linear space and we can show that it is *closed.* In fact, if

$$x_n = \sum_{k=1}^{m} z_{kn} \qquad \text{and} \qquad x_n' = \sum_{k=1}^{m} z_{kn}' \qquad z_{kn}, z_{kn}' \in N_k$$

then

$$\|x_n - x_n'\|^2 = \sum_{k=1}^{m} \|z_{kn} - z_{kn}'\|^2$$

by the orthogonality (∗) and hence $x_n \to x$ if and only if $\{z_{kn}\}$ is convergent for every k. Moreover, if $z_{kn} \to z_k$ then

$$x = \sum_{k=1}^{m} z_k \qquad z_k \in N_k .$$

Now, let $\mathcal{M}$ be a closed linear subspace of the Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$. Then by Theorem 2.5.1.1 there exists the projection x_p onto $\mathcal{M}$ and

$$x = x_p + (x - x_p) \qquad x_p \in \mathcal{M}; \quad x - x_p \in \mathcal{M}^{\perp} \tag{$*$}$$

by 2.4.1.2. Hence we obtain the following theorem.

2.5.4.2 Theorem. For every closed subspace $\mathcal{M}$ there is the direct sum decomposition

$$\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}.$$

Remark 1. It is easy to show that the decomposition

$$x = x_p + z \qquad x_p \in \mathcal{M}, \quad z \in \mathcal{M}^{\perp}$$

is unique.

Remark 2. It is also easy to show that 2.5.1.1 follows from 2.5.4.2.

Remark 3. It is not necessary for $\mathcal{H}$ to be a Hilbert space. As in Theorem 2.5.1.1, for the orthogonal decomposition of $\mathcal{H}$ it is sufficient that $\mathcal{M}$ is a complete linear subspace of a scalar product space $\mathcal{H}$.

The direct sum decomposition of $\mathcal{H}$ into more than two subspaces also follows from 2.5.4.2. In fact, if $\mathcal{N} \subset \mathcal{M}$ is a closed subspace, then, by Theorem 5.4.2,

$$\mathcal{M} = \mathcal{N} \oplus \mathcal{K}$$

where $\mathcal{K} = \{x\colon\ x \in \mathcal{M};\ (x|z) = 0 \text{ for } z \in \mathcal{N}\}$, i.e. the orthogonal complement of $\mathcal{N}$ in $\mathcal{M}$, and hence

$$\mathcal{H} = \mathcal{M}^{\perp} \oplus \mathcal{N} \oplus \mathcal{K}$$

is an orthogonal direct sum decomposition of $\mathcal{H}$ of three members.

Following this approach, for example, a direct sum decomposition of $\mathcal{H}$ of any finite number of members is obtained.

2.6 Some typical examples of the projection principle

First we shall show that the solution of the minimum problem posed in § 2.2.1 is a special case of the projection principle. After this we give examples for the important cases when $\mathscr{M}$ or $\mathscr{M}^\perp$ is finite dimensional, and the section concludes with a well-known optimum property of the solution of the pure boundary problem of the Laplace equation as an example of the projection theorem in the case of infinite-dimensional $\mathscr{M}$. Several applications of the projection theorem will also be found later in Chapter 2 and in Chapter 3.

2.6.1. The minimum problem of § 2.2.1 is a special case of the projection theorem when $\mathscr{M}$ is the n-dimensional subspace of a Hilbert space $\mathscr{H}$ spanned by an orthonormal sequence e_k; $k=1, 2, \ldots, n$. In this case the orthogonal projection in $\mathscr{M}$ of $x \in \mathscr{H}$ has the form

$$x_p = \sum_{k=1}^{n} \gamma_k e_k$$

where γ_n; $k=1, 2, \ldots, n$ is the solution of the system of equations

$$\sum_{k=1}^{n} \gamma_k (e_k|e_i) = (x|e_i) \qquad i = 1, 2, \ldots, n$$

by Theorem 2.4.1.4. In this case $\gamma_i = (x|e_i)$ since

$$(e_k|e_i) = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}$$

and so the result of § 2.2.1 is obtained in a simpler way.

2.6.2. Frequently we have to solve a system of linear algebraic equations with more equations than the number of unknowns and having no solution. This contradictory situation occurs, for example, when the linearity is only approximate and the measurements concerning the matrix of the system of equations are inaccurate. In this case *the solution is defined* as $\{x_k;\ k=1, 2, \ldots, m\}$ for which the sum

$$\sum_{i=1}^{n} \left|b_i - \sum_{k=1}^{m} a_{ik} x_k\right|^2 \qquad (*)$$

is minimal. (In the case of an exact solution this sum is of course 0.)

The solution of a system of linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m &= b_n \end{aligned}$$

in this generalised sense can be found by the following l^2-space model. Let

$$\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix}, \ldots, \begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix}, \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

be considered as vectors of the linear space of n-tuples with the l^2-norm and let $\mathscr{M}$ be the linear space generated by the m column vectors of the matrix of the system of linear equations to be solved. Now we have only to apply 2.4.1.4 when $\mathscr{H}=l^2$ and

$$x = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \qquad a_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{bmatrix} \qquad k = 1, 2, \ldots, n.$$

2.6.3. We shall now consider a problem from analytic function theory. Determine $f \in \mathrm{H}_0^2$ such that

$$f(z_k) = w_k \qquad k = 1, 2, \ldots, n$$

where z_k, w_k are given values, and

$$\int_0^{2\pi} |f(\mathrm{e}^{\mathrm{i}t})|^2 \,\mathrm{d}t$$

is minimal.

By the Cauchy integral formula and straightforward calculation,

$$f(z_k) = \frac{1}{2\pi \mathrm{i}} \oint_{|z|=1} \frac{f(z)}{z-z_k} \,\mathrm{d}z = \frac{1}{2\pi} \int_0^{2\pi} f(\mathrm{e}^{\mathrm{i}t}) \frac{\mathrm{e}^{\mathrm{i}t}}{\mathrm{e}^{\mathrm{i}t}-z_k} \,\mathrm{d}t$$

and hence the evaluation functionals in H_0^2 are represented by scalar products

$$f(z_k) = (f | g_k)$$

where

$$g_k(\mathrm{e}^{\mathrm{i}t}) = \frac{\mathrm{e}^{-\mathrm{i}t}}{\mathrm{e}^{-\mathrm{i}t} - \bar{z}_k}.$$

It is obvious that $g_k \in \mathrm{H}_0^2$, i.e. $g(\mathrm{e}^{\mathrm{i}t})$, is a continuous function and

$$g_k(z) = \frac{1}{1 - z\bar{z}_k}$$

is analytic for $|z| < 1$.

It follows that our problem is the special case of the problem of § 2.5.3 for $\mathscr{H} = \mathrm{H}_0^2$ and $a_k = g_k$.

2.6.4. One of the fundamental problems of system theory is to find a $u=u(t)$ control that transfers the state $y=y(t)$ of the system from the initial or zero state into a given state $y_1=y(t_1)$ such that the energy of the control is minimal.

The following example is closely connected to this problem. Consider the differential equation

$$y''(t)+y(t) = u(t) \tag{$*$}$$

with the boundary conditions

$$y(0) = 0 \qquad y(2\pi) = 1$$

$$y'(0) = 0 \qquad y'(2\pi) = 0$$

where y' means the derivative of y. The existence of the solution $y=y(t)$ is not guaranteed for every continuous $u=u(t)$. More precisely, it is known from the elementary theory of differential equations that the solution of the above differential equation is uniquely determined by $u=u(t)$ and the initial conditions $y(0)=y'(0)=0$. However, it is not guaranteed that the 'end conditions' $y(2\pi)=1$, $y'(2\pi)=0$ are also satisfied for this unique solution.

If a solution of the boundary problem $(*)$ exists *then* $u=u(t)$ *is called a control* of the system governed by the differential equation $(*)$ that transfers the state from $\begin{bmatrix}0\\0\end{bmatrix}$ into $\begin{bmatrix}1\\0\end{bmatrix}$.

As a further exercise, let us find the control $u=u(t)$ for which the energy integral

$$\int_0^{2\pi} u^2(t)\,dt$$

is minimal. It is known (and easy to verify) that

$$y(t) = \int_0^t \sin(t-\tau)u(\tau)\,d\tau$$

is the solution of the differential equation $(*)$ satisfying the initial conditions only and hence for the 'end conditions' to be satisfied also it is necessary and sufficient that

$$y(2\pi) = \int_0^{2\pi} \sin(2\pi-\tau)u(\tau)\,d\tau = 1$$

$$(**)$$

$$y'(2\pi) = \int_0^{2\pi} \cos(2\pi-\tau)u(\tau)\,d\tau = 0.$$

Considering $u(t)$, $\sin(2\pi-t)$, $\cos(2\pi-t)$ as elements of the real Hilbert space $L^2[0, 2\pi]$, we can write the system of equations $(**)$ in the form of scalar products

$$(\sin(2\pi-\tau)|u(\tau)) = 1$$
$$(\cos(2\pi-\tau)|u(\tau)) = 0.$$

We have thus obtained the problem of § 2.5.3 for $u=x_{\vartheta}$, $\mathscr{H}=L^2(0, 2\pi)$ and $a_1=\sin(2\pi-\tau)$; $a_2=\cos(2\pi-\tau)$. Consequently

$$u(t) = \alpha \sin t+\beta \cos t$$

and the corresponding system of linear equations with Gram matrix

$$-\alpha \int_0^{2\pi} \sin^2 t\, dt-\beta \int_0^{2\pi} \sin t \cos t\, dt = 1$$

$$\alpha \int_0^{2\pi} \cos t \sin t\, dt+\beta \int_0^{2\pi} \cos^2 t\, dt = 0;$$

hence

$$\alpha = -\frac{1}{\pi} \qquad \beta = 0.$$

Remark. It follows from this example that if a linear system is governed by the differential equation ($*$), then any minimal-energy control u that transfers the state y from $[0; 0]$ into $[y(2\pi); y'(2\pi)]$ has the sinusoidal form $u=A \sin(t+\varphi)$, and the amplitude A and phase φ are determined by the vector of the end state with components $y(2\pi)$; $y'(2\pi)$.

2.6.5. Let $\mathscr{D}$ be a convex domain in the (three-dimensional) space with smooth boundary $\mathscr{S}$ and $u=u(\boldsymbol{r})$ be a function of the space variable $\boldsymbol{r}$ with a continuous gradient in $\mathscr{D}$.

2.6.5.1 Definition. The volume integral

$$\int_{\mathscr{D}} |\operatorname{grad} u|^2\, dx\, dy\, dz$$

is called the *energy integral* of u in $\mathscr{D}$.

An important problem in potential theory is to find the function $u=u(\boldsymbol{r})$ with given boundary values on $\mathscr{S}$ and with a minimal energy integral. The answer to this problem lies in the following theorem.

2.6.5.2 Theorem. If

$$u(\boldsymbol{r}) = x_0(\boldsymbol{r}) \qquad \text{for } \boldsymbol{r}\in\mathscr{S}$$

and

$$\Delta u := u''_{xx}+u''_{yy}+u''_{zz}$$

has a meaning, then the energy integral of u is a minimum if u is the solution of the boundary problem

$$\Delta u = 0 \qquad u(\boldsymbol{r}) = x_0(\boldsymbol{r}) \qquad \text{if} \quad \boldsymbol{r}\in\mathscr{S}.$$

Conversely, if u is the solution of this boundary problem, then the energy integral of u is the minimum among functions with the same values on $\mathscr{S}$.
Proof. Let X be the linear space of real-valued functions with continuous gradient in $\mathscr{D}$. Then by the formula

$$(u|v) := \oint_{\mathscr{S}} u(\boldsymbol{r})v(\boldsymbol{r})\,\mathrm{d}\mathscr{S} + \int_{\mathscr{D}} \operatorname{grad} u \operatorname{grad} v \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z$$

a scalar product is defined in X. *Note that among $u\in X$ with the same boundary values, the energy integral is a minimum if and only if the scalar product $(u|u)$ is a minimum.*

Now, let $x_0\in X$ and $u(\boldsymbol{r})=x_0(\boldsymbol{r})$ if $\mathrm{r}\in\mathscr{S}$. Also,

$$\mathscr{N} = \{v\colon\ v(\boldsymbol{r}) = 0 \text{ if } \boldsymbol{r}\in\mathscr{S}\}.$$

It is obvious that $(x_0-u)\in\mathscr{N}$ and hence

$$u = x_0-(x_0-u) = x_0-v \qquad v\in\mathscr{N}.$$

It follows that $(u|u)$ *has the minimum value if and only if* $v\in\mathscr{N}$ *is the nearest element to* $x_0\in X$ *in the pre-Hilbert space thus obtained.*

Applying Theorem 2.4.1.3, we know that $v_0\in\mathscr{N}$ is the nearest element to x_0 if and only if

$$(x_0-v_0|v) = 0 \qquad v\in\mathscr{N}.$$

More particularly, for $u_0=x_0-v_0$,

$$\int_{\mathscr{D}} \operatorname{grad} u_0 \operatorname{grad} v \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = 0 \qquad v\in\mathscr{N}. \tag{$*$}$$

Our last step in the proof is an application of the Green formula. Applying the Divergence Theorem to v grad u_0, the *Green formula*

$$\int_{\mathscr{D}} \operatorname{grad} u_0 \operatorname{grad} v \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z + \int_{\mathscr{D}} v\Delta u \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = \int_{\mathscr{S}} v \operatorname{grad} u_0 n_0 \,\mathrm{d}\mathscr{S}$$

is obtained. It follows from $(*)$ that

$$\int_{\mathscr{D}} v\Delta u_0 \,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z = 0$$

for *every* $v\in\mathscr{N}$ and hence $\Delta u_0=0$.

Conversely, if $u, x\in X$, $u_0(\boldsymbol{r})=x(\boldsymbol{r})$ for $\boldsymbol{r}\in\mathscr{S}$ and $\Delta u_0=0$ then, by the Green formula, $(*)$ holds for every $v\in\mathscr{N}$. Again, $x-u_0\in\mathscr{N}$ and

$u_0 = x-(x-u_0)$; hence, applying Theorem 2.4.1.3, we obtain

$$(u_0|u_0) \leqslant (x|x)$$

and the energy integral of u_0 is smaller than the energy integral of x.

Remark. $\mathcal{N}$ contains a complete sequence in the sense of Definition 2.2.3.1 in $L^2(\mathscr{D})$ by the results of classical analysis.

*2.7 Controllability and optimal control of linear systems

One of the fundamental problems of linear system theory is the controllability and optimal control of a system. In this section a close connection between the projection theorem and the mathematical solution of these problems will be demonstrated.

2.7.1. The usual mathematical description of a linear continuous time system is the state, or dynamical, equation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}(t)\mathbf{x}(t)+\mathbf{B}(t)\mathbf{u}(t) \\ \dot{\mathbf{y}}(t) &= \mathbf{C}(t)\mathbf{x}(t) \end{aligned} \tag{$*$}$$

where $\mathbf{A}(t)$ is an $n\times n$ matrix, $\mathbf{B}(t)$ is an $n\times p$ matrix, $\mathbf{C}(t)$ is an $n\times q$ matrix, $p, q \leqslant n$ and their entries are bounded piecewise continuous functions.

$\mathbf{u}=\mathbf{u}(t)$ is called the *control* and $\mathbf{x}(t)$ the *state* of the system at time t. The basic problem concerning $(*)$ is to find, for a given state $\mathbf{x}_1$, a control $\mathbf{u}(t)$ and $t_1>0$ such that a solution $\mathbf{x}(t)$ of $(*)$ exists satisfying the boundary conditions $\mathbf{x}(t_1)=\mathbf{x}_1$ and $\mathbf{x}(0)=\theta$. If such a pair $(t_1, \mathbf{u})$ exists, then it is said that there exists a control $\mathbf{u}=\mathbf{u}(t)$ that transfers the system from the initial or zero state into the given state $\mathbf{x}_1$ during the period t_1.

The dynamical (or state) equation $(*)$ of a system is called *controllable* if for every state $\mathbf{x}_1$ there exists a control $\mathbf{u}$ and $t_1>0$ such that $\mathbf{u}$ transfers the system from the state θ into $\mathbf{x}_1$ during the period t_1. The *optimal energy control* $\mathbf{u}$ is the one for which

$$\|\mathbf{u}\|^2 = \sum_{k=1}^{p} \int_0^{t_1} |u_k(t)|^2 \, dt$$

is minimal (where u_k is the kth component of $\mathbf{u}$). The problem of finding an optimal energy control for the dynamical equation $(*)$ is the generalisation of the example in § 2.6.4.

It is obvious that controllability depends only on the first equation of (*). The matrix $\Phi(t, 0)$ is called the transfer matrix of (*) if the kth column of $\Phi(t, 0)$ is the solution of the equation

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) \qquad \mathbf{x}(0) = \mathbf{e}_k$$

where $\mathbf{e}_k$ is the column vector with 1 in the kth row and zero in all other rows.

It follows from the theory of matrix differential equations that

$$\mathbf{x}(t) = \int_0^t \Phi(\tau, 0)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau$$

in the solution of the first equation of (*) with $\mathbf{x}(0)=\theta$. Hence the problem of controllability can be formulated as follows.

For a given $\mathbf{x}_1$ *find* $\mathbf{u}$ *and* t_1 *such that*

$$\mathbf{x}_1 = \int_0^{t_1} \Phi(\tau, 0)\mathbf{B}(\tau)\mathbf{u}(\tau)\,d\tau.$$

Following these considerations we can state the fundamental controllability condition for the state equation (*).

2.7.1.1 Theorem. Let

$$\mathbf{W}(t, 0) = \int_0^t \Phi(\tau, 0)\mathbf{B}(\tau)\mathbf{B}^*(\tau)\Phi^*(\tau, 0)\,d\tau.$$

Then the system (or, more precisely, the state of dynamical equation (1) of a system) is controllable iff for every $\mathbf{x}_1$ there exists $t_1 > 0$ such that the system of linear algebraic equations

$$\mathbf{x}_1 = \mathbf{W}(t_1, 0)\mathbf{c}$$

(written in matrix form) has a solution $\mathbf{c}$. In this case, the optimal energy control is

$$\mathbf{u}(t) = \mathbf{B}^*(t)\Phi^*(t, 0)\mathbf{c}.$$

The proof of this theorem is based on the following lemma. *Let* $\mathbf{G}=\mathbf{G}(\tau)$ *be a matrix with n rows and entries from* $L^2(s, t)$. *Then for any n-dimensional column vector* $\mathbf{u}=\mathbf{u}(t)$ *there exists* $\mathbf{c}$ *such that*

$$\int_s^t \mathbf{G}(\tau)[\mathbf{u}(\tau)-\mathbf{G}^*(\tau)\mathbf{c}]\,d\tau = \theta. \qquad (**)$$

For the proof of this lemma, consider the finite-dimensional subspace $\mathscr{M}$ generated by the rows of $\mathbf{G}(\tau)$. Let $\mathbf{u}_p$ be the nearest element (orthogonal projection) of $\mathbf{u}$ in $\mathcal{M}$. Then $\mathbf{u}_p$ has the form

$$\mathbf{u}_p(t) = \mathbf{G}^*(t)\mathbf{c}$$

and $(**)$ holds, by Theorems 2.4.1.3 and 2.4.1.4. Now, if

$$\mathbf{G}(t) = \Phi(t, 0)\mathbf{B}(t)$$

and

$$\mathbf{W}(t, s) = \int_s^t \mathbf{G}(\tau)\mathbf{G}^*(\tau)\,d\tau$$

then Theorem 2.7.1.1. is obtained from this lemma.

A more detailed presentation and the physical background of the controllability problem can be found in Brockett (1970).

Remark. Some of the examples in this section are rather more concerned with linear algebra than with Hilbert space theory; however, in a Hilbert space setting they have a more general perspective.

2.8 Scalar product and bounded linear functionals

Recall that a linear functional is a special type of linear operator: if the values of an operator are complex numbers then it is called a linear functional. It follows from the results of § 1.4 that continuity and boundedness are the same for a linear functional, and for a continuous linear functional f of a Hilbert space $\mathcal{H}$:

$$|f(x)| \leqslant \|f\|\,\|x\| \qquad x \in \mathcal{H}$$

where $\|f\|$ is the least upper bound of

$$\{|f(x)|/\|x\|;\ x \neq \theta\}$$

and is called the *norm* of the functional f.

2.8.1. For any fixed $y_0 \in \mathcal{H}$, the mapping $f(x) = (x|y_0)$; $x \in \mathcal{H}$ is a bounded linear functional of $\mathcal{H}$, as can be seen from axioms (ii) and (iii) of the scalar product and from the Cauchy–Schwarz inequality. A surprising result discovered by F Riesz and M Fréchet in the early years of the twentieth century is that these are the only continuous linear functionals of a Hilbert space.

2.8.1.1 Theorem. For every continuous linear functional f of a Hilbert space $\mathcal{H}$ there exists a unique $y_f \in \mathcal{H}$ such that

$$f(x) = (x|y_f). \qquad (*)$$

Also, $\|f\| = \|y_f\|$.

Remark. We recall that

$$\|f\| = \sup\{|f(x)|: \|x\| = 1\}$$

by 1.4.3.2.

Proof. (*a*) *For separable* $\mathscr{H}$. In this case we have the expansion in an orthonormal sequence

$$x = \sum_{k=1}^{\infty} \xi_k e_k$$

for every $x \in \mathscr{H}$ and hence

$$f(x) = f\Big(\sum_{k=1}^{\infty} \xi_k e_k\Big) = \sum_{k=1}^{\infty} \xi_k f(e_k).$$

On the other hand, if

$$y_f = \sum_{k=1}^{\infty} \overline{f(e_k)}\, e_k$$

formally then

$$\sum_{k=1}^{\infty} \xi_k f(e_k) = (x|y_f).$$

We have 'only' to prove that the series defining y_f is convergent. Considering 2.14.10, what we have to prove is that

$$\sum_{k=1}^{\infty} |f(e_k)|^2 < \infty.$$

Estimating $\sum_{k=1}^{N} |f(e_k)|^2$, we have

$$f\Big(\sum_{k=1}^{N} \overline{f(e_k)}\, e_k\Big) = \sum_{k=1}^{N} |f(e_k)|^2$$

and

$$f\Big(\sum_{k=1}^{N} \overline{f(e_k)}\, e_k\Big) \leqslant \|f\| \Big\|\sum_{k=1}^{N} \overline{f(e_k)}\, e_k\Big\| = \|f\| \Big(\sum_{k=1}^{N} |f(e_k)|^2\Big)^{1/2}$$

since

$$\Big\|\sum_{k=1}^{N} \overline{f(e_k)}\, e_k\Big\|^2 := \Big(\sum_{k=1}^{N} \overline{f(e_k)}\, e_k \Big| \sum_{k=1}^{N} \overline{f(e_k)}\, e_k\Big).$$

Hence

$$\sum_{k=1}^{N} |f(e_k)|^2 \leqslant \|f\| \Big(\sum_{k=1}^{N} |f(e_k)|^2\Big)^{1/2}$$

and it follows that

$$\Big(\sum_{k=1}^{N} |f(e_k)|^2\Big)^{1/2} \leqslant \|f\|.$$

(*b*) *For any* $\mathscr{H}$. The existence of an orthonormal expansion for every $x \in \mathscr{H}$ is the characteristic property of a separable Hilbert space. Hence for non-separable Hilbert spaces the foregoing proof does not work. The usual proof of the Riesz Theorem in the absence of orthogonal expansions is based on Theorem 2.5.4.2.

Let f be a continuous linear functional of $\mathscr{H}$ and $\mathscr{N}=\{x\colon f(x)=0\}$; then $\mathscr{N}$ is a closed linear subspace of $\mathscr{H}$ and hence $\mathscr{H}=\mathscr{N}\oplus\mathscr{N}^{\perp}$ by Theorem 2.5.4.2. What does the decomposition

$$x = z+y \qquad z\in\mathscr{N},\quad y\in\mathscr{N}^{\perp}$$

look like for any $x\in\mathscr{H}$?

Observe that if $y_0\in\mathscr{N}^{\perp}$ then $x-f(x)/f(y_0)\, y_0\in\mathscr{N}$ for any $x\in\mathscr{H}$ since

$$f\left(x-\frac{f(x)}{f(y_0)}\,y_0\right) = f(x)-f(x) = 0$$

and we conclude that

$$z = x-\frac{f(x)}{f(y_0)}\,y_0 \qquad y = \frac{f(x)}{f(y_0)}\,y_0.$$

Hence $\mathscr{N}^{\perp}$ is one dimensional.

It follows from this orthogonal decomposition that

$$(x|y_0) = (z+y|y_0) = (y|y_0) = \frac{f(x)}{f(y_0)}\,\|y_0\|^2$$

and hence

$$f(x) = \frac{f(y_0)}{\|y_0\|^2}(x|y_0),$$

i.e. for $y_f=\overline{f(y_0)}y_0/\|y_0\|^2$ the representation (*) is valid.

If $f(x)=(x|y_f)$ then

$$|f(x)| \leqslant \|x\|\,\|y_f\|.$$

By the Cauchy–Schwarz inequality and hence by definition (see the beginning of this section),

$$\|f\| \leqslant \|y_f\|.$$

On the other hand, for $x=y_f/\|y_f\|$,

$$f(x) = \left(\frac{y_f}{\|y_f\|}\middle| y_f\right) = \|y_f\|.$$

The uniqueness of y_f is routine and is left to the reader.

2.8.2. Obviously the Riesz–Fréchet Theorem is not valid for a non-complete scalar product space. For example, $L_0^2(a, b)$ consists of continuous functions;

however, any step function induces a bounded linear functional on $L_0^2(a, b)$ via the scalar product. The Hilbert space $L^2(a, b)$, the completion of $L_0^2(a, b)$, can be considered as the space of continuous linear functionals of $L_0^2(a, b)$.

2.9 Bilinear functionals

A bilinear functional is a generalisation of the scalar product, and certain types of bilinear functional inherit several useful properties of the scalar product.

2.9.1. The mapping $\varphi=\varphi(x, y)$ from the ordered pairs of elements of a linear space X into the field of scalars (complex numbers) is called *bilinear* if

$$\begin{aligned} \varphi(\lambda x+\mu z, y) &= \lambda\varphi(x, y)+\mu\varphi(z, y) \\ \varphi(y, \lambda x+\mu z) &= \bar{\lambda}\varphi(y, x)+\bar{\mu}\varphi(y, z) \end{aligned} \qquad x, y, z\in X,\ \lambda, \mu\in\Phi.$$

φ is called *symmetric* or *Hermitian* if $\varphi(x, y)=\overline{\varphi(y, x)}$ and *positive* if $\varphi(x, x)\geqslant 0$.

Example 1. Let x and y be infinite sequences of complex numbers and $\mathbf{A}$ a square $n\times n$ matrix with elements $\{a_{ik}\}$; $i=1, 2, \ldots, n$, $j=1, 2, \ldots, n$. Then

$$\alpha(x, y) := \sum_{k=1}^{n}\sum_{i=1}^{n} a_{ik}\xi_i\overline{\eta_k}$$

is a bilinear functional on the linear space of infinite sequences of complex numbers, where

$$x = \{\xi_k;\ k = 1, 2, \ldots\} \qquad \text{and} \qquad y = \{\eta_k;\ k = 1, 2, \ldots\}.$$

The bilinear functional thus obtained is symmetric iff $\mathbf{A}$ is a symmetric matrix, i.e. if $a_{ik}=\bar{a}_{ki}$, and positive iff $\mathbf{A}$ is a positive definite matrix.

Example 2. If $\{a_{ik};\ i=1, 2, \ldots, k=1, 2, \ldots\}$ is a 'double sequence', called an infinite matrix, with the condition

$$\sum_{k=1}^{\infty}\sum_{i=1}^{\infty} |a_{ik}|^2 < \infty$$

then

$$\alpha(x, y) := \sum_{i=1}^{\infty}\sum_{k=1}^{\infty} a_{ik}\xi_k\bar{\eta}_i$$

is a bilinear functional on the l^2-space.

Example 3. If $k=k(s, t)$ is a continuous function on the closed finite two-dimensional interval $a \leqslant s,\ t \leqslant b$ then

$$\varphi(x, y) := \int_a^b \int_a^b k(s, t)x(t)\overline{y(s)}\,\mathrm{d}t\,\mathrm{d}s$$

is a bilinear functional on the linear space of continuous functions in $[a, b]$. The bilinear functional thus obtained is symmetric iff $k(s, t)=\overline{k(t, s)}$ and positive iff

$$\int_a^b \int_a^b k(s, t)x(t)\overline{x(s)}\,\mathrm{d}t\,\mathrm{d}s \geqslant 0$$

for every continuous function x. It can be proved that this inequality is valid iff for every finite sequence $\{s_i;\ i=1, 2, \ldots, n\}$ of points in $[a, b]$ the matrix with elements

$$a_{ik} := k(s_i, s_k)$$

is positive definite.

Example 4. If $b=b(t)$ is a piecewise continuous function in $[a, b]$ (i.e. the sum of a step function and a continuous function) then

$$\beta(x, y) := \int_a^b b(t)x(t)\overline{y(t)}\,\mathrm{d}t$$

is also a bilinear functional in the linear space of continuous functions in $[a, b]$. β is symmetric in the case of real-valued $b=b(t)$ and positive iff b has only non-negative values.

Example 5. If T is a linear operator of a Hilbert space then

$$\varphi_T(x, y) := (x|Ty) \qquad x, y \in \mathscr{H}$$

is a bilinear functional on $\mathscr{H}$.

For a symmetric real-valued φ,

$$\varphi(x+y, x+y) = \varphi(x, x)+\varphi(y, y)+2\varphi(x, y)$$
$$\varphi(x-y, x-y) = \varphi(x, x)+\varphi(y, y)-2\varphi(x, y)$$

and hence

$$\varphi(x, y) = \frac{1}{4}(\varphi(x+y, x+y)-\varphi(x-y, x-y)).$$

This means that *the symmetric bilinear functional φ is determined by the (quadratic) functional of a single variable*

$$\varphi = \varphi(z, z) \qquad z \in \mathscr{H}.$$

This is also valid for any bilinear functional since

$$\varphi(x, y) = \frac{1}{4}(\varphi(x+y, x+y) - \varphi(x-y, x-y) + i\varphi(x+iy, x+iy)$$
$$-i\varphi(x-iy, x-iy)) \qquad x, y \in \mathscr{H}. \qquad (*)$$

This can be verified by a somewhat longer but straightforward calculation. From the connection $(*)$ we may state the following theorem.

2.9.1.1 Theorem. φ is symmetric iff $\varphi(z, z)$ is real for every $z \in \mathscr{H}$. In particular, every positive bilinear functional φ is symmetric.
Proof. If φ is symmetric then, in particular,

$$\varphi(x, x) = \overline{\varphi(x, x)} \qquad x \in \mathscr{H}$$

and hence $\varphi(x, x)$ is real for every $x \in \mathscr{H}$.

For the converse statement, interchanging x and y in $(*)$,

$$\varphi(y, x) + \frac{1}{4}(\varphi(y+x, y+x) - \varphi(y-x, y-x)$$
$$+i\varphi(y+ix, y+ix) - i\varphi(y-ix, y-ix)).$$

Hence if $\varphi(y+x, y+x)$, $\varphi(y-x, y-x)$, $\varphi(y+ix, y+ix)$ and $\varphi(y-ix, y-ix)$ are real then $\varphi(y, x) = \varphi(x, y)$.

Note that if $\varphi(x, x) = 0$ implies $x = \theta$ for a positive (and hence also symmetric) bilinear functional then axioms (i)–(iv) of the scalar product are satisfied, i.e. a scalar product is defined by φ.

2.9.2. For a positive bilinear functional φ the Cauchy–Schwarz inequality is valid:

$$|\varphi(x, y)|^2 \leq \varphi(x, x)\varphi(y, y) \qquad x, y \in \mathscr{H}. \qquad (*)$$

Indeed, this inequality is obvious when $\varphi(x, y) = 0$. In the case when $\varphi(x, y) \neq 0$

$$\vartheta = \frac{\varphi(y, x)}{|\varphi(x, y)|}$$

is a real or complex number of modulus 1. Hence

$$0 \leq \varphi(\vartheta x + \lambda y, \vartheta x + \lambda y) = \varphi(x, x) + 2\lambda|\varphi(x, y)| + \lambda^2\varphi(y, y)$$

for any *real* λ, i.e. the real-valued polynomial thus obtained has at most one zero. Hence, for the discriminant,

$$4|\varphi(x, y)|^2 - 4\varphi(x, x)\varphi(y, y) \leq 0$$

which is equivalent to $(*)$.

2.9.3. If the bilinear functional φ is defined in a scalar product space $\mathscr{H}$, as in the case of Example 2 or Example 5, then it is natural to look for the connection between φ and the scalar product in $\mathscr{H}$.

2.9.3.1 Definition. If there exists $M>0$ such that

$$|\varphi(x, y)| \leq M\|x\|\,\|y\| \qquad x, y \in \mathscr{H}$$

then φ is called *bounded.*

If T is a bounded operator in Example 5, then

$$|\varphi_T(x, y)| \leq \|T\|\,\|x\|\,\|y\|$$

and hence φ_T is a bounded bilinear functional.

We shall show that these are the only bounded bilinear functionals.

2.9.3.2 Theorem. Let φ be a bounded bilinear functional in a scalar product space $\mathscr{H}$. Then there is a unique bounded linear operator T such that

$$\varphi(x, y) = (x|Ty) \qquad x, y \in \mathscr{H}.$$

Proof. The mapping

$$x \to \varphi(x, y)$$

is a bounded linear functional for any fixed y. Hence, by the Riesz–Fréchet Theorem, there exists $z \in \mathscr{H}$ such that

$$\varphi(x, y) = (x|z).$$

We define

$$Ty := z.$$

Then T is unique and $\varphi(x, y) = (x|Ty)$ for any pair $x, y \in \mathscr{H}$. T is a bounded linear operator. In fact, for every $x \in \mathscr{H}$,

$$(x|T(\lambda y_1+\mu y_2)) := \varphi(x, \lambda y_1+\mu y_2) = \bar{\lambda}\varphi(x, y_1)+\bar{\mu}\varphi(x, y_2)$$

$$:= \bar{\lambda}(x|Ty_1)+\bar{\mu}(x|Ty_2) = (x|\lambda Ty_1+\mu Ty_2) \qquad y_1, y_2 \in \mathscr{H}, \quad \lambda, \mu \in \Phi$$

and hence T is linear. From 2.8.1.1 and 2.9.3.1,

$$\|Ty\| = \sup_{\|x\|=1} |(x|Ty)| \qquad y \in \mathscr{H}$$

and

$$\|T\| := \sup\{\|Ty\|;\ \|y\| = 1\} \leq M.$$

2.9.3.3 Definition. The norm of a bounded bilinear functional φ is defined by

$$\|\varphi\| = \sup\{|\varphi(x, y)|;\ \|x\| = 1,\ \|y\| = 1\}.$$

It follows from this definition that if T is the bounded linear operator associated with φ by 2.9.3.2, then

$$\|\varphi\| = \|T\|.$$

2.10 First steps in the theory of linear operators on Hilbert spaces

The notion of the dual or adjoint operator T^* of a linear operator T in a Hilbert space is of crucial importance in Hilbert space theory. The main classification of Hilbert space operators is based on the connection between T and its dual T^*. In this section we shall investigate those parts of the theory that are related to Hilbert space geometry and will be needed in §§ 2.11 and 2.12.

2.10.1. Formally, the adjoint of a linear operator T that maps a Hilbert space $\mathscr{H}_1$ into a Hilbert space $\mathscr{H}_2$ is the operator T^* satisfying the condition

$$(Tx|y) = (x|T^*y) \qquad x\in\mathscr{H}_1, \quad y\in\mathscr{H}_2. \tag{$*$}$$

We now ask: does there exist such an operator T^* for every bounded linear operator?

2.10.1.1 Theorem. For every bounded linear operator T from $\mathscr{H}_1$ into $\mathscr{H}_2$ there exists a unique bounded linear operator T^* from $\mathscr{H}_2$ into $\mathscr{H}_1$ satisfying $(*)$; moreover,

$$\|T^*\| = \|T\|.$$

Proof. The mapping $x\to(Tx|y)$ is a bounded linear functional since T is linear and

$$|(Tx|y)| \leqslant \|Tx\|\,\|y\| \leqslant \|T\|\,\|x\|\,\|y\|.$$

Hence, by the Riesz–Fréchet Theorem, for every pair (y, T) there is a unique $z\in\mathscr{H}_1$ such that

$$(Tx|y) = (x|z).$$

Define $T^*y := z$; the operator T^* is obviously linear and, by 2.8.1.1,

$$\|T^*y\| = \sup\{|(x|T^*y)|;\ \|x\| = 1\} = \sup\{|(Tx|y)|:\ \|x\| = 1\}$$
$$\leqslant \sup\{\|Tx\|\,\|y\|;\ \|x\| = 1\} = \|T\|\,\|y\|.$$

Hence T^* is bounded, $\|T^*\|\leqslant\|T\|$ and

$$(Tx|y) = (x|T^*y) \qquad x\in\mathscr{H}_1, \quad y\in\mathscr{H}_2.$$

Now, if $T^{**} := (T^*)^*$ then $\|T^{**}\|\leqslant\|T^*\|$ from the above considerations.

We shall show that

$$T^{**} = T.$$

Indeed, by definition,

$$(T^*x|y) = (x|T^{**}y)$$

and hence

$$(Tx|y) = (x|T^*y) = \overline{(T^*y|x)} = \overline{(y|T^{**}x)} = (T^{**}x|y) \qquad x \in \mathscr{H}_1, \quad y \in \mathscr{H}_2.$$

2.10.1.2 Definition. If T is a bounded linear operator from $\mathscr{H}_1$ into $\mathscr{H}_2$ then the bounded linear operator T^* from $\mathscr{H}_2$ into $\mathscr{H}_1$ satisfying $(*)$ is called the *adjoint* or *dual* operator of T.

Remark. The adjoint is defined for a Banach space operator and also for certain unbounded Hilbert space operators.

Example 1. If T is the linear operator of the n-dimensional Euclidean space represented by the matrix $\mathbf{A}$ with elements $\{a_{ik}\}$ (see Example 1 in § 1.4.2) then the adjoint T^* is the linear operator represented by the adjoint (transposed) matrix $\mathbf{A}^*$ with elements $\{\bar{a}_{ki}\}$.

Example 2. If T is an integral operator of $L^2[a, b]$ with kernel $K=K(t, \tau)$, i.e.

$$Tf := \int_a^b K(t, \tau) f(\tau)\, d\tau \qquad f \in L^2[a, b]$$

where K is a continuous function in the square $[a, b] \times [a, b]$, then the adjoint T^* is the integral operator of the form

$$T^*f := \int_a^b \overline{K(\tau, t)}\, f(\tau)\, d\tau \qquad f \in L^2[a, b].$$

Indeed,

$$(Tf|g) = \int_a^b \int_a^b K(t, \tau) f(\tau) \overline{g(t)}\, d\tau\, dt = \int_a^b f(\tau) \Big(\int_a^b K(t, \tau) \overline{g(t)}\, dt \Big)\, d\tau$$

$$= \int_a^b f(\tau) \Big(\int_a^b \overline{\overline{K(t, \tau)} g(t)\, dt} \Big)\, d\tau = (f|T^*g).$$

Example 3. If F is the operator sending a function $f \in L^2[0, 2\pi]$ into the sequence of (complex) Fourier coefficients

$$c_k = \int_0^{2\pi} f(t) e^{-ikt}\, dt \qquad k = 0, \pm 1, \pm 2, \ldots$$

then for any $d=\{d_k\}\in l^2$,

$$(Ff|d) = \sum_{k=-\infty}^{+\infty} \bar{d}_k \int_0^{2\pi} f(t)\mathrm{e}^{-ikt}\,\mathrm{d}t = \int_0^{2\pi} f(t)\Big(\sum_{k=-\infty}^{+\infty} \bar{d}_k \mathrm{e}^{-ikt}\Big)\,\mathrm{d}t = \int_0^{2\pi} f(t)\overline{g(t)}\,\mathrm{d}t$$

where

$$g(t) = \sum_{k=-\infty}^{+\infty} d_k \mathrm{e}^{ikt}.$$

Hence

$$F^*\{d_k\} = \sum_{k=-\infty}^{+\infty} d_k \mathrm{e}^{ikt} \in \mathrm{L}^2[0, 2\pi]$$

for every $d\in l^2$. Hence, in this case, $F^*=F^{-1}$.

Remark. The interchange of the summation and the integration is justified since we have scalar products in l^2 and $\mathrm{L}^2[0, 2\pi]$, respectively, and for finite sums the equality holds.

Passing to the adjoint operator we have the following (algebraic) properties:

(i) $(\lambda T_1+\mu T_2)^* = \bar{\lambda}T_1^*+\bar{\mu}T_2^*$;

(ii) $(T_1T_2)^* = T_2^*T_1^*$ (if T_1T_2 exists); in particular,
$(T^{-1})^* = (T^*)^{-1}$ (if T^{-1} exists);

(iii) $T^{**} := (T^*)^* = T$.

The proofs are easy and are therefore omitted.

2.10.2. If T is a bounded linear operator of a Hilbert space (i.e. the range is also contained in $\mathscr{H}$) then it may happen that $T=T^*$. In this case T is called *self-adjoint* or *Hermitian.*

If T is the operator in § 2.10.1, represented by a matrix $\mathbf{A}$, then T is self-adjoint iff the matrix is self-adjoint, i.e. $a_{ik}=\bar{a}_{ki}$. If T is the integral operator of this subsection, then T is self-adjoint iff K is symmetric or Hermitian, i.e. $K(t, \tau)=\overline{K(\tau, t)}$.

It is natural to ask whether the sum and product of self-adjoint operators are self-adjoint. Applying 2.10.1 (i)–(ii), the answer is easy.

(i) If S, T are self-adjoint operators, then $\lambda S+\mu T$ is a self-adjoint operator if and only if λ, μ are *real* numbers.

(ii) If S, T are self-adjoint operators, then ST is self-adjoint if and only if $ST=TS$.

Moreover, the operators TT^* and $T+T^*$ are self-adjoint for any linear operator T.

In § 2.9.3. a 1–1 correspondence was established between bounded bilinear functionals and bounded operators. Based on this correspondence, if T is self-adjoint, then

$$\varphi_T(x, y) := (x|Ty) = \overline{(Ty|x)} = \overline{(y|Tx)} = \overline{\varphi_T(y, x)}$$

and hence φ_T is symmetric or Hermitian. Conversely, if φ_T is symmetric, then

$$(x|Ty) := \varphi_T(x, y) = \overline{\varphi_T(y, x)} = \overline{(y|Tx)} = (Tx|y)$$

and hence T is self-adjoint. This can be summarised as follows.

2.10.2.1 Theorem. The bounded bilinear functional φ_T corresponding to the operator T by 2.9.3.2 is symmetric if and only if T is self-adjoint.

By the considerations in § 2.9.3,

$$\|T\| = \sup \{(Tx|y): \|x\| = 1, \|y\| = 1\}$$

for a bounded linear operator T. If T is self-adjoint, then the norm is determined by the corresponding quadratic form.

2.10.2.2 Theorem. The operator T is self-adjoint if and only if $(Tx|x)$ is real for every $x \in \mathscr{H}$; moreover, in this case,

$$\|T\| = \sup \{|(Tx|x)|;\ \|x\| = 1\}.$$

Proof. The first part of the theorem follows from 2.10.2.1 and 2.9.1.1. For the second part we have only to prove that

$$\|T\| \leqslant \sup \{|(Tx|x)|;\ \|x\| = 1\}.$$

If

$$m = \sup \{|(Tx|x)|;\ \|x\| = 1\}$$

then

$$(T(x+y)|x+y) \leqslant m\|x+y\|^2$$

and

$$(T(x-y)|x-y) \leqslant m\|x-y\|^2.$$

Hence

$$4\mathrm{Re}\,(Tx|y) \leqslant m(\|x+y\|^2+\|x-y\|^2) \tag{$*$}$$

since

$$(T(x+y)|x+y) = (Tx|x)+2\mathrm{Re}\,(Tx|y)+(Ty|y)$$

and

$$(T(x-y)|x-y) = (Tx|x)-2\mathrm{Re}\,(Tx|y)+(Ty|y).$$

Applying the parallelogram law (see § 2.14.5) to the right-hand side of ($*$), we obtain

$$4\mathrm{Re}\,(Tx|y) \leqslant 2m(\|x\|^2+\|y\|^2). \tag{$**$}$$

Now, *if $\mathscr{H}$ is a real Hilbert space*, then $\operatorname{Re}(Tx|y)=(Tx|y)$ and hence

$$\|T\| = \sup\{(Tx|y);\ \|x\| = 1,\ \|y\| = 1\} \leqslant m.$$

In the case where $\{(Tx|y);\ x, y\in\mathscr{H}\}$ are complex numbers the end of the proof is more sophisticated. Then

$$(Tx|y) = |(Tx|y)|\,\mathrm{e}^{\mathrm{i}\varphi} \qquad x, y\in\mathscr{H}$$

and so

$$|(Tx|y)| = \mathrm{e}^{-\mathrm{i}\varphi}(Tx|y) = (T\mathrm{e}^{-\mathrm{i}\varphi}x|y) \qquad x, y\in\mathscr{H}$$

where φ is the argument of the complex number $(Tx|y)$. Hence, substituting $\mathrm{e}^{-\mathrm{i}\varphi}x$ for x in $(**)$, we have

$$|(Tx|y)| \leqslant \frac{m}{2}(\|x\|^2+\|y\|^2) \qquad x, y\in\mathscr{H}$$

since $\|\mathrm{e}^{-\mathrm{i}\varphi}x\|=\|x\|$ and $(T\mathrm{e}^{-\mathrm{i}\varphi}x|y)\geqslant 0$.

In particular, for

$$y = \frac{\|x\|}{\|Tx\|}Tx \qquad (Tx \neq \theta)$$

we obtain

$$\|Tx\|\,\|x\| \leqslant m\|x\|^2 \qquad x\in\mathscr{H}.$$

2.10.2.3 Definition. The linear operator T is called *positive* if $(Tx|x)\geqslant 0$ for every $x\in\mathscr{H}$. It is called *strictly positive* if $(Tx|x)=0$ only if $x=\theta$.

It is obvious that the positive linear operators form a subclass of self-adjoint operators and $\varphi_T(x, y):=(Tx|y)$ defines a scalar product (in the sense of the considerations at the end of § 2.9.1) if and only if T is strictly positive.

As for the real numbers, a natural order can be defined for self-adjoint linear operators by means of positivity.

2.10.2.4 Definition. For the self-adjoint operators A and B we write

$$A \leqslant B$$

if $B-A$ is a positive operator.

2.10.2.5 Definition. M is an *upper bound* and m is a *lower bound* of a self-adjoint operator T if

$$mE < T < ME$$

where E is the identity operator.

It is easy to verify that a bounded self-adjoint operator T always has (finite) upper and lower bounds; moreover, T is a positive operator if there exists a non-negative lower bound m.

2.10.3. It was shown in Theorem 2.5.4.2 that for any closed subspace $\mathcal{M}$ of a Hilbert space $\mathcal{H}$ there is the direct sum decomposition

$$\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^{\perp}$$

and if $x=y+z;\ y\in\mathcal{M},\ z\in\mathcal{M}^{\perp}$ is the direct sum decomposition of $x\in\mathcal{H}$ then y is the projection x_p of x onto $\mathcal{M}$.

The operator $P_{\mathcal{M}}$ which sends x into the projection x_p, i.e.

$$P_{\mathcal{M}}x = x_p \qquad x\in\mathcal{H},\quad x_p\in\mathcal{M}$$

is called the *projection operator*. It is obvious that

$$\|P_{\mathcal{M}}x\| \leqslant \|x\| \qquad x\in\mathcal{H}$$

for any $\mathcal{M}$, and $P_{\mathcal{M}}$ is a self-adjoint operator with $P^2_{\mathcal{M}}=P_{\mathcal{M}}$. In fact, if

$$x_1 = y_1+z_1 \qquad y_1\in\mathcal{M},\quad z_1\in\mathcal{M}^{\perp}$$

$$x_2 = y_2+z_2 \qquad y_2\in\mathcal{M},\quad z_2\in\mathcal{M}^{\perp}$$

then

$$(P_{\mathcal{M}}x_1|x_2) = (y_1|y_2+z_2) = (y_1|y_2)$$

$$(x_1|P_{\mathcal{M}}x_2) = (y_1+z_1|y_2) = (y_1|y_2).$$

Moreover,

$$P^2_{\mathcal{M}}x = P_{\mathcal{M}}x_p = x_p$$

for every $x\in\mathcal{H}$.

2.10.3.1 Theorem. If P is a self-adjoint operator with $P^2=P$, then there exists a closed linear subspace $\mathcal{M}$ of the Hilbert space $\mathcal{H}$ such that $P=P_{\mathcal{M}}$.
Proof. Let us consider

$$\mathcal{M} = \{Px;\ x\in\mathcal{H}\}.$$

$\mathcal{M}$ is a linear subspace since P is a linear operator; $\mathcal{M}$ is also closed since if $y_n=Px_n$ and $y_n\to y$, then

$$y = \lim_{n\to\infty} Px_n = \lim_{n\to\infty} P(Px_n) = Py$$

and hence $y\in\mathcal{M}$.

If $y\in\mathcal{M}$, then

$$Py = P(Px) = P^2x = Px = y.$$

If $z\in\mathcal{M}^{\perp}$, then

$$\|Pz\|^2 = (Pz|Pz) = (P^2z|z) = (Pz|z) = 0.$$

The one-to-one correspondence between projection operators and closed linear subspaces has the following properties.

2.10.3.2 Theorem. $P_{\mathcal{M}}P_{\mathcal{N}}$ is a projection operator if and only if $P_{\mathcal{M}}P_{\mathcal{N}}= =P_{\mathcal{N}}P_{\mathcal{M}}$; in this case

$$P_{\mathcal{M}}P_{\mathcal{N}} = P_{\mathcal{M}\cap\mathcal{N}}$$

where $\mathcal{M}\cap\mathcal{N}$ is the intersection of the sets $\mathcal{M}$ and $\mathcal{N}$.

$$P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}} = P_{\mathcal{M}\cup\mathcal{N}}$$

where $\mathcal{M}\cup\mathcal{N}$ is the closed linear space generated by the elements $m\in\mathcal{M}$, $n\in\mathcal{N}$.

Recall that the closed subspace generated by a set $\mathcal{S}$ is the closure in $\mathcal{H}$ of the elements in the form

$$\sum_{k=1}^{m} \alpha_k a_k \qquad \alpha_k\in\Phi, \quad a_k\in\mathcal{S}.$$

Proof. Since $(P_{\mathcal{M}}P_{\mathcal{N}})^*=P_{\mathcal{N}}^*P_{\mathcal{M}}^*=P_{\mathcal{N}}P_{\mathcal{M}}$, it is necessary that $P_{\mathcal{M}}P_{\mathcal{N}}= =P_{\mathcal{N}}P_{\mathcal{M}}$. If this commutation relation is satisfied then we also have

$$(P_{\mathcal{M}}P_{\mathcal{N}})^2 = P_{\mathcal{M}}P_{\mathcal{N}}P_{\mathcal{M}}P_{\mathcal{N}} = P_{\mathcal{M}}P_{\mathcal{N}}$$

and hence $P_{\mathcal{M}}P_{\mathcal{N}}$ is a projection operator.

If $z\in\mathcal{M}\cap\mathcal{N}$, then $P_{\mathcal{M}}P_{\mathcal{N}}z=z$. Conversely, if $P_{\mathcal{M}}P_{\mathcal{N}}z=z$, then

$$P_{\mathcal{M}}z = P_{\mathcal{M}}(P_{\mathcal{M}}P_{\mathcal{N}}z) = P_{\mathcal{M}}P_{\mathcal{N}}z = z$$

and

$$P_{\mathcal{N}}z = P_{\mathcal{N}}(P_{\mathcal{M}}P_{\mathcal{N}}z) = P_{\mathcal{N}}(P_{\mathcal{N}}P_{\mathcal{M}}z) = P_{\mathcal{N}}P_{\mathcal{M}}z = z$$

and hence $z\in\mathcal{M}\cap\mathcal{N}$.

$P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}}$ is a projection operator. In fact,

$$(P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}})^* = P_{\mathcal{M}}^*+P_{\mathcal{N}}^*-P_{\mathcal{N}}^*P_{\mathcal{M}}^* = P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}}$$

and

$$(P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}})^2 = P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}}.$$

If $z\in\mathcal{M}$, then

$$(P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}})z = z+P_{\mathcal{N}}z-P_{\mathcal{N}}z = z$$

and this is also the case when $z\in\mathcal{N}$.

For $(P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}})x=x$ it is necessary that x be the sum of $x_{\mathcal{M}}\in\mathcal{M}$ and $z\in\mathcal{N}$. In fact, for any $x\in\mathcal{H}$,

$$x = x_{\mathcal{M}}+z \qquad x_{\mathcal{M}}\in\mathcal{M}; \quad z\in\mathcal{M}^{\perp}$$

by 2.5.4.2 and

$$(P_{\mathcal{M}}+P_{\mathcal{N}}-P_{\mathcal{M}}P_{\mathcal{N}})x = [P_{\mathcal{M}}+P_{\mathcal{N}}(I-P_{\mathcal{M}})]x = x_{\mathcal{M}}+P_{\mathcal{N}}z.$$

2.10.4. A linear operator T from one Hilbert space $\mathscr{H}_1$ *onto* another Hilbert space $\mathscr{H}_2$ is called an *isomorphic* or *unitary* operator if

$$(x|y) = (Tx|Ty) \qquad x, y \in \mathscr{H}_1$$

so that a unitary operator 'preserves' the scalar product. This property can also be expressed by the connection between T and T^*. We begin with the following.

2.10.4.1 Theorem. The following conditions for an operator T mapping a Hilbert space $\mathscr{H}_1$ onto another Hilbert space $\mathscr{H}_2$ are equivalent:

(i) T is isometric, i.e. $\|Tx\| = \|x\|$ for every $x \in \mathscr{H}_1$;
(ii) $T^*T = E_1$ (identity operator in $\mathscr{H}_1$);
(iii) $(Tx|Ty) = (x|y)$; $x, y \in \mathscr{H}_1$.

Proof. It is easy to show that (iii)$\Rightarrow$(ii) and (ii)$\Rightarrow$(i); for the remaining part of the proof, apply 2.9.1 ($*$).

We can now characterise the isomorphic operators as follows.

2.10.4.2 Theorem. The bounded linear operator U is an isomorphic or unitary operator if

$$U^*Ux = x \qquad \text{for every } x \in \mathscr{H}_1$$

and

$$UU^*y = y \qquad \text{for every } y \in \mathscr{H}_2.$$

In other words, an isomorphic operator is an isometric operator mapping the Hilbert space $\mathscr{H}_1$ *onto* $\mathscr{H}_2$.
Proof. It follows from the previous theorem that U is an isometry if the first equality is satisfied. If U is *also onto*, then there exists the inverse operator U^{-1} from $\mathscr{H}_2$ onto $\mathscr{H}_1$ and

$$U^* = U^*(UU^{-1}) = (U^*U)U^{-1} = U^{-1}$$

and this is exactly what the two equalities say.
Conversely, if

$$U^* = U^{-1}$$

then U is an isometric operator (isometry) mapping $\mathscr{H}_1$ onto $\mathscr{H}_2$.

An important example of an isometric operator which is not an isomorphic operator is the *forward shift* U_τ; $\tau > 0$ in $L^2[0, \infty]$ defined by

$$[U_\tau f](t) = \begin{cases} 0 & \text{if } t < \tau \\ f(t-\tau) & \text{if } t \geqslant \tau. \end{cases}$$

In fact, $\|U_\tau f\| = \|f\|$ for every $f \in L^2[0, 2\pi]$ but U_τ is *not onto*.

2.11 Isomorphic Hilbert spaces and isomorphic operators

If there is a unitary operator from $\mathscr{H}_1$ onto $\mathscr{H}_2$, then the Hilbert spaces $\mathscr{H}_1$ and $\mathscr{H}_2$ are considered to be identical in a certain sense.

2.11.1. The Hilbert spaces $\mathscr{H}_1$ and $\mathscr{H}_2$ are called *isomorphic* or *congruent* if there exists a unitary operator U mapping $\mathscr{H}_1$ onto $\mathscr{H}_2$.

Example 1. $L^2[0, 2\pi]$ is isomorphic to l^2 via the Fourier series expansion. In fact,

$$Ff = \left\{ \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt} \, dt \right\} \qquad k = 0, \pm 1, \pm 2, \ldots$$

is an isomorphic operator from $L^2[0, 2\pi]$ onto l^2 by Example 3 in § 2.10.1.

Remark. If we take another complete orthonormal system in $L^2[0, 2\pi]$ (e.g. a suitably modified Walsh system) then, by 2.2.2.1, the operator sending $f \in L^2[0, 2\pi]$ into the sequence of the new Fourier coefficients is also a congruence operator from $L^2[0, 2\pi]$ onto l^2. Hence there may be many isomorphic operators between isomorphic Hilbert spaces (see also the proof of 2.11.1.1).

Example 2. It follows from Example 4 in § 2.1.2 that by the mapping

$$Lf(z) = f(e^{it})$$

H_0^2 is congruent to the closed subspace of $L_0^2[0, 2\pi]$ consisting of elements in the form

$$f = \sum_{k=0}^{\infty} \gamma_k e^{ikt}$$

i.e. whose Fourier coefficients are zero for $k < 0$.

Example 3. Consider the closed subspace

$$\mathscr{M} = \{y \colon y(a) = y'(a) = 0\}$$

of $\mathscr{H}_D$ (Example 5 in § 2.1.2). $\mathscr{M}$ is isomorphic to $L_0^2[a, b]$ and D is an isomorphic operator. Let us define

$$\mathscr{N} = \{y \colon Dy = \theta\}$$

in $\mathscr{H}_D$; $\mathscr{N}$ is isomorphic to the two-dimensional geometric vector space via the isomorphic operator

$$Ly = (y(a), y'(a)).$$

Example 4. The Fourier transform

$$Ff = \int_{-\infty}^{+\infty} e^{i\omega t} f(t)\,dt := \hat{f}(\omega) \qquad (*)$$

is defined for every $f \in L(-\infty, +\infty)$ and $Ff \in C_0$ (where C_0 is the class of continuous functions tending to zero at infinity). Moreover, if $f \in L^2(-\infty, +\infty)$ also, i.e. if $f \in L(-\infty, +\infty) \cap L^2(-\infty, +\infty)$, then

$$\int_{-\infty}^{+\infty} f(t)\overline{g(t)}\,dt = \int_{-\infty}^{+\infty} \hat{f}(\omega)\hat{g}(\omega)\,d\omega$$

by classical but non-trivial theorems of harmonic analysis. Hence the Fourier transform F is an isometry from $L \cap L^2$ into $C_0 \cap L^2$, both with norm $\|\,.\,\|_2$.

Since $L \cap L^2$ is a dense subspace of L^2, F can be defined for every $f \in L^2$ by

$$Ff := \lim_{n \to \infty} Ff_n$$

where $f_n \in L \cap L^2$ and $\lim f_n = f$ (obviously here the limit is understood in L^2-norm). Thus the operator F is extended to the whole of $L^2(-\infty, +\infty)$ as an isometry; however, the formula $(*)$ is *not* valid for every $f \in L^2$ (it is certainly valid for $f_n \in L \cap L^2$).

It can be proved that the extended operator F is not only an isometry but is also a unitary operator.

Remark 1. From a physical viewpoint, the domain and the range of the Fourier transform are not considered as the subspace of the *same* Hilbert space $L^2(-\infty, +\infty)$. In most cases the independent variable of the functions in the domain space is a 'time variable' and that in the range space is 'frequency'; however, from a pure-mathematical viewpoint we do not make this distinction.

Remark 2. It is very important in these examples that congruent Hilbert spaces, however, they may be considered equal from the Hilbert space point of view, could be very different. We can see this later also in § 3.3.

The most important theorem concerning isomorphic Hilbert spaces is the following.

2.11.1.1 Theorem. Every separable Hilbert space $\mathcal{H}$ is congruent to the l^2-space.

Proof. If $\{e_k\}$ is a complete orthonormal sequence and $x \in \mathscr{H}$ then, by 2.2.3.2.

$$x = \sum_{k=1}^{\infty} (x|e_k)e_k.$$

The operator

$$Tx = \{(x|e_k)\}; \; k = 1, 2, \ldots$$

is an isometry and maps $\mathscr{H}$ onto l^2. In fact,

$$\|x\|^2 = \sum_{k=1}^{\infty} |(x|e_k)|^2$$

by 2.2.2.1 (*b*) and, if $\{\xi_k;\; k=1, 2, \ldots\} \in l^2$, then

$$\sum_{k=1}^{\infty} \xi_k e_k \in \mathscr{H}$$

since

$$\Big(\sum_{k=m}^{n} \xi_k e_k \Big| \sum_{k=m}^{n} \xi_k e_k\Big) = \sum_{k=m}^{n} |\xi_k|^2$$

i.e.

$$\Big\{\sum_{k=1}^{n} \xi_k e_k\Big\} \qquad n = 1, 2, \ldots$$

is a Cauchy sequence.

2.11.2. If $\mathscr{H}_1$ and $\mathscr{H}_2$ are isomorphic Hilbert spaces with isomorphic operator U then a natural problem is to find for a bounded linear operator T of $\mathscr{H}_1$, an operator S of $\mathscr{H}_2$ with the following property: S sends Ux into UTx for every $x \in \mathscr{H}_1$. Formally,

$$SUx = UTx \qquad x \in \mathscr{H}.$$

Hence

$$U^{-1}SU = T \qquad \text{and} \qquad S = UTU^{-1}$$

or

$$U^*SU = T \qquad \text{and} \qquad S = UTU^*$$

since $U^{-1} = U^*$ by 2.10.4.2.

The bounded linear operator S is called *unitarily equivalent* to T.

The mapping $T \to UTU^{-1}$ preserves the main properties of T:

(i) $S^* = (UTU^{-1})^* = (UTU^*)^* = UT^*U^* = UT^*U^{-1}$ (see § 2.10.1(ii));

(ii) $S_1S_2 = (UT_1U^{-1})(UT_2U^{-1}) = UT_1T_2U^{-1}$;

(iii) $\alpha S_1 + \beta S_2 = U(\alpha T_1 + \beta T_2)U^{-1}$;

(iv) if T^{-1} exists then $S^{-1} = [UTU^{-1}]^{-1} = UT^{-1}U^{-1}$.

Remark. The structure of the operator S is shown in figure 2.10.

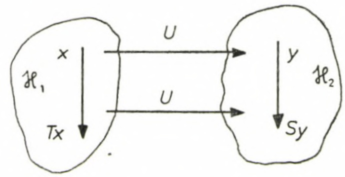

fig. 2.10

Example 1. Let

$$T_h f := hf \qquad f \in L^2[0, 2\pi]$$

be the operator of multiplication by $h \in C[0, 2\pi]$ and consider the isomorphic operator

$$Ff = \left\{ \frac{1}{2\pi} \int_0^{2\pi} f(t) e^{-ikt} \, dt \right\} \qquad k = 0, \pm 1, \pm 2, \ldots$$

from $L^2[0, 2\pi]$ onto l^2. Then T_h is a bounded linear operator of $L^2[0, 2\pi]$ and the operator in l^2, unitary equivalent to T_h, is

$$S_h\{c_k\} = \left\{ \sum_{n=-\infty}^{+\infty} d_{k-n} c_n \right\} \qquad k = 0, \pm 1, \pm 2, \ldots$$

where c_k is the sequence of the Fourier coefficients of f and d_k is the sequence of the Fourier coefficients of h. (Prove this!)

Example 2. Let T_h be the convolution operator

$$[T_h f](t) := \int_{-\infty}^{+\infty} h(t-\tau) f(\tau) \, d\tau \qquad f \in L^2(-\infty, +\infty)$$

by $h \in L^1(-\infty, +\infty)$. It can be proved that T_h is a bounded linear operator of $L^2(-\infty, +\infty)$.

If we consider the Fourier transform as an isomorphic operator (see Example 4 in § 2.11.1), then the multiplication operator

$$[S_h \hat{f}](\omega) = \hat{h}(\omega) \hat{f}(\omega) \qquad \hat{f} \in L^2(-\infty, +\infty)$$

is unitary equivalent to the convolution operator T_h since, by the Convolution Theorem of the Fourier transform,

$$F(h * f) = FhFf$$

where

$$[h * f](t) := \int_{-\infty}^{+\infty} h(t-\tau) f(\tau) \, d\tau.$$

Example 3. Consider the multiplication operator

$$[T_z f](z) = z f(z)$$

in H_0 and consider the congruence operator L (see Example 2 in § 2.11.1). Then the unitary equivalent operator is

$$S_z f = \sum_{k=0}^{\infty} c_{k+1} e^{ikt}$$

where

$$f = \sum_{k=0}^{\infty} c_k e^{ikt}$$

i.e. S_z is a right-hand shift in the Fourier coefficients of f. (Prove this!)

∗2.12 The conjugate gradient method

Let $\mathscr{H}$ be a *real* Hilbert space. In the following, we introduce a method for the solution of the equation

$$Tx = b$$

where $b \in \mathscr{H}$ is a given vector and T is a bounded strictly positive operator of $\mathscr{H}$. It will be shown that x_0 is the solution of the above equation if and only if

$$Q(x, x) := (Tx|x) - 2(b|x)$$

takes its minimum at x_0. Hence to seek the minimum of the functional $Q(x, x)$ called a quadratic from, is the same as to solve the equation $Tx=b$.

2.12.1. If the operator T has a positive lower bound m, i.e.

$$(Tx|x) > m(x|x) \qquad m > 0$$

then T is 1–1 and hence the solution of the above equation is unique; moreover, T is a strictly positive operator in this case and hence

$$(x|y)_T := (Tx|y)$$

is a scalar product. Comparing the norm

$$\|x\|_T := (x|x)_T^{1/2} = (Tx|x)^{1/2}$$

with the original norm in $\mathscr{H}$ we have

$$m\|x\|^2 \leqslant (Tx|x) \leqslant \|T\| \, \|x\|^2$$

and hence $\| \,.\, \|_T$ is equivalent to $\| \,.\, \|$. We conclude that if $\mathscr{H}_T$ is the scalar product space obtained from the scalar product $(\,.\, | \,.\,)_T$ then $\mathscr{H}_T$ is a Hilbert space consisting of the same elements as $\mathscr{H}$ and a sequence $\{x_n\}$ is convergent in $\mathscr{H}$ if and only if it is convergent in $\mathscr{H}_T$ with the same limit.

Following these considerations we seek the solution of the equation in the form

$$x_0 = x_1 + \sum_{j=1}^{\infty} \alpha_j e_j \tag{$*$}$$

where $\{e_j\}$ is a suitable *orthogonal system in* $\mathscr{H}_T$ and hence

$$\alpha_j = \frac{(x_0 - x_1|e_j)_T}{\|e_j\|_T^2} = \frac{(T(x_0 - x_1)|e_j)}{(Te_j|e_j)} = \frac{(b - Tx_1|e_j)}{(Te_j|e_j)}. \tag{$**$}$$

This orthogonal series method is called the conjugate gradient method if e_j; $j=0, 1, 2, \ldots$ is constructed in a particular way, to be described in the next sections.

2.12.2. We begin with the formal description of the conjugate gradient method. If x_k is the kth partial sum of a series in the form 2.12.1 ($*$) and

$$r_k = b - Tx_k \qquad k = 1, 2, \ldots$$

i.e. r_k is the error when the exact solution x_0 is replaced by x_k, then

$$r_{k+1} = r_k - \alpha_k T e_k \qquad k = 1, 2, \ldots \tag{$*$}$$

since

$$x_{k+1} = x_k + \alpha_k e_k$$

and hence

$$b - Tx_{k+1} = b - Tx_k - \alpha_k T e_k.$$

Now $\{e_k\}$; $k=1, 2, \ldots$ and the sequence $\{x_k\}$; $k=1, 2, \ldots$ of the approximate solutions are constructed in the following way. If $x_1 \in \mathscr{H}$ is considered as the first approximation of the solution, then the error is

$$r_1 = b - Tx_1$$

and $e_1 = r_1$. The next approximation is

$$x_2 = x_1 + \alpha_1 e_1$$

where

$$\alpha_1 = \frac{(r_1|e_1)}{(e_1|Te_1)}$$

and if the kth approximation x_k is obtained, then the kth error is

$$r_k = b - Tx_k$$

and the next approximation is

$$x_{k+1} = x_k + \alpha_k e_k \qquad \alpha_k = \frac{(r_k|e_k)}{(e_k|Te_k)}$$

where

$$e_k = r_k - \frac{(r_k|Te_{k-1})}{(e_{k-1}|Te_{k-1})} e_{k-1} \qquad k = 2, 3, \ldots.$$

Remark. Observe that we do not assume in the description of the method that $\{e_k\};\ k=1, 2, \ldots$ is an orthogonal system in $\mathscr{H}_T$. We shall prove later that $\{e_k\};\ k=1, 2, \ldots$ can also be obtained from $\{r_k\};\ k=1, 2, \ldots$ by the Gram–Schmidt process.

We shall show that x_k tends to the solution x_0 of the equation $b-Tx=\theta$ or, equivalently, $r_k \to \theta$.

2.12.3. For the proof of the convergence we need some preliminary results. For $n=1, 2, \ldots,$

$$(e_{n+1}|e_n)_T = (e_{n+1}|Te_n) = \left(r_{n+1} - \frac{(r_{n+1}|Te_n)}{(e_n|Te_n)} e_n \middle| Te_n\right) = 0. \qquad (*)$$

Moreover, from 2.12.2 (*),

$$(r_{n+1}|e_n) = (r_n|e_n) - \alpha_n(Te_n|e_n) = 0 \qquad (**)$$

since

$$\alpha_n = \frac{(r_n|e_n)}{(e_n|Te_n)}$$

and T is a positive operator.

If T has a positive lower bound m and an upper bound M, then T^{-1} has the lower bound $1/M$ and the upper bound $1/m$ (see § 4.13.24–25); hence

$$\|x\|_{T^{-1}} := (x|T^{-1}x)$$

is also a norm in $\mathscr{H}$, equivalent to the original norm. It follows that if

$$E(x_n) := (r_n|T^{-1}r_n)$$

then $E(x_n) \geq 0$ and $r_n \to \theta$ if and only if $E(x_n) \to 0$.

2.12.4. It follows from straightforward calculations (see, for example, 2.14.53) that

$$E(x_n) - E(x_{n+1}) = \alpha_n(r_n|e_n)$$

and, comparing the recursive formula for e_k in 2.12.2 with 2.12.3 (**), we obtain

$$(r_n|e_n) = (r_n|r_n) \qquad (*)$$

and hence

$$E(x_n) - E(x_{n+1}) = \alpha_n \frac{(r_n|r_n)}{(r_n|T^{-1}r_n)} E(x_n)$$

or

$$E(x_{n+1}) = \left(1-\alpha_n \frac{(r_n|r_n)}{(r_n|T^{-1}r_n)}\right) E(x_n).$$

It follows that $E(x_n)\to 0$ if we prove

$$1-\alpha_n \frac{(r_n|r_n)}{(r_n|T^{-1}r_n)} < 1$$

since $E(x_n)\geqslant 0$ by definition. In fact, from the recursive formula for e_k in § 2.12.2 and from 2.12.3 (∗),

$$(r_{n+1}|Tr_{n+1}) = \left(e_{n+1}+\frac{(r_{n+1}|Te_n)}{(e_n|Te_n)}e_n \middle| Te_{n+1}+\frac{(r_{n+1}|Te_n)}{(e_n|Te_n)}Te_n\right)$$

$$= (e_{n+1}|Te_{n+1})+\gamma^2(e_n|Te_n) \geqslant (e_{n+1}|Te_{n+1})$$

and hence, using (∗), we obtain

$$\alpha_n \geqslant \frac{(e_n|r_n)}{(r_n|Tr_n)} = \frac{(r_n|r_n)}{(r_n|Tr_n)}.$$

On the other hand,

$$\frac{(r_n|r_n)}{(r_n|Tr_n)} \geqslant \frac{1}{M}$$

since M is an upper bound for T; moreover, from the considerations in § 2.12.3,

$$\frac{(r_n|r_n)}{(r_n|T^{-1}r_n)} \geqslant m.$$

Summing up,

$$\alpha_n \frac{(r_n|r_n)}{(r_n|T^{-1}r_n)} \geqslant \frac{m}{M}$$

and hence

$$1-\alpha_n \frac{(r_n|r_n)}{(r_n|T^{-1}r_n)} \leqslant 1-\frac{m}{M} < 1.$$

2.12.5. The conjugate gradient method is partly motivated by the endeavour to find the solution in the form of an orthogonal series in $\mathscr{H}_T$ such that the successive errors $\{r_k;\ k=1, 2, \ldots\}$ are built into the partial sums $\{x_k;\ k=1, 2, \ldots\}$. Now we shall show some peculiar features of the conjugate gradient method in relation to this endeavour.

If $r_n=\theta$ for any n then x_n is the exact solution and so is obtained in the form of a finite sum. *We suppose in what follows that* $r_n\neq\theta$ $(n=1, 2, \ldots)$.

2.12.5.1 Theorem. If $r_n \neq \theta$ for $n=1, 2, \dots,$ then $\{r_k;\ k=1, 2, \dots\}$ are linearly independent and also

$$(r_n|e_j) = 0 \qquad \text{for} \quad j < n.$$

Proof. By immediate calculation, the theorem is valid for the pair r_1, r_2. Let us suppose that it holds for $r_k;\ k=1, 2, \dots, n;$ then from 2.12.2 (∗),

$$(r_{n+1}|e_j) = (r_n|e_j) - \alpha_n(Te_n|e_j) = 0$$

if $j<n$ and $(r_{n+1}|e_n)=0$ by 2.12.3 (∗∗).

If $r_1, r_2, \dots, r_n$ are linearly independent but this is not valid for $r_1, r_2, \dots$ $\dots, r_n, r_{n+1}$, then

$$r_{n+1} = \sum_{k=1}^{n} \alpha_k r_k.$$

Multiplying both sides by $e_j;\ j=1, 2, \dots, n$ successively, we obtain $\alpha_k=0;$ $k=1, 2, \dots, n$ and hence $r_{n+1}=\theta$.

If the Gram–Schmidt process is applied to $\{r_k;\ k=1, 2, \dots\}$ (without normed!), then

$$e_k = r_k - \sum_{j=1}^{k-1} \frac{(r_k|Te_j)}{(e_j|Te_j)} e_j. \tag{$*$}$$

We shall prove that by choosing $\{r_k;\ k=1, 2, \dots\}$ as an infinite basis for the Gram–Schmidt process, e_k is obtained in the simplest form:

$$(r_k|Te_j) = 0 \qquad \text{if} \quad j < k-1$$

in (∗) and hence e_k is the same as in the algorithm in § 2.12.2.

Remark 1. There is another advantage of choosing $\{r_n\}$: in this case the error is measured during the evaluation of $\{e_k\}$.

Remark 2. We conclude from the structure of the Gram–Schmidt process that $(r_n|e_j) = 0$ for $j < n$ implying that $(r_n|r_j) = 0$ for $j < n$ (prove this!).

2.12.5.2 Theorem. For $j<k-1$,

$$(r_k|Te_j) = 0.$$

Proof. It follows from 2.12.2 (∗) that

$$(r_{j+1}|r_k) = (r_j|r_k) - \alpha_j(Te_j|r_k)$$

and hence if $j<k-1$ then it follows from Remark 2 that

$$\alpha_j(Te_j|r_k) = 0.$$

If $\alpha_j=0$ then it also follows from 2.12.2 (*) that $r_{j+1}=r_j$ and, by induction, $r_n=r_j$ for $n>j$. But this implies $r_j=\theta$ since $r_k\to\theta$ by 2.12.4. Hence if $r_j\neq\theta$; $j=1, 2, \ldots,$ then

$$(r_k|Te_j) = (Te_j|r_k) = 0.$$

2.12.6. We shall show that x_0 is the solution of the equation $b-Tx=\theta$ if and only if $Q=Q(x, x)$ takes its minimum at x_0. In fact, for every fixed $c\in\mathscr{H}$,

$$(T(x-c)|x-c) = (Tx|x)-2(Tc|x)+(Tc|c)$$

since $\mathscr{H}$ is real and $T=T^*$. Comparing with $Q(x, x)$, if we substitute $b=Tc$, we obtain

$$Q(x, x) = (Tx-b|T^{-1}(Tx-b))-(b|T^{-1}b).$$

It follows that the minimum value of Q is $-(b|T^{-1}b)$ since T^{-1} is also a positive operator (see § 4.13.24).

*2.13 Construction of a separating hyperplane

The simplest case of classification of a finite set of data consisting of numbers, strings of numbers, functions etc is when the set is divided into only two classes:

$\mathscr{A}$:=the set of 'good' elements;

$\mathscr{B}$:=the set of 'wrong' elements.

In this case the classification should be done by means of a function f in such a way that $f(x)>0$ if $x\in\mathscr{A}$ and $f(x)<0$ if $x\in\mathscr{B}$. In the next section a Hilbert space model will be given for this type of classification.

2.13.1. Referring to § 2.5.2 a hyperplane $\mathscr{S}$ of a pre-Hilbert space $\mathscr{H}$ passing through the origin θ has the form

$$\mathscr{S} = \{y\colon (y|m) = 0\} \qquad m\in\mathscr{M}.$$

In this section $\mathscr{M}$ is a one-dimensional *subspace* and $\mathscr{H}$ is a *real* (pre-) Hilbert space.

The *distance* between a set $\mathscr{S}$ and $x\notin\mathscr{S}$ in a pre-Hilbert space is defined as

$$d = \inf_{y\in\mathscr{S}} \|x-y\|.$$

2.13.1.1 Theorem. If $\mathscr{M}$ is generated by a single vector z with $\|z\|=1$ then the distance of the hyperplane $\mathscr{S}$ and $x \notin \mathscr{S}$ is

$$d = |(x|z)|.$$

Proof. In this case, $\mathscr{S}$ is also a closed linear subspace of the pre-Hilbert space $\mathscr{H}$ and hence, by 2.5.1.1, there exists $y_0 \in \overline{\mathscr{S}}$, where $\overline{\mathscr{S}}$ is the completion of $\mathscr{S}$, such that

$$\inf_{y \in \mathscr{S}} \|x-y\| = \|x-y_0\|$$

and hence, by 2.4.1.3,

$$\|x-y_0\|^2 = (x-y_0|x)$$

from which we have

$$d = \|x-y_0\| = \left(\frac{x-y_0}{\|x-y_0\|}\middle| x\right).$$

On the other hand, $x-y_0 \in \mathscr{S}^\perp = \mathscr{M}^{\perp\perp} = \mathscr{M}$ and $\mathscr{M}$ is one dimensional; hence

$$z = \frac{x-y_0}{\|x-y_0\|}.$$

Remark. It turns out from the proof and especially from 2.14.20 that $y_0 \in \mathscr{S}$ if $\mathscr{M}$ is finite dimensional, i.e. there exists a projection in $\mathscr{S}$ for every $x \in \mathscr{H}$.

The pre-Hilbert space $\mathscr{H}$ is divided into three parts by the hyperplane $\mathscr{S}$: $\{y\colon (y|z) > 0\}$, called the positive halfspace of $\mathscr{S}$, $\{x\colon (x|z) < 0\}$, called the negative halfspace of $\mathscr{S}$ and $\{y\colon (y|z) = 0\}$, the hyperplane $\mathscr{S}$ itself.

2.13.2. Let $\mathscr{A}$ and $\mathscr{B}$ be two finite subsets of $\mathscr{H}$; our task is to find $z_0 \in \mathscr{H}$ such that

$$(x|z_0) > 0 \quad \text{if} \quad x \in \mathscr{A} \quad \text{and} \quad (x|z_0) < 0 \quad \text{if} \quad x \in \mathscr{B}. \qquad (*)$$

The hyperplane $\mathscr{S}_0 = \{y\colon (y|z_0) = 0\}$ is called a hyperplane separating $\mathscr{A}$ and $\mathscr{B}$. In what follows we suppose that there exists a separating hyperplane for $\mathscr{A}$ and $\mathscr{B}$, and an algorithm will be given for the construction of this separating hyperplane.

For each element $x_k \in \mathscr{A} \cup \mathscr{B}$ we construct ξ_k in such a way that $\xi_k = +1$ if $x_k \in \mathscr{A}$ and $\xi_k = -1$ if $x_k \in \mathscr{B}$, and hence $\mathscr{S}$ is a separating hyperplane for a sequence $\{x_k;\ k = 1, 2, \dots\}$ belonging to $\mathscr{A} \cup \mathscr{B}$ if and only if

$$\xi_k(x_k|z) > 0.$$

Now let

$$\mathscr{S}_0 = \{x\colon (x|z_0) = 0\};$$

if there exists n such that $\xi_n(x_n|z_0) \leq 0$ then $\mathscr{S}_0$ is not separating and a correction is needed. In this case let

$$z_1 = z_0 + \xi_n x_n$$

and we consider the hyperplane $\mathscr{S}_1 = \{y: (y|z_1) = 0\}$. Again, if $\xi_k(x_k|z_1) > 0$ for $k = 1, 2, \ldots$ then $\mathscr{S}_1$ is a separating hyperplane; if not — i.e. there exists m such that $\xi_m(x_m|z_1) \leq 0$ — then a correction is needed once more:

$$z_2 = z_1 + \xi_m x_m$$

and the trial will be continued by the hyperplane $\mathscr{S}_2 = \{y: (y|z_2) = 0\}$. The process is continued until $\mathscr{A}$ and $\mathscr{B}$ are separated by a hyperplane $\mathscr{S}$.

Thus the following algorithm is obtained: $z_0 = \theta$ and, after the jth correction,

$$z_{k+1} = \begin{cases} z_j & \text{if} \quad \xi_k(x_k|z_j) > 0 \\ z_j + \xi_k x_k & \text{if} \quad \xi_k(x_k|z_j) \leq 0 \end{cases} \qquad j \leq k. \qquad (**)$$

Our main subject in this section is an estimation of the numbers of hyperplanes $\mathscr{S}_k$ (or the number of corrections, which is the same) needed for a separating hyperplane to be obtained in the algorithm.

2.13.2.1 Theorem. Let $\mathscr{H}$ be a pre-Hilbert space and $\{x_k;\ k = 1, 2, \ldots, N\}$ a finite sequence of elements of $\mathscr{H}$ divided into classes $\mathscr{A}$ and $\mathscr{B}$. If there exists a hyperplane $\mathscr{S}$ with $\theta \in \mathscr{S}$, separating $\mathscr{A}$ and $\mathscr{B}$ in the stronger sense

$$|(x_k|z)| \geq d > 0 \qquad k = 1, 2, \ldots, N$$

then the algorithm $(**)$ leads to a hyperplane separating $\mathscr{A}$ and $\mathscr{B}$ *after a finite number of corrections.* More particularly, if n is the number of corrections then

$$n \leq M^2/d^2$$

where

$$M \geq \|x_k\| \qquad k = 1, 2, \ldots, N.$$

Remark 1. On the basis of 2.13.1.1. the condition $|(x_k|z)| \geq d > 0$ means exactly that the subsets $\mathscr{A}$ and $\mathscr{B}$ have a positive distance.

Remark 2. The number of corrections is in an inverse ratio to d^2 and proportional to M^2, *but it does not depend on the number of elements $\mathscr{A}$ and $\mathscr{B}$.*

Remark 3. In the most favourable case, when the separating hyperplane is given by the very first vector z_0, we need N scalar product tests to be assured that $\mathscr{S}_0$ is separating. Hence in this theorem we have no information about the number of scalar product tests.

Proof. After the nth correction,

$$z_n = \sum_{k=1}^{n} \xi_k' x_k'$$

where x_k' is the kth element for which the scalar product test is negative, i.e.

$$\xi_k'(x_k'|z_{k-1}) \leqslant 0.$$

Hence

$$\|z_n\|^2 = \|z_{n-1}+\xi_n' x_n'\|^2 = \|z_{n-1}\|^2+\|x_n'\|^2+2\xi_n'(x_n'|z_{n-1}) \leqslant \|z_{n-1}\|^2+M^2$$

since $\xi_n'(x_n'|z_{n-1})\leqslant 0$; more particularly,

$$\begin{aligned} \|z_1\|^2 &\leqslant M^2 \\ \|z_2\|^2 &\leqslant \|z_1\|^2+M^2 \leqslant 2M^2 \\ &\vdots \\ \|z_n\|^2 &\leqslant \|z_{n-1}\|^2+M^2 \leqslant nM^2. \end{aligned} \tag{1}$$

On the other hand, if $\mathscr{S}=\{y\colon (y|z)=0\}$ is a separating hyperplane with $\|z\|=1$, then

$$\|z_n\| \geqslant (z_n|z) = \sum_{k=1}^{n} \xi_k'(x_k'|z) = \sum_{k=1}^{n} |(x_k'|z)| \geqslant nd. \tag{2}$$

Comparing (1) and (2), we have

$$nd \leqslant \|z_n\| \leqslant n^{1/2}M$$

and hence

$$n \leqslant M^2/d^2.$$

2.14 Problems and notes

∘**2.14.1.** Find the condition for the Pythagorean law

$$\|x+y\|^2 = \|x\|^2+\|y\|^2 \qquad x, y\in\mathscr{H}$$

to be valid.

∘**2.14.2.** Prove that if

$$f(t) = A_1 \sin(t+\varphi_1) \qquad g(t) = A_2 \sin(t+\varphi_2)$$

then

$$\frac{1}{2\pi}\int_0^{2\pi} f(t)\overline{g(t)}\,\mathrm{d}t = \operatorname{Re} A_1 A_2 \mathrm{e}^{\mathrm{i}(\varphi_1-\varphi_2)}.$$

2.14.3. Prove, e.g. from the Cauchy inequality 2.1.2.1, that for every $z\in\mathscr{H}$,

$$\|z\| = \sup\{(x|z);\ \|x\| = 1\}.$$

∘**2.14.4.** Prove that for any scalar product space $\mathscr{H}$,

$$(x|y) = \frac{1}{2\pi} \int_0^{2\pi} \|x + e^{it} y\|^2 e^{it}\, dt.$$

∘**2.14.5.** Show that in any pre-Hilbert space $\mathscr{H}$,

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2 \qquad x, y \in \mathscr{H}.$$

Give a geometrical meaning to this equation, which is called the *parallelogram law.*

2.14.6. Prove that if the parallelogram law holds in a normed space $\mathscr{B}$, then $\mathscr{B}$ is a scalar product space, i.e. there is a scalar product in $\mathscr{B}$ such that

$$\|x\| = (x|x)^{1/2}.$$

2.14.7. If $\{x_n\}$ is a sequence in $\mathscr{H}$ and $\lim x_n = x$ then it follows from 2.1.2.3 that

$$(x_n|y) \to (x|y) \qquad y \in \mathscr{H}$$

but the converse does not hold. For example, for any orthonormal sequence $\{e_k\}$ we have $(e_k|y) \to 0$ for every $y \in \mathscr{H}$; however, $\|e_k\| = 1$ and $\{e_k\}$ is not a convergent sequence since

$$\|e_{k+1} - e_k\|^2 = 2$$

for any orthonormal $\{e_k\}$.

2.14.8. The function f, shown in figure 2.11, is an example of $f \in L^2(0, \infty)$ that does not converge to zero at infinity.

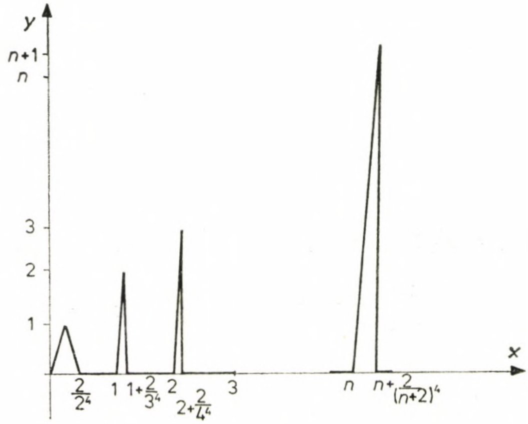

fig. 2.11

2.14.9. Show that if $e_k \neq 0$; $k=1, 2, \ldots, n$, and

$$(e_i|e_k) = 0 \qquad i \neq k$$

for $\{e_k;\ k=1, 2, \ldots, n\}$ then these are linearly independent vectors.

∘**2.14.10.** Let $\{e_k\}$ be an orthonormal system in a Hilbert space $\mathscr{H}$. Prove that $\sum_{k=1}^{\infty} \gamma_k e_k$ is convergent if and only if $\sum_{k=1}^{\infty} |\gamma_k|^2 < \infty$.

Is this true if $\mathscr{H}$ is only a scalar product space?

∘**2.14.11.** Prove that if $\{a_k\}$ is a complete sequence and $\{e_k\}$ is the orthonormal system obtained from $\{a_k\}$ by the Gram–Schmidt process, then $\{e_k\}$ is also complete.

2.14.12. Let $\{e_k(t);\ k=1, 2, \ldots\}$ be an orthonormal system in the *real* $L_0^2[a, b]$. Show that $\{e_k\}$ is complete if and only if

$$\sum_{k=1}^{\infty} \Big(\int_a^x e_k(t)\, dt \Big)^2 = x - a \qquad x \in (a, b).$$

2.14.13. Apply the Gram–Schmidt process to the sequence $1, z, z^2, \ldots, z^n, \ldots$ if

$$(f|g) := \iint_{|z|<R} f(z)\overline{g(z)}\, dx\, dy$$

in the linear space of functions analytic in the unit disc and continuous on $\{z:\ |z|=1\}$.

2.14.14. Show that $\{(2/\pi)^{1/2} \sin kt;\ k=1, 2, \ldots\}$ is a complete orthonormal system in $L^2[0, \pi]$.

2.14.15. If the Fourier coefficients $\{c_k\}$ and $\{d_k\}$ of the continuous functions f and g, respectively, are very close to each other in the sense

$$\sum_{k=-\infty}^{+\infty} |c_k - d_k|^2 < \varepsilon$$

then

$$\frac{1}{2\pi} \int_0^{2\pi} |f(t) - g(t)|^2\, dt < \varepsilon.$$

However, from the Fourier coefficients we cannot say anything about the magnitude of $|f(t) - g(t)|$; $t \in [0, 2\pi]$.

2.14.16. Show that the matrices $\mathbf{H}_N$, defined for $N=2^k$; $k=1, 2, \ldots$ in § 2.3.3 have an inverse. What is $\mathbf{H}_N^{-1}$?

2.14.17. Show that the step functions w_k^* defined in § 2.3.3 are Walsh functions.

2.14.18. In § 2.4.1.4 we found a formula for the projection of $x \in \mathscr{H}$ onto $\mathscr{M}$ in the case of finite-dimensional $\mathscr{M}$. Give a formula for the projection if $\mathscr{M}$ is a *separable* subspace of a Hilbert space $\mathscr{H}$.

2.14.19. If $\mathscr{M}$ is a complete linear subspace, then $\mathscr{M}^{\perp\perp} := (\mathscr{M}^\perp)^\perp = \mathscr{M}$. In fact, it is obvious that $\mathscr{M} \subseteq \mathscr{M}^{\perp\perp}$. Now let y be an element of $\mathscr{M}^{\perp\perp}$ such that $y \notin \mathscr{M}$ and let y_p be the projection of y onto $\mathscr{M}$; then $y - y_p \in \mathscr{M}^\perp$ by the projection theorem and hence $y - y_p \in \mathscr{M}^\perp \cap \mathscr{M}^{\perp\perp}$. But $\mathscr{M}^\perp \cap \mathscr{M}^{\perp\perp} = \{\theta\}$.
Find $\mathscr{M}^{\perp\perp}$ in the case where $\mathscr{M}$ is neither complete nor linear.

2.14.20. Let $\mathscr{M}$ be a complete subspace of a scalar product space $\mathscr{H}$ and $x_0 \notin \mathscr{M}$. Prove the following generalisation of 2.13.1.1:

$$\min\{\|x_0 - m\|;\ m \in \mathscr{M}\} = \max\{(x_0|y);\ y \in \mathscr{M}^\perp, \|y\| = 1\}.$$

2.14.21. Find the orthogonal complement in $L^2[0, 1]$ of the following sets:
(*a*) the polynomials in x;
(*b*) the polynomials in x^2;
(*c*) the polynomials with the *sum* of coefficients equal to zero.

∘**2.14.22.** Let

$$\mathscr{M} = \left\{z\colon \int_s^1 z(t)\,\mathrm{d}t = 0;\ z \in L_0^2[0, 1]\right\}$$

where $s \in (0, 1)$ is a fixed value depending on z. Show that $\mathscr{M}^\perp = \{\theta\}$.

2.14.23. Prove that the orthogonal complement of $\{e^{2\pi i n t};\ n = 0, \pm 1, \pm 2, \ldots\}$ in $L^2[a, b]$ is

(*a*) $\{\theta\}$ if $|b-a| \leq 1$;
(*b*) *Not* $\{\theta\}$ if $|b-a| > 1$.

∘**2.14.24.** Prove that $(AB)^{-1} = B^{-1}A^{-1}$, where A, B, A^{-1}, B^{-1} are bounded linear operators of $\mathscr{H}$.

2.14.25. If $\mathscr{H}$ is a Hilbert space and $\mathscr{M}$ is a closed linear subspace of $\mathscr{H}$ then there exists $\mathcal{N} \subset \mathcal{H}$ such that $\mathcal{M} = \mathcal{N}^{\perp}$. Is this true for any scalar product space?

∘**2.14.26.** Find the polynomial P_n of nth degree such that, for a given $x \in L^2[0, 1]$,

$$\int_0^1 |x(t) - P_n(t)|^2 \, dt$$

is minimal. (First solve the problem for $n=2$ and $x(t) = \sin 2\pi t$.)

2.14.27. For a fixed integer n,

$$e_k(t) = \begin{cases} 1 & \text{if } k/n \leq t < (k+1)/n \\ 0 & \text{elsewhere.} \end{cases}$$

Show that $\{e_k\}$ is a non-complete orthogonal system in $L^2[0, 1]$ and find the best approximation of $x \in L^2[0, 1]$ in the form

$$\sum_{k=0}^{n} \xi_k e_k(t).$$

2.14.28. Find the polynomial of the form

$$P_n(t) = t^n + \alpha_{n-1} t^{n-1} + \ldots + \alpha_1 t + \alpha_0$$

in $[-1, +1]$ with minimal square integral.

2.14.29. As we saw in § 2.8, the proof of the Riesz–Fréchet Theorem for non-separable $\mathcal{H}$ is based on 2.5.4.2. We shall now show that we can deduce 2.5.4.2 from the Riesz–Fréchet Theorem.

In fact, for any $x_0 \in \mathcal{H}$, $f(m) = (m|x_0)$; $m \in \mathcal{M}$ is a continuous linear functional of a closed linear subspace $\mathcal{M}$. Consider $\mathcal{M}$ as a Hilbert space by itself and supposing the Riesz–Fréchet Theorem to be valid; then there is a unique $m_0 \in \mathcal{M}$ such that

$$(m|m_0) = (m|x_0) \qquad m \in \mathcal{M}$$

and hence $x_0 - m_0 \in \mathcal{M}^{\perp}$ and

$$(x_0 - m_0) + m_0$$

is the orthogonal direct sum $\mathcal{M}^{\perp} \oplus \mathcal{M}$ decomposition of x_0.

2.14.30. Find $x = x(t)$ among the functions

$$\left\{x \colon \int_{1/n}^{1} x(t) \, dt = 1 - \frac{1}{n}\right\}$$

such that the integral

$$\int_0^1 \int_0^t (t-\tau)^2 x(\tau)\overline{x(t)}\, d\tau\, dt$$

is minimal.

2.14.31. Let $\{x_n\}$ be an infinite sequence of real numbers such that

$$\sum_{n=1}^{\infty} x_n^2 = \infty.$$

Show that there exists an infinite sequence $\{a_n\}$ of real numbers with

$$\sum_{k=1}^{\infty} a_k^2 < \infty$$

such that

$$\sum_{k=1}^{\infty} a_k x_k$$

does not converge.

2.14.32. We have the following generalisation of the previous observation. For every infinite sequence $\{x_n\}$ of a Hilbert space $\mathscr{H}$ with

$$\sum_{n=1}^{\infty} \|x_n\|^2 = \infty$$

there exists $\{a_n\}$ with

$$\sum_{n=1}^{\infty} \|a_n\|^2 < \infty$$

such that

$$\sum_{k=1}^{\infty} (a_k|x_k)$$

does not converge.

∘**2.14.33.** Prove that if $a_n, b_n \in \mathscr{H}$ for $n=1, 2, \ldots$ and

$$\sum_{k=1}^{\infty} \|a_n\|^2 < \infty \qquad \sum_{n=1}^{\infty} \|b_n\|^2 < \infty$$

then

$$\sum_{n=1}^{\infty} (a_n|b_n)$$

is a convergent series.

∘**2.14.34.** Prove that

$$(x|y) = \frac{1}{4}\{\|x+y\|^2 - \|x-y\|^2 + \mathrm{i}\,\|x+\mathrm{i}y\|^2 - \mathrm{i}\,\|x-\mathrm{i}y\|^2\}$$

and in particular, in a *real* Hilbert space,

$$(x|y) = \frac{1}{4}\{\|x+y\|^2-\|x-y\|^2\}.$$

2.14.35. Prove that a bounded bilinear functional is continuous in the following sense. If $x_n \to x$ and $y_n \to y$ then $\varphi(x_n, y_n) \to \varphi(x, y)$.

∘**2.14.36.** Let $\mathscr{H}_1$, $\mathscr{H}_2$ be separable Hilbert spaces and let U be a linear operator from $\mathscr{H}_1$ onto $\mathscr{H}_2$. Prove that the operator U is an isomorphic operator if and only if for any orthonormal system $\{e_k\}$ in $\mathscr{H}_1$, $\{Ue_k\}$ is an orthonormal system in $\mathscr{H}_2$.

∘**2.14.37.** Prove that for every bounded linear operator T, self-adjoint operators A and B can be found such that

$$T = A+\mathrm{i}B.$$

2.14.38. Let T be a positive operator. Is it true that

$$|(Tx|y)|^2 \leqslant (Tx|x)(Ty|y)?$$

What about the projection principle, if the scalar product is replaced by $(x|y)_T := (Tx|y)$?

2.14.39. If the equation

$$Tx = b \tag{$*$}$$

has no solution in a pre-Hilbert space $\mathscr{H}$, then $x_0 \in \mathscr{H}$ satisfying

$$\|b-Tx_0\| = \inf\{\|b-Tx\|;\ x\in\mathscr{H}\}$$

is called a *generalised solution* of the equation. Notice that x_0 is a generalised solution if and only if the quadratic form

$$(T^*Tx|x)-(x|T^*b)-(T^*b|x)+(b|b)$$

has a (relative) minimum in x_0, since this expression is equal to $\|b-Tx\|^2$. (Compare with 2.12.6.) Moreover, by 2.4.1.3, x_0 is a generalised solution if and only if

$$(b-Tx_0|Tz) = 0$$

for every $z\in\mathscr{H}$ and hence, if the inverse operator $(T^*T)^{-1}$ exists,

$$x_0 = (T^*T)^{-1}T^*b.$$

If the generalised solution is unique, then the operator $T^{\sim}$ defined by $T^{\sim}b = x_0$, i.e. the operator that maps $b \in \mathscr{H}$ into the generalised solution of (*), is called the *generalised inverse* of T.

It follows from the above considerations that if $(T^*T)^{-1}$ exists then $T^{\sim} = (T^*T)^{-1}T^*$. (Compare with 2.6.2.)

2.14.40. Considering a bilinear functional as a generalisation of the scalar product, as was established at the beginning of § 2.9, *formulate* the analogue of 2.4.1.3 for any bilinear functional φ.

What are the conditions for φ for 2.4.1.3 to be valid if the scalar product is replaced by φ?

∘**2.14.41.** Let T be a positive operator and let y_i; $i=1, 2, \ldots, n$ be linearly independent vectors of a pre-Hilbert space $\mathscr{H}$. Find $x_0 \in \mathscr{H}$ that satisfies the following conditions:

(i) $(x_0|y_i) = \eta_i \qquad i=1, 2, \ldots, n$

where η_i; $i=1, 2, \ldots, n$ are given;

(ii) the quadratic functional $x \to (Tx|x)$ is minimal if $x = x_0$.

Show that if the inverse T^{-1} exists then a solution of this problem is

$$x_0 = \sum_{k=1}^{n} \xi_k y_k$$

where $\{\xi_k;\ k=1, 2, \ldots, n\}$ is the solution of the following system of linear equations:

$$\sum_{k=1}^{n} \xi_k (y_k|Ty_i) = \eta_i \qquad i = 1, 2, \ldots, n.$$

2.14.42. If the linear space generated by $\mathscr{S} \subset \mathscr{H}$ is dense in $\mathscr{H}$, then from

$$(x_0|z) = 0 \qquad \text{for every } z \in \mathscr{S} \qquad (*)$$

it follows that $x_0 = \theta$. In fact, in this case there is a sequence $\{z_n\}$; $n=1, 2, \ldots$. of linear combinations of elements of $\mathscr{S}$ such that $z_n \to x_0$; moreover,

$$(x_0|z_n) = 0 \qquad n = 1, 2, \ldots.$$

Hence

$$(x_0|x_0) = 0.$$

Prove (e.g. from 2.5.4.2) the following converse statement: If from (*) it follows that $x_0 = \theta$, then the linear space generated by $\mathscr{S}$ is dense in $\mathscr{H}$.

∘**2.14.43.** Prove that $A \leqslant B$ implies $\|A\| \leqslant \|B\|$ for the self-adjoint operators A and B.

∘**2.14.44.** Prove that projection operators are positive.

2.14.45. Let P_k; $k=1, 2, \ldots, n$ be projection operators such that

$$P_i P_j = 0 \quad \text{if} \quad i \neq j \quad \text{and} \quad \sum_{k=1}^{n} P_k = E$$

where E is the identity operator. What can be said about the subspaces

$$\mathcal{M}_k = P_k \mathcal{H} := \{P_k x;\ x \in \mathcal{H}\}?$$

∘**2.14.46.** Let $P_{\mathcal{M}}$ and $P_{\mathcal{N}}$ be the projection operators onto $\mathcal{M}$ and $\mathcal{N}$, respectively, and $\mathcal{M} \cap \mathcal{N} = \{\theta\}$. Show that $P_{\mathcal{M}} + P_{\mathcal{N}}$ is also a projection. If

$$P = P_{\mathcal{M}} + P_{\mathcal{N}}$$

then find the closed linear subspace $\mathcal{K}$ such that $P = P_{\mathcal{K}}$.

2.14.47. Let $\{\varphi_n\}$; $n=1, 2, \ldots$ and $\{\psi_n\}$; $n=1, 2, \ldots$ be complete orthonormed systems in $L^2[a, b]$. Then

$$e_{kn}(t, \tau) := \varphi_k(t)\psi_n(\tau) \qquad k, n = 1, 2, \ldots$$

form a complete orthonormal system in the L^2-space of functions of two variables with scalar product

$$(f|g) := \int_a^b \int_a^b f(t, \tau)\overline{g(t, \tau)}\, dt\, d\tau.$$

2.14.48. (*a*) Prove that

$$TT^* \geqslant 0 \qquad \text{and} \qquad T^*T \geqslant 0$$

for any bounded linear T.

(*b*) Prove that if TT^* and T^*T have a lower bound $m>0$, then T has a bounded inverse T^{-1}. (Compare with 2.14.39.)

2.14.49. If T^*T has no positive lower bound then there exists $\{x_n\}$; $n=1, 2, \ldots$ such that

$$\|x_n\| = 1 \qquad \text{and} \qquad (T^*Tx_n|x_n) \to 0.$$

In this case there is no bounded inverse T^{-1} since if we suppose that bounded T^{-1} exists, then

$$\|x_n\| = \|T^{-1}Tx_n\| \leqslant \|T^{-1}\|\,\|Tx_n\|.$$

Moreover,

$$\|Tx_n\|^2 = (Tx_n|Tx_n) = (T^*Tx_n|x_n).$$

Hence our supposition led to a contradiction.

2.14.50. Prove (e.g. by applying 2.11.1.1) that every two separable Hilbert spaces $\mathscr{H}_1$ and $\mathscr{H}_2$ are isomorphic.

2.14.51. Show that the orthogonal direct sum decomposition

$$\mathscr{H} \supseteq \mathscr{M} = \bigoplus_{i=1}^{n} \mathscr{N}_i$$

is unique. (See 2.5.4.1.)

2.14.52. Show that the functions $a_i(t)$; $i=0, 1, 2, \ldots, N$ of Example 2 in § 2.4.1 form a basis in the linear space L of functions which plot as a broken line with nodes only at $\{t_i;\ i=1, 2, \ldots, N\}$.

More precisely, $f \in \mathrm{L}$ if and only if f is continuous and, in each of the intervals, (t_{i+1}, t_i); $i=0, 1, \ldots, N-1$ is a polynomial of first degree.

2.14.53. The calculation that led to the formula

$$E(x_n)-E(x_{n+1}) = \alpha_n(r_n|e_n)$$

in § 2.12.4 is as follows.

$$\begin{aligned}(r_{n+1}|T^{-1}r_{n+1}) &= (r_n-\alpha_n Te_n|T^{-1}(r_n-\alpha_n Te_n)) \\ &= (r_n|T^{-1}r_n)+\alpha_n^2(Te_n|e_n)-\alpha_n(Te_n|T^{-1}r_n)-\alpha_n(r_n|e_n)\end{aligned}$$

and hence

$$\begin{aligned}E(x_n)-E(x_{n+1}) &= (r_n|T^{-1}r_n)-(r_{n+1}|T^{-1}r_{n+1}) \\ &= \alpha_n(r_n|e_n)+\alpha_n(Te_n|T^{-1}r_n)-\alpha_n^2(Te_n|e_n) \\ &= \alpha_n(e_n|r_n)+\alpha_n[(e_n|r_n)-\alpha_n(Te_n|e_n)]\end{aligned}$$

since $T>0$ and $\mathscr{H}$ is a real Hilbert space. Moreover,

$$\alpha_n(Te_n|e_n) = (r_n|e_n).$$

2.14.54. Prove that every Gram matrix is positive definite in the sense that if a_{ik}; $i, k=1, 2, \ldots, n$ are the elements of the matrix then

$$\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i\bar{x}_k \geqslant 0$$

for every n-tuple $\{x_i\}$; $i=1, 2, \ldots, n$ of complex numbers.

Is the converse statement also true? Is it true that every positive definite matrix is a Gram matrix?

2.14.55. The Gram–Schmidt process can also be considered as a particular case of the projection principle. If we choose γ_k; $k=1, 2, \ldots, n$ such that

$$a_n - \sum_{k=1}^{n-1} \gamma_k e_k$$

is orthogonal to the vectors e_k; $k=1, 2, \ldots, n-1$, then

$$\sum_{k=1}^{n-1} \gamma_k e_k$$

is the projection of a_n onto the subspace generated by $\{e_k;\ k=1, 2, \ldots, n-1\}$.

2.14.56. Can it be proved that

$$\sup\{|\varphi(x, y)|;\ \|x\| \leqq 1, \|y\| \leqq 1\} = \sup\{\varphi(z, z);\ \|z\| \leqq 1\}$$

for a symmetric φ? (For example, another proof for 2.10.2.2 using 2.9.1.1.)

3

Reproducing Kernel Hilbert Spaces

We have shown in § 2.11 that every separable Hilbert space is isomorphic to the l^2 space, so that they may be considered as the same from a Hilbert space point of view; however, they could be very different. In this chapter we shall study Hilbert spaces of certain functions that also have interesting function theoretic properties.

3.1 Hilbert space and kernel

3.1.1. In a linear space B of complex or real-valued functions on a set $\mathscr{D}$, for every $t \in \mathscr{D}$ we have the linear functional

$$f \to f(t) \qquad f \in B$$

called the *evaluation functional.*

In many important Hilbert spaces of functions the evaluation functionals are continuous, but there are also important ones with non-continuous evaluation functionals. For example, the elements $\{x_k;\ k=1, 2, \ldots\}$ of the l^2-space are usually considered as functions on the positive integers and it is obvious that

$$|x_k| \leq \Big(\sum_{i=1}^{\infty} |x_i|^2\Big)^{1/2}.$$

Hence the evaluation functionals in l^2 are continuous.

Consider the functions $f_n = f_n(t)$:

$$f_n(t) = \begin{cases} n(t-t_0)+n^{1/2} & \text{if } t \in (t_0 - 1/n,\ t_0] \\ -n(t-t_0)+n^{1/2} & \text{if } t \in (t_0,\ t_0+1/n) \\ 0 & \text{if } t \notin (t_0-1/n,\ t_0+1/n). \end{cases}$$

It turns out that

$$f_n(t_0) = n \qquad \text{and} \qquad \int_{-\infty}^{+\infty} |f_n(t)|^2 \, dt < 3$$

and hence the evaluation functionals are unbounded (i.e. they are not continuous) in $L_0^2(-\infty, +\infty)$.

Consider the functions in $L^2(-\infty, +\infty)$ with a 'band-limited spectrum', i.e. the functions $f \in L^2(-\infty, +\infty)$ in the form

$$f(t) = \frac{1}{2A} \int_{-A}^{+A} e^{i\omega t} F(\omega)\, d\omega \tag{$*$}$$

where $F \in L^2(-A, +A)$. By the Cauchy–Schwarz inequality,

$$|f(t)| \leq \frac{1}{2A} \Big(\int_{-A}^{+A} d\omega \Big)^{1/2} \Big(\int_{-A}^{+A} |F(\omega)|^2\, d\omega \Big)^{1/2}$$

and hence, by Example 4 of § 2.11.1,

$$|f(t)| \leq \|f\|_2.$$

This means that in the subspace of $L^2(-\infty, +\infty)$ consisting of functions, of the form ($*$), the evaluation functionals are bounded by 1 and hence are continuous.

A Hilbert space $\mathscr{H}$ of functions with continuous evaluation functionals will be called a *reproducing kernel Hilbert space.*

3.1.2. In a Hilbert space (and also in any normed space) of functions the pointwise convergence can be expressed by the continuity of the evaluation functionals. More precisely, the following are equivalent for a normed space B of functions.

(i) If f_n, $f \in B$ and $\|f_n - f\| \to 0$ then

$$f_n(t) \to f(t) \qquad \text{for every } t \in \mathscr{D}.$$

(ii) The evaluation functionals are continuous.

(iii) For every $t \in \mathscr{D}$ there exists $K_t > 0$ such that

$$|f(t)| \leq K_t \|f\| \qquad f \in B.$$

Considering 1.4.1.5, the proof of (i)⇔(ii)⇔(iii) is easy.

3.1.2.1 Definition. A Hilbert space $\mathscr{H}$ is called a reproducing kernel Hilbert space (RKHS) if the following conditions are satisfied:

(*a*) the elements of $\mathscr{H}$ are (complex or real-valued) functions defined on any set $\mathscr{D}$;

(*b*) for every $t \in \mathscr{D}$ there exists $K_t > 0$ such that

$$|f(t)| \leq K_t \|f\| \qquad f \in \mathscr{H}.$$

The elements of an RKHS are also denoted by $f(\,.\,)$, $g(\,.\,)$, ..., indicating that the elements are functions, whereas $f(t)$ is the value of $f(\,.\,)$ at $t \in \mathscr{D}$.

In an RKHS $\mathscr{H}$, for every $t \in \mathscr{D}$ there is a function $R(\,.\,, t) \in \mathscr{H}$ such that

$$f(t) = \big(f(\,.\,)|R(\,.\,, t)\big) \qquad f \in \mathscr{H} \tag{$*$}$$

by the Riesz Representation Theorem (2.8.1.1) and hence the evaluation functionals are determined by the function $R = R(s, t)$ on $\mathscr{D} \times \mathscr{D}$, called the *kernel of the* RKHS $\mathscr{H}$; $(*)$ is sometimes called the *reproducing property of R.*

3.1.2.2 Definition. The (complex or real-valued) function $R = R(s, t)$ on $\mathscr{D} \times \mathscr{D}$ is called *symmetric* if

$$R(t, s) = \overline{R(s, t)}$$

and *positive definite* if for any finite set $\{s_i \in \mathscr{D};\ i = 1, 2, \ldots\}$ and complex numbers λ_i $(i = 1, 2, \ldots, n)$,

$$\sum_{i,j=1}^{n} \lambda_i \bar{\lambda}_j R(s_i, s_j) \geqslant 0.$$

3.1.2.3 Theorem. The kernel $R = R(s, t)$ of an RKHS is a symmetric and positive definite function.

Proof. If $f(\,.\,) = R(\,.\,, s)$ in the formula $(*)$, then we obtain

$$R(t, s) = \big(R(\,.\,, s)|R(\,.\,, t)\big)$$

and hence

$$\overline{R(s, t)} = \overline{\big(R(\,.\,, t)|R(\,.\,, s)\big)} = \big(R(\,.\,, s)|R(\,.\,, t)\big) = R(t, s).$$

Considering

$$\sum_{k=1}^{n} \lambda_k R(\,.\,, s_k)$$

we have

$$0 \leqslant \Big\|\sum_{k=1}^{n} \lambda_k R(\,.\,, s_k)\Big\|^2 = \Big(\sum_{i=1}^{n} \lambda_i R(\,.\,, s_i)\Big|\sum_{j=1}^{n} \lambda_j R(\,.\,, s_j)\Big)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \bar{\lambda}_j R(s_i, s_j)$$

and hence the kernel R is positive definite. We recall that *a normed space is generated by a subset $\mathscr{M}$ if the linear combinations of elements of $\mathscr{M}$ are dense in the normed space.*

3.1.2.4 Theorem. The RKHS $\mathscr{H}$ with kernel R is generated by $\{R(\,.\,, t);\ t \in \mathscr{D}\}$.

Proof. If for every $t \in \mathscr{D}$, $(f(\,.\,)|R(\,.\,,t))=0$, then it follows from the reproducing property $(*)$ that $f=\theta$. Hence the theorem follows from 2.14.42.

We have now arrived at the first serious problem of the RKHS theory. Is every symmetric positive definite $R=R(s,t)$ the kernel of an RKHS?

3.1.2.5 Theorem. For every symmetric positive definite $R=R(s,t)$ there is a unique RKHS with kernel R.

Proof. First we consider the linear space generated by $\{R(\,.\,,t);\ t \in \mathscr{D}\}$. If we define

$$(R(\,.\,,s)|R(\,.\,,t)) := R(t,s)$$

and

$$\Big(\sum_{j=1}^{m} \mu_j R(\,.\,,t_j)\Big|\sum_{i=1}^{n} \lambda_i R(\,.\,,t_i)\Big) := \sum_{j=1}^{m}\sum_{i=1}^{n} \mu_j \bar{\lambda}_i R(t_i,t_j)$$

then the axioms of the scalar product and the reproducing property $(*)$ are satisfied. The only non-trivial part of this assertion is that

$$\Big(\sum_{i=1}^{n} \lambda_i R(\,.\,,t_i)\Big|\sum_{j=1}^{n} \lambda_j R(\,.\,,t_j)\Big) = 0$$

implies

$$\sum_{k=1}^{n} \lambda_k R(t,t_k) = \theta$$

for every $t \in \mathscr{D}$. Indeed,

$$\sum_{k=1}^{n} \lambda_k R(t,t_k) = \Big(\sum_{k=1}^{n} \lambda_k R(\,.\,,t_k)|R(\,.\,,t)\Big)$$

and, applying the Cauchy–Schwarz inequality (by 2.9.2),

$$\Big|\Big(\sum_{k=1}^{n} \lambda_k R(\,.\,,t_k)|R(\,.\,,t)\Big)\Big|^2 \leq \Big(\sum_{i=1}^{n} \lambda_i R(\,.\,,t_i)\Big|\sum_{j=1}^{n} \lambda_j R(\,.\,,t_j)\Big)(R(\,.\,,t)|R(\,.\,,t))$$

the non-trivial part of the first assertion is also obtained.

Secondly we consider the completion $\mathscr{H}^c$ of this pre-Hilbert space $\mathscr{H}$. What we have to prove is that $\mathscr{H}^c$ is an RKHS.

Let $\{x_n\}$ be a Cauchy sequence in $\mathscr{H}$; then $\{x_n(t)\}$, for every $t \in \mathscr{D}$, is also a Cauchy sequence since

$$x_n(t) = (x_n|R(\,.\,,t)).$$

If $x \in \mathscr{H}^c$, $\lim x_n = x$ and $\lim x_n(t) = x(t)$, then

$$(x|R(\,.\,,t)) = \lim\,(x_n|R(\,.\,,t)) = x(t)$$

and the correspondence $x \to x(t)$ is 1–1. We conclude that the linear space of functions $\{x(t);\ t\in\mathscr{D}\}$ thus obtained is an RKHS with the scalar product

$$(x(\,.\,)|y(\,.\,)) = \lim_n\,(x_n(\,.\,)|y_n(\,.\,))$$

and kernel $R=R(s, t)$.

To summarise: for every RKHS the evaluation functionals are determined by the kernel R, which is a symmetric, positive definite function, and for every symmetric and positive definite R there is a unique RKHS with evaluation functionals represented by $\{R(\,.\,, t);\ t\in\mathscr{D}\}$. *The* RKHS *belonging to the kernel R is denoted by $\mathscr{H}(R)$.*

3.1.3. Some consequences of the foregoing theorems are as follows.

3.1.3.1 Proposition.

$$\|R(., t)\|^2 = (R(., t)|R(., t)) = R(t, t)$$

$$|x(t)| \leqslant \|R(., t)\|\,\|x\| = (R(t, t))^{1/2}\|x\|.$$

In particular, for any pair $x_n(\,.\,),\ x_m(\,.\,)\in\mathscr{H}(R)$,

$$|x_n(t)-x_m(t)| \leqslant (R(t, t))^{1/2}\|x_n-x_m\|.$$

Based on Proposition 3.1.3.1, we have the following connections between the kernel R and the elements of $\mathscr{H}(R)$.

3.1.3.2 Corollary. If $R(t, t)$ is a bounded function, then every $x(\,.\,)\in\mathscr{H}(R)$ is also bounded.

3.1.3.3 Corollary. If $R(t, t)$ is bounded on a subset $\mathscr{D}'\subseteq\mathscr{D}$ and $\{x_n\}$ is convergent in $\mathscr{H}(R)$, then $\{x_n(\,.\,)\}$ is uniformly convergent on $\mathscr{D}'$, i.e. *in this case the uniform convergence is implied by the norm convergence.*

3.1.3.4 Theorem. If $\mathscr{D}=\mathbf{R}^n$ and $R(t, s)$ is continuous then every $x(\,.\,)\in\mathscr{H}(R)$ is a continuous function.

Proof. If $x(\,.\,)$ is a linear combination of functions $R(\,.\,, t);\ t\in\mathscr{D}$ then there is nothing to prove. Otherwise x is the limit of such functions. $R(t, t)$ is bounded on every bounded subset of $\mathbf{R}^n$; hence the uniform convergence on every bounded subset of $\mathbf{R}^n$ is implied by the convergence in $\mathscr{H}(R)$ by Corollary 3.1.3.3. Hence $x=x(\cdot)$ is the uniform limit of continuous functions on every bounded subset and so it is also continuous.

It is also important that, in most cases, an RKHS is a *separable* Hilbert space.

3.1.3.5 Theorem. If $\mathscr{D}\subseteq\mathbf{R}^n$ and R is a continuous function, then $\mathscr{H}(R)$ is separable.

Proof. In this case $\mathscr{H}(R)$ is generated by the countable subset $\{R(\,.\,, t); t \in \mathscr{D}\}$ with rational coordinates.

Example 1. The elements of the Hilbert space l^2 are sequences, i.e. functions defined on the natural numbers, and

$$|x(k)|^2 := |x_k|^2 \leqslant \sum_{k=1}^{\infty} |x_k|^2 = \|x\|_2^2 \qquad k = 1, 2, \ldots .$$

Hence it is obvious that l^2 is an RKHS. The kernel is

$$R(m, n) = \begin{cases} 1 & \text{if } m = n \\ 0 & \text{if } m \neq n. \end{cases}$$

We shall see that l^2 is uninteresting as an RKHS.

Remark. $L^2[a, b]$ is not an RKHS, as we have seen in 3.1.1. *The* RKHS *property is not invariant for Hilbert space isomorphisms.*

Example 2. Let $\mathscr{T}_n$ be the linear space of trigonometric polynomials of degree n considered as a $(2n+1)$-dimensional (and hence closed) subspace of $L^2[-\pi, +\pi]$. In this case,

$$\Big\|\sum_{k=-n}^{+n} \gamma_k e^{ikt}\Big\|^2 = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \sum_{k=-n}^{+n} \gamma_k e^{ikt} \sum_{m=-n}^{+n} \gamma_m e^{-imt}\, dt = \sum_{k=-n}^{+n} |\gamma_k|^2$$

and, applying the Cauchy–Schwarz inequality

$$\Big|\sum_{k=-n}^{+n} \gamma_k e^{ikt}\Big|^2 \leqslant \sum_{k=-n}^{+n} |\gamma_k|^2 \sum_{k=-n}^{+n} |e^{ikt}|^2 = (2n+1) \sum_{k=-n}^{+n} |\gamma_k|^2$$

we obtain

$$\Big|\sum_{k=-n}^{+n} \gamma_k e^{ikt}\Big| \leqslant (2n+1)^{1/2} \Big\|\sum_{k=-n}^{+n} \gamma_k e^{ikt}\Big\|$$

i.e. each of the evaluation functionals is bounded by $(2n+1)^{1/2}$.

Remark. Although $L^2[-\pi, +\pi]$ is not an RKHS, the closed linear subspace $\mathscr{T}_n$ of $L^2[-\pi, +\pi]$ is.

Example 3. The linear space of functions represented in the form of a finite Fourier transform

$$f(t) = \frac{1}{2A} \int_{-A}^{+A} e^{i\omega t} F(\omega)\, d\omega$$

is also considered as a closed subspace L_A of an L^2-space. It is, in fact, a closed subspace of $L^2(-\infty, +\infty)$. If A is a fixed number, $F \in L^2[-A, +A]$, then it is known from Fourier transform theory that

$$\int_{-\infty}^{+\infty} |f(t)|^2 \,\mathrm{d}t = \int_{-A}^{+A} |F(\omega)|^2 \,\mathrm{d}\omega. \qquad (*)$$

It is easy to prove that the elements of L_A are continuous functions (this will also be done by an RKHS method in § 3.3), and

$$|f(t)| = \frac{1}{2A}\left|\int_{-A}^{+A} \mathrm{e}^{\mathrm{i}\omega t} F(\omega)\,\mathrm{d}\omega\right|$$

$$\leqslant \left(\frac{1}{2A}\int_{-A}^{+A} |\mathrm{e}^{\mathrm{i}\omega t}|^2 \,\mathrm{d}\omega\right)^{1/2}\left(\frac{1}{2A}\int_{-A}^{+A} |F(\omega)|^2 \,\mathrm{d}\omega\right)^{1/2} = \|F\|.$$

Hence, bearing in mind expression $(*)$, the evaluation functionals in L_A are bounded.

Example 4. Let us consider the linear space $\{f\colon f' \in L^2[0, 1]\}$, i.e. the linear space $\mathscr{H}_D$ of (completely continuous) functions on $[0, 1]$ with derivatives in $L^2[0, 1]$. It is easy to show that

$$(f|g) := f(0)\overline{g(0)} + \int_0^1 f'(t)\overline{g'(t)}\,\mathrm{d}t \qquad (*)$$

is a scalar product for these functions and clearly

$$f(s) = f(0) + \int_0^s f'(t)\,\mathrm{d}t.$$

It follows that if

$$g_s\colon \frac{\mathrm{d}}{\mathrm{d}t} g_s(t) = \begin{cases} 1 & \text{for } t \leqslant s \\ 0 & \text{elsewhere} \end{cases} \qquad g_s 0 = 1$$

then for $s \in [0, 1]$,

$$(f|g_s) = f(s).$$

It is known that $\mathscr{H}_D$ with the scalar product $(*)$ is complete (as will be shown by an RKHS method in § 3.3), and hence $\mathscr{H}_D$ is an RKHS. It is easy to verify that the kernel is given by

$$R(s, t) = g_s(t) = 1 + \min(s, t)$$

where

$$\min(s, t) = \begin{cases} t & \text{if } s > t \\ s & \text{if } s \leqslant t. \end{cases}$$

Remark. A Hilbert space $\mathscr{H}_D$ can also be constructed in a similar way for more general linear differential operators D. (See e.g. Example 5 in § 2.1.2.) All of these Hilbert spaces, called Sobolev spaces, are RKHS.

Example 5. The important property of H^2-spaces shown in Example 4 of § 2.1.2 implies that the H^2-space is an RKHS. It follows from 2.6.3 that the kernel is

$$R(s, t) = \frac{1}{1-t\bar{s}}$$

where t, s are complex variables; this is called the *Szegő kernel.*

Example 6. If $\mathscr{D}$ is the open unit disc in the complex plane, i.e. $\mathscr{D}=\{z\colon |z|<1\}$, and $A(\mathscr{D})$ is the linear space of functions analytic in $\mathscr{D}$ and satisfying the condition

$$\int_0^{2\pi}\int_0^1 |f(re^{it})|^2 r\,dr\,dt < \infty$$

where the integral is to be understood as

$$\lim_{R\to 1-0} \int_0^{2\pi}\int_0^R |f(re^{it})|^2 r\,dr\,dt$$

then $A(\mathscr{D})$ is a Hilbert space with the scalar product

$$(f|g) := \int_0^{2\pi}\int_0^1 f(re^{it})\overline{g(re^{it})}\,r\,dr\,dt \qquad f, g \in A(\mathscr{D}).$$

If

$$f = \sum_{k=0}^{\infty} a_k z^k \qquad \text{and} \qquad g = \sum_{k=0}^{\infty} b_k z^k$$

then by a straightforward calculation,

$$(f|g) = \pi \sum_{n=0}^{\infty} \frac{1}{n+1} a_n \bar{b}_n. \tag{$*$}$$

Moreover, $f \in A(\mathscr{D})$ if and only if $a_n = (n+1)^{1/2} c_n$ with $c_n \in l^2$.
(We shall not go into the detailed proof of these assertions.)

It will be shown that $A(\mathscr{D})$ is an RKHS by constructing the kernel. If the kernel $R(z, \xi)$ exists, then the function $R(\,.\,, z_0)$ depending on the parameter z_0 has the form

$$R(z, z_0) = \sum_{k=0}^{\infty} b_k(z_0) z^k$$

and by (*),

$$(f(.)|R(., z_0)) = \pi \sum_{n=0}^{\infty} \frac{1}{n+1} a_n \overline{b_n(z_0)}. \qquad (**)$$

On the other hand,

$$(f(.)|R(., z_0)) = f(z_0) = \sum_{n=0}^{\infty} a_n z_0^n. \qquad (***)$$

Comparing (**) and (***), we have

$$\pi \frac{1}{n+1} \overline{b_n(z_0)} = z_0^n \qquad n = 0, 1, 2, \ldots.$$

We conclude that

$$b_n(z_0) = \frac{1}{\pi}(n+1)\bar{z}_0^n \qquad n = 0, 1, 2, \ldots$$

and so

$$R(z, z_0) = \frac{1}{\pi} \sum_{k=0}^{\infty} (k+1)\bar{z}_0^n z^n$$

which can easily be verified.

It follows from the identity

$$\sum_{k=0}^{\infty} (k+1) q^k = \frac{1}{(1-q)^2} \qquad \text{for } |q| < 1$$

that $R(\,.\,, z_0) \in \mathrm{A}(\mathscr{D})$ and

$$R(z, z_0) = \frac{1}{\pi} \frac{1}{(1-\bar{z}_0 z)^2}$$

which is called the *Bergman kernel.*

3.2 Kernels in the form of an infinite series

3.2.1. The construction of a kernel for a Hilbert space is a direct demonstration that the Hilbert space is an RKHS; moreover, the kernel gives the major information about the Hilbert space. In the previous section *ad hoc* methods were used for the construction of kernels or we did not find a kernel at all, as was the case in Examples 2 and 3.

In this section a general method will be given for the construction of kernels in the form of an infinite series. This is the oldest method; a modern approach for constructing the RKHS, the kernel and the scalar product simultaneously, will be the subject of the next section.

3.2.1.1 Theorem. If $\{e_k(\,.\,);\ k=1, 2, \ldots\}$ is a complete orthonormal sequence in the separable space $\mathscr{H}(R)$, then the kernel has the form

$$R(s, t) = \sum_{k=1}^{\infty} e_k(s)\overline{e_k(t)}. \qquad (*)$$

Proof. The series expansion of $R(\,.\,, t)$ is

$$R(.\,, t) = \sum_{k=1}^{\infty} \big(R(.\,, t)|e_k(.)\big)\, e_k(.)$$

and $(*)$ follows from

$$\big(R(.\,, t)|e_k(.)\big) = \overline{\big(e_k(.)|R(.\,, t)\big)} = \overline{e_k(t)}.$$

Remark. Setting $s=t$ in $(*)$ we obtain

$$\sum_{k=1}^{\infty} |e_k(t)|^2 < \infty \qquad t \in \mathscr{D}$$

for any orthonormal sequence $\{e_k;\ k=1, 2, \ldots\}$ in an RKHS.

Example 1. It is easy to verify that

$$e_k(z) = \frac{1}{\pi^{1/2}}(k+1)^{1/2} z^k$$

is a complete orthonormal sequence in $\mathrm{A}(\mathscr{D})$. Hence the kernel is

$$R(\xi, z) = \frac{1}{\pi} \sum_{k=0}^{\infty} (k+1)(\bar{z}\xi)^k$$

in accordance with Example 6 in § 3.1.3.

Example 2. By Example 4 in § 2.2.1 the sequence $\{z^k;\ k=1, 2, \ldots\}$ is a complete orthonormal system in H_0^2. Hence the kernel is

$$R(\xi, z) = \sum_{k=0}^{\infty} \xi^k \bar{z}^k = \frac{1}{1-\xi\bar{z}}$$

in accordance with Example 5 in § 3.1.3.

Example 3. In the linear space $\mathscr{T}_n$ of trigonometric polynomials of nth degree, considered as a closed subspace of $\mathrm{L}^2[-\pi, +\pi]$, the sequence $\{\mathrm{e}^{ikt};\ k=0, \pm 1, \ldots, \pm n\}$ is complete and orthonormal. Hence

$$R(s, t) = \sum_{k=-n}^{+n} \mathrm{e}^{iks}\mathrm{e}^{-ikt} = \sum_{k=-n}^{+n} \mathrm{e}^{ik(s-t)}$$

is the kernel of $\mathscr{T}_n$. The sum of the right-sided geometrical series

$$\sum_{k=-n}^{+n} e^{ik(s-t)} = \frac{\sin\left(n+\frac{1}{2}\right)(s-t)}{\sin\frac{1}{2}(s-t)}$$

is called the *Fejér kernel.*

3.2.2. If $\{a_k(t);\ k=1, 2, \ldots, t\in\mathscr{D}\}$ is a sequence of functions such that

$$\sum_{k=1}^{\infty} |a_k(t)|^2 < \infty$$

for every $t\in\mathscr{D}$, then an RKHS can be constructed from this sequence.

3.2.2.1 Theorem. If

$$\sum_{k=1}^{\infty} |a_k(t)|^2 < \infty$$

then the linear space $\mathscr{H}$ generated by the set

$$\{\sum_{k=1}^{\infty} c_k a_k(.);\ \sum_{k=1}^{\infty} |c_k|^2 < \infty\}$$

is an RKHS with the kernel

$$R(s, t) = \sum_{k=1}^{\infty} a_k(s)\overline{a_k(t)}$$

and with the scalar product

$$(f|g) := \sum_{k=1}^{\infty} a_k \bar{b}_k$$

where

$$f(.) = \sum_{k=1}^{\infty} c_k a_k(.) \qquad \text{and} \qquad g(.) = \sum_{k=1}^{\infty} b_k a_k(.).$$

Moreover, $\{a_k(\,.\,);\ k=1, 2, \ldots\}$ is a complete orthonormal sequence in $\mathscr{H}(R)$.
Proof. If

$$\sum_{k=1}^{\infty} |c_k|^2 < \infty$$

then the series

$$\sum_{k=1}^{\infty} c_k a_k(t)$$

is convergent for every $t \in \mathscr{D}$ since

$$\Big|\sum_{k=m}^{n} c_k a_k(t)\Big|^2 \leqslant \sum_{k=m}^{n} |c_k|^2 \sum_{k=m}^{n} |a_k(t)|^2$$

via the Cauchy–Schwarz inequality.

Let us now consider the linear space $\mathscr{H}$ of the functions in the form of (pointwise convergent) infinite series

$$\sum_{k=1}^{\infty} c_k a_k(.) \qquad \sum_{k=1} |c_k|^2 < \infty.$$

If the scalar product of

$$f(.) = \sum_{k=1}^{\infty} c_k a_k(.) \qquad \text{and} \qquad g(.) = \sum_{k=1}^{\infty} b_k a_k(.)$$

is defined as

$$(f|g) := \sum_{k=1}^{\infty} c_k \bar{b}_k$$

then a Hilbert space is obtained since the mapping

$$\{c_k\} \to \sum_{k=1}^{\infty} c_k a_k(.)$$

is a Hilbert space isomorphism from l^2 onto $\mathscr{H}$ in this case. Moreover, $R(\,.\,, t) \in \mathscr{H}$ and

$$(f(.)|R(., t)) = \sum_{k=1}^{\infty} c_k a_k(t) = f(t)$$

for every $f \in \mathscr{H}$.

3.3 A modern approach to the RKHS model

3.3.1. Let $\mathscr{H}$ be a Hilbert space, $\mathscr{D}$ a set and $h=h(t)$ a mapping from $\mathscr{D}$ into $\mathscr{H}$. A method will be shown for the construction of an RKHS isomorphic with the closed subspace $\mathscr{H}_1$ of $\mathscr{H}$ generated by $\{h(t);\ t \in \mathscr{D}\}$.

3.3.1.1 Definition. For $x \in \mathscr{H}_1$, the function

$$\hat{x}(t) := (x|h(t))$$

is called the *Loéve transform* of x.

3.3.1.2 Theorem. The Loéve transforms form an RKHS with kernel

$$R(s, t) = (h(t)|h(s)) \qquad s, t \in \mathscr{D}.$$

Moreover, $\mathscr{H}(R)$ is isomorphic with $\mathscr{H}_1$.
Proof. $R(s, t)$ is symmetric since

$$R(s, t) := (h(t)|h(s)) = \overline{(h(s)|h(t))} = \overline{R(t, s)}.$$

$R(s, t)$ is positive definite since

$$\sum_{i,j=1}^{n} \lambda_i \bar{\lambda}_j R(t_i, t_j) := \sum_{i,j=1}^{n} \lambda_i \bar{\lambda}_j (h(t_j)|h(t_i))$$

$$= \Big(\sum_{i=1}^{n} \lambda_i h(t_i) \Big| \sum_{j=1}^{n} \lambda_j h(t_j)\Big) \geqslant 0$$

for any finite $\{t_i \in \mathscr{D};\ i=1, 2, \ldots, n\}$ and complex numbers $\{\lambda_i;\ i=1, 2, \ldots\}$.
If $x=h(t_0)$, then $\hat{x}(t)=(h(t_0)|h(t))=R(t, t_0)$ and for $x=h(s_1)$, $y=h(s_2)$,

$$(x|y) = (h(s_1)|h(s_2)) = R(s_2, s_1) := (R(., s_1)|R(., s_2))$$

by the reproducing property 3.1.2 $(*)$ of the kernel.

$\mathscr{H}_1$ is generated by $\{h(t);\ t\in\mathscr{D}\}$ and $\mathscr{H}(R)$ by $\{R(\,.\,, t);\ t\in\mathscr{D}\}$. Thus by the above considerations we have proved that $\mathscr{H}_1$ and $\mathscr{H}(R)$ are isomorphic Hilbert spaces.

Example 1. For any finite interval $[-A, +A]$ let us consider the subspace of $L^2[-A, +A]$ generated by

$$h(t) = e^{-i\omega t} \qquad -\infty < t < +\infty.$$

In this case,

$$(h(t)|h(s)) = \frac{1}{2A} \int_{-A}^{+A} e^{i\omega t} e^{-i\omega s}\, d\omega = \frac{\sin A(t-s)}{A(t-s)}$$

and the Loéve transform of $F \in L^2[-A, +A]$ is the finite Fourier transform

$$f(t) = \frac{1}{2A} \int_{-A}^{+A} e^{i\omega t} F(\omega)\, d\omega.$$

Hence we have shown that the finite Fourier transforms (with *fixed* band limit A) form an RKHS, called the *Hilbert space* L_A *of band-limited signals*, with the scalar product

$$(f|g) = \int_{-A}^{+A} F(\omega)\overline{G(\omega)}\, d\omega$$

where

$$f(t)=\frac{1}{2A}\int_{-A}^{+A} e^{i\omega t}F(\omega)\,d\omega \qquad g(t)=\frac{1}{2A}\int_{-A}^{+A} e^{i\omega t}G(\omega)\,d\omega$$

and with the kernel

$$R(s,t)=\frac{\sin A(t-s)}{A(t-s)}.$$

Every function represented by a finite Fourier transform is continuous since the kernel is a continuous function. Thus by constructing the kernel in L_A the assertions of Example 3 in § 3.1.3 are also completed.

The kernel $R(s, t)$ has the following important property: $\{R(\,.\,, t_k);\ t_k=k\pi/A\}$ form a *complete orthonormal sequence*. Indeed,

$$R(s,t_k)=\frac{\sin(k\pi-As)}{k\pi-As}=\frac{1}{2A}\int_{-A}^{+A} e^{i(k\pi/A)\omega}e^{-i\omega s}\,d\omega;$$

on the other hand $\{e^{i(k\pi/A)\omega};\ k=0, \pm 1, \pm 2, \ldots\}$ form a complete orthonormal sequence in $L^2[-A, +A]$ and the Loéve transform is a Hilbert space isomorphism. It follows also that L_A is isomorphic with $L^2[-A, +A]$.

Example 2. The finite and discrete analogue of the above example is the RKHS $\mathscr{T}_n$ of trigonometric polynomials introduced in Examples 3.1.3(2) and 3.2.1(3). In this case ω takes the discrete values $k=0, \pm 1, \pm 2, \ldots, \pm n$ only. Hence

$$h(t)=\{e^{-ikt};\ k=0, \pm 1, \pm 2, \ldots, \pm n\}$$

are linearly independent elements of the $(2n+1)$-dimensional Euclidean space, l_n^2 (instead of $L^2[-A, +A]$). In this case the kernel is

$$(h(t)|h(s))=\sum_{k=-n}^{+n} e^{ikt}e^{-iks}=\sum_{k=-n}^{+n} e^{ik(t-s)}=\frac{\sin\left(n+\frac{1}{2}\right)(t-s)}{\sin\frac{1}{2}(t-s)}$$

corresponding to Example 3 in § 3.2.1.

Example 3. The space $L_0^2[0, 1]$ is also generated by the functions

$$h(t)=(t-\tau)_+ \qquad 0<t<1$$

shown in figure 3.4(*a*) on p. 158. In fact, every piecewise linear function $l(t)$ in [0, 1] with $l(1)=0$ is a linear combination of these functions and these piecewise linear functions form a dense subspace of $L_0^2[0, 1]$. In this case,

$$(h(t)|h(s)) = \int_0^1 (t-\tau)_+(s-\tau)_+ \, d\tau \qquad (*)$$

and the Loéve transform of $F \in L_0^2[0, 1]$ is

$$f(t) = \int_0^1 (t-\tau)_+ F(\tau)\, d\tau = \int_0^t (t-\tau)F(\tau)\, d\tau. \qquad (**)$$

The kernel of the RKHS thereby obtained,

$$R(t, s) = \int_0^1 (t-\tau)_+(s-\tau)_+ \, d\tau$$

is piecewise polynomial for fixed s. More particularly,

$$R(t, s) = \frac{1}{6}(s-t)_+^3 - \frac{1}{2}s^2 t - \frac{1}{6}s^3$$

and hence $R(\,.\,, s)$ is a polynomial of third degree on $[0, s)$ and a first-degree polynomial on $(s, 1]$ so that the second derivative $R''(\cdot, s)$ is a continuous function. *Such a kernel will be denoted by* $S(\cdot, s)$ *and the corresponding* RKHS *by* $\mathscr{H}(S)$.

Let (a, b) be a finite interval and

$$a = s_0 < s_1 < \ldots < s_{n-1} < s_n = b.$$

If $s = s(t)$ is a function in (a, b) with a continuous second derivative and there exist polynomials $\{p_k;\ k=1, 2, \ldots, n\}$ of at most third degree such that

$$s(t) = p_k(t) \qquad \text{for} \quad t \in (s_{k-1}, s_k)$$

then s is called a *cubic spline* with nodes $\{s_k;\ k=1, 2, \ldots, n-1\}$. The kernel $S(t, s)$ of $\mathscr{H}(S)$ is a cubic spline for every fixed s with only the node $t=s$.

Considering $(**)$, the RKHS with kernel $(*)$ consists of the functions

$$\{f\colon f'' = F \in L^2[0, 1];\ f(0) = f'(0) = 0\}$$

with scalar product

$$(f|g) := \int_0^1 f''(t)\overline{g''(t)}\, dt \qquad f, g \in \mathscr{H}(s)$$

since the Loéve transform is an isomorphic operator.

Example 4. Let $\{X(t);\ 0 \le t < \infty\}$ be a stochastic process with random variables in $L^2(\Omega, A, P)$ defined in Example 6 of § 2.1.2. Then the Loéve transform of a random variable $\xi \in L^2(\Omega, A, P)$ is $M(\xi X(t))$, a deterministic time function, and the kernel of the RKHS of Loéve transforms,

$$M(X(s)X(t))$$

is the covariance function of the process. The RKHS with kernel

$$R(s, t) = M(X(t)X(s))$$

has an important role in the investigations of Gaussian stochastic processes, as will be seen in § 3.8.

To summarise, in the method demonstrated in the above examples, the essential point is to choose the appropriate $\{h(t);\ t \in \mathscr{D}\}$ for the Loéve transform. The kernel, the elements and the values of scalar products are then determined in the RKHS thereby obtained. Thus,

$$R(s, t) = (h(t)|h(s))$$

$$x(t) = (x|h(t))$$

and

$$(x(\,.\,)|y(\,.\,)) = (x|y)$$

since the Loéve transform is an isomorphic operator.

In the light of the general approach to forming RKHS introduced in this section it is clear that we can find many RKHS that are isomorphic to a given Hilbert space. Among them are RKHS with an 'interesting' kernel, some of which were introduced in the above examples. Referring to the beginning of § 3.2, the major information about the RKHS is given by its kernel. One can say that *the theory of reproducing kernel Hilbert spaces is the theory of Hilbert spaces with 'interesting' kernels.*

In spite of this, in this chapter we shall also deal with the applications of certain Hilbert spaces with continuous evaluation functionals but with useless kernels. For example, in § 3.9 it is important that the Hilbert space convergence implies the pointwise convergence in Sobolev spaces; however, the kernel is uninteresting in this case.

3.4 The projection principle in RKHS

The most important theorem in Hilbert space geometry is the projection principle. There is a special constructive method for giving the projection *onto an* RKHS *subspace* $\mathscr{M}$ of a Hilbert space $\mathscr{H}$.

3.4.1. It is obvious that every closed subspace $\mathscr{M}$ of an RKHS is also an RKHS.

We begin by showing the connection between the kernel R of the RKHS $\mathscr{H}(R)$ and the kernel $R_{\mathscr{M}}$ of $\mathscr{M}$.

3.4.1.1 Theorem. If $\mathscr{M}$ is a closed linear subspace of an RKHS $\mathscr{H}(R)$ and

$$R(.,t) = R_1(.,t)+R_2(.,t) \qquad R_1 \in \mathscr{M}, \quad R_2 \in \mathscr{M}^{\perp}$$

is the decomposition of the kernel R into the corresponding direct sum, then R_1 is the kernel of $\mathscr{M}$ and R_2 is the kernel of $\mathscr{M}^{\perp}$.
Proof. In this case,

$$y(t) = \big(R(.,t)|y(.)\big) = \big(R_1(.,t)|y(.)\big)+\big(R_2(.,t)|y(.)\big)$$

$$\big(R_2(.,t)|y(.)\big) = 0 \qquad \text{if} \quad y \in \mathscr{M}$$

$$\big(R_1(.,t)|y(.)\big) = 0 \qquad \text{if} \quad y \in \mathscr{M}^{\perp}.$$

Hence

$$y(t) = \big(R_1(.,t)|y(.)\big) \qquad \text{for} \quad y \in \mathscr{M}$$

and

$$y(t) = \big(R_2(.,t)|y(.)\big) \qquad \text{for} \quad y \in \mathscr{M}^{\perp}.$$

The projection operator $P_{\mathscr{M}}$ is represented by the kernel of $\mathscr{M}$ as follows.

3.4.1.2 Theorem. If $\mathscr{H}$ is a Hilbert space and $\mathscr{M}$ is an RKHS subspace of $\mathscr{H}$, then for every $x \in \mathscr{H}$,

$$P_{\mathscr{M}}x = \big(R_{\mathscr{M}}(.,t)|x\big)$$

where $P_{\mathscr{M}}$ is the projection operator and $R_{\mathscr{M}}$ is the kernel of $\mathscr{M}$.
Proof. If

$$x = y+z \qquad y \in \mathscr{M}, \quad z \in \mathscr{M}^{\perp}$$

is the decomposition of $x \in \mathscr{H}$ into the corresponding direct sum, then

$$\big(R_{\mathscr{M}}(.,t)|x\big) = \big(R_{\mathscr{M}}(.,t)|y\big)+\big(R_{\mathscr{M}}(.,t)|z\big) = y(t)$$

since $\big(R_{\mathscr{M}}(.,t)|z\big)=0$ by Theorem 3.4.1.1.

Example 1. The projection of $L^2[-\pi, +\pi]$ onto the subspace $\mathscr{T}_n$ has the form

$$s_n(t) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f(t)\frac{\sin\left(n+\frac{1}{2}\right)(s-t)}{\sin\frac{1}{2}(s-t)}\,\mathrm{d}t \qquad f \in L^2[-\pi, +\pi]$$

which is called the representation of the nth partial sum of the Fourier series by the Fejér kernel.

Example 2. The projection of $L^2(-\infty, +\infty)$ onto L_A has the form

$$P_A f(s) = \int_{-\infty}^{+\infty} f(t) \frac{\sin A(s-t)}{A(s-t)} \, dt$$

and this is also a frequently used formula in Fourier transform theory. (The usual treatment is as follows. If the Fourier transform of f is $F=F(\omega)$ then, by the convolution theorem, the Fourier transform of $P_A f$ is

$$\mathscr{F} P_A f = \begin{cases} F(\omega) & \text{if } |\omega| < A \\ 0 & \text{elsewhere} \end{cases}$$

since

$$\mathscr{F} \frac{\sin At}{At} = \begin{cases} 1 & \text{if } |\omega| < A \\ 0 & \text{elsewhere.} \end{cases}$$

One can then verify that the truncated $F=F(\omega)$ is the projection of F.)

3.4.2. A more general example is the projection of an arbitrary $f \in \mathscr{H}(R)$ onto the n-dimensional subspace generated by

$$\{R(.,s_k);\ k=1,2,\dots,n\} \qquad s_k \in \mathscr{D}.$$

In this case the usual method for constructing a projection is as follows. The projection has the form

$$c_1 R(.,s_1) + c_2 R(.,s_2) + \dots + c_n R(.,s_n)$$

and

$$\Big(f(.) - \sum_{k=1}^{n} c_k \big(R(.,s_k) | R(.,s_j)\big)\Big) = 0 \qquad j=1,2,\dots,n$$

by the projection principle in the case of a finite-dimensional subspace. Hence $\{c_k, k=1,2,\dots,n\}$ is the solution of the following system of linear equations:

$$\sum_{k=1}^{n} c_k R(s_j; s_k) = f(s_j) \qquad j=1,2,\dots,n \qquad (*)$$

and so *we can compute* $\{c_k;\ k=1,2,\dots,n\}$ *merely by the evaluation of the functions* $f(\,.\,)$, $R(\,.\,,s_k)$ $(k=1,2,\dots,n)$ *at the points* $s_1, s_2, \dots, s_n$ (instead of using product integrals, which are the usual form of scalar product in Hilbert spaces).

The formula $(*)$ for computing $\{c_k;\ k=1,2,\dots,n\}$ tells us that the projection of $f \in \mathscr{H}(R)$ onto the n-dimensional subspace generated by

$$\{R(.,s_k);\ k=1,2,\dots,n\} \qquad s_k \in \mathscr{D}$$

is the interpolation of $f=f(t)$ with nodes $\{s_k;\ k=1,2,\dots,n\}$ by the functions in the form

$$c_1 R(.\,, s_1)+c_2 R(.\,, s_2)+\ldots+c_n R(.\,, x_n).$$

Remark. For computational purposes it is more convenient to adopt the following RKHS method. Applying Theorem 3.4.1.1, we compute the kernel $R_{\mathcal{N}}$ of the n-dimensional subspace generated by

$$R(.\,, s_k) \qquad k = 1, 2, \ldots, n, \quad s_k \in \mathcal{D}$$

and the projection is constructed by this kernel according to Theorem 3.4.1.2. Hence

$$R_{\mathcal{N}}(.\,, t) = \sum_{k=1}^{n} c_k(t) R(.\,, s_k)$$

such that

$$(R_{\mathcal{N}}(.\,, t) - \sum_{k=1}^{n} c_k(t) R(.\,, s_k) | R(.\,, s_j)) = 0 \qquad j = 1, 2, \ldots, n.$$

It follows that

$$R_{\mathcal{N}}(s_j, t) = \sum_{k=1}^{n} c_k(t) R(s_j, s_k)$$

where $\{c_k(t);\ k=1, 2, \ldots, n\}$ is the solution of the system of linear equations

$$\sum_{k=1}^{n} c_k(t)\, R(s_j, s_k) = R_{\mathcal{N}}(s_j, t) \qquad j = 1, 2, \ldots, n$$

and hence each $c_k(t)$ is a linear combination of the functions

$$R(s_j, t) = \overline{R(t, s_j)}; \quad j = 1, 2, \ldots, n.$$

We now apply Theorem 3.4.1.2:

$$Pf(t) = (R_{\mathcal{N}}(.\,, t) | f(.)) = \sum_{k=1}^{n} c_k(t)(R(.\,, s_k) | f(.))$$

$$= \sum_{k=1}^{n} c_k(t) f(s_k) \qquad f \in \mathcal{H}(R).$$

3.5 Quadrature formulae and splines

3.5.1. Certain formulae for the approximate evaluation of definite integrals are called quadrature formulae. In the simplest case they have the form

$$\sum_{k=1}^{n} c_k f(t_k) \tag{*}$$

where f is a continuous function and $\{t_k;\ k=1, 2, \ldots, n\}$ are given points in the interval $[0, 1]$ and we require $\{c_k;\ k=1, 2, \ldots, n\}$ such that

$$r(f) := \Big|\int_0^1 f(t)\,\mathrm{d}t - \sum_{k=1}^{n} c_k f(t_k)\Big|$$

is small. $r=r(f)$ is called the *remainder functional* and the quadrature formula is called *exact for a class* $\mathcal{M}$ *of continuous functions* if

$$\int_0^1 f(t)\,\mathrm{d}t = \sum_{k=1}^{n} c_k f(t_k) \qquad \text{for } f\in\mathcal{M}.$$

The simplest examples of quadrature formulae are the trapezium formula and Simpson's rule. In both cases $t_k=k/n$ and

$$c_k = \begin{cases} 1/n & \text{for } 1<k<n+1 \\ 1/2n & \text{for } k=1 \text{ and } k=n+1 \end{cases}$$

for the trapezium formula and

$$c_k = \begin{cases} 1/3n & \text{for odd } k;\ 1<k<2n+1 \\ 2/3n & \text{for even } k \\ 1/6n & \text{for } k=1 \text{ and } k=2n+1 \end{cases}$$

for Simpson's rule.

The trapezium formula is exact for every spline of first degree and Simpson's rule is still exact for the quadratic splines.

3.5.2. Now let f be a twice-differentiable function such that $f''\in L^2[0, 1]$ and $f(0)=f'(0)=0$; then

$$f(t) = \int_0^1 (t-\tau)_+ f''(\tau)\,\mathrm{d}\tau$$

and hence

$$r(f) := \Big|\int_0^1 f(t)\,\mathrm{d}t - \sum_{k=1}^{n} c_k f(t_k)\Big|$$

$$= \Big|\int_0^1 \Big(\int_0^1 (t-\tau)_+ f''(\tau)\,\mathrm{d}\tau\Big)\mathrm{d}t - \sum_{k=1}^{n} c_k \int_0^1 (t_k-\tau)_+ f''(\tau)\,\mathrm{d}\tau\Big|.$$

By changing the order of integration we obtain

$$r(f) = \int_0^1 \Big(\int_0^1 (t-\tau)_+\,\mathrm{d}t - \sum_{k=1}^{n} c_k (t_k-\tau)_+\Big) f''(\tau)\,\mathrm{d}\tau.$$

This means that r is a bounded linear functional of the RKHS $\mathcal{H}(S)$ introduced in Example 3 of § 3.3. The norm of this functional is

$$\|r\| = \Big\| \int_0^1 (t-\tau)_+ \, dt - \sum_{k=1}^{n} c_k (t_k - \tau)_+ \Big\|_2$$

by the Riesz–Fréchet Theorem and the definition of $\mathscr{H}(S)$ ($\| \, . \, \|_2$ means L^2-norm). Our purpose in the next section is to determine $\{c_k;\ k=1, 2, \ldots, n\}$ in such a way that $\|r\|$ is minimal.

3.5.3. Applying the Projection Theorem for

$$\int_0^1 (t-\tau)_+ \, dt \in L_0^2[0, 1]$$

and the n-dimensional subspace $\mathscr{M}$ of $L_0^2[0, 1]$ generated by

$$\{(t_k - \tau)_+;\ k = 1, 2, \ldots, n\}$$

we find that $\|r\|$ is minimal if and only if

$$\Big(\int_0^1 (t-\tau)_+ \, dt - \sum_{k=1}^{n} c_k (t_k - \tau)_+ \Big| (t_j - \tau)_+ \Big) = 0 \qquad j = 1, 2, \ldots, n$$

where $(\, . \mid . \,)$ is the scalar product in $L^2[0, 1]$. That is,

$$\int_0^1 \int_0^1 (t-\tau)_+ (t_j - \tau)_+ \, dt \, d\tau$$

$$= \sum_{k=1}^{n} c_k \int_0^1 (t_k - \tau)_+ (t_j - \tau)_+ \, d\tau \qquad j = 1, 2, \ldots, n.$$

Interchanging the order of integration and considering Example 3 in § 3.3 once more, we obtain the following system of linear equations for $\{c_k;\ k=1, 2, \ldots, n\}$:

$$\int_0^1 S(t, t_j) \, dt = \sum_{k=1}^{n} c_k S(t_k, t_j) \qquad j = 1, 2, \ldots, n. \qquad (*)$$

What we have obtained is the following result.

3.5.3.1 Theorem. Let us consider the class of functions

$$\{f\colon f'' \in L^2[0, 1];\ f(0) = f'(0) = 0\}.$$

Then the remainder functional $r = r(f)$ of the quadrature formula 3.5.1 (*) has the minimal norm in $\mathscr{H}(S)$ if and only if the quadrature formula is exact for the cubic splines $S(\, . \, , t_j);\ j=1, 2, \ldots, n$.

In this case, $\{c_k;\ k=1, 2, \ldots, n\}$ is the unique solution of the system $(*)$ of linear equations.

Remark 1. For more general applications of the RKHS theory to quadrature formulae see § 3.11.31.

Remark 2. The functions

$$\{(t_k-\tau)_+;\ k = 1, 2, \ldots, n\}$$

are linearly independent and hence the solution of $(*)$ is unique.

Remark 3. Every cubic spline has a continuous second derivative and hence it is considered to be an element of $\mathscr{H}(S)$; it is particularly interesting to find quadrature formulae that are exact for the B-splines (see § 3.10.1).

3.6 Sampling

3.6.1. A fundamental problem in communication theory relates to how a 'signal' $f=f(t)$ can be reconstructed from the sampled values $f(t_k)$ with good accuracy. The basic theorem relating to this problem is as follows. If the function $f \in \mathrm{L}^2[-\infty, +\infty]$ is represented in the form

$$f(t) = \frac{1}{2A} \int_{-A}^{+A} \mathrm{e}^{\mathrm{i}\omega t} F(\omega)\, \mathrm{d}\omega$$

where $F \in \mathrm{L}^2[-A, +A]$ then

$$f(t) = \sum_{k=-\infty}^{+\infty} f\left(\frac{k\pi}{A}\right) \frac{\sin A(t-(k\pi/A))}{A(t-(k\pi/A))}$$

for all t. This is the Sampling Theorem.

The popular formulation of this theorem is that a band-limited signal f can be completely characterised from the samples $f(k\pi/A)$; $k=1, 2, \ldots$. Moreover, in every finite time interval only a finite sample is needed and the necessary number of samples is in an inverse ratio to the bandwidth.

The usual proof of the Sampling Theorem is based on the Fourier transform technique; however, the simplest proof is based on an RKHS method. It was shown in Example 1 of § 3.3 that the finite Fourier transforms form an RKHS with kernel

$$R(s, t) = \frac{\sin A(s-t)}{A(s-t)}$$

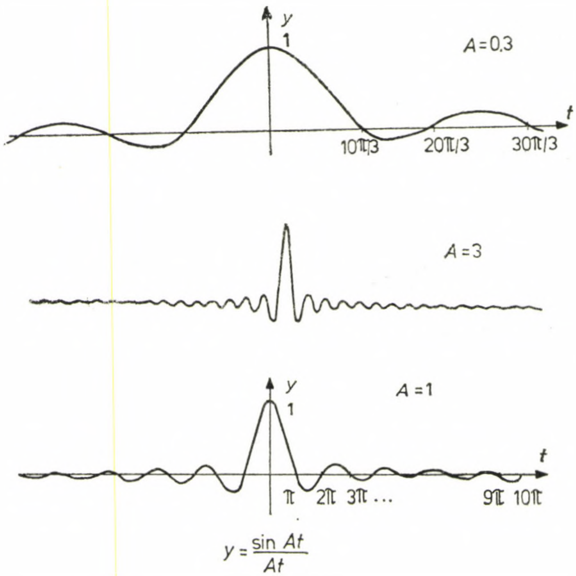

fig. 3.1

(see figure 3.1) and $\{R(\,.\,,t_k);\ t_k=k\pi/A\}$ is a complete orthonormal sequence. In this context we have only to prove for the Sampling Theorem that the Fourier coefficient with respect to $R(\,.\,,t_k)$ is $f(t_k)$ for any $f\in\mathscr{H}(R)$. Indeed,

$$(f|R(.,t_k)) = f(t_k).$$

3.6.2. The following generalisation is indicated by this simple proof (Gulyás, 1967).

3.6.2.1 Theorem. Let $R(s,t)$ be a symmetric, positive definite function and let t_k; $k=1,2,\ldots$ be a sequence of points such that

(*a*) $\mathscr{H}(R)$ is generated by $\{R(\,.\,,t_k);\ k=1,2,\ldots\}$;

(*b*) $$R(t_i,t_j)=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j.\end{cases}$$

Then for every $f\in\mathscr{H}(R)$,

$$f(t)=\sum_{k=1}^{\infty} f(t_k)R(t,t_k)$$

for every $t\in\mathscr{D}$.

If $\mathscr{D}\subseteq R^n$ and $R(s,t)$ is a continuous function then the series is uniformly convergent on every bounded subset of $\mathscr{D}$. The proof is left to the reader.

It is interesting to study sampling pairs $\{t_k, R(t, t_k);\ k=1, 2, \ldots\}$ besides the traditional

$$\left\{\frac{k\pi}{A}, \frac{\sin A\big(t-(k\pi/A)\big)}{A\big(t-(k\pi/A)\big)}\right\}.$$

$*$3.7 Conformal mappings and kernels

If $\mathscr{D}$ is a simple connected bounded region with a smooth boundary in the complex plane then, by Riemann's Conformal Mapping Theorem, there is a unique holomorphic function $f=f(z)$ that maps $\mathscr{D}$ one-one onto the unit disc $\{z\colon |z|<1\}$ such that $f(z_0)=0$ and $f'(z_0)>0$ (for a certain interior point $z_0 \in \mathscr{D}$).

We now give a generalisation of $\mathrm{A}(\mathscr{D})$, introduced in Example 6 of § 3.1.3 for any simply connected bounded region $\mathscr{D}$, and it will be shown that the mapping $f(z)$ in Riemann's Conformal Mapping Theorem has the form

$$f(z) = \left(\frac{\pi}{R(z_0, z_0)}\right)^{1/2} \int_{z_0}^{z} R(\zeta, z_0)\, \mathrm{d}\xi$$

where R is the kernel of $\mathrm{A}(\mathscr{D})$.

3.7.1. Our considerations are based on the complex form of the Divergence Theorem on the plane, which may be stated as follows. Let $\mathscr{D}$ be a simple connected bounded region with a smooth boundary, let $\mathrm{A}(\mathscr{D})$ be the linear space of functions analytic on $\mathscr{D}$ and

$$\iint_{\mathscr{D}} |f(z)|^2\, \mathrm{d}x\, \mathrm{d}y < \infty.$$

If $g, h \in \mathrm{A}(\mathscr{D})$, $\mathscr{D}' \subset \mathscr{D}$ and C is the boundary of $\mathscr{D}'$, then

$$\frac{1}{2\mathrm{i}} \int_C g(z)\overline{h(z)}\, \mathrm{d}z = \iint_{\mathscr{D}} g(z)\overline{h'(z)}\, \mathrm{d}x\, \mathrm{d}y.$$

Proof. For the pair $P=P(x, y)$, $Q=Q(x, y)$, the Divergence Theorem is as follows:

$$\int_C P\, \mathrm{d}y - Q\, \mathrm{d}x = \iint_{\mathscr{D}'} (P'_x + Q'_y)\, \mathrm{d}x\, \mathrm{d}y. \tag{1}$$

On the other hand, for any $f=f(z)$,

$$\int_C f(z)\, \mathrm{d}z = \int_C u\, \mathrm{d}x - v\, \mathrm{d}y + \mathrm{i}\Big(\int_C v\, \mathrm{d}x + u\, \mathrm{d}y\Big) \tag{2}$$

where $u=u(x, y)=\operatorname{Re} f(z)$ and $v=v(x, y)=\operatorname{Im} f(z)$. Comparing (1) and (2), we obtain

$$\int_C f(z)\,\mathrm{d}z = \iint_{\mathscr{D}'} (-v_x - u_y)\,\mathrm{d}x\,\mathrm{d}y + \mathrm{i} \iint_{\mathscr{D}'} (u_x - v_y)\,\mathrm{d}x\,\mathrm{d}y \tag{3}$$

and, applying (3) for $f(z) = g(z)\overline{h(z)}$, the complex form of the Divergence Theorem is obtained.

3.7.2. It is easy to show that

$$(f|g) := \iint_{\mathscr{D}} f(z)\overline{g(z)}\,\mathrm{d}x\,\mathrm{d}y \qquad f, g \in \mathrm{A}(\mathscr{D})$$

is also a scalar product in this case, and it turns out that $\mathrm{A}(\mathscr{D})$ is a Hilbert space.

$\mathrm{A}(\mathscr{D})$ is an RKHS. In fact, for $h(z) = z - z_0$, $\mathscr{D}' = \{z\colon |z - z_0| < r\}$, the Divergence Theorem is as follows:

$$\frac{1}{2\mathrm{i}} \int_{|z-z_0|=r} g(z)\overline{(z-z_0)}\,\mathrm{d}z = \iint_{\mathscr{D}'} g(z)\,\mathrm{d}x\,\mathrm{d}y. \tag{4}$$

Moreover,

$$\frac{1}{2\mathrm{i}} \int_{|z-z_0|=r} g(z)\overline{(z-z_0)}\,\mathrm{d}z = \frac{r^2}{2\mathrm{i}} \int_{|z-z_0|=r} \frac{g(z)}{z-z_0}\,\mathrm{d}z = \pi r^2 g(z_0) \tag{5}$$

by applying the Cauchy integral formula for this case. Comparing (4) and (5), we obtain

$$|g(z_0)| = \frac{1}{\pi r^2} \left| \iint_{\mathscr{D}'} g(z)\,\mathrm{d}x\,\mathrm{d}y \right|$$

and from the Cauchy–Schwarz inequality in $\mathrm{A}(\mathscr{D})$,

$$\left| \iint_{\mathscr{D}'} g(z)\,\mathrm{d}x\,\mathrm{d}y \right|^2 \leqslant r^2 \pi \iint_{\mathscr{D}'} |g(z)|^2\,\mathrm{d}x\,\mathrm{d}y.$$

We conclude that the evaluation functionals are bounded by $1/r\pi^{1/2}$.

3.7.3. We now turn to the construction of $f = f(z)$, the conformal mapping from $\mathscr{D}$ onto the unit disc.

If $\mathscr{D}' \subset \mathscr{D}$ with boundary C and $f = f(z)$ is the conformal mapping function from $\mathscr{D}$ onto the unit disc, then for any $g \in \mathrm{A}(\mathscr{D})$,

$$\frac{1}{2\pi\mathrm{i}} \int_C \frac{g(z)}{f(z)}\,\mathrm{d}z = \operatorname*{Res}_{z=z_0} \frac{g(z)}{f(z)} = \frac{g(z_0)}{f'(z_0)} \tag{6}$$

by a well-known theorem of analytic function theory, since $f(z_0) = 0$ and $f'(z_0) > 0$.

Let $\mathscr{D}'=\mathscr{D}_r$ be a subdomain of $\mathscr{D}$ such that f maps C_r, the boundary of $\mathscr{D}_r$, onto $\{z\colon |z|=r<1\}$. Then for any $g\in A(\mathscr{D})$,

$$\frac{1}{2\pi i}\int_{C_r}\frac{g(z)}{f(z)}\,dz=\frac{1}{2\pi r^2 i}\int_{C_r} g(z)\overline{f(z)}\,dz \tag{7}$$

since $f(z)\overline{f(z)}=r^2$ in this case, and applying the Divergence Theorem from 3.7.1,

$$\frac{1}{2\pi r^2 i}\int_{C_r} g(z)\overline{f(z)}\,dz=\frac{1}{\pi r^2}\iint_{\mathscr{D}_r} g(z)\overline{f'(z)}\,dx\,dy. \tag{8}$$

Comparing (6), (7) and (8), we obtain, for any $z_0\in\mathscr{D}$,

$$g(z_0)=\frac{1}{2\pi r^2 i}f'(z_0)\int_{C_r} g(z)\overline{f(z)}\,dz=\frac{1}{\pi r^2}\,f'(z_0)\iint_{\mathscr{D}_r} g(z)\overline{f'(z)}\,dx\,dy$$

for any $0<r<1$ and hence also

$$g(z_0)=\frac{1}{\pi}\iint_{\mathscr{D}} f'(z_0)\overline{f'(z)}g(z)\,dx\,dy$$

which means that the kernel is given by

$$R(z,\zeta)=\frac{1}{\pi}\overline{f'(\zeta)}f'(z). \tag{9}$$

3.7.4. It follows that

$$R(z_0,z_0)=\frac{1}{\pi}|f'(z_0)|^2$$

and hence

$$f'(z_0)=(\pi R(z_0,z_0))^{1/2}$$

since it is assumed that $f'(z_0)>0$. Now, from (9) and the above considerations it is clear that

$$f(z)=\left(\frac{\pi}{R(z_0,z_0)}\right)^{1/2}\int_{z_0}^{z} R(\zeta,z_0)\,d\zeta.$$

Remark 1. Since $\{z^k;\ k=0,1,2,\dots\}$ is a complete system of functions, a complete orthonormal sequence can be obtained if the Gram–Schmidt process (§ 2.2.4) is applied to $\{z^k;\ k=0,1,2,\dots\}$. The kernel $R=R(z,\xi)$ can then be constructed e.g. in the manner described in § 3.2.

Remark 2. It is easy to show that Example 6 in § 3.1.3 can be considered as a special case of this result. In fact, if

$$R(z,z_0)=\frac{1}{\pi}\frac{1}{(1-\bar{z}_0 z)^2}\qquad |z_0|<1$$

then

$$f(z) = (1-|z_0|^2) \int_{z_0}^{z} \frac{d\zeta}{(1-\bar{z}_0\zeta)^2} = \frac{z-z_0}{1-\bar{z}_0 z}$$

which gives a conformal mapping of the unit disc onto itself.

*3.8 Gaussian processes

The (joint) probability law is completely determined by the mean and covariance function for a Gaussian process. On the other hand, there is a 1–1 correspondence between kernel and RKHS. Hence it is natural to attempt to formulate the connections between Gaussian processes as the connections between $\mathscr{H}(R_1)$ and $\mathscr{H}(R_2)$, where $R_1 = R_1(s, t)$ and $R_2 = R_2(s, t)$ are the covariance functions of the corresponding Gaussian process.

More particularly, let two probability measures P and Q be considered on a measure space and let $\{X(t);\ 0 \leqslant t < \infty\}$ be a Gaussian stochastic process with respect to both measures. Furthermore, let R_P and R_Q be the covariance functions and m and 0 the means of the processes.

P and Q are called equivalent if the same subsets have measure zero with respect to P and Q; in other words, if the same events have a zero probability. What are the conditions, in terms of $m(t)$, $R_P(s, t)$ and $R_Q(s, t)$, for the equivalence of P and Q?

The answer is as follows. For the equivalence of P and Q it is necessary and sufficient that

(*a*) $m(\,.\,) \in \mathscr{H}(R_Q)$;
(*b*) R_P has a representation of the form

$$R_P(s, t) = \sum_{k=1}^{\infty} \alpha_k e_k(s) e_k(t)$$

with

$$\sum_{k=1}^{\infty} (1-\alpha_k)^2 < \infty \qquad \text{and} \qquad \alpha_k \geqslant c > 0 \qquad k = 1, 2, \ldots$$

where $\{e_k\}$ is a complete orthonormal sequence in $\mathscr{H}(R_Q)$.

A more detailed discussion of the equivalence and singularity of the Gaussian measures is beyond the scope of this book.

∗3.9 Sobolev spaces and generalised derivative

Here we present a short review of the generalisation of the derivative given by S L Sobolev and L Schwartz. This general notion of derivative also led to the weak solution of differential equations and Hilbert spaces with continuous evaluation functionals.

3.9.1. First we shall consider functions on the real line. We shall use the notation C_{00}^{∞} for the linear space of infinitely differentiable functions with compact (i.e. bounded) support. Recall that the support of a function f is the closure of

$$\{t\colon f(t) \neq 0;\ t \in \mathcal{D}\}$$

where $\mathcal{D}$ is the domain of the function f. The elements of C_{00}^{∞} are often called *very good functions.*

A function f is called *locally integrable* if

$$\int_a^b f(t)\,\mathrm{d}t$$

exists for every finite interval $[a, b]$.

The locally integrable functions $\{h_\delta\}$ are called an *approximate identity* if

(*a*) $\{\delta\}$ is a set of non-negative numbers and 0 is contained in the closure of $\{\delta\}$.

(*b*) $\operatorname{supp} h_\delta \subseteq \{t\colon |t| \leqslant \delta\}$ and $h_\delta(t) \geqslant 0$.

(*c*) $\int_{-\infty}^{+\infty} h_\delta(t)\,\mathrm{d}t = 1.$

It is easy to construct an approximate identity. In fact, if $h=h(t)$ is any locally integrable function, then the functions

$$h_\delta(t) = \begin{cases} (1/K_\delta)\,|h(t)| & \text{if } |t| < \delta \\ 0 & \text{otherwise} \end{cases}$$

(where $K_\delta = \int_{|t| \leqslant \delta} |h(t)|\,\mathrm{d}t \neq 0$) form an approximate identity, as is easily verified.

To construct an approximate identity belonging to C_{00}^{∞} is more difficult. If

$$h_\delta(t) = \begin{cases} \exp(t^2 \quad \delta^2)^{-1} & \text{if } |t| < \delta \\ 0 & \text{otherwise} \end{cases} \tag{$*$}$$

then

$$\lim_{|t|\to\delta} \frac{d^n}{dt^n} \exp(t^2-\delta^2)^{-1} = 0 \qquad n = 0, 1, 2, \ldots$$

and hence $h_\delta = h_\delta(t)$ are infinitely differentiable functions with compact support.

It can also be proved that

$$g*f := \int_{-\infty}^{+\infty} g(\tau)f(t-\tau)\,d\tau$$

which is called the convolution of the functions g and f, is an infinitely differentiable function for any locally integrable f if $g \in C_{00}^{\infty}$.

3.9.1.1 Theorem. If $\{h_\delta\}$ is the approximate identity defined by $(*)$ then

$$h_\delta * f \to f$$

uniformly on compact subsets for any continuous function f and $h_\delta * f$ tends to f in the L^2-norm if $f \in L^2$ as $\delta \to 0$.

For the *proof* see, for example, Showalter (1977), § IX.3.1.

The linear functionals on C_{00}^{∞} are called *distributions.*

Remark. Here we have given an oversimplified notion of distribution. Later we shall give a more rigorous definition for those classes of distributions that are connected with Hilbert space structures.

Example 1. For every locally integrable function g we have the linear functional

$$G(\varphi) := \int_{-\infty}^{+\infty} g(t)\varphi(t)\,dt \qquad \varphi \in C_{00}^{\infty}.$$

G is well defined since for every $\varphi \in C_{00}^{\infty}$ there is a finite interval I such that $\varphi(t) = 0$ if $t \notin I$ (i.e. φ has compact support). Moreover, it can be proved on the basis of the previous theorem that if

$$\int_{-\infty}^{+\infty} g(t)\varphi(t)\,dt = 0$$

for every $\varphi \in C_{00}^{\infty}$, then $g(t) = 0$. In the sense described in this example, therefore, *each locally integrable function is a distribution.*

Example 2. The evaluation functionals

$$\delta_c(\varphi) := \varphi(c)$$

are linear functionals on C_{00}^{∞} called *Dirac delta distributions.*

Example 3. For every locally integrable function g,

$$Dg(\varphi) := -\int_{-\infty}^{+\infty} g(t)\varphi'(t)\,\mathrm{d}t \qquad \varphi \in \mathrm{C}_{00}^{\infty}$$

where φ', the derivative of φ, is a distribution called the *generalised derivative* of g. Similarly,

$$D^n g(\varphi) := (-1)^n \int_{-\infty}^{+\infty} g(t)\varphi^{(n)}(t)\,\mathrm{d}t$$

is called the nth generalised derivative of g.

If the function g has a continuous derivative, then

$$\int_{-\infty}^{+\infty} g'(t)\varphi(t)\,\mathrm{d}t = -\int_{-\infty}^{+\infty} g(t)\varphi'(t)\,\mathrm{d}t \qquad \varphi \in \mathrm{C}_{00}^{\infty}$$

by integration by parts, since φ has compact support. Hence we also have $Dg = \mathrm{d}g/\mathrm{d}t$ in terms of distributions if g' exists. Let

$$l_+(t_0 - t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\int_{-\infty}^{+\infty} l_+(t_0 - t)\varphi'(t)\,\mathrm{d}t = \int_{-\infty}^{t_0} \varphi'(t)\,\mathrm{d}t = \varphi(t_0)$$

by integration by parts and hence δ_c is the generalised derivative of $l_+ = l_+(c - t)$. Moreover,

$$D\delta_c(\varphi) = -\varphi'(c)$$

i.e. the second derivative of $l_+(c-t)$ is the functional whose value is $-\varphi'(c)$ for every $\varphi \in \mathrm{C}_{00}^{\infty}$ since

$$\int_{-\infty}^{+\infty} l_+(t_0 - t)\varphi''(t)\,\mathrm{d}t = -\int_{-\infty}^{t_0} \varphi''(t)\,\mathrm{d}t = -\varphi'(t_0).$$

3.9.2. It is easy to extend the considerations that led to the concepts of distribution and generalised derivative to the case of functions of several variables.

C_{00}^{∞} is the linear space of functions with compact (i.e. bounded) support. A function f is called *locally integrable* if the volume integral

$$\int_{\mathscr{D}} f(t)\,\mathrm{d}t$$

exists for any compact $\mathscr{D}$ (there and in the sequel $t := (x_1, x_2, \ldots, x_n)$ and $\mathrm{d}t := \mathrm{d}x_1\,\mathrm{d}x_2 \ldots \mathrm{d}x_n$). The definition of *approximate identity* $\{h_\delta\}$ is the same with

$$|t| := (x_1^2+x_2^2+\ldots+x_n^2)^{1/2}$$

and t^2 replaced by $|t|^2$ in 3.9.1 (*).

The *convolution* of the functions f and g is defined by the volume integral

$$g*f := \int_{R^n} g(\tau)f(t-\tau)\,\mathrm{d}t$$

where R^n is the n-dimensional Euclidean Space and Theorem 3.9.1.1 is also valid. In general, $\int_{-\infty}^{+\infty}$ is replaced by $\int_{R^n}$ in the case of several variables.

$$D_i f(\varphi) := -\int_{R^n} f(t)\frac{\partial}{\partial x_i}\varphi(t)\,\mathrm{d}t$$

is defined as the generalised partial derivative. In particular, if

$$E_+(t) = \begin{cases} 1 & \text{if } x_i > 0;\; i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

then

$$D_i E_+(\varphi) = -\int_{R^n} E_+(t)\frac{\partial}{\partial x_i}\varphi(t)\,\mathrm{d}t$$

$$= -\int_0^\infty \int_0^\infty \ldots \int_0^\infty \frac{\partial}{\partial x_i}\varphi(t)\,\mathrm{d}t = -\int_0^\infty \int_0^\infty \ldots \int_0^\infty \varphi(0, x_2, \ldots, x_n)\,\mathrm{d}x_2\ldots\mathrm{d}x_n$$

and

$$D_1 D_2 \ldots D_n E_+(\varphi) = \varphi(0, 0, \ldots, 0).$$

3.9.3. An idea intermediate between the derivative and the generalised derivative of a function is the weak derivative. If $Dg = f \in \mathrm{L}^2$, i.e. there exists $f \in \mathrm{L}^2$ such that

$$-\int_{-\infty}^{+\infty} g(t)\varphi'(t)\,\mathrm{d}t = \int_{-\infty}^{+\infty} f(t)\varphi(t)\,\mathrm{d}t \qquad \varphi \in \mathrm{C}_{00}^{\infty}$$

then $f \in \mathrm{L}^2$ is called the *weak derivative* of g. A weak derivative is a function and also a generalised derivative. However, there are generalised derivatives that are not L^2-functions and hence are not weak derivatives, (e.g. DI_+ or $D_i E_+$, as we have seen).

Now, let L be a *linear differential form*, i.e.

$$Lf := \sum_{k=0}^{n} c_k \frac{\mathrm{d}^k}{\mathrm{d}t^k} f$$

where $c_k = c_k(t)$ are appropriate functions. Then the formal adjoint L^* is defined as

$$\int_{-\infty}^{+\infty} Lf\varphi \,\mathrm{d}t = \int_{-\infty}^{+\infty} fL^*\varphi \,\mathrm{d}t \qquad f\in \mathrm{L}_0^2,\ \varphi\in \mathrm{C}_{00}^{\infty}.$$

The definition is similar in the case of functions of several variables.

Example 1. If

$$Lf := \frac{\mathrm{d}}{\mathrm{d}t} f$$

then

$$\int_{-\infty}^{+\infty} f'\varphi \,\mathrm{d}t = -\int_{-\infty}^{+\infty} f\varphi' \,\mathrm{d}t.$$

Hence

$$L^* = -\frac{\mathrm{d}}{\mathrm{d}t}.$$

Example 2. If

$$Lf := \frac{\mathrm{d}}{\mathrm{d}t}\left[c\frac{\mathrm{d}}{\mathrm{d}t} f\right]$$

then

$$\int_{-\infty}^{+\infty} Lf\varphi \,\mathrm{d}t = -\int_{-\infty}^{+\infty} cf'\varphi' \,\mathrm{d}t.$$

Moreover,

$$\int_{-\infty}^{+\infty} cf'\varphi' \,\mathrm{d}t = -\int_{-\infty}^{+\infty} f(c\varphi')' \,\mathrm{d}t$$

by integration by parts. Hence $L=L^*$ in this case.

Example 3. If u is a function of two variables and $Lu=\Delta u$, where Δ is the Laplace operator, then

$$\int_{R^n} \Delta u\varphi \,\mathrm{d}t = \int_{R^n} \left(\frac{\partial^2}{\partial x_1^2} u + \frac{\partial^2}{\partial x_2^2} u\right) \varphi \,\mathrm{d}t$$

$$= -\int_{R^n} \frac{\partial}{\partial x_1} u \frac{\partial}{\partial x_1} \varphi \,\mathrm{d}t - \int_{R^n} \frac{\partial}{\partial x_2} u \frac{\partial}{\partial x_2} \varphi \,\mathrm{d}t$$

$$= -\int_{R^n} \operatorname{grad} u \operatorname{grad} \varphi \,\mathrm{d}t.$$

Moreover (again by integration by parts),

$$\int_{R^n} \operatorname{grad} u \operatorname{grad} \varphi \,\mathrm{d}t = -\int_{R^n} u\Delta\varphi \,\mathrm{d}t$$

and hence $L=L^*$ in this case, too.

Let us consider the differential equation

$$Lu = f \qquad f \in L^2.$$

Then u is a *weak solution* if

$$\int_{R^n} uL^*\varphi \, dt = \int_{R^n} f\varphi \, dt.$$

Remark. Note the connection between the weak solution and the weak derivative. On the basis of the foregoing examples, if

$$-\int_{-\infty}^{+\infty} c(t)y'(t)\varphi'(t)\, dt = \int_{-\infty}^{+\infty} f(t)\varphi(t)\, dt \qquad \varphi \in C_{00}^{\infty}$$

for $y = y(t)$, then it is a weak solution of the differential equation

$$\frac{d}{dt}\left(c\frac{d}{dt}y\right) = f$$

however, in the integral formula it is assumed only that the *first derivative* of y exists. If

$$-\int_{R^n} \operatorname{grad} u \operatorname{grad} \varphi \, dt = \int_{R^n} f\varphi \, dt \qquad \varphi \in C_{00}^{\infty}$$

for the function $u = u(t)$ of several variables, then it is a weak solution of the Laplace equation

$$\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} u = f$$

however, in the integral formula it is assumed only that the *gradient of u* exists.

3.9.4. Let

$$W^m = \{f\colon D^m f \in L^2(\Omega)\}$$

where $\Omega \subseteq R^n$, i.e. W^m is the linear space of functions with mth weak derivatives. W^m is a Hilbert space with the scalar product

$$(f|g) := \sum_{k=0}^{m} \int_{\Omega} D^k f D^k g \, dt$$

where m means the indices $(k_1, k_2, \ldots, k_n)$ whose sum is m.

It is easy to verify that $(\,.\,|\,.\,)$, defined above, is indeed a scalar product. For the completeness, let $\{f_n\}$ be a Cauchy sequence in W^m. Then $\{D^k f_n\}$ is a Cauchy sequence in $L^2(\Omega)$ for every $|k| \leq |m|$ and hence there exist $f^{(k)} \in L^2$ such that

$$D^k f_n \to f^{(k)} \qquad |k| \leq |m|$$

in L^2-norm. We assert that $f^{(k)}=D^k f$. Indeed,

$$\int_{R^n} D^k f_n \varphi \, dt = (-1)^{|k|} \int_{R^n} f_n \varphi^{(k)} \, dt$$

and

$$\int_{R^n} |f_n - f| \, \varphi^{(k)} \, dt \leq \|f_n - f\|_2 \|\varphi^{(k)}\|_2 .$$

Moreover,

$$\int_{R^n} |D^k f_n - f^{(k)}| \, \varphi \, dt \leq \|D^k f_n - f^{(k)}\|_2 \|\varphi\|_2 .$$

It follows that

$$\int_{R^n} f \varphi^{(k)} \, dt = \int_{R^n} f^{(k)} \varphi dt.$$

The Hilbert spaces W^m; $m=1, 2, \ldots$ are the Sobolev spaces in the strict sense.

Let W_0^m be the closure of C_{00}^{∞} in W^m. Then it can be shown that, in general, $W_0^m \neq W^m$. Let u be a continuous function, $u \in W^m$, and let $\partial\Omega$ be the boundary of the domain Ω. Moreover,

$$Bu := u|_{\partial\Omega}$$

i.e. the restriction operator. It can then be shown that B is *bounded* in the norm of W^m. Hence there is a unique extension of B to a bounded linear operator from $W^m(\Omega)$ into $L^2(\partial\Omega)$. For a noncontinuous $u \in W^m(\Omega)$, $Bu \in L^2(\partial\Omega)$ is called the *generalised boundary value*. (For more about this concept see Showalter 1977 § II.3.)

3.10 The finite-element method

The finite-element method can be considered as a particular application of the projection principle for a finite-dimensional subspace to the approximate solution of functional differential, integral, etc equations. Again, certain reproducing kernel Hilbert spaces will be applied without using the kernel.

3.10.1. The simplest cubic splines with nodes $\{s_k;\ k=1, 2, \ldots, n\}$ are

$$(s_k - t)_+^3 \qquad k = 1, 2, \ldots, n \qquad (*)$$

(see figure 3.2). They are linearly independent functions (see, for example, § 3.11.23). Now let $s_{i+1}-s_i=h$, i.e. constant for $i=3, 4, \ldots, n-2$, and

$$B_i(t) = \frac{1}{h^3} (t-s_{i-2})_+^3 - 4(t-s_{i-1})_+^3 + 6(t-s_i)_+^3$$
$$-4(t-s_{i+1})_+^3 + (t-s_{i+2})_+^3 \qquad i = 3, 4, \ldots, n-2.$$

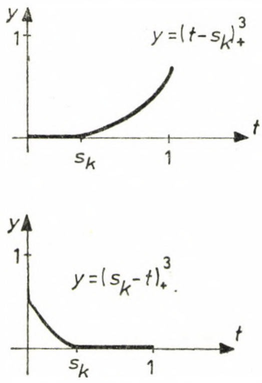

fig. 3.2

It can be proved that

$$B_i(t) = 0 \qquad \text{if} \quad t \neq [s_i-2h,\ s_i+2h]$$

i.e. the support of each function $B_i=B_i(t)$ is of length $4h$. Moreover, $\{B_i;\ i=3, 4, \ldots, n-2\}$ are the splines of minimal support since it can also be proved that every spline with possible nodes $\{s_k;\ k=1, 2, \ldots, n\}$ and support smaller than $4h$ is identically zero. $B_i(t)$; $i=3, 4, \ldots, n-2$ are called *B-splines.*

Remark. We can also give $B_i(t)$ as a linear combination of functions in the form $(*)$ instead of $(t-s_k)^3_+$.

3.10.2. Let us now consider the differential equation

$$y''-cy=f \qquad c=c(t)\geqslant 0$$

where $f\in L^2[0, 1]$ and c is a continuous function, with the boundary conditions $y(0)=y(1)=0$. We seek an approximate solution of the form

$$y_s(t) = \sum_{k=1}^{n} \alpha_k B_k(t)$$

where $\{B_k;\ k=1, 2, \ldots, n\}$ are the B-splines (figure 3.3). In what follows the notation $Dy:=y''-cy$ will be used and $y_s=y_s(t)$ will be called the best approximation of the solution of the differential equation if

$$\|f-Dy_s\|_2 \qquad (*)$$

is minimal.

Applying 2.4.1.4, $(*)$ is minimal if and only if

$$\int_0^1 (f(t)-Dy_s(t))B_k(t)\,dt = 0 \qquad k=1, 2, \ldots, n$$

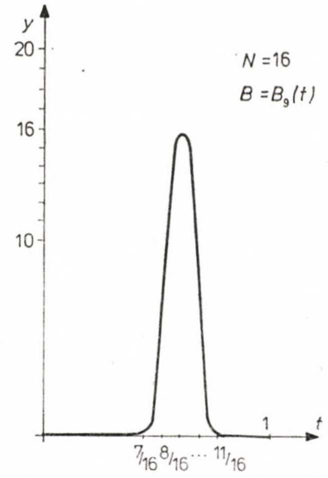

fig. 3.3

and we can compute the coefficients $\{\alpha_k;\ k=1, 2, \ldots, n\}$ from the system of linear equations

$$\sum_{i=1}^{n} \alpha_i (B_i|B_k)_D = (y|B_k)_D \qquad k = 1, 2, \ldots, n \qquad (**)$$

where

$$(h|g)_D := -\int_0^1 Dh(t)\overline{g(t)}\,dt$$

and it is easy to verify that $(\,.\,|\,.\,)_D$ is a scalar product for functions with square integrable second derivatives satisfying the boundary conditions.

Notice that *in each row of the $n \times n$ matrix of this system of linear equations there are at most four non-zero elements,* i.e.

$$(B_i|B_j)_D = 0 \qquad \text{if } |i-j| > 3$$

since $\{B_i;\ i=1, 2, \ldots, n\}$ are B-splines.

Remark. The boundary value problem investigated in this subsection is the simplest that does not have an exact solution.

3.10.3. We shall now give a more useful form of the scalar product $(\,.\,|\,.\,)_D$:

$$(f|g)_D := -\int_0^1 (f''(t) - c(t)f(t))\overline{g(t)}\,dt = \int_0^1 f'(t)\overline{g'(t)}\,dt + \int_0^1 c(t)f(t)\overline{g(t)}\,dt$$

using integration by parts. But the form

$$(f|g)_\delta := \int_0^1 f'(t)\overline{g'(t)}\,\mathrm{d}t + \int_0^1 c(t) f(t)\overline{g(t)}\,\mathrm{d}t \qquad (*)$$

of the scalar product can also be applied to the functions $\{f\colon f(0)=0;\ f' \in \mathrm{L}_0^2[0, 1]\}$.

The completion $\mathscr{H}_\delta$ of $\{f\colon f(0)=0;\ f' \in \mathrm{L}_0^2[0, 1]\}$ in the norm generated by the scalar product $(*)$ is an RKHS. In fact, using the Cauchy inequality in $\mathrm{L}_0^2[0, 1]$,

$$|f(t_0)| = \left|\int_0^{t_0} f'(t)\,\mathrm{d}t\right| \leq \left(\int_0^1 |f'(t)|^2\,\mathrm{d}t\right)^{1/2} \leq \|f\|_\delta$$

if $t_0 \in [0, 1]$ and $f(0)=0$ and hence the evaluation functionals are continuous in $\mathscr{H}_\delta$.

We now have a modified finite-element method for the approximate solution of the differential equation

$$y'' - cy = f \qquad c = c(t) \geqslant 0$$
$$y(0) = y(1) = 0$$

in which we apply $\mathscr{H}_\delta$ (instead of $\mathscr{H}_D$) and have more simple calculations but less accuracy than in § 3.10.2.

Let us consider the functions

$$\{(t - s_k)_+;\ k = 1, 2, \ldots, n\}$$

with $s_{k+1} - s_k = h$ as in § 3.10.1 (figure 3.4). Then

$$L_i(t) = \frac{1}{h}[(t - s_{i-1})_+ - 2(t - s_i)_+ + (t - s_{i+1})_+] \qquad i = 1, 2, \ldots, n-1$$

form a basis for piecewise linear functions with nodes $\{s_i;\ i=1, 2, \ldots, n-1\}$

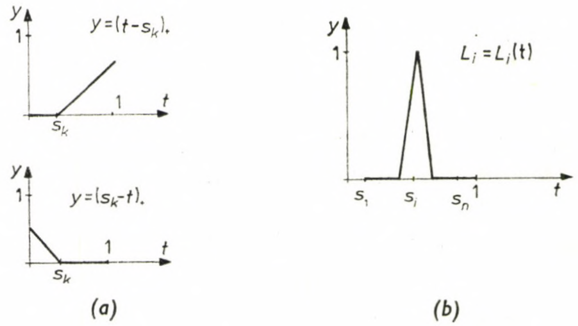

fig. 3.4

satisfying the boundary conditions $y(0)=y(1)=0$ (see § 3.11.17). Moreover,

$$L_i(t) = 0 \qquad \text{if} \quad t \notin [s_i - h,\ s_i + h] \qquad (**)$$

and $L_i = L_i(t)$ is the only piecewise linear function with

$$L_i(s_k) = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{if} \quad i \neq k. \end{cases}$$

Hence, using $\{L_k;\ k=1, 2, \ldots, n-1\}$ as basis functions and the scalar product $(\,.\,|\,.\,)_\delta$ (instead of the B-splines and $(\,.\,|\,.\,)_D$), a simpler system of linear equations will be obtained with less computation. In fact, in computing the coefficients we have to deal with piecewise linear functions instead of cubic splines and

$$(L_i|L_j)_\delta = 0 \qquad \text{if} \quad |i-j| > 1$$

by $(**)$.

The formula 3.10.2 $(*)$, the measure of the approximation, is meaningless for

$$y_s = \sum_{k=1}^{n-1} \alpha_k L_k.$$

However, it is the best approximation in the sense that

$$\|y - y_s\|_\delta$$

is minimal, where $y=y(t)$ is the exact solution. In fact, by 2.4.1.4, this is the case if

$$(y - y_s|L_i)_\delta = 0 \qquad i = 1, 2, \ldots, n-1$$

which means that

$$\sum_{k=1}^{n-1} \alpha_k (L_k|L_i)_\delta = (y|L_i)_\delta \qquad i = 1, 2, \ldots, n-1$$

where

$$(y|L_i)_\delta = \int_0^1 y'(t) L_i'(t)\,\mathrm{d}t + \int_0^1 c(t) y(t) L_i(t)\,\mathrm{d}t = \int_0^1 f(t) L_i(t)\,\mathrm{d}t.$$

The best approximation in $\|\,.\,\|_\delta$ is called a *weak approximate solution.*

Remark. y_s is an approximation of the weak solution in the sense of § 3.9.3. In fact,

$$\left| \int_0^1 y_s(t) D\varphi(t)\,\mathrm{d}t - \int_0^1 f(t)\varphi(t)\,\mathrm{d}t \right|$$

can be as small as we like if $s_{i+1} - s_i = h \to 0$.

3.10.4. The two-dimensional analogue of the boundary value problem in § 3.10.2 is

$$\Delta u - cu = f \qquad (*)$$

$$u(\mathscr{S}) = 0$$

where $\mathscr{S}$ is the boundary of a 'nice' domain $\mathscr{D}$ of the two-dimensional Euclidean space, $c=c(x, y)\geqslant 0$, $f\in L^2(\mathscr{D})$ and

$$\Delta = \frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}.$$

If $\{B_k=B_k(x, y);\ k=1, 2, \ldots, n\}$ are linearly independent functions and

$$u_s = \sum_{k=1}^{n} \alpha_k B_k$$

then u_s is called the best approximation of the solution of the differential equation $(*)$ if

$$\|\Delta u_s - cu_s - f\|_2$$

is minimal. Again, by 2.4.1.4, this is the case if and only if

$$\iint_{\mathscr{D}} (\Delta u_s - cu_s)B_i \,\mathrm{d}x\,\mathrm{d}y = \iint_{\mathscr{D}} fB_i \,\mathrm{d}x\,\mathrm{d}y \qquad i = 1, 2, \ldots, n. \qquad (*)$$

Applying the Green formula (see § 2.6.5) or the considerations in § 3.9.3, based on integration by parts for functions satisfying the boundary condition $u(\mathscr{S})=0$, we have

$$\iint_{\mathscr{D}} \operatorname{grad} u\, \overline{\operatorname{grad} v} \,\mathrm{d}x\,\mathrm{d}y = -\iint_{\mathscr{D}} \bar{v}\Delta u \,\mathrm{d}x\,\mathrm{d}y.$$

Hence using the scalar product

$$(u|v)_\delta := \iint_{\mathscr{D}} \operatorname{grad} u\, \overline{\operatorname{grad} v} \,\mathrm{d}x\,\mathrm{d}y + \iint_{\mathscr{D}} cu\bar{v} \,\mathrm{d}x\,\mathrm{d}y$$

we obtain that $(*)$ has the form

$$\sum_{k=1}^{n} \alpha_k (B_k|B_i)_\delta = (u|B_i)_\delta \qquad i = 1, 2, \ldots, n$$

where

$$(u|B_i)_\delta := \iint_{\mathscr{D}} \operatorname{grad} u \operatorname{grad} B_i \,\mathrm{d}x\,\mathrm{d}y + \iint_{\mathscr{D}} cuB_i \,\mathrm{d}x\,\mathrm{d}y$$

$$= -\iint_{\mathscr{D}} B_i(\Delta u - cu) \,\mathrm{d}x\,\mathrm{d}y = \iint_{\mathscr{D}} fB_i \,\mathrm{d}x\,\mathrm{d}y.$$

It follows that we have the same method for the approximate solution u_s in the form

$$\sum_{k=1}^{n} \alpha_k B_k$$

as we had in § 3.10.3 for the one-dimensional case.

For the base functions $\{B_k\}$ we can choose the series of equidistance points s_{ij}; $i=1, 2, \ldots, n$, $j=1, 2, \ldots, n$ and piecewise linear functions L_{mk} such that

$$L_{mk}(s_{ij}) = \begin{cases} 1 & \text{if} \quad i = m \quad \text{and} \quad k = j \\ 0 & \text{for the remaining nodes.} \end{cases}$$

Remark. The Hilbert space with scalar product $(\,.\,|\,.\,)_\delta$ is the analogue of $\mathscr{H}_\delta$ for functions with two variables, and it is also an RKHS (see Shapiro 1971, § II.4).

3.11 Problems and notes

3.11.1. Let $R=R(s, t)$ be a symmetric positive definite function and $\mathbf{A}$ be the matrix with

$$a_{ik} = R(t_i, t_k) \qquad i, k = 1, 2, \ldots, n$$

where $t_k \in \mathscr{D}$. Show that $R(s, t_k)$; $k=1, 2, \ldots, n$ are linearly independent functions if and only if the inverse matrix $\mathbf{A}^{-1}$ exists.

3.11.2. Is it true that if the kernel $R(\,.\,, t)$ is an analytic function on the simply connected region $\mathscr{D}$ of the complex plane for every fixed $t \in \mathscr{D}$, then every $f(\,.\,) \in \mathscr{H}(R)$ is analytic on $\mathscr{D}$?

3.11.3. Prove that $\{f\colon f'' \in \mathrm{L}^2[0, 1]\}$ is an RKHS if

$$(f|g) := f(0)\overline{g(0)} + f(1)\overline{g(1)} + \int_0^1 f''(t)\overline{g''(t)}\,\mathrm{d}t.$$

What is the kernel?

3.11.4. A set $\mathscr{M}$ of continuous functions is called *equicontinuous* if for every $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that $|f(t) - f(s)| < \varepsilon$ is implied by $|s-t| < \delta(\varepsilon)$ *for every* $f \in \mathscr{M}$, i.e. the pair $\varepsilon, \delta(\varepsilon)$ is common for $\mathscr{M}$.

One of the important features of equicontinuity is that a closed subset $\mathscr{M} \subset \mathrm{C}[0, 1]$ is compact if and only if $\mathscr{M}$ is an equicontinuous set of functions (1.6.2.1).

Show that if R is a continuous kernel in $[0, 1] \times [0, 1]$, then the elements of $\mathscr{H}(R)$ form an equicontinuous subset of C[0, 1]. What is the connection between the uniform norm $\|\cdot\|_\infty$ and the RKHS norm in this case?

∘**3.11.5.** Find the Loéve transform that maps $L^2[0, 1]$ onto the subspace $\{f: f(0)=0\}$ of $\mathscr{H}_D$.

∘**3.11.6.** Prove Theorem 3.2.2.1 by applying the considerations in § 3.3.1.

3.11.7. Using the complex form of the Divergence Theorem in § 3.7.1, compute the kernel of H^2 from the kernel of $A(\mathscr{D})$ in the case of $\mathscr{D} = \{z: |z| < 1\}$ and vice versa.

∘**3.11.8.** Consider the following subset $\mathscr{M}$ of an RKHS $\mathscr{H}(R)$:

$$\mathscr{M} = \{f: f(t_0) = 1;\ t_0 \in \mathscr{D},\ f \in \mathscr{H}(R)\}.$$

What is the element in $\mathscr{M}$ of minimum norm?

∘**3.11.9.** Consider the functions f, analytic on the open unit disc and with

$$f(z_k) = \eta_k \qquad k = 1, 2, \ldots, n$$

where $|z_k| < 1$ and η_k are prescribed values. Choose f such that

$$\frac{1}{2\pi} \int_{-\pi}^{+\pi} |f(e^{it})|^2 \, dt$$

is minimal.

∘**3.11.10.** Let $f = f(z)$ be analytic in the open unit disc with the only possible singular point at $z=0$. Moreover, let $f(e^{it})$ be a continuous function. What is the connection between the Fourier series of $f(e^{it})$ and the Laurent series of $f(z)$ at $z=0$?

3.11.11. If x^* is a continuous linear functional of a Hilbert space $\mathscr{H}$, then there exists $h \in \mathscr{H}$, called the *representative of* x^*, such that

$$x^*(f) = (f|h) \qquad f \in \mathscr{H}$$

by the Riesz Theorem. How can the representative h of a given continuous linear functional x^* be found in an RKHS? (For example, for the case where $\mathscr{D}$ is a bounded closed domain of the real line, the integration is a continuous linear functional if the kernel is continuous. What is the representative of the integration?)

3.11.12. (*a*) $Q(s, t)=f(s)\overline{f(t)}$ is a kernel for any function $f=f(\,.\,)$. Describe $\mathscr{H}(Q)$.

(*b*) Prove that for any kernel $R=R(s, t)$, $f\in\mathscr{H}(R)$ if and only if

$$R(s, t)-f(s)\overline{f(t)}$$

is positive definite.

3.11.13. Let the bounded linear operator T in $\mathscr{H}(R)$ be defined by

$$Tf(t) := \varphi(t)f(t)$$

i.e. T is defined as multiplication with a function φ.

(*a*) Is the function φ necessarily bounded?

(*b*) Does the function φ necessarily belong to $\mathscr{H}(R)$?

3.11.14. Show that in $\mathscr{H}_D$, $\varphi f\in\mathscr{H}_D$ for every $f\in\mathscr{H}_D$ if and only if $\varphi\in\mathscr{H}_D$.

o**3.11.15.** Compute the first and second generalised derivatives of the functions

(*a*) $f(t) = \begin{cases}\cos t & \text{if } t\geqslant 0\\ 0 & \text{otherwise.}\end{cases}$

(*b*) $g(t) = \begin{cases}\sin t & \text{if } t\geqslant 0\\ 0 & \text{otherwise.}\end{cases}$

(c) $h(t) = \begin{cases}at & \text{if } t\geqslant 0\\ 0 & \text{otherwise.}\end{cases}$

(*d*) What are the first and second generalised derivatives of a continuous piecewise linear function in [0, 1] with nodes $0=s_0<s_1<\ldots<s_{n-1}<s_n=1$?

3.11.16. Prove that every continuous function f on [0, 1] with an initial value $f(0)=0$ satisfying the condition 'f is a polynomial of first degree in $(t_k,\ t_{k+1})$; $k=1, 2, \ldots, n$ where

$$0\leqslant t_1 < t_2 <\ldots< t_{n-1} < t_n \leqslant 1\text{'}$$

has the form

$$f(t) = \sum_{k=1}^{n} c_k(t-t_k)_+ .$$

o**3.11.17.** We now give another basis for piecewise linear functions. Let I be the identity operator and $U_h f:=f(t-h)$; then

$$L_i(t) := (I-U_h)(t-s_{i-1})_+ = (t-s_{i-1})_+ -2(t-s_i)_+ +(t-s_{i+1})_+ \qquad i = 1, 2, \ldots, n-1$$

(figure 3.4 (b)) are a basis for the piecewise linear functions with nodes $\{s_k;\ k=0, 1, 2, \ldots, n\}$ and zero at the endpoints $s_0=0$; $s_n=1$. (Prove this!)

Moreover, adding two extra points $s_{-1}=-h$ and $s_{n+1}=1+h$ to $\{s_k;\ k=0, 1, 2, \ldots, n\}$, we see that

$$(I-U_h)(t-s_{i-1})_+ \qquad i = 0, 1, \ldots, n, \quad t\in[0, 1]$$

form a basis for *every* piecewise linear function in [0, 1] with nodes $\{s_k;\ k=0, 1, 2, \ldots, n\}$.

∘**3.11.18.** Prove that

$$(f|g)_D := \int_a^b (f''-cf)g\,dt \qquad c = c(t) \geqslant 0$$

is a scalar product if $f'', g''\in L^2[a, b]$ with $f(a)=f(b)=g(b)=g(a)=0$.

3.11.19. It is easy to show that the evaluation functionals are continuous in W^1 (and consequently also in W^m; $m>1$). In fact,

$$\int_a^t g'(\tau)\overline{g(\tau)}\,d\tau = |g(\tau)|^2-\int_a^t g(\tau)\overline{g'(\tau)}\,d\tau \qquad g\in W^1(a, b)$$

by integration by parts and hence

$$|g(t)|^2 < 2\int_a^t |g'(\tau)\overline{g(\tau)}|\,d\tau \leqslant 2\|g'\|_2\|g\|_2.$$

On the other hand,

$$2\|g'\|_2\|g\|_2 \leqslant \|g\|_2^2+\|g'\|_2^2$$

since

$$\|g\|_2^2+\|g'\|_2^2-2\|g\|_2\|g'\|_2 = (\|g\|_2-\|g'\|_2)^2 \geqslant 0.$$

∘**3.11.20.** Do the cubic splines $S(\,.\,, t_k)$; $k=1, 2, \ldots, n$, where $S(s, t)$ is the kernel of the RKHS $\mathscr{H}(S)$ in Example 3 of § 3.3.1, form a basis for the cubic splines with nodes t_k; $k=1, 2, \ldots, n$? (Compare with § 3.11.16.)

3.11.21. It was shown in § 3.4.2 that the interpolation of $f\in\mathscr{H}(R)$ with a linear combination of

$$R(t, s_k);\ k = 1, 2, \ldots, n$$

is the projection of f onto the subspace generated by $\{R(\,.\,, s_k);\ k=1, 2, \ldots, n\}$. Hence the RKHS with spline kernel, i.e. with kernel $S(t, s)$ which is a cubic spline for any fixed s, is a natural Hilbert space model for spline interpolation.

○**3.11.22.** Show on the basis of § 3.11.17 that

$$B_i(t) = (I-U_h)^4(t-s_{i-1})^3_+ \qquad i = 3, 4, \ldots, n-2$$

where $B_i(t)$ are the B-splines defined in § 3.10.1.

3.11.23. Prove that if the support of $f=f(t)$; $t \in R$ is *bounded below*, i.e. $f(t)=0$ if $t<t_0$, then the right translations

$$f_\alpha(t) := f(t-\alpha) \qquad \alpha > 0$$

are linearly independent.

3.11.24. Prove that

$$\operatorname{supp} B_i = [t_{i-2}, t_{i+2}] \qquad i = 3, 4, \ldots, n-2$$

and hence the B-splines are linearly independent.

3.11.25. The cubic splines with nodes s_k; $k=1, 2, \ldots, n$ form a finite-dimensional linear space. Find the dimension of this space. Do the B-splines of 3.11.22 form a basis for this linear space?

3.11.26. Give the two-dimensional version of § 3.11.19.

3.11.27. If $F=F(\lambda)$ is a non-decreasing function on the real line R and the Stieltjes integral

$$K(t) = \int_{-\infty}^{+\infty} e^{it\lambda}\, dF(\lambda) \qquad (*)$$

exists, then

$$R(t, s) := K(t-s)$$

is a symmetric, positive definite function and hence it is the kernel of an RKHS. In fact,

$$\sum_{k=1}^{n}\sum_{j=1}^{n} t_j t_k K(t_j-t_k) = \sum_{k=1}^{n}\sum_{j=1}^{n} t_j t_k \int_{-\infty}^{+\infty} e^{i(t_j-t_k)\lambda}\, dF(\lambda) = \int_{-\infty}^{+\infty} |t_k e^{it_k}|^2\, dF(\lambda) \geqslant 0$$

for any n-tuples $\{t_k;\ k=1, 2, \ldots, n\}$ of real numbers.

The Bochner–Khinchin Theorem states that for *every* continuous positive definite function in the form

$$K(t, s) = K(t-s)$$

$K=K(t)$ is the Stieltjes integral $(*)$ with a non-decreasing $F=F(\lambda)$; $\lambda \in R$ called the *Fourier–Stieltjes transform* of F.

3.11.28. It is obvious that every spline $s=s(t)$ with nodes in [0, 1] and with $s(0)=s'(0)=0$ can be considered as an element of $\mathscr{H}(S)$ defined in Example 3 of § 3.3.1.

Complete $\{S(\,.\,, t_j);\ j=1, 2, \dots, n\}$ to a basis for every spline with nodes $\{t_j;\ j=1, 2, \dots, n\}$ and contained in $\mathscr{H}(S)$. (See § 3.11.17.)

3.11.29. For $f\in\mathscr{H}(S)$ we can find an interpolation and hence a best approximation in terms of B-splines with given nodes $\{s_k;\ k=1, 2, \dots, n\}$ by solving the system of linear equations

$$f(s_j) = \sum_{k=3}^{n-2} c_k B_k(j) \qquad j = 3, 4, \dots, n-2.$$

Compute the matrix of this system of linear equations and find the connections with the best approximation (in L^2-norm) of $f''\in L^2[0, 1]$ by piecewise polynomials of first degree with nodes $\{s_k;\ k=1, 2, \dots, n\}$.

3.11.30. By adding two extra nodes $s_{n+1}=1+h$, $s_{n+2}=1+2h$ to $\{s_k;\ k=1, 2, \dots, n\}$, with $s_{k+1}-s_k=h$, construct B_{n-1} and B_n.

Prove that the B-splines defined in § 3.10.1 together with B_{n-1} and B_n form a basis for the splines with nodes $\{s_n;\ k=1, 2, \dots, n\}$ and contained in $\mathscr{H}(S)$.

3.11.31. Prove that if $R=R(s, t)$ is a continuous kernel on the finite two-dimensional interval $[a, b]\times[a, b]$, then a quadrature formula can be obtained for

$$\int_0^1 f(t)\,dt \qquad f\in\mathscr{H}(R)$$

such that

$$\int_0^1 R(t, s_i)\,dt = \sum_{k=1}^{n} c_k R(s_k, s_i).$$

On the basis of the results in § 3.5, find a method for computing $\{c_k;\ k=1, 2, \dots, n\}$. Find examples of such quadrature formulae.

4

Operator Theory

4.1 Background from linear algebra

In the same way that the Hilbert space geometry is connected with properties of geometric vector space, as we have seen in Chapter 2, the operator theory of Hilbert spaces is connected with the properties of matrices.

4.1.1. If T is a linear operator of a finite-dimensional Hilbert space, then T can be represented by matrix multiplication by means of an orthonormal basis $\{e_k\}$ of $\mathscr{H}$.

4.1.1.1 Theorem. Let $\mathbf{T}$ be the matrix with entries $t_{ik}=(Te_k|e_i)$ and let $\mathbf{x}, \mathbf{y}$ be column matrices with entries $x_k=(x|e_k)$, $y_k=(Tx|e_k)$, where $x \in \mathscr{H}$; then

$$\mathbf{y} = \mathbf{Tx}.$$

Proof. We have

$$x = \sum_{k=1}^{n} (x|e_k)e_k = \sum_{k=1}^{n} x_k e_k$$

and

$$Te_k = \sum_{i=1}^{n} (Te_k|e_i)e_i = \sum_{i=1}^{n} t_{ik} e_i$$

(see 2.2.3.2). Moreover,

$$Tx = \sum_{k=1} x_k Te_k$$

since T is a *linear* operator. It follows that

$$Tx = \sum_{k=1}^{n} x_k \Big(\sum_{i=1}^{n} t_{ik} e_i\Big) = \sum_{i=1}^{n} \Big(\sum_{k=1}^{n} t_{ik} x_k\Big) e_i$$

and hence

$$y_i = (Tx|e_i) = \sum_{k=1} t_{ik} x_k.$$

With this theorem, many problems of finite-dimensional Hilbert spaces connected with linear mappings may lead to matrix problems.

Similarly we can prove that the mapping $T \to \mathbf{T}$ from the linear operators of the n-dimensional Hilbert space $\mathscr{H}$ onto the set (algebra) of $n \times n$ matrices has the following properties. If $T_1 \to \mathbf{T}_1$ and $T_2 \to \mathbf{T}_2$ then

(i) $T_1 = T_2$ if and only if $\mathbf{T}_1 = \mathbf{T}_2$;

(ii) $\alpha T_1 + \beta T_2 \to \alpha \mathbf{T}_1 + \beta \mathbf{T}_2$, where α and β are scalars;

(iii) $T_1 T_2 \to \mathbf{T}_1 \mathbf{T}_2$;

(iv) $T_1^* \to \mathbf{T}_1^*$.

(v) The inverse operator T^{-1} exists if and only if the inverse matrix $\mathbf{T}^{-1}$ exists, and then

$$T^{-1} \to \mathbf{T}^{-1}.$$

Every linear operator of a finite-dimensional Hilbert space is continuous (see Theorem 1.7.3.1).

4.1.1.2 Definition. λ is called a *regular value* of the operator T if the inverse operator $(\lambda I - T)^{-1}$ exists. Here I is the identity operator (corresponding to the unit matrix $\mathbf{I}$).

If λ is *not* a regular value then it belongs to the *spectrum* $\sigma(T)$ of T.

The most important problems of operator theory lead to the investigation of the spectrum. Obviously, if λ is a regular value of T, then $x = (\lambda I - T)^{-1} f$ is the unique solution of the equation

$$\lambda x - Tx = f \qquad f \in \mathscr{H}.$$

However, for the case $\lambda \in \sigma(T)$ the situation is more complicated.

4.1.1.3 Theorem. If λ belongs to the spectrum of T, then there exists a solution of the equation

$$\lambda x - Tx = \theta \tag{$*$}$$

that is different from θ.

Proof. In this case there is no inverse matrix $(\lambda \mathbf{I} - \mathbf{T})^{-1}$ and hence

$$\det(\lambda \mathbf{I} - \mathbf{T}) = 0.$$

This means that there are solutions $\mathbf{x} \neq \boldsymbol{\theta}$ to the system of homogenous linear equations corresponding to

$$(\lambda \mathbf{I} - \mathbf{T})\mathbf{x} = \boldsymbol{\theta}.$$

It follows from the considerations in the proof of 4.1.1.1 that in this case

$$x = \sum_{k=1}^{n} x_k e_k$$

is a solution of (*), where x_k is the nth element of the column vector **x**.

If there are solutions $x \neq \theta$ for (*), λ is called an *eigenvalue* of T and the solutions are the corresponding *eigenvectors.*

It follows from the considerations in the proof of 4.1.1.3 that in a finite-dimensional Hilbert space every linear operator has an eigenvalue.

Let us introduce the following notation for an operator A:

$$N(A) := \{x\colon\ Ax = \theta\} \qquad R(A) := \{f\colon f = Ax\}.$$

$N(A)$ is called the *null-space* of A and $R(A)$ is called the *range*. For a linear A, $N(A)$ and $R(A)$ are linear subspaces and for a continuous A, $N(A)$ is closed.

In this notation, if λ is an eigenvalue of T, then

$$N(\lambda I - T) \neq \{\theta\}$$

and it is a closed linear space, sometimes called the *eigenspace of* λ.

4.1.1.4 Theorem. If λ is an eigenvalue of T, then the equation

$$\lambda x - Tx = f$$

has solutions if and only if $f \in N(\bar{\lambda} I - T^*)^\perp$, i.e. if f is orthogonal to every element of $N(\bar{\lambda} I - T^*)$.

Proof. If the equation is solvable for $f \in \mathscr{H}$, i.e. $f \in R(\lambda I - T)$, then for every $z \in N(\bar{\lambda} I - T^*)$,

$$(f|z) = (\lambda x - Tx|z) = (x|\bar{\lambda} z - T^* z) = 0$$

and hence $f \in N(\bar{\lambda} I - T^*)^\perp$.

For the converse statement we need to prove

$$R(\lambda I - T)^\perp = N(\bar{\lambda} I - T^*). \qquad (**)$$

But this we can see from the identity

$$(z|\lambda x - Tx) = (\bar{\lambda} z - T^* z|x) \qquad z, x \in \mathscr{H}.$$

Remark. For the proof it is enough to verify (**). In fact, it follows from (**) that

$$R(\lambda I - T)^{\perp\perp} = N(\bar{\lambda} I - T^*)^\perp$$

and, by 2.14.19,

$$R(\lambda I - T)^{\perp\perp} = R(\lambda I - T)$$

since R is a finite-dimensional linear subspace and hence is closed. But

$$R(\lambda I-T) = N(\bar{\lambda} I-T^*)^{\perp}$$

is precisely the first part of the theorem.

4.1.2. The operator T can be represented in an orthonormal basis by a diagonal matrix with only real elements if and only if *T is self-adjoint.*

The operator T can be represented in an orthonormal basis by a diagonal matrix if and only if

$$T^*T = TT^*.$$

T is then called a *normal operator.*

We shall now prove the statement for self-adjoint operators. From Theorem 4.1.1.1 it is clear that the statement is equivalent to the following characterisation of self-adjoint operators by eigenvalues and eigenvectors.

4.1.2.1 Theorem. In a finite-dimensional space T is a self-adjoint operator if and only if every eigenvalue λ of T is real and there is an orthonormal basis of $\mathscr{H}$ formed of eigenvectors of T.

Remark. We have to choose the eigenvectors as a basis of $\mathscr{H}$ and T is represented, by Theorem 4.1.1.1 as a diagonal matrix with real elements.

Proof. If λ is an eigenvalue with eigenvector x, then

$$(Tx|x) = \lambda(x|x)$$

and

$$(x|Tx) = (x|\lambda x) = \bar{\lambda}(x|x).$$

Since $T=T^*$, it follows that $\lambda=\bar{\lambda}$. It is obvious that if $y \in N(\lambda I-T)$ then $Ty \in N(\lambda I-T)$ also. Furthermore, we shall show that

$$y \in N(\lambda I-T)^{\perp} \Rightarrow Ty \in N(\lambda I-T)^{\perp}.$$

In fact, if $x \in N(\lambda I-T)$ and $(y|x)=0$, then

$$-(Ty|x) = (\lambda y-Ty|x) = (y|\lambda x-Tx) = 0.$$

We conclude that T is a self-adjoint operator on the subspace

$$N(\lambda I-T)^{\perp} := N_{\lambda} \ominus \mathscr{H}.$$

(Here and later on we use the abbreviation $N_{\lambda} := N(\lambda I-T)$ and also the more precise notation for the orthogonal complement introduced in § 4.13.41.)

Now let λ_1 be an eigenvalue with eigenspace N_1; then T_1, the operator T restricted to $\mathscr{H}_1 := N_1 \ominus \mathscr{H}$, is a self-adjoint operator of $\mathscr{H}_1$ and hence there is a real eigenvalue λ_2 with eigenspace $N_2 \subset \mathscr{H}_1$.

Again, T_2, the operator T restricted to $\mathscr{H}_2 := N_2 \ominus \mathscr{H}_1$, is a self-adjoint operator of $\mathscr{H}_2$ and hence there is a real eigenvalue with eigenspace $N_3 \subset \mathscr{H}_2$. We continue this procedure until $\mathscr{H}_m := N_m \ominus \mathscr{H}_{m-1}$ is one dimensional.

It follows from this procedure that

$$\mathscr{H} = \bigoplus_{k=1}^{m+1} N_k \qquad m+1 \leqslant n$$

where n is the dimension of $\mathscr{H}$. Now we can *choose an orthonormal basis in each* N_k and the proof is complete. In fact, we have to choose the eigenvectors as a basis of $\mathscr{H}$ and T is represented by Theorem 4.1.1.1 as a diagonal matrix with real elements.

Remark. If the self-adjoint operator T is positive and x is an eigenvector of T, then

$$0 \leqslant (Tx|x) = \lambda(x|x)$$

and hence every eigenvalue λ is non-negative.

Some further properties connected with eigenvalues are as follows.

4.1.2.2 Theorem. If T is a normal operator, i.e. $T^*T = TT^*$, then

$$Tx = \lambda x \quad \text{iff} \quad T^*x = \bar{\lambda}x.$$

Proof. It is easy to check that $\lambda I - T$ is also normal; moreover, for a normal operator A,

$$\|Ax\|^2 = (A^*Ax|x) = (AA^*x|x) = \|A^*x\|^2$$

and hence

$$(\lambda I - T)x = \theta \quad \text{iff} \quad (\bar{\lambda}I - T^*)x = \theta.$$

4.1.2.3 Theorem. If T^{-1} exists and $(\lambda I - T)x = \theta$ for $x \neq \theta$, then $(1/\lambda I - T^{-1})x = \theta$.

Proof. In this case, for $\lambda \neq 0$,

$$(\lambda T^{-1} - I)x = T^{-1}(\lambda I - T)x = \theta.$$

4.1.3. If the operator T is not normal, then it may happen that there is only a single eigenvalue of T.

Example. The operator of a three-dimensional Hilbert space with the matrix representation

$$\begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

in a basis e_1, e_2, e_3 is called a *translation* operator. It is easy to check that the only eigenvalue is $\lambda=0$ and the eigenvectors are the scalar multiples of e_3.

Thus T cannot be represented by a diagonal matrix. However, we shall show that every linear operator T can be represented by a lower triangular matrix in a certain orthonormal basis. Bearing in mind Theorem 4.1.1.1, this statement is equivalent to the next theorem on the invariant subspaces of a linear operator of a finite-dimensional Hilbert space.

4.1.3.1 Definition. The closed linear subspace $\mathscr{M}$ of $\mathscr{H}$ is called an invariant subspace of the operator T if

$$x \in \mathscr{M} \Rightarrow Tx \in \mathscr{M}$$

or, in an abbreviated notation, $T\mathscr{M} \subseteq \mathscr{M}$.

Example 1. The eigenspaces $N(T-\lambda I)$; $\lambda \in \sigma(T)$ are invariant subspaces. In particular, if x is an eigenvector then the scalar multiples of x form a one-dimensional invariant subspace of T.

Example 2. Consider the translation operator represented by the matrix

$$\mathbf{U} = \begin{pmatrix} 0 & 0 & 0 \dots & 0 \\ 1 & 0 & 0 \dots & 0 \\ 0 & 1 & 0 \dots & 0 \\ 0 & 0 & 1 \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

in a basis $e_1, e_2, \dots, e_n$; i.e., for the elements of $\mathbf{U}$,

$$u_{ik} = \begin{cases} 1 & \text{if } i = k+1 \\ 0 & \text{elsewhere.} \end{cases}$$

Then each of the linear subspaces

$$\{e_i, e_{i+1}, \dots, e_n\} \qquad i = 2, 3, \dots, n$$

i.e. the set of linear combinations of the elements in the bracket, form an $(n+1-i)$-dimensional invariant subspace for $\mathbf{U}$.

4.1.3.2 Theorem. There is a chain

$$\mathscr{M}_k \subset \mathscr{M}_{k+1} \qquad k = 1, 2, \dots, n-2$$

of invariant subspaces for every linear operator of an n-dimensional Hilbert space.

Proof. Every matrix $\mathbf{T}$ has at least one eigenvalue since

$$\det(\mathbf{T}-\lambda\mathbf{I})$$

is a polynomial of degree n. Moreover, $N_\lambda := N(\lambda I - T)$ is an invariant subspace, as we saw in Example 1. If N_λ is one-dimensional, then $\mathcal{M}_1 = N_\lambda$. If the dimension of N_λ is higher than one, say m, then N_λ contains m linearly independent eigenvectors e_k; $k=1, 2, \ldots, m$ and $\mathcal{M}_1$, the scalar multiple of e_1, $\mathcal{M}_2$, the linear span of $\{e_1, e_2\}$, $\mathcal{M}_3$, the linear span of $\{e_1, e_2, e_3\}$ etc are the first m members of the chain.

If $\mathcal{M}_m \ominus \mathcal{H}$ is one-dimensional then the proof is complete. If $\mathcal{M}_m \ominus \mathcal{H}$ is higher than one-dimensional, then we can find an invariant subspace $\mathcal{M}'$ such that

$$\mathcal{M}_m \subset \mathcal{M}' \subset \mathcal{H}.$$

In fact, if $(y|x)=0$ for every $x \in \mathcal{M}_m$, then

$$-(T^* y|x) = (\bar{\lambda} y - T^* y|x) = (y|\lambda x - Tx) = 0$$

and hence $\mathcal{M}_m \ominus \mathcal{H}$ is an invariant subspace for T^*. There is an eigenvalue μ of T^* restricted to $\mathcal{M}_m \ominus \mathcal{H}$ and obviously $N_\mu \subset \mathcal{M}_m \ominus \mathcal{H}$ since $\mathcal{M}_m \ominus \mathcal{H}$ is a finite-dimensional Hilbert space in itself and every linear operator of a finite-dimensional space has at least one eigenvalue, as was established at the beginning of the proof.

By the same reasoning as above, the orthogonal complement of N_μ is an invariant subspace of $(T^*)^* = T$. On the other hand,

$$N_\mu^\perp := N_\mu \ominus \mathcal{H} \supset \mathcal{M}_m$$

(see, for example, § 2.14.19). We conclude that $\mathcal{M}' = N_\mu^\perp$.

If $\mathcal{M}_m \ominus \mathcal{M}'$ is one-dimensional, then $\mathcal{M}' = \mathcal{M}_{m+1}$, i.e. they are consecutive members of the chain.

If $\mathcal{M}_m \ominus \mathcal{M}'$ is higher than one-dimensional then we can find an invariant subspace $\mathcal{M}''$ such that

$$\mathcal{M}_m \subset \mathcal{M}'' \subset \mathcal{M}'$$

by the same process, with $\mathcal{H} = \mathcal{M}'$, as we found an invariant subspace between $\mathcal{M}_m$ and $\mathcal{H}$.

Remark 1. The triangular matrix representation is obtained from the chain of invariant subspaces $\mathcal{M}_k$ as follows. Starting from $\mathcal{M}_1$, composed of the scalar multiples of an eigenvector, an orthonormal basis can be formed in every $\mathcal{M}_k$. It is easy to verify that the subspace $\mathcal{M}_k$ is k dimensional and the matrix representation of the operator in this basis is lower triangular.

Remark 2. Reversing the order of the basis, we obtain an upper triangular matrix respresentation.

Remark 3. The triangular matrix representation consists of blocks with particular structure, called *Jordan forms,* but in the next section we shall not need this particular structure and hence its investigation is omitted.

4.2 Uniform operator norm and the Neumann series expansion for inverse operators

In this section we turn to the case of an infinite-dimensional $\mathscr{H}$. The usual norm — sometimes called the uniform norm — for operators and its basic properties were introduced in § 1.4. We now continue these investigations. (Recall that the domain of a bounded linear operator is supposed to be all of $\mathscr{H}$.)

4.2.1. The set of bounded linear operators from one Hilbert space $\mathscr{H}_1$ into another $\mathscr{H}_2$ form a Banach space with the norm

$$\|T\| := \sup \{\|Tx\| : \|x\| = 1\}$$

by 1.4.3.2 and 1.8.15. However, this usual norm is not a Hilbert space norm. We shall show that the parallelogram law fails to be valid and hence, by 2.1.3, our statement will be proved. Let P_1, P_2 be projection operators with $P_1P_2=0$; then P_1+P_2 is also a projection operator and hence

$$\|P_1+P_2\| = \|P_1\| = \|P_2\| = 1$$

by 2.10.3. In this case,

$$\|P_1+P_2\|^2+\|P_1-P_2\|^2 \leqslant 1+2$$

and

$$2\|P_1\|^2+2\|P_2\|^2 = 4.$$

4.2.2. The bounded linear operators of a Hilbert space $\mathscr{H}$ form an algebra $B(\mathscr{H})$. That is, if $T_1, T_2 \in B(\mathscr{H})$ then $\alpha T_1+\beta T_2 \in B(\mathscr{H})$ for any pair α, β of scalars and also $T_1T_2 \in B(\mathscr{H})$. The operator multiplication has the following properties:

(i) $T_1(T_2+T_3) = T_1T_2+T_1T_3$

(ii) $T_1(T_2T_3) = (T_1T_2)T_3$

$T_1, T_2, T_3 \in B(\mathscr{H})$

and we showed in 1.4.3.4 that

$$\|T_1T_2\| \leqslant \|T_1\|\,\|T_2\|.$$

A Banach space with these properties is called a *Banach algebra.*

It is important for the solution of various types of functional equation to find the inverse operator in the form of a particular infinite series called the *Neumann series,* as we saw in Example 2 of § 1.3.2 and in § 1.8.25. We now approach this problem from a different direction.

4.2.2.1 Theorem. If the power series

$$|\alpha_0|+|\alpha_1|\,\|T\|+|\alpha_2|\,\|T\|^2+\ldots+|\alpha_n|\,\|T\|^n+\ldots$$

is convergent for $T\in B(\mathscr{H})$, then the series of bounded operators

$$\alpha_0 I+\alpha_1 T+\alpha_2 T^2+\ldots+\alpha_n T^n+\ldots$$

is also convergent in the Banach space $B(\mathscr{H})$.

Proof. It follows from the properties of the operator norm that for every $\varepsilon>0$,

$$\Big\|\sum_{k=m}^{n} \alpha_k T^k\Big\| \leqslant \sum_{k=m}^{n} |\alpha_k|\,\|T\|^k < \varepsilon$$

if $n, m>N(\varepsilon)$ since

$$\sum_{k=0}^{\infty} |\alpha_k|\,\|T\|^k <\infty.$$

Hence by the Cauchy Convergence Theorem there exists $A\in B(\mathscr{H})$ such that

$$A = \sum_{k=0}^{\infty} \alpha_k T^k \qquad T^0 := I.$$

Example.

$$e^{Bt} := \sum_{k=0}^{\infty} \frac{t^k}{k!} B^k$$

is a bounded linear operator for every real (or complex) t and $B\in B(\mathscr{H})$.

4.2.3. The most important application of the foregoing theorem is the following.

4.2.3.1 Theorem. If $|\lambda|>\|T\|$, then

$$R(\lambda;\, T) := \sum_{k=0}^{\infty} \frac{1}{\lambda^{k+1}} T^k$$

is a bounded operator of $\mathscr{H}$ and

$$R(\lambda;\, T) = (\lambda I-T)^{-1}.$$

Proof. Considering 4.2.2.1, we have only to prove the second part of the theorem.

$$(\lambda I-T)\sum_{k=0}^{n}\frac{1}{\lambda^{k+1}}T^k = \sum_{k=0}^{n}\frac{1}{\lambda^{k+1}}T^k(\lambda I-T)$$

and

$$(\lambda I-T)\sum_{k=0}^{n}\frac{1}{\lambda^{k+1}}T^k = I-\frac{1}{\lambda^{n+1}}T^{n+1} \to I.$$

Example 1. Let $K=K(t, \tau)$ be a continuous function on $[a, b]\times[a, b]$ and

$$\int_a^b\int_a^b |K(t, \tau)|^2\,dt\,d\tau < \infty;$$

find the solution of the integral equation

$$\lambda y(t)-\int_a^b K(t, \tau)y(t)\,d\tau = f(t) \tag{$*$}$$

in $L^2[a, b]$.

Let

$$Ty := \int_a^b K(t, \tau)y(\tau)\,d\tau.$$

Then

$$\|Ty\|^2 = \int_a^b\left|\int_a^b K(t, \tau)y(\tau)\,d\tau\right|^2 dt$$

and

$$\left|\int_a^b K(t, \tau)y(\tau)\,d\tau\right|^2 \leqslant \int_a^b |K(t, \tau)|^2\,d\tau \int_a^b |y(\tau)|^2\,d\tau$$

by the Cauchy inequality; hence

$$\|Ty\|^2 \leqslant \int_a^b\int_a^b |K(t, \tau)|^2\,d\tau\,dt \int_a^b |y(\tau)|^2\,d\tau$$

which means that

$$\|T\|^2 \leqslant \int_a^b\int_a^b |K(t, \tau)|^2\,d\tau\,dt.$$

It follows from Theorem 4.2.3.1 that for

$$|\lambda| > \int_a^b\int_a^b |K(t, \tau)|^2\,d\tau\,dt$$

the Neumann series $R(\lambda;\, T)$ is convergent. In this case,

$$y(t) = R(\lambda;\, T)f(t) = \frac{1}{\lambda}f(t)+\sum_{n=1}^{\infty}\frac{1}{\lambda^{n+1}}\int_a^b K_n(t, \tau)f(\tau)\,d\tau \tag{$**$}$$

where

$$K_1(t, \tau) = K(t, \tau) \qquad K_{n+1}(t, \tau) = \int_a^b K(t, s) K_n(s, \tau)\, ds \qquad n = 1, 2, \ldots.$$

Remark 1. In Example 2 of § 1.3.2 we solved a similar problem for the Banach space C[0, T].

Remark 2. Based on 1.8.25 and 1.8.24 we can also find the solution $y=y(t)$ as the limit of the recursive sequence

$$y_0(t) = f(t) \qquad y_{n+1}(t) = \frac{1}{\lambda} f(t) + \frac{1}{\lambda} \int_a^b K(t, \tau) y_n(\tau)\, d\tau$$

in $L^2[a, b]$.

Example 2. The integral equation (∗) is called the *Volterra equation* if it has a lower triangular kernel, which means that

$$K(t, \tau) = 0 \qquad \text{if} \quad t < \tau.$$

In the case of a Volterra kernel, (∗) will become

$$\lambda y(t) - \int_a^t K(t, \tau) y(\tau)\, d\tau = f(t).$$

It turns out that for a Volterra equation the Neumann series $R(\lambda; T)$ is convergent for every $\lambda \neq 0$.

Here we restrict ourselves to a finite interval $[a, b]$ and continuous $K=K(t, \tau)$; however, our statement is valid for any $[a, b]$ and any square integrable K. That is,

$$K\colon \int_a^b \int_a^b |K(t, \tau)|^2\, dt\, d\tau < \infty.$$

In our case, $|K(t, \tau)| < M$ and, by induction,

$$|K_n(t, \tau)| \leqslant M^{n+1} \frac{(t-\tau)^n}{(n)!} \leqslant M^{n+1} \frac{(b-a)^n}{n!} \qquad t, \tau \in [a, b].$$

Indeed, suppose that

$$|K_{n-1}(t, \tau)| \leqslant M^n \frac{(t-\tau)^{n-1}}{(n-1)!}$$

where $K_0(t, \tau) := K(t, \tau)$; then

$$|K_n(t, \tau)| = \left|\int_\tau^t K(t, s) K_{n-1}(s, \tau)\, ds\right| \leqslant \int_\tau^t MM^n \frac{(s-\tau)^{n-1}}{(n-1)!}\, ds$$

$$= M^{n+1} \frac{(t-\tau)^n}{n!} < M^{n+1} \frac{(b-a)^n}{n!}.$$

It follows that

$$\sum_{n=1}^{\infty} \frac{1}{\lambda^{n+1}} K_n(t, \tau)$$

is uniformly convergent on $[a, b] \times [a, b]$ for every $\lambda \neq 0$ since

$$\left|\sum_{n=1}^{N} \frac{1}{\lambda^{n+1}} K_n(t, \tau)\right| \leqslant \sum_{n=1}^{N} M^{n+1} \frac{(b-a)^n}{n!} < M e^{M(b-a)}$$

and so

$$y(t) = \frac{1}{\lambda} f(t) + \sum_{n=1}^{\infty} \frac{1}{\lambda^{n+1}} \int_a^b K_n(t, \tau) f(\tau)\, d\tau$$

by (**).

Example 3. If

$$(U_\tau f)(t) = f(t-\tau) \qquad \tau > 0$$

then the Neumann series is convergent for $|\lambda| > 1$ since $\|U_\tau\| = 1$ in $L^2[a, b]$. It follows that the solution $x = x(t)$ of the equation

$$\lambda x(t) - x(t-\tau) = f$$

in $L^2[a, b]$ is

$$x(t) = \sum_{k=0}^{\infty} \frac{1}{\lambda^{k+1}} f(t-k\tau)$$

and in recursive form,

$$x_0(t) = f(t) \qquad x_{n+1}(t) = \frac{1}{\lambda} f(t) + \frac{1}{\lambda} x_n(t-\tau)$$

i.e. $\{x_n(t)\}$ tends to the solution in $L^2[a, b]$.

4.3 The spectrum of an operator

4.3.1. The main result of the previous section was that for $|\lambda| > \|T\|$ there is a *bounded inverse* $(\lambda I - T)^{-1}$, everywhere defined in $\mathscr{H}$, and hence a unique solution

$$y = (\lambda I - T)^{-1} f$$

for the equation

$$y - Ty = f \qquad f \in \mathscr{H}$$

in the form of an infinite series or a recursive sequence.

As in the finite-dimensional case, we have the following definition.

4.3.1.1 Definition. λ is called a *regular value* of the linear operator T if there is an everywhere-defined bounded inverse $(\lambda I - T)^{-1}$.

Remark 1. In a finite-dimensional normed space, if $Tx = \theta$ only for $x = \theta$, then T is *onto* and every linear operator is bounded. Hence in the finite-dimensional case, 4.3.1.1 is the same as 4.1.1.2.

Remark 2. Here we do not assume that T is bounded.

4.3.1.2 Definition. If λ is not a regular value then it belongs to the *spectrum* $\sigma(T)$ *of the operator T.*

Contrary to the finite-dimensional case, the spectrum $\sigma(T)$ *does not consist only of eigenvalues.*

Example 1. For the multiplication operator

$$Ty = ty$$

in $L^2[0, 1]$ each $\lambda \in [0, 1]$ belongs to the spectrum. In fact, multiplication by the function $\lambda - t$ is a 1–1 operator, but the inverse, multiplication by $(\lambda - t)^{-1}$, is a bounded operator only if $\lambda \notin [0, 1]$.

Example 2. The spectrum of the integration operator

$$Ty = \int_0^t y(\tau)\,\mathrm{d}\tau$$

in $L^2[0, 1]$ contains the single point $\lambda = 0$, which is not an eigenvalue. In fact, for every $\lambda \neq 0$,

$$(\lambda I - T)y = \lambda y(t) - \int_0^t y(\tau)\,\mathrm{d}\tau$$

is a Volterra equation with kernel

$$K(t, \tau) = \begin{cases} 1 & \text{if } t > \tau \\ 0 & \text{if } t \leqslant \tau \end{cases}$$

and hence, by Example 2 in § 4.2.3, each $\lambda \neq 0$ is a regular value.

If $\lambda=0$, the inverse is the operator of differentiation, an unbounded operator in $L^2[0, 1]$.

Remark. We also have a direct calculation for the spectrum. If $\lambda\neq0$, then for every $f\in L^2[0, 1]$ the unique solution of

$$\lambda y(t)-\int_0^t y(\tau)\,d\tau = f(t) \tag{$*$}$$

via the substitution

$$z(t) = \int_0^t y(\tau)\,d\tau$$

is

$$y(t) = \frac{1}{\lambda}\int_0^t e^{\lambda(t-\tau)}f(\tau)\,d\tau$$

and obviously the operator

$$R(\lambda;\, T)f := \frac{1}{\lambda}\int_0^t e^{\lambda(t-\tau)}f(\tau)\,d\tau$$

is a bounded one.

If $\lambda=0$, then the solution of $(*)$ is defined only for differentiable $f=f(t)$ since $f(0)=0$ and

$$y(t) = \frac{d}{dt}f(t).$$

Example 3. Let $k=k(t)$ be a continuous function and

$$Kf := \frac{1}{2\pi}\int_0^{2\pi} k(t-\tau)f(\tau)\,d\tau$$

for $f\in L^2[0, 2\pi]$. It is easy to check that if the coefficients of the (complex) Fourier series of f are $\{\hat{f}(n);\ n=0, \pm1, \pm2, \ldots\}$ then

$$[Kf]\hat{}\,(n) = \hat{k}(n)\hat{f}(n) \qquad n = 0, \pm1, \pm2, \ldots$$

where $\hat{k}(n)$ is the nth Fourier coefficient of $k=k(t)$.

The mapping

$$f \to \{\hat{f}(n);\ n = 0, \pm1, \pm2, \ldots\}$$

is an isomorphic operator (§ 2.11.1, Example 1) and hence K is unitarily equivalent to the multiplication operator

$$\hat{K}\hat{f}(n) = \hat{k}(n)\hat{f}(n).$$

Hence the operator $\lambda I-K$ has a bounded inverse if and only if the multiplication operator $\lambda\hat{I}-\hat{K}$ has a bounded inverse (see §§ 2.11.2 and 4.13.3).

We conclude, as in Example 1, that the spectrum is

$$\lambda = \{\hat{k}(n);\ n = 0, \pm 1, \pm 2, \ldots\}.$$

However, contrary to the previous examples, the spectrum consists only of eigenvalues in this case. In fact, if $\lambda = \hat{k}(n_0)$, then

$$\lambda e^{-n_0 it} - \frac{1}{2\pi} \int_0^{2\pi} k(t-\tau) e^{-n_0 i\tau}\, d\tau = 0$$

since for the Fourier coefficients,

$$\left(\lambda e^{-n_0 it} - \frac{1}{2\pi} \int_0^{2\pi} k(t-\tau) e^{-n_0 i\tau}\, d\tau\right)^{\wedge} = \begin{cases} \lambda - \hat{k}(n) & \text{if } n = n_0 \\ 0 & \text{if } n \neq n_0. \end{cases}$$

Example 4. Let $k = k(t)$ be a continuous function again and also let

$$\int_{-\infty}^{+\infty} |k(t)|\, dt < \infty$$

be satisfied. Then the convolution of k and any $f \in L^2(-\infty, +\infty)$,

$$K * f := \int_{-\infty}^{+\infty} K(t-\tau) f(\tau)\, d\tau$$

belongs to $L^2(-\infty, +\infty)$ by a well-known theorem of Fourier transform theory and

$$[Kf]^{\wedge}(\omega) = \hat{k}(\omega) \hat{f}(\omega) \qquad (*)$$

where $\hat{}$ denotes the Fourier transform. The mapping $f \to \hat{f}$ is an isomorphic operator (§ 2.11.1, Example 4) and hence the operator K is unitarily equivalent to the multiplication operator defined by $(*)$. By the same reasoning as in Examples 1 and 3, the spectrum is

$$\lambda = \{\hat{k}(\omega);\ -\infty < \omega < +\infty\}$$

but *none of the values λ is an eigenvalue*. In fact, if

$$(\lambda - \hat{k}(\omega)) \hat{f}(\omega) = 0$$

then for any ω, either $\lambda - \hat{k}(\omega) = 0$ or $\hat{f}(\omega) = 0$. If $f \neq \theta$ then $\hat{f}(\omega) = 0$ only at a finite number of points $\{\omega_k\}$ for every finite interval. But it is impossible that

$$\hat{k}(\omega) = \lambda (\text{constant})$$

except at most countable $\{\omega_k\}$, by Fourier transform theory.

Example 5. Let us consider the (right) shift operator U_r in $l^2[0, \infty]$. If $x=\{x_k;\ k=1, 2, \ldots\}$ then

$$U_r x := \{0, x_1, x_2, \ldots, x_{k-1}, \ldots\}.$$

If $|\lambda|>1$, then λ is a regular value since $\|U_r\|=1$. It follows that there is a unique solution of the equation

$$(\lambda I - U_r)x = f \tag{$*$}$$

for every $f \in l^2$ in the form of a Neumann series if $|\lambda|>1$.

More particularly, by direct computation we can check that $(*)$ is equivalent to

$$\lambda x(n) - x(n-1) = f(n) \qquad n = 1, 2, \ldots.$$

That is,

$$\begin{aligned} &\lambda x(1) - x(0) = f(1) \\ &\lambda x(2) - x(1) = f(2) \\ &\vdots \\ &\lambda x(k) - x(k-1) = f(k) \\ &\vdots \end{aligned}$$

and if the initial condition is $x(0)=0$, then

$$x(1) = \frac{1}{\lambda} f(1)$$

$$x(2) = \frac{1}{\lambda}(f(2)+x(1)) = \frac{1}{\lambda} f(2) + \frac{1}{\lambda^2} f(1)$$

$$x(3) = \frac{1}{\lambda}(f(3)+x(2)) = \frac{1}{\lambda} f(3) + \frac{1}{\lambda^2} f(2) + \frac{1}{\lambda^3} f(1)$$

$$\vdots$$

$$x(k) = \frac{1}{\lambda} f(k) + \frac{1}{\lambda^2} f(k-1) + \ldots + \frac{1}{\lambda^{k-1}} f(2) + \frac{1}{\lambda^k} f(1).$$

Thus we have obtained the solution of $(*)$ in an explicit form for $|\lambda|>1$. However, from this form of $(\lambda I - U_r)^{-1} f$ we cannot see immediately that $(\lambda I - U_r)^{-1}$ is a bounded operator for $|\lambda|>1$.

The result of the direct computation is also valid for $|\lambda|<1$. In this case,

$$\lambda^{k+1} x(k) = \lambda^k f(k) + \lambda^{k-1} f(k-1) + \ldots + \lambda f(1)$$

and the right-hand side tends to the scalar product of $\{\lambda^k\}$ and $\{f(k)\}$, while the left-hand side tends to zero for every $\{x(k)\} \in l^2$.

We conclude that the range of $\lambda I - U_r$ is orthogonal to $\{\lambda^k;\ k=1, 2, \ldots\} \in l^2$ for $|\lambda| < 1$. In other words, $(*)$ has a solution if and only if $\{f(k)\}$ is orthogonal to $\{\lambda^k\}$; moreover, $\{\lambda^k\}$ is the solution of

$$(\lambda I - U_r^*)x = \theta$$

where

$$U_r^* x := \{x_2, x_3, \ldots, x_{k+1}, \ldots\}.$$

If $\lambda = 0$, then there is a solution if and only if $f(1)=0$. Hence the range of U_r is not dense, either.

For $\lambda = \pm 1$ the situation is quite different. It can be shown that the range of $I \pm U_r$ is dense in l^2 and $I \pm U_r$ is 1–1; however, $(I \pm U_r)^{-1}$ is unbounded. (For more details see 4.13.43.)

To summarise, there exists an inverse $(\lambda I - U_r)^{-1} \in B(\mathscr{H})$ if and only if $|\lambda| > 1$ and hence every λ: $|\lambda| \leqslant 1$ belongs to the spectrum of U_r.

Example 6. For the spectrum of an unbounded operator let us consider the operator

$$Dy := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\, y$$

from $\{y\colon y'' \in \mathrm{L}^2[0, 1];\ y(0) = y(1) = 0\} \subset \mathrm{L}^2[0, 1]$ into $\mathrm{L}^2[0, 1]$. It is easy to verify that

$$\lambda y - Dy = \theta$$

for $y \neq \theta$ if and only if $\lambda = -k^2\pi^2$ $(k=0, 1, 2, \ldots)$ and for $\lambda \neq k^2\pi^2$ the equation

$$\lambda y - Dy = f \qquad f \in \mathrm{L}^2[0, 1]$$

has the unique solution

$$y(t) = -\frac{\sin \lambda^{1/2} t}{\sin \lambda^{1/2}} \int_0^1 \sin \lambda^{1/2}(1-\tau) f(\tau)\, \mathrm{d}\tau + \int_0^t \sin \lambda^{1/2}(t-\tau) f(\tau)\, \mathrm{d}\tau.$$

Thus we conclude that the spectrum of D consists of eigenvalues $\{k^2\pi;\ k=0, 1, 2, \ldots\}$ and every $\lambda \neq k^2\pi^2$ is a regular value, since the operator that sends f into y is an integral operator with kernel

$$K(t, \tau) = \begin{cases} -\dfrac{\sin \lambda^{1/2} t}{\sin \lambda^{1/2}} \sin \lambda^{1/2}(1-\tau) + \sin \lambda^{1/2}(t-\tau) & \text{for} \quad \tau < t \\[2ex] -\dfrac{\sin \lambda^{1/2} t}{\sin \lambda^{1/2}} \sin \lambda^{1/2}(1-\tau) & \text{for} \quad \tau \geqslant t \end{cases}$$

and this is a bounded operator by Example 2 in § 4.2.3.

4.3.2. The possible situations in which the linear operator $\lambda I - T$ might not have an inverse $(\lambda I - T)^{-1} \in B(\mathscr{H})$ are as follows:

(*a*) $\lambda I - T$ is not 1–1;

(*b*) $\lambda I - T$ is 1–1 but its range is not dense;

(*c*) $\lambda I - T$ is 1–1 with dense range but the inverse operator is not continuous.

Parts (*a*), (*b*) and (*c*) of the spectrum are pairwise disjoint subsets.

4.3.2.1 Definition. (*a*) The *point spectrum* of a linear operator T is the subset of all λ for which the operator $I - T$ is *not* 1–1, i.e. the elements of the point spectrum are exactly the eigenvalues.

(*b*) The *residual spectrum* of a linear operator T is the subset of all λ for which the operator $\lambda I - T$ is 1–1 but the range is not dense in $\mathscr{H}$, i.e.

$$\overline{(\lambda I - T)\mathscr{H}} \neq \mathscr{H}.$$

(*c*) The *continuous spectrum* of a linear operator T is the subset of all λ for which the operator $\lambda I - T$ is 1–1 and has its range dense in $\mathscr{H}$, i.e.

$$\overline{(\lambda I - T)\mathscr{H}} = \mathscr{H}$$

but the inverse $(\lambda I - T)^{-1}$ is *not continuous*.

We gave examples of each kind of spectrum in § 4.3.1.

4.3.2.2 Theorem. The spectrum $\sigma(T)$ of a *bounded* linear operator T is a bounded closed subset of the complex plane and for $\lambda \in \sigma(T)$,

$$|\lambda| \leqslant \limsup_n \|T^n\|^{1/n} \leqslant \|T\|.$$

Proof. We shall show that the set of regular values, the complement of the spectrum, is open. If $\lambda \notin \sigma(T)$ then

$$(\lambda + \mu) I - T = (\lambda I - T)[I + \mu(\lambda I - T)^{-1}] \qquad (*)$$

and it follows from 4.2.3.1 that $I + \mu(\lambda I - T)^{-1}$ has a bounded inverse if

$$|\mu| < \|(\lambda I - T)^{-1}\|^{-1}.$$

In this case there is also a bounded inverse for $(\lambda + \mu) I - T$ by $(*)$.

A more precise estimate of the radius of convergence of

$$R(\lambda;\, T) := \sum_{k=0}^{\infty} \frac{1}{\lambda^{k+1}} T^k$$

is

$$\|R(\lambda;\, T)\| \leqslant \sum_{k=0}^{\infty} \frac{1}{|\lambda|^{k+1}} \|T^k\| = \sum_{k=0}^{\infty} a_{k+1} x^{k+1}$$

with $x=1/|\lambda|$ and $a_k=\|T^{k-1}\|$; $k=1, 2, \ldots$. It follows from the elementary theory of power series that this is convergent for

$$|\lambda| > \limsup_n \|T^n\|^{1/n}.$$

4.4 Operators with finite-dimensional range

As we saw in § 4.1, the linear operators of a finite-dimensional space have a standard form of matrix multiplication. Similarly, the bounded linear operators T in a Hilbert space $\mathscr{H}$ with the property that the range

$$\{Tx;\; x\in\mathscr{H}\}$$

is finite dimensional have a common standard form.

4.4.1. We begin with some important examples.

Example 1. Let h_k, g_k; $k=1, 2, \ldots, N$ be piecewise continuous functions in the closed finite interval $[a, b]$ and

$$k(t, \tau) = \sum_{k=1}^{N} h_k(t) g_k(\tau).$$

Then the integral operator

$$Kf := \int_a^b k(t, \tau) f(\tau)\, d\tau$$

has an N-dimensional range. In this case $K=K(t, \tau)$ is called a *degenerate kernel.*

Example 2. Let $\{e_k;\; k=1, 2, \ldots\}$ be an orthonormal system in a scalar product space $\mathscr{H}$. Then the operator

$$F_N f := \sum_{k=1}^{N} (f|e_k) e_k \qquad f\in\mathscr{H}$$

has an N-dimensional range.

Example 3. In a reproducing kernel Hilbert space $\mathscr{H}$, if we order the sample $\{f(s_k);\; k=1, 2, \ldots, n\}$ to every $f\in\mathscr{H}$, then a bounded linear operator is obtained with n-dimensional range.

The common structure of the operators in these examples is as follows. Let $\mathscr{H}_1, \mathscr{H}_2$ be Hilbert spaces and let

$$\{a_k \in \mathscr{H}_1;\ k = 1, 2, \ldots, N\} \qquad \{b_k \in \mathscr{H}_2;\ k = 1, 2, \ldots, N\}$$

be linearly independent elements. Define

$$\Big(\sum_{k=1}^{N} a_k \otimes b_k\Big) f := \sum_{k=1}^{N} (f \mid a_k) b_k \qquad f \in \mathscr{H}_1.$$

Then

$$\sum_{k=1}^{N} a_k \otimes b_k$$

is a bounded linear operator from $\mathscr{H}_1$ into $\mathscr{H}_2$ with finite-dimensional range.

We shall show that this is a common representation for *every* operator with finite-dimensional range.

4.4.1.1 Theorem. If T is a bounded linear operator from $\mathscr{H}_1$ into $\mathscr{H}_2$ with finite-dimensional range, then T has the form

$$\sum_{k=1}^{N} a_k \otimes b_k \qquad a_k \in \mathscr{H}_1, \quad b_k \in \mathscr{H}_2.$$

Remark. This representation is not unique, as can be seen from the following proof.

Proof. If T has a finite-dimensional range and $\{b_k;\ k=1, 2, \ldots, N\}$ is an orthonormal basis in the range of T, then

$$Tf = \sum_{k=1}^{N} \alpha_k b_k \qquad f \in \mathscr{H}_1$$

where $\alpha_k = (Tf \mid b_k)$ are continuous linear functionals of $\mathscr{H}_1$ since T is a continuous linear operator. Hence by the Riesz–Fréchet Theorem (2.8.1.1) there is a unique $a_k \in \mathscr{H}_1$ such that

$$\alpha_k = (Tf \mid b_k) = (f \mid a_k).$$

Remark. In §§ 4.7 and 4.8 we shall define operators in the form of *infinite sum*

$$\sum_{k=1}^{\infty} \lambda_k a_k \otimes b_k$$

with $\|a_k\| = \|b_k\| = 1$ and $\lambda_k \to 0$ as important classes of operators.

4.4.2. Let us assume that $\lambda \neq 0$ and $f \in H$ are given; our next task object is then to solve the equation

$$\lambda y - Ty = f \tag{$*$}$$

in the case of T with finite-dimensional range. If

$$T = \sum_{k=1}^{N} a_k \otimes b_k$$

where $\{a_k;\ k=1, 2, \ldots, N\}$ and $\{b_k;\ k=1, 2, \ldots, N\}$ are linearly independent vectors of the Hilbert space $\mathscr{H}$, then equation $(*)$ has the form

$$\lambda y - \sum_{k=1}^{N} (y|a_k) b_k = f \tag{$**$}$$

and hence

$$y = \frac{1}{\lambda} f + \sum_{k=1}^{N} \eta_k b_k$$

and we have to find $\{\eta_k;\ k=1, 2, \ldots, N\}$ such that $(*)$ is satisfied.

Substituting this form (of y) into $(**)$, we obtain

$$f + \lambda \sum_{k=1}^{N} \eta_k b_k - \sum_{k=1}^{N} \left(\frac{1}{\lambda} f + \sum_{i=1}^{N} \eta_i b_i \middle| a_k \right) b_k = f$$

and hence

$$\lambda \eta_k - \sum_{i=1}^{N} (b_i|a_k) \eta_i = \frac{1}{\lambda} (f|a_k) \qquad k = 1, 2, \ldots, N, \tag{$***$}$$

since $\{b_k;\ k=1, 2, \ldots, N\}$ are linearly independent vectors.

It follows from these considerations that $(***)$ and $(*)$ are equivalent equations in the sense that $y \in \mathscr{H}$ is a solution of $(*)$ if and only if

$$y = \frac{1}{\lambda} f + \sum_{k=1}^{N} \eta_k b_k \qquad \lambda \neq 0$$

where $\{\eta_k;\ k=1, 2, \ldots, N\}$ is a solution of $(***)$.

We have also obtained the counterparts of 4.1.1.3 and 4.1.1.4:

4.4.2.1 Theorem. If $\lambda \neq 0$ belongs to the spectrum $\sigma(T)$ of T, then λ is an eigenvalue. T has at most N different eigenvalues.

$\lambda = 0$ always belongs to the spectrum, since the finite-dimensional range of T cannot be dense in the infinite-dimensional $\mathscr{H}$.

If λ is an eigenvalue, then $(*)$ has a solution if and only if $f \in N(\bar{\lambda} I - T^*)^{\perp}$.

Remark. For the proof we have to take into consideration also 1.8.21 and that 4.1.1.4 is valid for every linear operator with finite-dimensional range.

4.4.3. It is natural to ask how we can generalise the matrix representation of the finite-dimensional operators to the case of an infinite-dimensional Hilbert space.

Let T be a linear operator of the infinite-dimensional separable Hilbert space $\mathscr{H}$, let $\{e_k\}$ be an orthonormal basis in $\mathscr{H}$ and

$$a_{ik} = (Te_i|e_k) \qquad x_k = (x|e_k).$$

Then the matrix representation of T is the matrix multiplication

$$\begin{pmatrix} a_{11} & a_{12} \dots & a_{1k} & \dots \\ a_{21} & a_{22} \dots & a_{2k} & \dots \\ \vdots & & & \\ a_{i1} & a_{i2} \dots & a_{ik} & \dots \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ \vdots \end{pmatrix} \tag{$*$}$$

and this means that Ax is equal to the unique $y \in \mathscr{H}$ satisfying

$$(y|e_k) = \sum_{j=1}^{\infty} a_{kj} x_j.$$

Example 1. If U_r is the shift operator given in Example 5 of § 4.3.1 and $\{e_k\}$ is the standard orthonormal basis then

$$(U_r e_i|e_k) = \begin{cases} 1 & \text{if } k = i+1 \\ 0 & \text{if } k \neq i+1. \end{cases}$$

Hence the matrix representation of U_r is the matrix multiplication

$$\begin{pmatrix} 0 & 0 & 0 \dots \\ 1 & 0 & 0 \dots \\ 0 & 1 & 0 \dots \\ \vdots & & \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{pmatrix}.$$

Example 2. Let K be an integral operator on $L^2(a, b)$ with kernel $k \in L^2([a, b] \times [a, b])$. Then

$$a_{ik} = (Ke_i|e_k) = \int_a^b \int_a^b k(t, \tau) e_i(\tau) e_k(t)\, d\tau\, dt$$

and a_{ik} are the Fourier coefficients of $k = k(t, \tau)$ with respect to the orthonormal basis in $L^2([a, b] \times [a, b])$,

$$\psi_{ik}(t, \tau) = e_i(\tau) e_k(t)$$

(see 2.14.47). It follows that the matrix representation has the property

$$\sum_{k=1}^{\infty} \sum_{i=1}^{\infty} |a_{ik}|^2 = \int_a^b \int_a^b |K(t, \tau)|^2\, d\tau\, dt < \infty.$$

4.5 Compact operators

Several operator-theoretic properties remain valid if we turn to the uniform limit of operators with finite-dimensional range. These operators will be called *compact operators.*

4.5.1. The compact operators of a Hilbert space $\mathscr{H}$ form a closed subalgebra of $B(\mathscr{H})$. In more detail, this statement has the following meaning.

4.5.1.1 Theorem. (*a*) If T_1, T_2 are compact operators then T_1T_2 and $\alpha T_1+\beta T_2$ ($\alpha, \beta\in \Phi$) are also compact operators.

(*b*) If $\{T_n\}$ is a sequence of compact linear operators and $\|T_n-T\|\to 0$ then T is also compact.

Proof. (*a*) If A is an operator with finite-dimensional range then AL and LA also have finite-dimensional range for every bounded operator L. In fact, by 4.4.1.1,

$$A = \sum_{k=1}^{N} a_k \otimes b_k$$

and hence, for every $f\in\mathscr{H}$,

$$LAf = L \sum_{k=1}^{N} (f|a_k)b_k = \sum_{k=1}^{N} (f|a_k)Lb_k := \Big(\sum_{k=1}^{N} a_k\otimes Lb_k\Big)f.$$

If T is compact, then there is a sequence $\{T_n\}$ of operators with finite-dimensional range tending to T and

$$\|LT_n-LT\| \leqslant \|L\|\,\|T_n-T\|.$$

Hence LT is compact. A similar argument shows that TL is compact. It is obvious that if T_1, T_2 are compact, so also is $\alpha T_1+\beta T_2$.

Part (*b*) of the theorem follows from the fact that the 'double' closure of a subspace is identical to the closure. More precisely, for every $\varepsilon>0$ there exists $N=N(\varepsilon/2)$ such that $\|T_n-T\|<\varepsilon/2$ if $n>N(\varepsilon/2)$. But for every T_n there is an operator A_{nm} with finite-dimensional range such that $\|A_{nm}-T_n\|<\varepsilon/2$ if $m>M(\varepsilon/2, n)$; hence $\|A_{nm}-T\|<\varepsilon$. It follows that T is the (uniform) limit of operators A_{nm} with finite-dimensional range and hence T is compact.

Remark. A subalgebra A of $B(\mathscr{H})$ with the property

$$T\in A,\ L\in B(\mathscr{H}) \Rightarrow TL\in A \quad \text{and} \quad LT\in A$$

is called a (two-sided) *ideal* of $B(\mathscr{H})$. Thus we have also proved that the compact operators form not only a subalgebra but a two-sided ideal of $B(\mathscr{H})$.

If T is a bounded linear operator with finite-dimensional range, then T sends a bounded set into a pre-compact set. Indeed, $\{Tx;\ x\in\mathscr{H}\}$ is finite-dimensional and in a finite-dimensional subspace every bounded set is pre-compact, by 1.6.2.4, 1.7.1.3 and 1.7.2. We shall show that this is a characteristic property of compact operators.

4.5.1.2 Theorem. If T is a compact operator, i.e. the uniform limit of operators with finite-dimensional range, and $B\subset\mathscr{H}$ is a bounded set, then $\{Tx;\ x\in B\}$ is pre-compact.
Proof. We have to prove that if $\|x_n\|\leq 1$ then there is a convergent subsequence of $\{Tx_n\}$.

Let $\{T_k\}$ be a sequence of operators with finite-dimensional range and $T_k\to T$. Then there is a subsequence $\{x_n^{(1)}\}$ such that $\{T_1 x_n^{(1)}\}$ is convergent. Similarly, there is a subsequence $\{x_n^{(2)}\}$ of $\{x_n^{(1)}\}$ such that $\{T_2 x_n^{(2)}\}$ is convergent. Continuing recursively, we have a subsequence $\{x_n^{(k)}\}$ of $\{x_n^{(k-1)}\}$ such that $\{T_m x_n^{(k)}\}$ is convergent for $m=1, 2, ..., k$.

The $\{x_n^{(k)};\ n, k=1, 2, ...\}$ may be arranged in a rectangular array. Consider the 'diagonal sequence'

$$x_1^{(1)}, x_2^{(2)}, ..., x_k^{(k)}, ...$$

For each k, the sequence $\{x_n^{(n)};\ n=k, k+1, ...\}$ is a subsequence of the kth row $\{x_n^{(k)};\ n=1, 2, ...\}$ and hence $\{T_k x_n^{(n)}\}$ is convergent; let us say $\lim_{n\to\infty} T_k x_n^{(n)}=x^{(k)}$. We shall show that $Tx_n^{(n)}$ is a Cauchy sequence and thus the proof is complete.

For all m and n,

$$\|Tx_m^{(m)}-Tx_n^{(n)}\| \leq \|(T-T_k)x_m^{(m)}\|+\|T_k x_m^{(m)}-T_k x_n^{(n)}\|+\|T_k-T)x_n^{(n)}\|.$$

Since $T_k\to T$ and $\{T_k x_n^{(n)}\}$ is convergent for every k, it follows from a standard estimation that the right-hand side of the above inequality is less then any $\varepsilon>0$ if $n, m>N(\varepsilon)$.

Remark. The usual definition of compact operators is as follows.

The linear operator T is called compact if the range $\{Tx;\ x\in B\}$ of any bounded set B is pre-compact.

Later, in § 4.7.2, we shall show that the operators in a Hilbert space that send bounded sets into pre-compact sets are exactly the uniform limit of operators with finite-dimensional range.

4.5.2. Several linear operators, connected with the solution of important differential and integral equations, are the uniform limit of operators with finite-dimensional range, i.e. compact operators of a Hilbert space.

Example 1. The integral operator

$$Kx := \int_a^b K(t, \tau)x(\tau)\,d\tau$$

in $L^2[a, b]$ is compact if

$$\int_a^b \int_a^b |K(t, \tau)|^2\,d\tau\,dt < \infty.$$

In fact, if $\{e_k(t)\}$ is a complete orthonormal system in $L^2[a, b]$, then $e_k(t)e_i(\tau)$; $k=1, 2, \ldots,\ i=1, 2, \ldots$ is a complete orthonormal system in $L^2([a, b]\times[a, b])$ (2.14.47). Hence if

$$S_n(t, \tau) = \sum_{k=1}^{n} \sum_{i=1}^{n} c_{ik} e_i(t) e_k(t)$$

then

$$\int_a^b \int_a^b |S_n(t, \tau) - K(t, \tau)|^2\,dt\,d\tau \to 0$$

if c_{ik} are the Fourier coefficients with respect to the orthonormal system $\{e_k(t)e_i(\tau);\ k=1, 2, \ldots,\ i=1, 2, \ldots\}$. That is,

$$c_{ik} = \int_a^b \int_a^b K(t, \tau)e_k(\tau)e_i(t)\,d\tau\,dt.$$

Moreover,

$$\|Kx\|_2^2 := \int_a^b \Big|\int_a^b K(t, \tau)x(\tau)\,d\tau\Big|^2\,dt$$

$$\leqslant \int_a^b \Big(\int_a^b |K(t, \tau)|^2\,d\tau \int_a^b |x(\tau)|^2\,d\tau\Big)\,dt$$

$$= \int_a^b \int_a^b |K(t, \tau)|^2\,d\tau\,dt \int_a^b |x(\tau)|^2\,d\tau$$

and so

$$\|S_n - K\|^2 \leqslant \int_a^b \int_a^b |S_n(t, \tau) - K(t, \tau)|^2\,dt\,d\tau.$$

On the other hand, the integral operators with kernel $S_n(t, \tau)$ are operators with finite-dimensional range, by Example 1 in § 4.4.1.

Example 2. Let us consider the functions on [0, 1],

$$\mathscr{H} = \{y:\ y'' \in L^2[0, 1];\ y(0) = y(1) = 0\}$$

and the differential operator $Dy := y''$ from $\mathscr{H}$ to $L^2[0, 1]$. The inverse operator D^{-1} is compact. In fact, if

$$G(t,\tau)=\begin{cases}(t-1)\tau & 0\leqslant\tau<t\\(\tau-1)t & t\leqslant\tau\leqslant 1\end{cases}$$

then it is easily verified that the integral operator

$$Gf:=\int_0^1 G(t,\tau)f(\tau)\,\mathrm{d}\tau$$

is the inverse D^{-1} and

$$\int_0^1\int_0^1 G(t,\tau)^2\,\mathrm{d}t\,\mathrm{d}\tau<\infty.$$

Hence G is compact, by Example 1.

Example 3. Let us consider a Sturm–Liouville differential operator

$$Dy=\frac{\mathrm{d}}{\mathrm{d}t}(a(t)y')+c(t)y$$

with 'smooth' $a(t)$, $c(t)$ and let $G(t,\tau)$ be the Green function of D. A slight generalisation of Example 2 gives us a continuous Green function $G(t,\tau)$ of D, i.e. the kernel of D^{-1}. Then, by Example 1, the operator

$$Gf:=\int_a^b G(t,\tau)f(\tau)\,\mathrm{d}\tau$$

is a compact operator since

$$\int_a^b\int_a^b G(t,\tau)^2\,\mathrm{d}t\,\mathrm{d}\tau<\infty$$

in this case.

Remark. The *Green function* $G(t,\tau)$ of D is a continuous function satisfying the equation $DG(t,\tau)=0$ if $\tau\neq t$ for any fixed t with certain boundary conditions. It can be proved that the integral operator with kernel $G(t,\tau)$ is the inverse operator of D (Gohberg and Goldberg 1981, § IV.5).

Example 4. Let us consider the *Green function* G of the Dirichlet problem

$$\frac{\partial^2u}{\partial x^2}+\frac{\partial^2u}{\partial t^2}=0 \qquad u(S)=g$$

where S is the boundary of an appropriate bounded domain of the (t,x)

plane. Then the integral operator with kernel G is compact (Riesz and Sz Nagy 1955, Nr 81).

Example 5. Let $h=h(t)$ be a real-valued continuous function on [0, 1]. Then the multiplication operator

$$T_h f := h(t)\cdot f(t)$$

is a bounded but *non-compact* operator in $L^2[0, 1]$. In fact, there then exists a subinterval I such that $|h(t)|>\alpha>0$ for $t\in I$. For any orthonormal system $\{e_k\}$ in $L^2(\bar{I})$,

$$\int_0^1 |h(t)e_k(t)-h(t)e_m(t)|^2\,dt > \alpha^2 \int_I |e_k(t)-e_m(t)|^2\,dt \geqslant 2\alpha^2$$

$$k, m = 1, 2, \ldots .$$

It follows that the range $\{T_h e_k;\ k=1, 2, \ldots\}$ of the bounded set $\{e_k;\ k=1, 2, \ldots\}$ is not pre-compact.

Example 6. It follows from the previous example that the convolution operator

$$Kf := \int_0^t K(t-\tau)f(\tau)\,d\tau$$

is not compact in $L^2[0, \infty)$ for any

$$k:\ \int_0^\infty |k(t)|\,dt < \infty .$$

Indeed, by the convolution theorem,

$$\Big(\int_0^t k(t-\tau)f(\tau)\,d\tau\Big)^{\wedge}(\omega) = \hat{k}(\omega)\hat{f}(\omega)$$

where $\hat{}$ denotes the Fourier transform, which is an isomorphic operator (§ 2.11.1, Example 4).

Example 7. The translation operator

$$[U_\tau f](t) = f(t-\tau) \qquad \tau > 0$$

is not compact in $L^2[0, \infty)$. In fact,

$$(U_\tau f|U_\tau g) = (f|g) \qquad f, g\in L^2[0, \infty)$$

and hence $\{U_\tau e_k;\ k=1, 2, \ldots\}$ is an orthonormal system for any orthonormal system $\{e_k;\ k=1, 2, \ldots\}$. But any orthonormal system is a bounded non-compact set since

$$\|e_k\| = 1 \qquad \|e_k - e_m\| = 2 \qquad k, m = 1, 2, \ldots.$$

Example 8. The identity operator I is a non-compact operator in an infinite-dimensional Hilbert space (consider the range of a bounded non-compact subset).

It follows from Example 8 that an operator T and its inverse T^{-1} cannot be compact at the same time in an infinite-dimensional Hilbert space. In fact, if both T and T^{-1} are compact operators, then $T^{-1}T = TT^{-1} = I$ is also compact, by 4.5.1.1 (*a*) which is impossible according to Example 8. Thus we have also obtained the following.

4.5.2.1 Theorem. $\lambda = 0$ belongs to the continuous spectrum of every compact operator T.

4.5.2.2 Theorem. If T is an operator in a Hilbert space with compact inverse T^{-1}, then T is necessarily an unbounded operator.

We have seen these phenomena in Examples 2–4; later, in § 4.13.11, we shall use them for the investigation of spectral properties of unbounded operators.

4.5.3. We now consider the operator equation

$$\lambda y - Ty = f \qquad \lambda \neq 0 \tag{$*$}$$

in a Hilbert space $\mathscr{H}$ for compact linear T.

Let A be an operator with finite-dimensional range such that $\|T-A\| < |\lambda|$ and $B = T - A$; then there exists $(\lambda I - B)^{-1}$ and

$$(\lambda I - B)^{-1} = \frac{1}{\lambda I} + \sum_{k=1}^{\infty} \frac{1}{\lambda^{k+1}} B^k$$

by 4.2.3.1; moreover, $\lambda I - T = (\lambda I - B) - A$ and hence

$$(\lambda I - T)(\lambda I - B)^{-1} = I - A(\lambda I - B)^{-1}$$

where $A(\lambda I - B)^{-1}$ is an operator with finite-dimensional range (see the proof of 4.4.1.1(*a*)).

We can therefore solve the operator equation

$$x + A(\lambda I - B)^{-1} x = f \qquad \lambda \neq 0 \tag{$**$}$$

as in § 4.4.2 and there is a 1–1 correspondence between (∗∗) and the original operator equation (∗). If x is a solution of (∗∗), then

$$y = (\lambda I - B)^{-1}x$$

is a solution of (∗) and vice versa.

Thus we have obtained a method for solving equation (∗) and, together with 4.4.2.1 and 4.5.2.1, a constructive proof of the following theorem.

4.5.3.1 Theorem. If $\lambda \neq 0$ belongs to the spectrum of the compact operator T, then λ is an eigenvalue. In this case a solution of the equation (∗) exists if and only if

$$f \in N(\bar{\lambda} I - T^*)^{\perp}.$$

$\lambda = 0$ is contained in the continuous spectrum.

4.5.4. A more precise description of the spectrum can be obtained from the following considerations. Let λ_0 be fixed and let us construct (∗∗) for $\|T - A\| = \|B\| < |\lambda_0|$; then (∗) and (∗∗) are also equivalent for $|\lambda| > |\lambda_0|$ in the sense described in § 4.5.3. But, by 4.4.2.1, (∗∗) has only a finite number of eigenvalues. Hence we obtain the following.

4.5.4.1 Theorem. Let $0 < r < \|T\|$; then

$$\{\lambda \colon |\lambda| > r;\ \lambda \in \sigma(T)\}$$

consists of a finite number of eigenvalues.

There is one more common property with the finite-dimensional case.

4.5.4.2 Theorem. If T is compact then

$$N_\lambda := \{y \colon (\lambda I - T)y = \theta\}$$

is a finite-dimensional subspace of $\mathscr{H}$ for every $\lambda \neq 0$.

Proof. If $N_\lambda = \{\theta\}$, then the theorem is obvious. Now let $N_\lambda \neq \{\theta\}$; then the restriction of T to N_λ is λI since $Ty = \lambda y$ for $y \in N_\lambda$. Since the restriction of a compact operator remains compact, λI is compact on N_λ and, by Example 8, this is impossible if N_λ is infinite dimensional.

4.6 Self-adjoint compact operators

The simplest operators are the projection operators since a projection operator P is equal to the identity on a closed subspace $\mathscr{M}$ and zero on the orthogonal complement $\mathscr{M}^{\perp}$. In this section we shall show that the building blocks of a self-adjoint operator T are projection operators and thus even more information is obtained on the spectrum of T, and hence on the solution of $\lambda y - Ty = f$, than in the general case of compact operators.

4.6.1. We shall show that the spectrum of a self-adjoint compact T contains 'enough' eigenvalues in the sense that the linearly independent eigenvectors form a basis for the range of T as in the finite-dimensional case.

If $\mathbf{A}$ is a symmetric matrix, i.e. $\mathbf{A}^* = \mathbf{A}$, then

$$\lambda = \sup \{\mathbf{x}^* \mathbf{A} \mathbf{x};\ \mathbf{x}^* \mathbf{x} = 1\}$$

($\mathbf{x}$ is column matrix) is an eigenvalue. The same is essentially true for a compact self-adjoint T.

4.6.1.1 Theorem. For a compact self-adjoint operator T, there exists an eigenvalue λ with $|\lambda| = \|T\|$.

Proof. For any bounded self-adjoint T,

$$\|T\| = \sup \{|(Tx|x)| : \|x\| = 1\}$$

as we saw in 2.10.2.2. Hence there exists

$$\{x_n\}: |(Tx_n|x_n)| \to \|T\|$$

(and $\|x_n\| = 1$); moreover, $\|Tx_n\| \leqslant \|T\|$.

Since T is compact, there exists a convergent subsequence $\{Tx_{n_i}\}$.

Let $y = \lim_{i\to\infty} Tx_{n_i}$ and $\lambda \neq 0$. If

$$Tx_{n_i} - \lambda x_{n_i} \to \theta,$$

then $x_{n_i} \to (1/\lambda) y$ and hence $T(1/\lambda) y = y$, which means that λ is an eigenvalue with eigenvector y.

We shall show that $Tx_n - \lambda x_n \to \theta$ if

$$\lambda = \lim_{n\to\infty} (Tx_n|x_n).$$

In fact,

$$0 \leqslant \|Tx_n - \lambda x_n\|^2 = \|Tx_n\|^2 + \|\lambda x_n\|^2 - 2\lambda (Tx_n|x_n)$$

$$\leqslant (\|T\|^2 + |\lambda|^2) \|x_n\|^2 - 2\lambda (Tx_n|x_n) = 2|\lambda|^2 - 2\lambda (Tx_n|x_n)$$

and

$$2|\lambda|^2 - 2\lambda (Tx_n|x_n) \to 0$$

since $(Tx_n|x_n) \to \lambda$.

Remark 1. λ is real since $(Tx_n|x_n)$ is real for a self-adjoint T.

Remark 2. It can be supposed that $\{(Tx_n|x_n)\}$ is convergent since if this is not the case then we replace $\{(Tx_n|x_n)\}$ by a convergent subsequence at the very beginning of the proof.

Now, the counterparts of Theorem 4.1.2.1 are the following.

4.6.1.2 Theorem. If T is a compact self-adjoint operator on a separable Hilbert space $\mathscr{H}$ then there exists a sequence $\{\lambda_k\}$ of real eigenvalues

$$\|T\| = |\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda_k| \geqslant \ldots$$

with the only possible accumulation point $\lambda=0$.

Proof. By the previous theorem, there is a real eigenvalue λ_1 with $|\lambda_1|=\|T\|$. By the same procedure as we adopted in the proof of Theorem 4.1.2.1, we can find a sequence of eigenvalues $\{\lambda_k\}$ with

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \ldots \geqslant |\lambda|_k \geqslant \ldots$$

since for the norm of the restricted operators,

$$\|T\| \geqslant \|T_k\| \geqslant \|T_{k+1}\| \qquad k = 1, 2, \ldots.$$

There are two possible cases. If $T_k=0$ for a certain k, then T has only a finite number of eigenvalues. If $T_k\neq 0$ for all k, then $\{|\lambda_k|\}$ is a convergent infinite sequence and

$$\|Ty_m-Ty_n\|^2 = \|\lambda_m y_m-\lambda_n y_n\|^2 = \lambda_m^2+\lambda_n^2$$

if $y_k\in N_k$; $k=1, 2, \ldots$, where N_k is the eigenspace corresponding to the eigenvalue λ_k.

$\{y_k\}$ is orthogonal, by Theorem 4.1.2.1, T is compact and we can suppose that $\|y_k\|=1$. Hence there is a convergent subsequence $\{Ty_k\}$, which implies that the limit of $\{|\lambda_k|\}$ cannot be different from 0.

4.6.1.3 Theorem. Every compact self-adjoint T can be represented in the form

$$Tx = \sum_{k=1}^{\infty} \lambda_k(x|y_k)y_k \qquad x\in\mathscr{H} \qquad (*)$$

where $\{\lambda_k, y_k;\ k=1, 2, \ldots\}$ are related pairs of eigenvalues and eigenvectors with $\|y_k\|=1$.

Proof. By the procedure adopted in the proof of 4.1.2.1,

$$x-\sum_{k=1}^{m}(x|y_k)y_k\in\mathscr{H}_m \qquad \text{and} \qquad Tx-\sum_{k=1}^{m}\lambda_k(x|y_k)y_k\in\mathscr{H}_m.$$

Moreover,

$$Tx-\sum_{k=1}^{m}\lambda_k(x|y_k)y_k=T\Big(x-\sum_{k=1}^{m}(x|y_k)y_k\Big)=T_m\Big(x-\sum_{k=1}^{m}(x|y_k)y_k\Big).$$

Hence

$$Tx=\sum_{k=1}^{m}\lambda_k(x|y_k)\,y_k$$

if $T_m=0$. For the case where $T_m\neq 0$ for all m,

$$\Big\|Tx-\sum_{k=1}^{m}\lambda_k(x|y_k)y_k\Big\|\leq \|T_m\|\,\Big\|x-\sum_{k=1}^{m}(x|y_k)y_k\Big\|\leq \|T_m\|\,\|x\|=|\lambda_m|\,\|x\|$$

by 4.6.1.1 and 2.2.1 $(*)$. Moreover, by Theorem 4.6.1.2, $\lambda_m\to 0$.

4.6.2. Theorem 4.6.1.3 can be formuled as the decomposition of a self-adjoint compact operator into projection operators. First we shall show that there is an orthogonal decomposition of $\mathcal{H}$ into the eigenspaces of T if T is self-adjoint.

4.6.2.1 Definition. Let $\mathcal{M}_k$; $k=1,2,\ldots$ be pairwise orthogonal subspaces of $\mathcal{H}$, i.e. $\mathcal{M}_k\subset\mathcal{H}$ and if $z_i\in\mathcal{M}_i$, $z_j\in\mathcal{M}_j$, $i\neq j$ then $(z_i|z_j)=0$. Then

$$\bigoplus_{k=1}^{\infty}\mathcal{M}_k$$

is defined as the linear space of the sums

$$\sum_{k=1}^{\infty}z_k:\ \sum_{k=1}^{\infty}\|z_k\|^2<\infty \qquad z_k\in\mathcal{M}_k.$$

4.6.2.2 Theorem. If

$$\mathcal{N}_0=\{x\colon Tx=\theta\}$$

and $\mathcal{N}_k$ is the eigenspace corresponding to the kth eigenvalue λ_k, then

$$\mathcal{H}=\bigoplus_{k=0}^{\infty}\mathcal{N}_k.$$

Remark. A similar decomposition to a finite number of eigenspaces can be seen in the proof of 4.1.2.1. Moreover, a more detailed form of 4.6.1 $(*)$ is

$$Tx=\sum_{k=1}^{\infty}\sum_{i=1}^{k(i)}\lambda_k(x|y_{ki})\,y_{ki}$$

where $\{y_{ki};\ i=1, 2, \ldots, k(i)\}$ is an orthonormal basis for $\mathcal{N}_k$ (see also the proof of 4.1.2.1).

Proof. For the decomposition

$$x = \Big(x - \sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} (x|y_{ki})\, y_{ki}\Big) + \sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} (x|y_{ki})\, y_{ki} \qquad (*)$$

where $y_{ki} \in \mathcal{N}_k$, we have, by 4.6.1.3,

$$T\Big(x - \sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} (x|y_{ki})\, y_{ki}\Big) = Tx - \sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} \lambda_k (x|y_{ki})\, y_{ki} = 0.$$

Moreover,

$$\Big(\sum_{k=1}^{\infty} \sum_{i=1}^{k(t)} (x|y_{ki})\, y_{ki} \,\Big|\, \Big(x - \sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} (x|y_{ki})\, y_{ki} | y_{jl}\Big)\Big) = 0 \qquad l = 1, 2, \ldots, k(i).$$

Hence $(*)$ is the unique decomposition

$$\mathcal{H} = \mathcal{N}_0 \oplus \mathcal{N}_0^{\perp}$$

where $\mathcal{N}_0^{\perp} = \bigoplus_{k=1}^{\infty} \mathcal{N}_k$ since, by definition,

$$\sum_{k=1}^{\infty} \sum_{i=1}^{k(i)} (x|y_{ki})\, y_{ki} \in \bigoplus_{k=1}^{\infty} \mathcal{N}_k.$$

4.6.2.3 Theorem. Let P_k be the projection onto $\mathcal{N}_k$; $k=1, 2, \ldots$ and let P_0 be the projection onto the null space $\mathcal{N}_0$ of T. Then

$$T = \sum_{k=1}^{\infty} \lambda_k P_k.$$

Proof. We can write

$$P_k = \sum_{i=1}^{k(i)} P_{ki}$$

where P_{ki} is the projection onto the one-dimensional subspace generated by y_{ki}, and hence

$$P_{ki}\, x = (x|y_{ki})\, y_{ki} \qquad x \in \mathcal{H}.$$

Now it is obvious from 4.6.1.3 that

$$Tx = \sum_{k=1}^{\infty} \lambda_k P_k x \qquad x \in \mathcal{H}$$

and we have only to show that

$$\sum_{k=1}^{\infty} \lambda_k P_k$$

is a convergent series in $B(\mathcal{H})$ (i.e. that it is convergent in the operator norm).

In fact, by 4.6.2.2,

$$\|x\|^2 = \Big\|\sum_{k=0}^{\infty} P_k x\Big\|^2 = \sum_{k=0}^{\infty} \|P_k x\|^2$$

and so

$$\Big\|Tx - \sum_{k=1}^{n-1} \lambda_k P_k x\Big\|^2 = \Big\|\sum_{k=n}^{\infty} \lambda_k P_k x\Big\|^2 = \sum_{k=n}^{\infty} |\lambda_k|^2 \|P_k x\|^2$$

$$\leqslant |\lambda_n|^2 \sum_{k=n}^{\infty} \|P_k x\|^2 \leqslant |\lambda_n|^2 \|x\|^2.$$

Applying 4.6.1.2, we obtain

$$\Big\|T - \sum_{k=1}^{n-1} \lambda_k P_k\Big\| \to 0.$$

Remark. The decomposition P_k into projections onto one-dimensional subspaces is not necessary. In fact, it follows from 4.6.2.2 that

$$x = P_0 x + \sum_{k=1}^{\infty} P_k x \qquad x \in \mathscr{H}$$

where P_k is the projection onto N_k and hence $P_i P_j = 0$ for $i \neq j$. It follows that

$$Tx = \sum_{k=1}^{\infty} \lambda_k P_k x \qquad \text{and} \qquad \|x\|^2 = \sum_{k=0}^{\infty} \|P_k x\|^2$$

and we can continue the proof as above.

The form

$$\sum_{k=1}^{\infty} \lambda_k P_k$$

of a self-adjoint operator T is called the *spectral decomposition* of T.

Remark. The result of the previous theorem is essentially the same as 4.6.1.3 but in a more elaborate form. However, the representation of T given in 4.6.2.3 will be the starting point for the spectral representation of the non-compact case.

4.6.3. We now pose the converse question: when will

$$\sum_{k=1}^{\infty} \lambda_k P_k$$

be a compact self-adjoint operator?

4.6.3.1 Theorem. If

(i) P_k; $k=1, 2, \ldots$ are projection operators with finite-dimensional range;
(ii) $P_i P_j = 0$ for $i \neq j$;
(iii) $\{\lambda_k;\ k=1, 2, \ldots\}$ is a sequence of real numbers and $\lambda_k \to 0$;

then

$$T = \sum_{k=1}^{\infty} \lambda_k P_k$$

is a compact self-adjoint operator of $\mathscr{H}$.
Proof. In this case,

$$\Big\|\sum_{k=m}^{n} \lambda_k P_k x\Big\|^2 = \sum_{k=m}^{n} |\lambda|_k^2 \|P_k x\|^2 \leq |\lambda_m|^2 \|x\|^2$$

since

$$\|x\|^2 \geq \Big\|\sum_{k=1}^{\infty} P_k x\Big\|^2 = \sum_{k=1}^{\infty} \|P_k x\|^2$$

and hence

$$\sum_{k=1}^{\infty} \lambda_k P_k$$

is convergent in the operator norm.

$$T = \sum_{k=1}^{\infty} \lambda_k P_k$$

is a compact self-adjoint operator since the partial sums

$$T_n = \sum_{k=1}^{n} \lambda_k P_k$$

are self-adjoint operators with finite-dimensional range.

Remark. Later on we shall see that for complex $\{\lambda_k;\ k=1, 2, \ldots\}$ the above series defines a compact normal operator.

4.6.4. By applying the results of § 4.6.2 to the solution of the equation

$$(\lambda I - T)y = f \qquad \lambda \neq 0 \tag{$*$}$$

a simpler method can be obtained for a self-adjoint T than for the general case in § 4.5.3. When T is self-adjoint, equation $(*)$ can be given in the form

$$\lambda P_0 y + \sum_{k=1}^{\infty} (\lambda - \lambda_k) P_k y = \sum_{k=0}^{\infty} P_k f$$

since

$$x = \sum_{k=0}^{\infty} P_k x.$$

Multiplying both sides by the projection P_n, we have

$$(\lambda-\lambda_n)P_n y = P_n f \qquad n = 1, 2, \ldots.$$

If λ is a regular value, then there is a unique solution y of $(*)$ for every $f \in \mathscr{H}$. In this case,

$$P_n y = \frac{1}{\lambda-\lambda_n} P_n f \qquad n = 1, 2, \ldots$$

and hence

$$y = \sum_{k=0}^{\infty} P_k y = \sum_{k=0}^{\infty} \frac{1}{\lambda-\lambda_k} P_k f.$$

If λ is an eigenvalue, i.e. $\lambda=\lambda_i$, then there exists a solution y of $(*)$ if and only if $f \in \mathscr{N}_{\lambda_i} \ominus \mathscr{H}$, by 4.5.3.1. In this case $P_i f = \theta$ and, on the same grounds as in the case of a regular value,

$$y = \sum_{k \neq i} \frac{1}{\lambda-\lambda_k} P_k f$$

is a solution of $(*)$.

4.7 Compact normal operators and the first step towards the representation of non-normal operators

In this section we shall extend the representation 4.6.1 $(*)$ to compact normal operators, and we shall also derive a useful representation for any compact operator.

4.7.1. Every bounded linear operator has the following decomposition into self-adjoint operators:

$$T = \frac{T+T^*}{2} + \mathrm{i}\frac{T-T^*}{2}.$$

If T is also a normal operator, then this decomposition has the following properties in common with complex numbers.

4.7.1.1 Theorem. The bounded linear operator is normal, i.e.

$$TT^* = T^*T$$

if and only if $T=A+\mathrm{i}B$, where A and B are self-adjoint operators with

$AB=BA$. Then

$$TT^* = A^2+B^2.$$

Proof.

$$TT^* = (A+iB)(A-iB) = A^2+B^2+i(BA-AB)$$

$$T^*T = (A-iB)(A+iB) = A^2+B^2+i(AB-BA)$$

and hence the following are equivalent:

(*a*) $TT^*=T^*T$;

(*b*) $AB=BA$;

(*c*) $TT^*=A^2+B^2$.

4.7.1.2 Theorem. If λ is an eigenvalue of T, with $\alpha=\operatorname{Re}\lambda$, $\beta=\operatorname{Im}\lambda$, then α is an eigenvalue of A and β is an eigenvalue of B.

Proof. If $x\in\mathscr{H}$ is an eigenvector of T belonging to λ, then (applying also 4.1.2.2)

$$Ax = \frac{1}{2}(T+T^*)x = \frac{1}{2}(\lambda+\bar{\lambda})x = \alpha x$$

and,

$$Bx = \frac{1}{2i}(T-T^*)x = \frac{1}{2i}(\lambda-\bar{\lambda})x = \beta x.$$

Based on the connections between A, B and T the following can be proved.

4.7.1.3 Theorem. Let T be a compact normal operator of a Hilbert space $\mathscr{H}$. Then

$$T = \sum_{k=1}^{\infty} \lambda_k P_k$$

where $\{\lambda_k;\ k=1, 2, \ldots\}$ is the sequence of all eigenvalues of T and P_k is the projection onto $\mathscr{N}_{\lambda_k}$. (See, for example, Naylor and Sell 1982, §§ 6.10–11.)

4.7.2. T^*T is a positive operator for every linear operator T since

$$(T^*Tx|x) = (Tx|Tx) \geqslant 0$$

and hence, by 4.6.1 (*) and the Remark following 4.1.2.1, we have, for every compact operator T,

$$T^*Tx = \sum_{k=1}^{\infty} \mu_k^2(x|x_k)x_k$$

where $\{\mu_k^2,\ x_k\}$ are the related pairs of eigenvalues and eigenvectors of T^*T.

4.7.2.1 Theorem. For every compact operator T,

$$Tx = \sum_{k=1}^{\infty} \mu_k(x|y_k)x_k \qquad x \in \mathscr{H} \qquad (*)$$

where x_k is the eigenvector of T^*T and y_k is the eigenvector of TT^*, both belonging to the eigenvalue μ_k^2.

Proof. Let y_k be the eigenvector of TT^* and

$$x_k = \frac{1}{\mu_k} T^* y_k \qquad \mu_k \neq 0$$

then

$$(x_k|x_m) = \frac{1}{\mu_k \mu_m}(T^* y_k|T^* y_m) = \frac{1}{\mu_k \mu_m}(y_k|TT^* y_m)$$

$$= \frac{\mu_m}{\mu_k}(y_k|y_m) = 0 \qquad \text{if} \quad k \neq m.$$

Hence $\{x_k\}$ is also an orthogonal system.

If $\{x_k\}$ is not complete, then we add $\{z_k\}$ by the Gram–Schmidt process so that $\{x_k\}$ and $\{z_k\}$ together form an orthonormal basis for $\mathscr{H}$. Obviously, $\{z_k\} \subset (T^*\mathscr{H})^\perp$ and hence $Tz_k = 0$ since

$$R(T^*)^\perp = N(T^{**}) \qquad \text{and} \qquad T^{**} = T$$

can be proved as $(**)$ in 4.1.1.4.

It follows that

$$Tx = T\Big(\sum_k (x|x_k)x_k + \sum_k (x|z_k)z_k\Big) = \sum_k (x|x_k)Tx_k$$

$$= \sum_k \frac{1}{\mu_k}(x|x_k)TT^* y_k = \sum_k \mu_k(x|x_k)y_k.$$

Remark. It follows from the representation $(*)$ that every compact operator T in a Hilbert space $\mathscr{H}$ (or mapping a Hilbert space $\mathscr{H}_1$ into another Hilbert space $\mathscr{H}_2$) is the uniform limit of operators with finite-dimensional range.

4.8 Hilbert-Schmidt operators

In this section we shall define operators in an arbitrary separable Hilbert space that are the counterparts of integral operators in $L^2[a, b]$ with square integrable kernel.

4.8.1. Besides the (uniform) operator norm there is the natural norm

$$\|T\|_2 := \Big(\int_a^b \int_a^b |k(t,\tau)|^2 \, d\tau \, dt\Big)^{1/2}$$

for integral operators with square integrable kernel k. We learned in Example 1 of § 4.2.3 that $\|T\|_2 \geq \|T\|$.

Let us denote the set of such integral operators with norm $\|\,.\,\|_2$ by $HS(L^2)$.

4.8.1.1 Theorem. (*a*) $HS(L^2)$ forms a subalgebra of $B(L^2)$; i.e. if $T_1, T_2 \in HS(L^2)$ and α, β are scalars, then $\alpha T_1 + \beta T_2 \in HS(L^2)$ and $T_1 T_2 \in HS(L^2)$.

(*b*) $\|T_1 T_2\|_2 \leq \|T_1\|_2 \|T_2\|_2$.

(*c*) If $T \in HS(L^2)$, then $T^* \in HS(L^2)$ and $\|T\|_2 = \|T^*\|_2$.

(*d*) $HS(L^2)$ is a Banach space.

Proof. For any $f \in L^2[a, b]$,

$$T_1 T_2 f := \int_a^b k_1(t,s)\Big(\int_a^b k_2(s,\tau) f(\tau)\, d\tau\Big)\, ds$$

$$= \int_a^b \Big(\int_a^b k_1(t,s) k_2(s,\tau)\, ds\Big) f(\tau)\, d\tau.$$

Hence $T_1 T_2$ is also an integral operator with kernel

$$k(t,\tau) = \int_a^b k_1(t,s) k_2(s,\tau)\, ds.$$

Moreover, by the Cauchy inequality,

$$\Big|\int_a^b k_1(t,s) k_2(s,\tau)\, ds\Big|^2 \leq \int_a^b |k_1(t,s)|^2\, ds \int_a^b |k_2(s,\tau)|^2\, ds$$

and hence $k = k(t,\tau)$ is also a square integrable kernel and (*b*) is also satisfied. (*c*) follows from Example 2 of § 2.10.1 since

$$\int_a^b \int_a^b |k(t,\tau)|^2\, d\tau\, dt = \int_a^b \int_a^b |\overline{k(\tau,t)}|^2\, dt\, d\tau.$$

Finally, $L^2([a,b]\times[a,b])$ is a Banach space (see, for example, Example 6 of § 1.6.1) and hence, from the definition of $\|\,.\,\|_2$, $HS(L^2)$ is also a Banach space, since $\|\,.\,\|_2$ is equal to the norm of the kernel in $L^2([a,b]\times[a,b])$.

An algebra that is also a Banach space with a norm satisfying inequality (*b*) of the previous theorem is called a *Banach algebra*. The Banach algebra $HS(L^2)$ can be described completely in Hilbert space terms.

4.8.1.2 Theorem. The operator K of $L^2[a, b]$ has the form

$$Kf = \int_a^b k(t, \tau) f(\tau)\, d\tau$$

with square integrable kernel $k=k(t, \tau)$ if and only if

$$\sum_{k=1}^{\infty} \|Ke_k\|^2 < \infty \qquad (*)$$

for every orthonormal basis $\{e_k\}$.

The sum $(*)$ is independent of the choice of the orthonormal basis and

$$\|K\|_2 = \Big(\sum_{k=1}^{\infty} \|Ke_k\|^2\Big)^{1/2}.$$

Proof. For any orthonormal basis $\{e_k\}$,

$$\|Ke_k\|^2 = \sum_{j=1}^{\infty} |(Ke_k|e_j)|^2$$

since $(Ke_k|e_j)$ is the jth Fourier coefficient of Ke_k with respect to the orthonormal system $\{e_k\}$ and hence

$$\sum_{k=1}^{\infty} \|Ke_k\|^2 = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} |(Ke_k|e_j)|^2. \qquad (**)$$

Now let K be an integral operator. Then

$$(Ke_k|e_j) = \int_a^b \int_a^b K(t, \tau) e_k(\tau) e_j(t)\, d\tau\, dt$$

so that $(Ke_k|e_j)$ are the Fourier coefficients of $k=k(t, \tau)$ with respect to the complete orthonormal system $e_k(\tau)e_j(t)$; $k=1, 2, \ldots, j=1, 2, \ldots$ and hence

$$\sum_{k=1}^{\infty} \sum_{j=1}^{\infty} |(Ke_k|e_j)|^2 = \int_a^b \int_a^b |K(t, \tau)|^2\, d\tau\, dt < \infty.$$

Conversely, if $(**)$ is finite, i.e.

$$\sum_{k=1}^{\infty} \|Ke_k\|^2 < \infty$$

then by the Riesz–Fischer Theorem (see, for example, 1.6.1, Example 6) there exists $k \in L^2([a, b] \times [a, b])$ such that

$$(Ke_k|e_j) = \int_a^b \int_a^b K(t, \tau) e_k(\tau) e_j(t)\, d\tau\, dt$$

and hence

$$[Ke_k](t) = \int_a^b K(t, \tau)e_k(\tau)\, d\tau$$

for the orthonormal basis $\{e_k\}$. It follows that

$$[Kf](t) = \int_a^b K(t, \tau)f(\tau)\, d\tau$$

for every $f \in L^2[a, b]$ since

$$f = \sum_{k=1}^{\infty} (f|e_k)e_k$$

and the operator K is continuous.

4.8.2. We now define the subalgebra $HS(\mathscr{H})$ of $B(\mathscr{H})$ for any separable Hilbert space $\mathscr{H}$.

4.8.2.1 Definition. The linear operator T of $\mathscr{H}$ is called a *Hilbert–Schmidt operator* if

$$\sum_{k=1}^{\infty} \|Te_k\|^2 < \infty$$

for an orthonormal basis $\{e_k\}$.

If T is a Hilbert–Schmidt operator, then

$$\|T\|_2 := (\sum_{k=1}^{\infty} \|Te_k\|^2)^{1/2}$$

is called the Hilbert–Schmidt norm. The set of Hilbert–Schmidt operators of $\mathscr{H}$ is denoted by $HS(\mathscr{H})$.

4.8.2.2 Theorem. The Hilbert–Schmidt norm is independent of the choice of the orthonormal basis $\{e_k\}$ and

$$\|T\|_2 = (\sum_{j=1}^{\infty} \sum_{k=1}^{\infty} |(Te_k|e_j)|^2)^{1/2} \geqslant \|T\|.$$

Proof. Let $\{e_k\}$ and $\{\varphi_k\}$ be two different orthonormal bases in $\mathscr{H}$. Then

$$\|Te_k\|^2 = \sum_{j=1}^{\infty} |(Te_k|\varphi_j)|^2$$

since $(Te_k|\varphi_j)$ is the jth Fourier coefficient of Te_k with respect to the orthonormal basis $\{\varphi_j\}$. Hence

$$\sum_{k=1}^{\infty} \|Te_k\|^2 = \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} |(Te_k|\varphi_j)|^2$$

$$= \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} |(e_k|T^*\varphi_j)|^2 = \sum_{j=1}^{\infty} \|T^*\varphi_j\|^2$$

and for $e_k=\varphi_k$; $k=1, 2, \ldots,$

$$\sum_{k=1}^{\infty} \|T\varphi_k\|^2 = \sum_{k=1}^{\infty} \|T^*\varphi_k\|^2.$$

We conclude that

$$\sum_{k=1}^{\infty} \|Te_k\|^2 = \sum_{k=1}^{\infty} \|T\varphi_k\|^2.$$

It follows from the definition of the operator norm that for every $\varepsilon>0$ there exists $e_0\in\mathscr{H}$ with $\|e_0\|=1$ such that

$$\|T\| < \|Te_0\| + \varepsilon.$$

If we choose an orthonormal basis $\{e_k\}$ containing $e_0\in\mathscr{H}$, then

$$\|Te_0\|^2+\varepsilon < \sum_{k=1}^{\infty} \|Te_k\|^2 = \|T\|_2^2.$$

The important property of $\|\,.\,\|_2$ is that it is a *Hilbert space norm.*

4.8.2.3 Theorem. If $T_1, T_2\in \mathrm{HS}(\mathscr{H})$, then

$$(T_1|T_2) := \sum_{k=1}^{\infty} (T_2^*T_1e_k|e_k)$$

is a scalar product. In particular,

$$\|T\|_2 = \Big(\sum_{k=1}^{\infty} (T^*Te_k|e_k)\Big)^{1/2} = \Big(\sum_{k=1}^{\infty} \|Te_k\|^2\Big)^{1/2}.$$

Proof.

$$(T_1|T_2) := \sum_{k=1}^{\infty} (T_2^*T_1e_k|e_k) = \sum_{k=1}^{\infty} (e_k|T_1^*T_2e_k)$$

$$= \sum_{k=1}^{\infty} \overline{(T_1^*T_2e_k|e_k)} = \overline{(T_2|T_1)}.$$

$(T|T)\geqslant 0$ and $(T|T)=0$ if and only if $T=0$ since

$$(T|T) := \sum_{k=1}^{\infty} (T^*Te_k|e_k) = \sum_{k=1}^{\infty} \|Te_k\|^2 = \|T\|_2^2$$

and it is obvious that the remaining axioms of the scalar product in § 2.1.1 are also satisfied.

4.8.3. We now consider the representation of a Hilbert–Schmidt operator T by an infinite matrix, as defined in § 4.4.3. It follows from the second part of Theorem 4.8.2.2 that $T \in \mathrm{HS}(\mathscr{H})$ if and only if

$$\sum_{k=1}^{\infty} \sum_{i=1}^{\infty} |a_{ik}|^2 < \infty .$$

In fact, it can be proved that these are the Hilbert–Schmidt operators in l^2 and every $T \in \mathrm{HS}(\mathscr{H})$ is unitarily equivalent to such an operator.

Every $T \in \mathrm{HS}(\mathscr{H})$ is compact and hence it also has a representation in the form

$$Tx = \sum_{k=1}^{\infty} \lambda_k (x|e_k) \varphi_k \qquad x \in \mathscr{H} \qquad (*)$$

where $\{e_k\}$ and $\{\varphi_n\}$ are orthonormal bases and $\lambda_k \to 0$ as was shown in § 4.7.2. In fact, if T_n is the operator

$$T_n e_k = \begin{cases} T e_k & \text{if } k \leqslant n \\ 0 & \text{if } k > n \end{cases}$$

then T_n is an operator with finite-dimensional range and

$$\|T - T_n\| \leqslant \|T - T_n\|_2 = \Big(\sum_{k=n+1}^{\infty} \|Te_k\|^2 \Big)^{1/2}.$$

We conclude that $\|T - T_n\| \to 0$ and hence T is compact. It follows from the representation $(*)$ that

$$\sum_{n=1}^{\infty} \|Te_n\|^2 = \sum_{n=1}^{\infty} \|\lambda_n \varphi_n\|^2 = \sum_{n=1}^{\infty} |\lambda_n|^2$$

and hence a compact operator T is a Hilbert–Schmidt operator if and only if

$$\sum_{k=1}^{\infty} |\lambda_k|^2 < \infty$$

in the representation $(*)$.

Based on these representations, the following can be proved.

4.8.3.1 Theorem. (*a*) $\mathrm{HS}(\mathscr{H})$ forms a subalgebra of $B(\mathscr{H})$.

(*b*) $\mathrm{HS}(\mathscr{H})$ with the Hilbert–Schmidt norm $\| \, . \, \|_2$ is a Banach algebra.

(*c*) If $T \in \mathrm{HS}(\mathscr{H})$ then $T^* \in \mathrm{HS}(\mathscr{H})$ and $\|T\|_2 = \|T^*\|_2$.

Remark. If $T \in \mathrm{HS}(\mathscr{H})$ then $LT \in \mathrm{HS}(\mathscr{H})$ also for $L \in B(\mathscr{H})$ and $\|LT\|_2 \leqslant \|L\| \|T\|_2$, i.e. $\mathrm{HS}(\mathscr{H})$ is not only a subalgebra but a (two sided) ideal in $B(\mathscr{H})$.

4.9 Positive operators

We defined positive operators in § 2.10.2. and also a natural ordering of self-adjoint operators similar to the real numbers. We now continue to demonstrate the similarities between real numbers and self-adjoint operators, as we did in §§ 4.1.2 and 4.7.2.

4.9.1. Referring to 2.10.2.3, a self-adjoint operator A is said to be positive if

$$(Ax|x) \geqslant 0 \qquad x \in H.$$

In this case we also write $A \geqslant 0$, and for the self-adjoint operators A, B we define

$$B \geqslant A \quad \text{if} \quad B-A \geqslant 0.$$

Throughout this section an operator means a self-adjoint operator. It is obvious that $B \geqslant A$ and $0 \leqslant \lambda$ imply

$$\lambda B \geqslant \lambda A \quad \text{and} \quad B+C \geqslant A+C$$

for every operator C. Moreover, if $\lambda < 0$ and $B \geqslant A$ then $\lambda B \leqslant \lambda A$ since in this case,

$$\lambda((B-A)x|x) \leqslant 0.$$

The most important positive operators are the *projectors (projection operators)*. For further examples see 4.13.7–8.

First we shall give the analogue of the well-known theorem that a monotonically decreasing sequence $\{a_n\}$ of real numbers bounded from below is convergent.

4.9.1.1 Theorem. If $\{T_n\}$ is a sequence of self-adjoint operators such that

$$T_1 \geqslant T_2 \geqslant \ldots \geqslant T_n \geqslant \ldots \geqslant B$$

then there exists $T \in B(\mathscr{H})$ such that

$$\lim T_n x = Tx \qquad x \in \mathscr{H}.$$

Proof. By subtracting B and dividing by $\|T_1 - B\|$, we may assume that

$$0 \leqslant T_n \leqslant I \qquad n = 1, 2, \ldots$$

which means that

$$0 \leqslant (T_n x|x) \leqslant (x|x) \qquad x \in \mathscr{H} \qquad (*)$$

and so

$$\|T_n\| = \sup\{(T_n x|x);\ \|x\| = 1\} \leqslant 1.$$

Moreover, also for $n<m$,

$$0 \leqslant (T_n x|x)-(T_m x|x) \leqslant (x|x) \tag{$**$}$$

and so

$$\|T_n-T_m\| \leqslant 1. \tag{$***$}$$

We assert that $\{T_n x\}$ is a Cauchy sequence for every $x\in\mathscr{H}$. Applying 2.14.38 to $T=T_n-T_m$ and $y=(T_n-T_m)x$, we have

$$((T_n-T_m)x|(T_n-T_m)x)^2 \leqslant ((T_n-T_m)x|x)((T_n-T_m)^2x|(T_n-T_m)x).$$

Moreover, by $(**)$ and $(***)$,

$$((T_n-T_m)^2x|(T_n-T_m)x) \leqslant ((T_n-T_m)x|(T_n-T_m)x) \leqslant (x|x).$$

Thus we have

$$\|T_n x-T_m x\|^4 \leqslant ((T_n x|x)-(T_m x|x))\,\|x\|^2$$

and so $\{T_n x\}$ is a Cauchy sequence for every $x\in\mathscr{H}$ (see also 4.13.18). Let

$$Tx = \lim T_n x \qquad x\in\mathscr{H}.$$

Then

$$(Tx|y) = \lim (T_n x|y) = \lim (x|T_n y) = (x|Ty) \qquad x, y\in\mathscr{H}$$

and hence T is self-adjoint. Moreover, it follows from $(*)$ that

$$0 \leqslant (Tx|x) \leqslant (x|x)$$

and so $\|T\|\leqslant 1$. It is obvious that T is a linear operator.

Remark 1. There is a similar theorem for an increasing sequence $\{T_n\}$ of operators, with the same proof.

Remark 2. If $T_n x\to Tx$ for every $x\in\mathscr{H}$ then we say that $T_n\to T$ *strongly (or pointwise)*. If $\|T_n-T\|\to 0$ then $T_n\to T$ strongly also, but the converse is not true.

4.9.2. Our main result in this section is the following.

4.9.2.1 Theorem. Let T be a positive operator and

$$T_0 = I \qquad T_{n+1} = T_n+\frac{1}{2}(T-T_n^2) \qquad n = 0, 1, \dots.$$

Then the sequence $\{T_n\}$ is strongly convergent.

If $\lim T_n x:=Sx$; $x\in\mathscr{H}$, then $S>0$ and

$$S^2 = T.$$

S is called the *square root* of T.

Proof. If $T_n \to S$, then

$$S = S + \frac{1}{2}(T - S^2)$$

and hence $S^2 = T$ (see also 4.13.19). Moreover, T_m; $m = 1, 2, \dots$ are polynomials of T and hence

$$T_n T_k = T_k T_n \qquad n, k = 1, 2, \dots .$$

We may also suppose that $T < I$ since for any self-adjoint T, $T < \|T\| I$.

We shall prove that

$$T_n \geqslant T_{n+1} \geqslant 0 \qquad n = 1, 2, \dots$$

and the convergence will follow from 4.9.1.1.

First we shall prove that $I - T_n$ is *a polynomial of* $I - T \geqslant 0$ *with positive coefficients.* In fact,

$$I - T_1 = \frac{1}{2}(I - T) \tag{$*$}$$

and

$$\begin{aligned} I - T_{n+1} &= I - T_n - \frac{1}{2}(T - T_n^2) \\ &= I - T_n - \frac{1}{2}[(I - T_n^2) - (I - T)] \\ &= (I - T_n)[I - \frac{1}{2}(I + T_n)] + \frac{1}{2}(I - T) \\ &= \frac{1}{2}(I - T_n)^2 + \frac{1}{2}(I - T). \end{aligned} \tag{$**$}$$

Now, if we suppose that $I - T_n$ is a polynomial of $I - T$ with positive coefficients, then obviously $(I - T_n)^2$ also has this property and hence $I - T_{n+1}$ is also a polynomial of $I - T$ with positive coefficients, by induction.

It also follows from $(*)$ and $(**)$, by induction, that

$$I - T_{n+1} \leqslant I \qquad n = 0, 1, \dots$$

and hence $T_{n+1} \geqslant 0$. Finally, also from $(**)$,

$$\begin{aligned} T_n - T_{n+1} = (I - T_{n+1}) - (I - T_n) &= \frac{1}{2}[(I - T_n)^2 - (I - T_{n-1})^2] \\ &= \frac{1}{2}[(I - T_n) + (I - T_{n-1})][(I - T_n) - (I - T_{n-1})]. \end{aligned}$$

In view of $(*)$, it follows that $T_n - T_{n+1}$ is also a polynomial of $I - T$ with positive coefficients and so

$$T_n - T_{n+1} \geqslant 0$$

(see 4.13.20).

Remark. It can also be proved that the square root S is unique if it is required to be positive.

The operator S is called a *commutant* of T if

$$ST = TS.$$

4.9.2.2 Definition. An operator S is called a *bicommutant* of T if

$$TL = LT \qquad \text{implies} \qquad SL = LS \qquad L \in B(\mathscr{H}).$$

We note that if S is a bicommutant, then it is also a commutant of T.

Remark. The usual definition of the commutant is a little different. The commutant of T is called the set $\mathscr{S}$ of all operators such that

$$ST = TS \qquad S \in \mathscr{S}$$

and there is a similar difference between the usual definition of the bicommutant and that given above.

It follows from the recurrence formula for the construction of the square root that if $S^2 = T$ then

$$ST = TS$$

and we also have the following theorem.

4.9.2.3 Theorem. The square root S of a positive operator T is a bicommutant of T.

Proof. If $L \in B(\mathscr{H})$ and $LT = TL$ then

$$LT_{n+1} = LT_n + \frac{1}{2}(LT + LT_n^2)$$

and hence $LT_{n+1} = T_{n+1}L$ if $LT_n = T_nL$. But it is obvious that $LT_0 = LI = IL = T_0L$. We conclude that each T_n; $n = 1, 2, \ldots$ is a bicommutant of T and hence we also have

$$S\colon\ Sx = \lim T_n x.$$

4.9.3. As an application of the results of the previous subsection, we give the analogue of the multiplication law of inequalities for positive operators.

4.9.3.1 Theorem. If $AB=BA$ then

$$A \geqslant 0, \quad B \geqslant 0 \Rightarrow AB \geqslant 0.$$

Proof. Let S be the square root of B; then

$$(ABx|x) = (ASSx|x) = (SASx|x) = (ASx|Sx) \geqslant 0.$$

4.10 Invariant subspaces and projection operators

To every invariant subspace $\mathcal{M}$ of an operator T we can construct a projector $P_{\mathcal{M}}$, as we saw in § 2.10.3. There is an intimate connection between T and $P_{\mathcal{M}}$, the family of projection operators of all invariant subspaces of T.

4.10.1. There is a certain commutation relation between T and the projector $P_{\mathcal{M}}$ of an invariant subspace $\mathcal{M}$ of T.

4.10.1.1 Theorem. $P_{\mathcal{M}}$ is the projector of an invariant subspace $\mathcal{M}$ of T if and only if

$$TP_{\mathcal{M}} = P_{\mathcal{M}}TP_{\mathcal{M}}. \qquad (*)$$

Proof. It follows by definition (see § 2.10.3) that

$$\mathcal{M} = \{P_{\mathcal{M}}x;\ x \in \mathcal{H}\}.$$

If $\mathcal{M}$ is an invariant subspace of T then $TP_{\mathcal{M}}x \in \mathcal{M}$ and hence $P_{\mathcal{M}}TP_{\mathcal{M}}x = TP_{\mathcal{M}}x$ for every $x \in \mathcal{H}$. If $\mathcal{M}$ is *not* an invariant subspace, then there exists $x \in \mathcal{H}$ such that $TP_{\mathcal{M}}x \notin \mathcal{M}$ and hence $TP_{\mathcal{M}}x \neq P_{\mathcal{M}}TP_{\mathcal{M}}x$.

4.10.1.2 Theorem. If $\mathcal{M}$ is an invariant subspace of T, then the orthogonal complement $\mathcal{M}^{\perp}$ is an invariant subspace of T^*.
Proof. By the previous theorem, $\mathcal{M}$ is an invariant subspace of T if and only if $TP_{\mathcal{M}} = P_{\mathcal{M}}TP_{\mathcal{M}}$ and, passing to the adjoint operators, we obtain

$$P_{\mathcal{M}}T^* = P_{\mathcal{M}}T^*P_{\mathcal{M}}. \qquad (**)$$

From § 2.10.3, we see that

$$P_{\mathcal{M}} = I - P_{\mathcal{M}^{\perp}}$$

and substituting into $(**)$ we obtain

$$(I-P_{\mathcal{M}^{\perp}})T^* = (I-P_{\mathcal{M}^{\perp}})T^*(I-P_{\mathcal{M}^{\perp}})$$

and hence

$$0 = P_{\mathcal{M}^{\perp}}T^*P_{\mathcal{M}^{\perp}} - T^*P_{\mathcal{M}^{\perp}}.$$

Now let T be a self-adjoint operator and let $\mathscr{M}$ be an invariant subspace of T. Then it follows from the previous theorem that the orthogonal complement $\mathscr{M}^{\perp}$ is also an invariant subspace of T and hence we have

$$TP_{\mathscr{M}} = P_{\mathscr{M}}T$$

instead of (*).

4.10.2. There is a natural partial ordering of projection operators:

$$P_{\mathscr{M}} < P_{\mathscr{N}} \qquad \text{iff} \qquad P_{\mathscr{M}}\mathscr{H} \subset P_{\mathscr{N}}\mathscr{H}$$

i.e. the range of $P_{\mathscr{M}}$ is a subset of the range of $P_{\mathscr{N}}$.

In 2.10.2.4 we defined an ordering for *arbitrary* self-adjoint operators T, S:

$$T \leqslant S \qquad \text{iff} \qquad S-T \geqslant 0.$$

We shall show that these two kinds of ordering are the same for projection operators.

4.10.2.1 Theorem. For the projection operators $P_{\mathscr{M}}$, $P_{\mathscr{N}}$ the following are equivalent:

(i) $P_{\mathscr{N}}\mathscr{H} \supset P_{\mathscr{M}}\mathscr{H}$

(ii) $P_{\mathscr{N}}P_{\mathscr{M}} = P_{\mathscr{M}}$

(iii) $P_{\mathscr{N}} - P_{\mathscr{M}} \geqslant 0$.

Proof. It is obvious that (i)⇔(ii). If $P_{\mathscr{N}}P_{\mathscr{M}} = P_{\mathscr{M}}$ then

$$(P_{\mathscr{N}} - P_{\mathscr{M}})^2 = P_{\mathscr{N}} - P_{\mathscr{M}}$$

and hence $P_{\mathscr{N}} - P_{\mathscr{M}} \geqslant 0$. Conversely, if $P_{\mathscr{N}} - P_{\mathscr{M}} \geqslant 0$, then

$$I - P_{\mathscr{N}} \leqslant I - P_{\mathscr{M}}$$

and hence

$$\|(I-P_{\mathscr{N}})P_{\mathscr{M}}x\|^2 = ((I-P_{\mathscr{N}})P_{\mathscr{M}}x|P_{\mathscr{M}}x) \leqslant ((I-P_{\mathscr{M}})P_{\mathscr{M}}x|P_{\mathscr{M}}x) = 0.$$

Hence we conclude that

$$(I-P_{\mathscr{N}})P_{\mathscr{M}} = 0$$

i.e. $P_{\mathscr{M}} = P_{\mathscr{N}}P_{\mathscr{M}}$.

4.10.3. The most important and frequently used properties of projections, connected with their ordering, are as follows.

4.10.3.1. Suppose that $\{P_{\mathscr{M}}\}$ form an ordered set. Then

$$(I-P_{\mathscr{M}})P_{\mathscr{N}} = \begin{cases} P_{\mathscr{N}} - P_{\mathscr{M}} & \text{if } P_{\mathscr{N}} \geqslant P_{\mathscr{M}} \\ 0 & \text{otherwise.} \end{cases}$$

4.10.3.2. If $P_a \leqslant P_b \leqslant P_c \leqslant P_d$ then

$$(P_d - P_c)(P_b - P_a) = \begin{cases} P_b - P_a & \text{if } P_a = P_c \text{ and } P_b = P_d \\ 0 & \text{otherwise.} \end{cases}$$

The proofs of these theorems are obvious. A more serious property is the following.

4.10.3.3. If $P_{\mathcal{N}} \geqslant P_{\mathcal{M}}$ then $P_{\mathcal{N}} - P_{\mathcal{M}}$ is a projector onto

$$P_{\mathcal{N}} \mathcal{H} \ominus P_{\mathcal{M}} \mathcal{H}$$

i.e. onto the orthogonal complement of $P_{\mathcal{M}} \mathcal{H}$ in $P_{\mathcal{N}} \mathcal{H}$.
Proof. By 4.10.2.1 (see also 4.13.42),

$$(P_{\mathcal{N}} - P_{\mathcal{M}})^2 = P_{\mathcal{N}} + P_{\mathcal{M}} - 2P_{\mathcal{N}} P_{\mathcal{M}} = P_{\mathcal{N}} - P_{\mathcal{M}}$$

and $P_{\mathcal{N}} - P_{\mathcal{M}}$ is self-adjoint. Hence $P_{\mathcal{N}} - P_{\mathcal{M}}$ is a projector. Let $\mathcal{S}$ be the subspace corresponding to $P_{\mathcal{N}} - P_{\mathcal{M}}$ by 2.10.3.1; then

$$P_{\mathcal{N}} \mathcal{H} = P_{\mathcal{M}} \mathcal{H} \oplus \mathcal{S}$$

since

$$P_{\mathcal{N}} = (P_{\mathcal{N}} - P_{\mathcal{M}}) + P_{\mathcal{M}} \qquad \text{and} \qquad (P_{\mathcal{N}} - P_{\mathcal{M}}) P_{\mathcal{M}} = 0.$$

∗4.11 Non-compact self-adjoint operators

The spectral representation of a compact self-adjoint operator T is based on the existence of an orthonormal basis in $\mathcal{H}$ consisting of eigenvectors of T. However, in the case of a non-compact self-adjoint operator T it may occur that there is no eigenvalue at all, as we have seen in Example 1 of § 4.3.1 and Example 5 of § 4.5.2.

In this section we shall introduce an operator-valued Riemann integral and show how the spectral representation of compact self-adjoint operators can be extended to the non-compact case.

4.11.1. The spectrum of a compact self-adjoint operator T consists of $\{0\}$ and eigenvalues in the interval $(-\|T\|, +\|T\|)$ of the real line. Moreover, the spectral radius is $\|T\|$ since there is an eigenvalue λ with $|\lambda| = \|T\|$ for a compact self-adjoint T. The spectrum of a non-compact self-adjoint T is also a non-void subset of $(-\|T\|, +\|T\|)$; however, it may happen that there are no eigenvalues.

An introductory result is the following.

4.11.1.1 Theorem. If T is a self-adjoint operator of $\mathscr{H}$, then λ is a regular value of T if and only if there exists $(K>0)$

$$\|(T-\lambda I)x\| \geqslant K\|x\| \qquad x\in\mathscr{H}.$$

The proof can be found as exercises in 4.13.22 and 4.13.23.

4.11.1.2 Theorem. Let m and M be the lower and upper bounds of T (see 2.10.2.5); then $\sigma(T)$ is a non-empty subset of the interval $[m, M]$. Moreover, $m, M\in\sigma(T)$.*

Proof. If $\lambda=\alpha+i\beta$, then

$$\|(T-\lambda I)x\|^2 = (Tx-\alpha x-i\beta x|Tx-\alpha x-i\beta x) = \|Tx-\alpha x\|^2+\beta^2\|x\|^2 \geqslant \beta^2\|x\|^2$$

and hence λ is a regular value of T if $\beta\neq 0$. If $\lambda>M$, then

$$\|(T-\lambda I)x\|\,\|x\| \geqslant |(Tx-\lambda x|x)| = |\lambda(x|x)-(Tx|x)| \geqslant (\lambda-M)\|x\|^2$$

and hence λ is a regular value of T. Similarly, if $\lambda<m$ then λ is also a regular value. $m, M\in\sigma(T)$ can be proved as in Theorem 4.6.1.1.

4.11.2. In the case of a compact T, the spectral representation has the form

$$T=\sum_{k=1}^{\infty}\lambda_k P_k \qquad \lambda_k\in[-\|T\|;\ +\|T\|]. \tag{$*$}$$

In the non-compact case the spectrum may fill the whole of $[m, M]$ as, for example, in the case of multiplication operators, and the analogue of $(*)$ should be

$$T=\int_m^M \lambda\,\mathrm{d}P(\lambda).$$

What is the meaning of this integral when $P(\lambda)$ is a projection operator, also including O and I, for every λ and how should we choose $\{P(\lambda);\ \lambda\in[m, M]\}$ for T?

Let us consider a partition

$$\lambda_0 = m \leqslant \lambda_1 < \lambda_2 < \ldots < \lambda_{n-1} \leqslant M = \lambda_n$$

of the interval $[m, M]$. For every such partition we have the 'Riemann sum'

$$\sum_{k=1}^{n}\lambda_k(P(\lambda_k)-P(\lambda_{k-1}))$$

which is a well-defined self-adjoint operator.

The usual definition for a *finer partition* is as follows:

$$\lambda_0 = m \leqslant \lambda_1' < \lambda_2' < \ldots < \lambda_{m-1}' \leqslant M = \lambda_n$$

* Obviously, M is the supremum and m is the infimum of the corresponding bounds.

is finer than the original one if $\{\lambda_k\}$ is a subset of $\{\lambda_k'\}$. Now we can give the following definition.

4.11.2.1 Definition. The Riemann sum

$$\sum_{k=1}^{n} \lambda_k (P(\lambda_k) - P(\lambda_{k-1}))$$

converges to the operator T if for every $\varepsilon > 0$ there is a partition of an interval $[m, M]$ such that for any finer partition $\{\lambda_k';\ k=0, 1, \ldots, m\}$,

$$\left\| T - \sum_{k=1}^{m} \lambda_k' (P(\lambda_k') - P(\lambda_{k-1}')) \right\| < \varepsilon.$$

In this case we define

$$T := \int_m^M \lambda \, \mathrm{d}P(\lambda).$$

Remark. There is also a strong version of this operator-valued integral:

$$Tx = \int_a^b \lambda \, \mathrm{d}P(\lambda) x$$

if for every $\varepsilon > 0$ and $x \in \mathscr{H}$ there is a division of $[m, M]$ such that for any finer division $\{\lambda_k'\}$,

$$\left\| Tx - \sum_{k=1}^{m} \lambda_k' (P(\lambda_k')x - P(\lambda_{k-1}')x) \right\| < \varepsilon.$$

In the case of compact T,

$$P(\lambda) = \sum_{k \leqq \lambda} P_k$$

where P_k is the projection onto the eigenspace N_{λ_k}, and it follows from the definition 4.11.2.1 of the operator-valued Riemann integral that

$$\sum_{k=1}^{\infty} \lambda_k P_k = \int_m^M \lambda \, \mathrm{d}P(\lambda)$$

i.e. in this case,

$$T = \int_m^M \lambda \, \mathrm{d}P(\lambda).$$

Thus, the problem of spectral representation for a non-compact self-adjoint operator T reads as follows.

Find $\{P(\lambda);\ -\infty<\lambda<+\infty\}$ such that

(i) if $\lambda<\mu$ then $P(\lambda)\leqslant P(\mu)$, and $P(m)=0$, $P(M)=I$;

(ii) $P(\lambda)T=TP(\lambda)$;

(iii) $T=\int\limits_{-\infty}^{+\infty}\lambda\,\mathrm{d}P(\lambda)$.

4.11.3. In the case of a compact self-adjoint T,

$$(T-\lambda I)P(\lambda)\leqslant 0\qquad \lambda\in[m,M].$$

In fact, in this case,

$$(T-\lambda I)P(\lambda)=\sum_{i=1}^{\infty}(\lambda_i-\lambda)P_i\sum_{\lambda_k\leqq\lambda}P_k=\sum_{\lambda_i\leqq\lambda}(\lambda_i-\lambda)P_i\leqslant 0$$

where $\{\lambda_i\}$ are the eigenvalues of T (with eigenspaces N_{λ_k}) and hence $P_iP_k=0$ for $i\neq k$. Similarly,

$$(T-\lambda I)(I-P(\lambda))=\sum_{\lambda_i>\lambda}(\lambda_i-\lambda)P_i\geqslant 0$$

since every projection operator is positive.

4.11.3.1 Theorem. If 4.11.2 (i) is satisfied and

$$(T-\lambda I)P(\lambda)\leqslant 0\qquad (T-\lambda I)(I-P(\lambda))\geqslant 0$$

for every $\lambda,\mu\in[m,M]$, then

$$P(\lambda)T=TP(\lambda)\qquad\text{and}\qquad T=\int\limits_{m}^{M}\lambda\,\mathrm{d}P(\lambda).$$

Proof. The commutation relation is an immediate consequence of

$$(T-\lambda I)P(\lambda)\leqslant 0.$$

In fact, the product $(T-\lambda I)P(\lambda)$ is then self-adjoint and the product of the self-adjoint operators $T-\lambda I$ and $P(\lambda)$ is self-adjoint if and only if

$$(T-\lambda I)P(\lambda)=P(\lambda)(T-\lambda I)$$

(see 2.10.2 (ii)); hence $P(\lambda)$ and T are also commutable.

For every partition of the interval $[m,M]$,

$$(T-\lambda_{i-1}I)(P(\lambda_i)-P(\lambda_{i-1}))=(T-\lambda_{i-1}I)(I-P(\lambda_{i-1}))(P(\lambda_i)-P(\lambda_{i-1}))\geqslant 0$$

and

$$(\lambda_iI-T)(P(\lambda_i)-P(\lambda_{i-1}))=(\lambda_iI-T)P(\lambda_i)(P(\lambda_i)-P(\lambda_{i-1}))\geqslant 0$$

since $P(\lambda_i)-P(\lambda_{i-1})$ is a projection operator, by 4.10.3.3, and the product of two commuting positive operators is also positive (see 4.9.3.1).

It follows that

$$T-\sum_{i=1}^{m}\lambda_{i-1}(P(\lambda_i)-P(\lambda_{i-1})) = \sum_{i=1}^{m}(T-\lambda_{i-1}I)(P(\lambda_i)-P(\lambda_{i-1})) \geqslant 0$$

$$\sum_{i=1}^{m}\lambda_i(P(\lambda_i)-P(\lambda_{i-1}))-T = \sum_{i=1}^{m}(\lambda_i I-T)(P(\lambda_i)-P(\lambda_{i-1})) \geqslant 0$$

and hence

$$\sum_{i=1}^{m}\lambda_{i-1}(P(\lambda_i)-P(\lambda_{i-1})) \leqslant T \leqslant \sum_{i=1}^{m}\lambda_i(P(\lambda_i)-P(\lambda_{i-1})).$$

Subtracting $\sum_{i=1}^{m}\ldots$ from each side, we obtain

$$0 \leqslant T-\sum_{i=1}^{m}\lambda_{i-1}(P(\lambda_i)-P(\lambda_{i-1}))$$

$$\leqslant \sum_{i=1}^{m}(\lambda_i-\lambda_{i-1})(P(\lambda_i)-P(\lambda_{i-1})).$$

Finally we shall show that

$$\Big\|\sum_{i=1}^{m}(\lambda_i-\lambda_{i-1})(P(\lambda_i)-P(\lambda_{i-1}))\Big\| \leqslant \max_i(\lambda_i-\lambda_{i-1})$$

and thus the theorem will be proved (see 2.14.43). In fact,

$$\Big(\sum_{i=1}^{m}(\lambda_i-\lambda_{i-1})(P(\lambda_i)-P(\lambda_{i-1}))x|x\Big) = \sum_{i=1}^{m}(\lambda_i-\lambda_{i-1})((P(\lambda_i)-P(\lambda_{i-1}))x|x)$$

$$\leqslant \max_i(\lambda_i-\lambda_{i-1})\Big(\sum_{i=1}^{m}(P(\lambda_i)-P(\lambda_{i-1}))x|x\Big) = \max_i(\lambda_i-\lambda_{i-1})\|x\|^2$$

since each member of the sum is positive and

$$\sum_{i=1}^{m}(P(\lambda_i)-P(\lambda_{i-1})) = I. \qquad (*)$$

Finally, we apply 2.14.43 once more.

Remark. It also turns out in the proof that the 'upper sum' and the 'lower sum' have the same limit if the Riemann sums converge to an operator.

4.11.4. It follows from the results of the previous subsection, for the spectral representation of a self-adjoint T we have to find a set of projectors $\{P(\lambda);$

$\lambda \in R\}$ with the property 4.11.2 (i), that the operator $T - \lambda I$ is negative on the subspace $P(\lambda)\mathscr{H}$ and positive on the orthogonal complement $(I - P(\lambda))\mathscr{H}$.

For this purpose we shall define the positive and negative parts of a self-adjoint operator. If $|T|$ is the square root of T^2, then

$$T^+ := \frac{1}{2}(T + |T|) \qquad T^- := \frac{1}{2}(T - |T|)$$

are the *positive* and *negative* parts of T.

Now let us introduce the notation

$$T_\lambda := T - \lambda I$$

and let P_λ be the projection onto the null space N_λ^+ of T_λ^+, i.e. onto $N_\lambda^+ = \{x\colon T_\lambda^+ x = \theta\}$; then 4.11.2 (i) is satisfied (see 4.13.26). We also have the following theorem.

4.11.4.1 Theorem. If P_λ is the projection onto

$$N_\lambda^+ = \{x\colon T_\lambda^+ x = \theta\}$$

then

$$T_\lambda P_\lambda \leqslant 0 \qquad T_\lambda(I - P_\lambda) \geqslant 0$$

and hence we have the spectral representation

$$T = \int_{-\infty}^{+\infty} \lambda \, \mathrm{d}P(\lambda).$$

Proof. It follows from 4.9.2.1 that $|T_\lambda|$ is the strong limit of polynomials of T^2 and hence

$$T_\lambda |T_\lambda| = |T_\lambda| T_\lambda.$$

First we shall show that

$$T_\lambda P_\lambda = T_\lambda^- \qquad \text{and} \qquad T_\lambda(I - P_\lambda) = T_\lambda^+$$

and secondly that $T_\lambda^- \leqslant 0$ and $T_\lambda^+ \geqslant 0$.

$$T_\lambda^+ T_\lambda^- = \frac{1}{4}(T + |T|)(T - |T|) = \frac{1}{4}(T^2 - |T|^2) = 0$$

and hence $T_\lambda^- x \in P_\lambda \mathscr{H}$ for every $x \in \mathscr{H}$, i.e.

$$P_\lambda T_\lambda^- x = T_\lambda^- x \qquad x \in \mathscr{H}. \tag{$*$}$$

It follows that $P_\lambda T_\lambda^-$ is self-adjoint since T_λ^- is self-adjoint and, from 2.10.2 (ii), that

$$P_\lambda T_\lambda^- = T_\lambda^- P_\lambda.$$

We conclude that

$$T_\lambda P_\lambda = (T_\lambda^+ + T_\lambda^-)P_\lambda = T_\lambda^- P_\lambda = P_\lambda T_\lambda^- = T_\lambda^-$$

and

$$T_\lambda(I-P_\lambda) = T_\lambda^+ + T_\lambda^- - T_\lambda^- P_\lambda = T_\lambda^+.$$

For the second part of the statement to be proved

$$P_\lambda |T_\lambda| = P_\lambda(T_\lambda^+ - T_\lambda^-) = -P_\lambda T_\lambda^- = -T_\lambda^-$$

and

$$|T_\lambda| P_\lambda = (T_\lambda^+ - T_\lambda^-)P_\lambda = -T_\lambda^- P_\lambda = -P_\lambda T_\lambda^- = -T_\lambda^-.$$

Hence $T_\lambda^- \leqslant 0$ since from 4.9.3.1 it follows that $P_\lambda |T_\lambda| \geqslant 0$. Similarly, $T_\lambda^+ \geqslant 0$ since

$$0 \leqslant |T_\lambda|(I-P_\lambda) = (T_\lambda^+ - T_\lambda^-)(I-P_\lambda) = T_\lambda^+ - T_\lambda^- + T_\lambda^- P_\lambda = T_\lambda^+.$$

$*$4.12 Functional calculus

The spectral representation discussed in the previous section can be extended to real-valued polynomials of a self-adjoint T and more generally to all bicommutants of T, and thus a useful isomorphism is established between the real-valued continuous functions on the spectrum $\sigma(T)$ and certain subalgebras of $B(\mathscr{H})$.

4.12.1. Let T be a bounded self-adjoint operator of $B(\mathscr{H})$ and if

$$q(\lambda) = \sum_{k=0}^{n} \alpha_k \lambda^k$$

then

$$q(T) := \sum_{k=0}^{n} \alpha_k T^k$$

where $T^0 := I$. Then we have the following.

4.12.1.1 Theorem.

$$T^k = \int_m^M \lambda^k \, \mathrm{d}P(\lambda) \qquad k = 0, 1, 2, \ldots.$$

More generally,

$$q(T) = \int_m^M q(\lambda) \, \mathrm{d}P(\lambda) \tag{$*$}$$

where $\{P(\lambda)\}$ is the set of projection operators defined in 4.11.4.1. Moreover,

$$\|q(T)\| \leqslant \sup\{|q(\lambda)|;\ \lambda \in [m, M]\}. \tag{$**$}$$

Proof. Since $P(\lambda_i)-P(\lambda_{i-1})$; $i=1, 2, \dots, m$ are projection operators,

$$\Big(\sum_{i=1}^{m} \lambda_i(P(\lambda_i)-P(\lambda_{i-1}))\Big)^k = \sum_{k=1}^{m} \lambda_i^k(P(\lambda_i)-P(\lambda_{i-1}))$$

and hence this is also valid for the limit. Expression (∗) then follows from the linearity of the integral.

$$\Big(\sum_{i=1}^{m} q(\lambda_i)(P(\lambda_i)-P(\lambda_{i-1}))x|x\Big) = \sum_{i=1}^{m} q(\lambda_i)((P(\lambda_i)-P(\lambda_{i-1}))x|x)$$

$$\leqslant \sup\{|q(\lambda)|;\ \lambda\in\sigma(T)\} \sum_{i=1}^{m} ((P(\lambda_i)-P(\lambda_{i-1}))x|x) = \|x\|^2 \sup|q(\lambda)|$$

(see also 4.11.3 (∗)). It follows that

$$\Big\|\sum_{i=1}^{m} q(\lambda_i)(P(\lambda_i)-P(\lambda_{i-1}))\Big\| \leqslant \sup\{|q(\lambda)|;\ \lambda\in\sigma(T)\}$$

by 2.14.43, and hence this is also valid for the limit of Riemann sums.

Now let $f=f(\lambda)$ be a continuous real-valued function on $[m, M]$. Then there exists a sequence of polynomials $q_n(\lambda)$ such that $q_n \to f$ uniformly on $[m, M]$, i.e.

$$\|q_n-f\|_\infty \to 0$$

and hence, by the inequality (∗∗), there is a self-adjoint operator $f(T)$ such that

$$\|q_n(T)-f(T)\| \to 0.$$

4.12.1.2 Theorem.

$$f(T) = \int_m^M f(\lambda)\, \mathrm{d}P(\lambda).$$

Remark. For any function f on $[m, M]$ this integral is defined as the limit of Riemann sums

$$\sum_{i=1}^{m} f(\lambda_i)(P(\lambda_i)-P(\lambda_{i-1}))$$

described in 4.11.2.1.

Proof. As in the previous theorem, we can show that

$$\Big|\sum_{i=1}^{m} (q(\lambda_i)-f(\lambda_i))((P(\lambda_i)-P(\lambda_{i-1}))x|x)\Big| \leqslant \|x\|^2 \|q-f\|_\infty$$

and hence

$$q_n(T) \to \int_m^M f(\lambda)\,\mathrm{d}P(\lambda)$$

in the operator norm.

By methods very similar to what has been used in the previous theorems, it is easy to prove the following.

4.12.1.3 Theorem. If $T \in B(\mathscr{H})$ is a self-adjoint operator with upper bound M and lower bound m, f and g are real-valued continuous functions on $[m, M]$ and α is a real number, then

(i) If $f(\lambda) \geqslant 0$ on $[m, M]$, then $f(T) \geqslant 0$;

(ii) $[\alpha f](T) = \alpha f(T)$;

(iii) $(f+g)(T) = f(T) + g(T)$;

(iv) $(fg)(T) = f(T)\, g(T)$;

(v) $f(T)$ is a bicommutant of T.

The functional calculus above can be extended to certain measurable functions and to complex-valued functions on $[m, M]$ (Riesz and Sz Nagy 1955, Nr 129). One of the important features of this extension is that every bicommutant of T can be given in the form $f(T)$ with such functions f.

4.12.2. The spectral representation of a self-adjoint operator can be applied to the solution of the equation

$$\lambda x - Tx = f \qquad (*)$$

also in the case of non-compact T.

If $p_n = p_n(t)$ is a sequence of polynomials uniformly convergent to $y = 1/t$ in $[m, M]$, then by the result of the previous subsection, $p_n(A)f$ tends to the solution of the equation $(*)$ when $A = \lambda I - T$ and λ is a regular value.

Example. Since

$$\frac{1}{t} = \frac{1}{1-(1-t)} = r\frac{1}{1-(1-rt)} \qquad r > 0 \qquad (**)$$

we have

$$\frac{1}{t} = r\sum_{k=0}^{\infty}(1-rt)^k \qquad \text{if } |1-rt| < 1$$

and hence the sequence of polynomials

$$p_n(t) := r \sum_{k=0}^{n} (1-rt)^k \qquad n = 0, 1, 2, \dots$$

tends to $1/t$ uniformly in every closed subinterval of $(0, 2/r)$.

By the result of the previous subsection, if $m>0$ and $M<2/r$ (i.e. $r<2/M$) then $\{p_n(A)f\}$ tends to $A^{-1}f$, the solution of $(*)$ if $A=\lambda I-T$ and λ is a regular value.

Remark. It is easy to verify that there is also a recursive form of the sequence $\{A_n\}$ of operators tending to $(\lambda I-T)^{-1}$:

$$A_0 = rI \qquad A_{n+1} = [I-r(\lambda I-T)]A_n + rI.$$

4.13 Problems and notes

∘**4.13.1.** Prove properties (i)–(v) of § 4.1.1.

∘**4.13.2.** Prove that if $\mathscr{H}$ is finite dimensional, $T\in B(\mathscr{H})$ and $Tx=\theta$ only if $x=\theta$, then

(*a*) T is onto

(*b*) T^{-1} is bounded.

4.13.3. It follows from the considerations in § 2.11.2 that if U is a unitary operator from one Hilbert space $\mathscr{H}_1$ onto another $\mathscr{H}_2$, then

(i) $\{x_n\}$ is convergent in $\mathscr{H}_1$ if and only if $\{Ux_n\}$ is convergent in $\mathscr{H}_2$;

(ii) $\mathscr{M}\subset\mathscr{H}_1$ is compact if and only if $U\mathscr{M}$ is compact;

(iii) $T\in B(\mathscr{H}_1)$ has an inverse $T^{-1}\in B(\mathscr{H}_1)$ if and only if

$$UTU^{-1} = UTU^* \in B(\mathscr{H}_2)$$

has an inverse

$$(UTU^*)^{-1} = UT^{-1}U^* \in B(\mathscr{H}_2).$$

∘**4.13.4.** Show that a linear operator of l^2 is a Hilbert–Schmidt operator if and only if there exists an infinite matrix $\mathbf{A}$ with elements a_{ik} obeying the condition

$$\sum_{i=1}^{\infty}\sum_{k=1}^{\infty} |a_{ik}|^2 < \infty$$

such that $Tx=\mathbf{A}\mathbf{x}$ in the sense of § 4.4.3.

∘**4.13.5.** Find an infinite matrix which represents a bounded linear operator of l^2 that is *not a Hilbert–Schmidt operator.*

∘**4.13.6.** Does there exist a compact unitary operator other than $T=0$?

∘**4.13.7.** Show that a linear operator T of a finite-dimensional space $\mathscr{H}$ is *positive* if and only if its matrix representation is *positive definite,* i.e.

$$\sum_{i=1}^{n}\sum_{k=1}^{n} a_{ik}x_i\bar{x}_k \geqslant 0$$

where $\{x_j\}$; $j=1, 2, ..., n$ is an n-tuple of complex numbers.

Give the condition for the kernel of an integral operator in $L^2[a, b]$ to be a positive operator.

∘**4.13.8.** Show that the multiplication operator

$$[Tf](t) := h(t)f(t) \qquad f\in L^2[a, b]$$

where $h=h(t)$ is a continuous function, is positive if and only if $h(t)\geqslant 0$ for every $t\in[a, b]$.

∘**4.13.9.** Show that Theorem 4.7.2.1 is also valid if T is a compact operator from $\mathscr{H}_1$ into another Hilbert space $\mathscr{H}_2$.

∘**4.13.10.** Show that every operator T in the form

$$Tx := \sum_{k=1} \alpha_k(x|x_k)\,y_k$$

where $\|x_k\|=\|y_k\|=1$ $(k=1, 2, ...)$ and x_k, y_k are given vectors and $\alpha_k\to 0$ is a compact operator.

4.13.11. Let D be an (unbounded) linear operator and let $\lambda_0\neq 0$ such that $(\lambda_0 I-D)^{-1}$ exists as a compact operator of $\mathscr{H}$. This is the case for many 'symmetric' differential operators. Prove that

$$Dy = \sum_{k=1}^{\infty} \alpha_k P_k y \qquad y\in\mathscr{D}(D)$$

where P_k; $k=1, 2, ...$ are the projectors onto the eigenspaces of D. How do you find $\{\alpha_k;\ k=1, 2, ...\}$?

∘**4.13.12.** Prove that every compact projector P has a finite-dimensional range.

4.13.13. Prove that the range of a compact operator is a separable subspace of $\mathscr{H}$ (in the case where $\mathscr{H}$ is non-separable).

4.13.14. Let $\mathscr{M}$ be a subset of a Hilbert space $\mathscr{H}$. Then $\mathscr{M}_\varepsilon \subset \mathscr{M}$ is called the *ε-net of* $\mathscr{M}$ if

(*a*) $\mathscr{M}_\varepsilon$ consists of a finite number of elements;

(*b*) for every $x \in \mathscr{M}$ there is an element $z \in \mathscr{M}_\varepsilon$ such that $\|x-z\| < \varepsilon$.

Prove that if $\mathscr{M}$ is compact, then we can find an ε-net $\mathscr{M}_\varepsilon \subset \mathscr{M}$ for every $\varepsilon > 0$.

4.13.15. Prove that the range of a compact operator T is 'almost finite dimensional' in the sense:

'For every $\varepsilon > 0$ there is a finite-dimensional subspace $\mathscr{M} \subset \mathscr{H}$ such that for every $x \in \mathscr{H}$,

$$\inf\{\|m - Tx\|;\ m \in \mathscr{M}\} < \varepsilon\|x\|.\text{'}$$

∘**4.13.16.** Prove that if $(AB)^{-1}$ and B^{-1} exist then A^{-1} also exists. Give A^{-1} in terms of B and $(AB)^{-1}$.

4.13.17. (*a*) Give the square root of a multiplication operator

$$[Tx](t) := h(t)x(t) \qquad x \in L^2(-\infty, +\infty)$$

where $h(t)$ is a continuous function with $h(t) \geq 0$ for every t.

(*b*) Give the square root of a positive definite matrix $\mathbf{A}$ (i.e. find the matrix $\mathbf{B}$ such that $\mathbf{B}^2 = \mathbf{A}$).

(*c*) Give the square root of a Hilbert–Schmidt operator in $L^2[a, b]$.

∘**4.13.18.** Show that if

$$T_1 \geq T_2 \geq \ldots \geq T_n \geq \ldots \geq B$$

for the self-adjoint operators B, T_n; $n = 1, 2, \ldots$ then the (numerical) sequence

$$\{(T_n x|x)\};\ x \in \mathscr{H} \qquad n = 1, 2, \ldots$$

is convergent.

4.13.19. If T, T_n, S, S_n; $n = 1, 2, \ldots$ are bounded linear operators of a Hilbert space $\mathscr{H}$ and

$$T_n x \to Tx \qquad S_n x \to Sx \qquad x \in \mathscr{H}$$

then

$$T_n S_n x \to TSx \qquad x \in \mathscr{H}.$$

This theorem follows from the inequality

$$\|T_n S_n x - TSx\| \leqslant \|T_n S_n x - T_n Sx\| + \|T_n Sx - TSx\|$$
$$\leqslant \|T_n\| \, \|S_n x - Sx\| + \|T_n Sx - TSx\|$$

since $\|T_n\|$ is bounded by the uniform boundedness principle (see the Appendix § A.2.2.1).

How can we prove the theorem without referring to the uniform boundedness?

4.13.20. If $T \geqslant 0$, then $T^n \geqslant 0$; $n=2, 3, \ldots$ also. In fact, for $n=2k$,

$$(T^{2k}x|x) = (T^k x|T^k x) \geqslant 0$$

and for $n=2k+1$,

$$(T^{2k+1}x|x) = (TT^k x|T^k x) \geqslant 0$$

since $T \geqslant 0$. It is easy to show that the sum of positive operators is positive. Hence if $T \geqslant 0$ then $p(T) \geqslant 0$ for every polynomial p *with positive coefficients.*

4.13.21. Prove that $T \leqslant I$ implies $T^2 \leqslant I$. Is it also true that in this case,

$$T^n \leqslant 1 \qquad n = 3, 4, \ldots?$$

∘**4.13.22.** Show that if there exists an inverse $S^{-1} \in B(\mathscr{H})$ for $S \in B(\mathscr{H})$ then there exists $M > 0$ such that

$$\|Sx\| > M\|x\|. \tag{$*$}$$

Find such an M.

4.13.23. Show that if 4.13.22 $(*)$ holds for a *self-adjoint* S, then the range of S is $\mathscr{H}$ (i.e. S is onto).

4.13.24. If

$$mI \leqslant A \leqslant MI \qquad m > 0$$

for the operator A, then $A^{-1} \in B(\mathscr{H})$ and $A^{-1} \geqslant 0$.

Proof. By 4.11.1.1, $A^{-1} \in B(\mathscr{H})$ exists since $A \geqslant mI$ $(m > 0)$. Moreover, *for* $\alpha > M$,

$$(\alpha I - A)^n \geqslant 0 \qquad n = 1, 2, \ldots$$

by 4.13.20 (or 4.9.3.1) and, applying the Neumann series expansion,

$$A^{-1} = [\alpha I - (\alpha I - A)]^{-1} = \sum_{k=0}^{\infty} \frac{1}{\alpha^{k+1}} (\alpha I - A)^k \geqslant 0.$$

∘**4.13.25.** A more precise statement about the operator A in 4.13.24 is the following:

$$\frac{1}{M} I \leqslant A^{-1} \leqslant \frac{1}{m} I.$$

Prove this!

4.13.26. Prove that if $P(\lambda)$ is the projection onto

$$\mathcal{N}_\lambda^+ := \{x\colon T_\lambda^+ x = \theta\}$$

then $P(\lambda) \leqslant P(\mu)$ for $\lambda < \mu$. (See 4.11.4.1.)

4.13.27. Prove that if $\mathbf{A}$ is an $n \times n$ matrix and $\lambda \in \sigma(\mathbf{A})$ then

$$|\lambda - a_{kk}| \leqslant \sum_{i \neq k} |a_{ik}| \qquad k = 1, 2, \ldots$$

(see 1.8.26).

∘**4.13.28.** Find the eigenvalues and the corresponding eigenvectors of a projector. Find the spectrum of a projector.

4.13.29. Let T be an operator with finite-dimensional range, i.e. T has the form

$$Tx := \sum_{k=1}^{n} (x|a_k)\, b_k$$

where $\{a_k, b_k;\ k=1, 2, \ldots, n\}$ are given vectors of $\mathcal{H}$. Find $a_k, b_k;\ k=1, 2, \ldots$ such that the spectrum $\sigma(T) = \{0\}$.

∘**4.13.30.** Does the Volterra operator

$$[Tf](t) := \int_0^t \Big(\sum_{k=1}^{n} a_k(t)\, b_k(\tau)\Big) f(\tau)\, d\tau$$

have a finite-dimensional range?

∘**4.13.31.** Let $\mathbf{A}$ be an $n \times n$ matrix with elements a_{ik} $(i, k=1, 2, \ldots, n)$. Let us consider the Hilbert space $\mathcal{H}_n$ of n-tuples of complex numbers with scalar product

$$(\mathbf{x}|\mathbf{y}) := \sum_{k=1}^{n} x_k \bar{y}_k$$

and the operator defined by

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \qquad (*)$$

Write the operator ($*$) in the form

$$\sum_{k=1}^{n} a_k \otimes b_k$$

where $a_k, b_k \in \mathscr{H}_n$.

○**4.13.32.** Prove that the operators in $B(\mathscr{H})$ with finite-dimensional range form a subalgebra of $B(\mathscr{H})$. Moreover, show that they form a (two-sided) ideal of $B(\mathscr{H})$.

○**4.13.33.** Prove that for any operator with finite-dimensional range,

$$\left\| \sum_{k=1}^{N} a_k \otimes b_k \right\| \leqslant \sum_{k=1}^{N} \|a_k\| \, \|b_k\|.$$

Moreover,

$$\|a \otimes b\| = \|a\| \, \|b\|$$

where the norm on the left-hand side is the operator norm.

4.13.34. Let T be defined as

$$T\{x_k\} = \{s_k x_k\} \qquad \{x_k\} \in l^2$$

where $\{s_k\}$ is a (finite or infinite) sequence of real or complex numbers.

(*a*) What is the condition for $\mathscr{D}(T) = l^2$?

(*b*) When is T bounded?

(*c*) When is T compact?

(*d*) Give the spectrum of T.

○**4.13.35.** Let T be a bounded self-adjoint operator. Prove that

(*a*) $T - \lambda I \leqslant 0$ if $\lambda > \|T\|$;

(*b*) $T - \lambda I \geqslant 0$ if $\lambda < -\|T\|$;

(*c*) $\limsup_n \|T^n\|^{1/n} = \|T\|$.

○**4.13.36.** Give a sufficient condition for C such that

$$A \geqslant B \Rightarrow AC \geqslant BC.$$

○**4.13.37.** If $B_n \geqslant 0$ and $\|B_n - B\| \to 0$, then $B \geqslant 0$. Prove this!

4.13.38. Find $g=g(\lambda)$ such that

$$T^{+} = \int_{m}^{M} g(\lambda)\,\mathrm{d}P(\lambda)$$

where $P(\lambda)$, T^{+}, m, M are as defined in § 4.11.

4.13.39. The operator T is called *closed* if $x_n \to x$ and $Tx_n \to y$ imply $x \in \mathscr{D}(T)$ and $Tx=y$.

Examples.

(i) Every $T \in B(\mathscr{H})$ is closed.

(ii) The differentiation operator in $L^2[0, 1]$ is closed. This follows from the following sequence of inequalities:

$$\int_0^1 \left|x_n(t)-x_n(0)-\int_0^t y(\tau)\,\mathrm{d}\tau\right|^2 \mathrm{d}t \leqslant \max\left\{\left|x_n(t)-x_n(0)-\int_0^t y(\tau)\,\mathrm{d}\tau\right|^2;\ t\in[0, 1]\right\}$$

$$= \max\left\{\left|\int_0^t \left(-\frac{\mathrm{d}}{\mathrm{d}\tau}x_n(\tau)-y(\tau)\right)\mathrm{d}\tau\right|^2;\ t\in[0, 1]\right\}$$

$$\leqslant \int_0^1 \left|\frac{\mathrm{d}}{\mathrm{d}\tau}x_n(\tau)-y(\tau)\right|^2 \mathrm{d}\tau$$

∘**4.13.40.** Let L be an *unbounded* linear operator with domain in the Hilbert space $\mathscr{H}$. Also, let $T \in B(\mathscr{H})$ be a right inverse of L, i.e.

$$LTx = x \qquad x \in \mathscr{H}.$$

Prove that L is a closed operator if the operator TL is bounded.

This is the case for many differential operators.

4.13.41. If $\mathscr{M}=\mathscr{S}\oplus\mathscr{N}$ for the linear subspaces $\mathscr{M}$, $\mathscr{S}$ and $\mathscr{N}$ of $\mathscr{H}$, then we write

$$\mathscr{S} = \mathscr{N}\ominus\mathscr{M} \qquad \text{and} \qquad \mathscr{N} = \mathscr{S}\ominus\mathscr{M}.$$

This means that

$x\in\mathscr{S}$ iff $x\in\mathscr{M}$ and $(x|n)=0$ for every $n\in\mathscr{N}$;
$x\in\mathscr{N}$ iff $x\in\mathscr{M}$ and $(x|s)=0$ for every $s\in\mathscr{S}$.

∘**4.13.42.** Prove that if $P_{\mathscr{N}}\mathscr{H}\supset P_{\mathscr{M}}\mathscr{H}$, then $P_{\mathscr{N}}P_{\mathscr{M}}=P_{\mathscr{M}}P_{\mathscr{N}}$, where $P_{\mathscr{N}}$, $P_{\mathscr{M}}$ are projection operators.

4.13.43. Let U_r be the right shift in l^2:

$$U_r\{\xi_k\} = \{0, \xi_1, \xi_2, \dots, \xi_{k+1}, \dots\}.$$

We shall show that $\lambda=1$ is in the spectrum of U_r. For every $x:=\{\xi_k\}\in l^2$,

$$(I-U_r)x = \{\xi_1, \xi_2-\xi_1, \dots, \xi_k-\xi_{k-1}, \dots\}$$

and hence, if $y:=\{\eta_k\}$ is in the range of $I-U_r$, then

$$\sum_{k=1}^{\infty} \eta_k = 0 \tag{$*$}$$

and so *the range is a proper subspace of* l^2. Moreover, if $(I-U_r)x=\theta$ then $x=\theta$ and hence the inverse $(I-U_r)^{-1}$ exists as a possibly unbounded, *not* everywhere defined operator. If

$$\xi_k = \begin{cases} 1 & \text{for } k \leqslant N \\ 0 & \text{for } k > N \end{cases}$$

then

$$\eta_k = \begin{cases} 1 & \text{for } k=1 \\ -1 & \text{for } k = N+1 \\ 0 & \text{elsewhere} \end{cases} \tag{$**$}$$

for $y=(I-U_r)x$. It follows that $(I-U_r)^{-1}$ is unbounded since $(I-U_r)^{-1}y=x$, $\|y\|=\sqrt{2}$ and $\|x\|=N^{1/2}$, i.e. $(I-U_r)^{-1}$ sends a bounded set into an unbounded one.

Finally, we shall show that the range of $I-U_r$ is dense in l^2. Let $y\in l^2$ satisfying ($*$) and $y_N\in l^2$ satisfying ($**$) for $N=1, 2, \dots$. Then, by easy calculation,

$$\|y+\sum_{k=2}^{N} \eta_k y_k\|^2 = |\sum_{k=1}^{N} \eta_k|^2 + \sum_{k=N+1}^{\infty} |\eta_k|^2$$

and the right-hand side can be as small as we please if N is large enough since ($*$) holds and $y\in l^2$.

Thus we have proved that the range of $I-U_r$ is dense in the subspace satisfying ($*$).

Now let $z\in l^2$ be arbitrary, and let $\varepsilon>0$ and $M>0$ be sufficiently large that

$$\frac{1}{M}|\sum_{k=1}^{N} \zeta_k|^2 < \varepsilon$$

where $z:=\{\zeta_k\}$. Then $y:=\{\eta_k\}$, where

$$\eta_k = \begin{cases} \zeta_k & \text{for } k \leqslant N \\ -\alpha/M & \text{for } N < k \leqslant M \\ 0 & \text{for } M \leqslant k \end{cases}$$

has the property (*) if

$$\alpha = \sum_{k=1}^{N} \xi_k.$$

Moreover,

$$\|z-y\|^2 = \frac{M-N}{M^2}|\alpha|^2 + \sum_{k>M} |\zeta_k|^2$$

and the right-hand side can be as small as we please if M is large enough.

It can be shown by similar calculations that $\lambda=-1$ is also in the spectrum of U_r.

○**4.13.44.** Let P^t, P^r, P^s be projection operators. Prove that

$$P^t(P^r-P^s) = \begin{cases} P^r-P^s & \text{if } P^t \geqslant P^r \geqslant P^s \\ 0 & \text{if } P^t \leqslant P^r \leqslant P^s. \end{cases}$$

What happens when $P^r \geqslant P^t \geqslant P^s$?

4.13.45. A linear operator T of a finite-dimensional Hilbert space always has an eigenvalue and hence it has a (one-dimensional) invariant subspace (see § 4.1.1). The existence of a closed invariant subspace for every linear operator of an *infinite-dimensional Hilbert space* $\mathscr{H}$ has not yet been proved.

Now, let us suppose that a linear operator T of the infinite-dimensional Hilbert space $\mathscr{H}$ has a property which implies the existence of a (closed) invariant non-trivial subspace $\mathscr{M} \subset \mathscr{H}$ for T. Moreover, suppose that this property is inherited by the restriction of T onto $\mathscr{M}$. (Such a property might be, for example, that T is compact or that T is self-adjoint.) Let $P_{\mathscr{M}}$ be the projector onto the (closed) invariant subspace $\mathscr{M}$. Then the set of invariant subspaces has the following properties.

(i) For every $\mathscr{M}$ there exists an invariant subspace $\mathscr{N}$ such that

$$I-P_{\mathscr{M}} \leqslant I-P_{\mathscr{N}}.$$

(ii) If $(I-P_{\mathscr{M}})z=\theta$ for every $\mathscr{M}$, then $z=\theta$.

(Property (ii) is equivalent to saying that the common part of the invariant subspaces is $\{\theta\}$.)

5

Causal Operators

A fundamental principle of input–output systems is that the output depends only on the 'past' of the input. This is called the causality principle. The transition operators of systems satisfying the causality principle are called causal operators. The object of this chapter is to present a mathematical theory for causal operators. We shall investigate the fundamental properties of causal operators and the connections with other possible properties such as time invariance, passivity and stability.

5.1 Causal operators in L^2-spaces

The values of a 'time function' $x=x(t)$ could be real or complex numbers, matrices, functions etc, but in this section the values are supposed to be real or complex numbers only, since the formalism is simplified but all the main problems remain the same in this case.

5.1.1. The mathematical expression of the causality principle is as follows. Let Ω be one of the following 'time structures':

(*a*) the real line;
(*b*) the $t \geqslant 0$ half of the real line;
(*c*) the natural numbers;
(*d*) $0, \pm 1, \pm 2, \ldots$.

5.1.1.1 Definition. Let X be a linear space of functions defined on Ω and let T be an operator of X. Then T is causal if from

$$f(\tau) = g(\tau) \qquad \text{for} \quad \tau < t$$

it follows that

$$[Tf](\tau) = [Tg](\tau) \qquad \text{for} \quad \tau < t$$

where t is any fixed point in Ω.

Notice that causality is defined also for unbounded or even non-linear operators.

Let us define

$$E^t f := \begin{cases} f(\tau) & \text{for} \quad \tau < t \\ 0 & \text{for} \quad \tau \geqslant t \end{cases} \tag{$*$}$$

and $E_t = I - E^t$. $\{E^t;\ t \in \Omega\}$ are called *truncation operators.*

Remark. E^t is also represented by multiplication with

$$e_t(\tau) = \begin{cases} 1 & \text{for} \quad \tau < t \\ 0 & \text{for} \quad \tau \geqslant t. \end{cases}$$

It is easy to show that $\{E^t;\ t \in \Omega\}$ and $\{E_t;\ t \in \Omega\}$ are *projection operators.* Moreover, T is causal if and only if from

$$E^t f = E^t g$$

it follows that

$$E^t T f = E^t T g.$$

5.1.1.2 Theorem. Each of the following properties for a linear operator T of $\mathscr{H} = \mathrm{L}^2(\Omega)$ is equivalent to the causality of T:

(i) $E^t T = E^t T E^t$;

(ii) $T E_t = E_t T E_t$;

(iii) $E_t \mathscr{H} := \{E_t x;\ x \in \mathscr{H}\}$ is an invariant subspace of T for every $t \in \Omega$.

Proof. Let T be causal. It follows from the definition of E^t that $E^t(I - E^t) = 0$, which means that

$$E^t(I - E^t)x = 0 \qquad x \in \mathscr{H}$$

and hence

$$E^t T(I - E^t)x = 0 \qquad x \in \mathscr{H}$$

since T is causal. But the latter equality is exactly the same as (i).

Conversely, if (i) is satisfied and $E^t x = E^t y$, then

$$E^t T x = E^t T E^t x = E^t T E^t y = E^t T y$$

which means exactly that T is causal.

Thus we have proved that (i) is satisfied if and only if T is a causal operator.

For the equivalence of (i) and (ii), substitute $E^t = I - E_t$ in (i) and $E_t = I - E^t$ in (ii). For the equivalence of (ii) and (iii) we refer to 4.10.1.1.

The most important examples of causal operators are the following.

Example 1. In the case where Ω is $\{t: t \geqslant 0\}$ or the real line, every Volterra operator in $L^2(\Omega)$ is causal. Indeed, if $f(\tau)=0$ for $\tau<t$ then

$$\int_0^t K(t,\tau)f(\tau)\,d\tau = 0.$$

Example 2. The Volterra operators in the discrete case, i.e. when Ω is a set of integers, are those operators that are represented by a lower triangular matrix

$$a_{ij} = 0 \qquad \text{for} \quad i < j.$$

Indeed, in this case,

$$y_i = \sum_{k=-\infty}^{i} a_{ik}x_k = 0 \qquad \text{if} \quad x_k = 0, \quad \text{for} \quad k \leqslant i.$$

Example 3. The (right) translation operators

$$U_{t_0}f := f(\tau - t_0) \qquad t_0 > 0$$

and hence the operators in the form

$$\sum_{k=1}^{N} \alpha_k U_{t_k}$$

are also causal. In fact, if $f(\tau)=0$ for $\tau \leqslant t$ then

$$f(t_0-\tau) = 0 \qquad \text{for} \quad \tau - t_0 \leqslant t, \quad \text{i.e.}$$

$$f(t_0-\tau) = 0 \qquad \text{for} \quad \tau < t + t_0.$$

Example 4. For any bounded function $a=a(t)$, the multiplication operator

$$[T_a f](t) := a(t)f(t)$$

is obviously causal.

Most of the above examples are characteristic in the sense that an operator represented by a finite or infinite matrix is causal if and only if the matrix is lower triangular and an integral operator is causal if and only if it is a Volterra operator. (As an easy exercise, prove these!)

5.1.2. Some of the important classes of operators are 'automatically' causal.

5.1.2.1 Definition. The operator T is called *time invariant* if

$$U_t T = T U_t \qquad t \in \Omega$$

where U_t is the translation operator

$$[U_t x](\tau) := x(\tau - t).$$

U_t is an isometry, i.e. $\|U_t x\| = \|x\|$ for every $x \in L^2(\Omega)$ and, in the cases 5.1.1(*a*) and (*d*), U_t is a unitary operator. A fundamental connection between the truncation operators $\{E_t;\ t \in \Omega\}$ and $\{U_t;\ t \in \Omega\}$ is

$$E_{t+\tau} U_t = U_t E_\tau \qquad t, \tau \in \Omega \tag{$*$}$$

which is easy to verify.

5.1.2.2 Theorem. If T is time invariant and

$$\{E_{t_0} x;\ x \in L^2(\Omega)\}$$

is an invariant subspace of T, then $\{E_t x;\ x \in L^2(\Omega)\}$ is also an invariant subspace for $t < t_0$.

If $\{U_t;\ t \in \Omega\}$ consists of unitary operators and hence U_t^{-1} exists (in this case $U_t^{-1} = U_{-t}$), then $\{E_t x;\ x \in L^2(\Omega)\}$ is an invariant subspace for every $t \in \Omega$ and hence T is causal.

Proof. It follows from the (ii)$\Leftrightarrow$(iii) part of Theorem 5.1.1.2 that

$$T E_{t_0} = E_{t_0} T E_{t_0}.$$

Multiplying both sides *from the right* by U_s and applying ($*$) for E_{t_0} and U_s, we obtain

$$U_s T E_{t_0 - s} = U_s E_{t_0 - s} T E_{t_0 - s}$$

if we are careful enough to ensure that $t_0 - s \in \Omega$. Since U_s has a left inverse in any case, the proof is complete.

5.1.2.3 Definition. The operator T in $\mathscr{H} = L^2(\Omega)$ is *passive* if

$$(E^t x | T x) + (T x | E^t x) \geqslant 0.$$

Remark. For the case where Ω is the real line, this means that

$$\operatorname{Re}\Big(\int_{-\infty}^{t} x(\tau)[Tx](\tau)\, d\tau \Big) \geqslant 0$$

and, in any case, the above definition has the following meaning: 'the energy supplied by the transition $x \to Tx$ is positive'.

5.1.2.4 Theorem. If the linear operator T of $\mathscr{H} = L^2(\Omega)$ is passive then it is also causal.

Proof. In this case the bilinear functional

$$B_T(f, g) := (E^t f|Tg)+(Tf|E^t g)$$

is positive and hence, by 2.9.2, the Cauchy inequality

$$|B_T(f, g)|^2 \leq B_T(f, f)B_T(g, g) \qquad (*)$$

is valid. Thus if $E^t f=\theta$, then $B_T(f,f)=0$ and hence $B_T(f, g)=0$ for every $g \in L^2(\Omega)$. This means that

$$(E^t Tf|g) = (Tf|E^t g) = (E^t f|Tg)+(Tf|E^t g) = 0$$

for every $g \in L^2(\Omega)$ and hence $E^t Tf=\theta$, i.e. T is causal.

5.1.3. The set of causal operators in $L^2(\Omega)$ form a closed subalgebra of $B(\mathscr{H})$. In fact, if K and L are causal operators then

$$E^t (KL) = (E^t K) L = (E^t KE^t) L = E^t K (E^t LE^t) = (E^t KE^t) LE^t = E^t KLE^t$$

and hence KL is also causal. Similarly, $\alpha K+\beta L$ is causal for any scalars α, β and, if $\{K_n;\ n=1, 2, \ldots\}$ is a sequence of scalar operators and $K_n \to K$ in operator norm (or even strongly!), then

$$E^t K = E^t KE^t \qquad t \in \Omega$$

since in this case

$$E^t K_n x \to E^t Kx \qquad \text{and} \qquad E^t K_n E^t x \to E^t KE^t x$$

for every $x \in L^2(\Omega)$.

However, the adjoint T^* of a causal operator is not causal in general and the same holds, if it exists, for the inverse operator T^{-1}. Some of the important problems of causal operators originate from these 'instabilities'.

Let us begin with the adjoint operator T^*. The adjoint of a lower triangular matrix is an upper triangular one. Moreover, we proved in 4.10.1.2 that if $\mathscr{M}$ is an invariant subspace of T then the orthogonal complement of $\mathscr{M}$ is an invariant subspace of T^*. Motivated by these considerations, we have the following definition.

5.1.3.1 Definition. Let X be a linear space of functions defined on Ω and let T be an operator of X. Then T is called *anticausal* if from

$$f(\tau) = g(\tau) \qquad \text{for} \quad \tau > t$$

it follows that

$$[Tf](\tau) = [Tg](\tau) \qquad \text{for} \quad \tau > t$$

where t is any fixed point in Ω.

The counterpart of Theorem 5.1.1.2 is the following.

5.1.3.2 Theorem. Each of the following properties for a linear operator T of $\mathscr{H} = L^2(\Omega)$ is equivalent to the anticausality of T:

(i) $E_t T = E_t T E_t$;
(ii) $TE^t = E^t TE^t$;
(iii) $E^t \mathscr{H} := \{E^t x;\ x \in \mathscr{H}\}$ is an invariant subspace of T for every $t \in \Omega$.

If we interchange E_t and E^t in the proof of 5.1.1.2 then we obtain the proof of the above theorem.

If T is causal, then T^* is anticausal and vice versa. The diagonal matrices represent those operators for a 'discrete' Ω (i.e. if $\Omega = \{0, \pm 1, \pm 2, \ldots\}$ or a subset of it) that are both causal and anticausal. These operators are called *memoriless*. The multiplication operator of Example 4 is also memoriless.

It follows from these definitions that the *causal self-adjoint operators* are memoriless operators.

5.1.4. It is easy to verify by an algorithm giving the inverse matrix that the inverse of a lower triangular matrix is also lower triangular. The situation is quite different in the general L^2-case. The most simple example of this is the translation operator U_t in $L^2(-\infty, +\infty)$, when $U_t^{-1} = U_{-t} = U_t^*$ is an anticausal operator.

The block diagram of the simplest feedback system is shown in figure 5.1; it is characterised by the following system of equations:

$$y = Kh$$
$$h = Fy + u$$

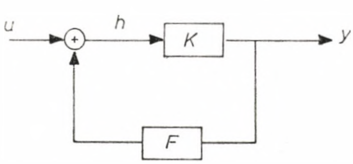

fig. 5.1

where $y, h, u \in L^2(\Omega)$ and K, F are bounded linear operators of $L^2(\Omega)$. It follows that

$$(I - FK)h = u$$
$$(I - KF)y = Ku$$

and hence, h and y are uniquely determined by the input u if and only if $I - FK$ and $I - KF$, respectively, have a left inverse. By the causality principle we claim that the transition operators $u \to y$ and $u \to h$ should be also causal, and this is satisfied when $I - FK$ and $I - KF$, respectively, have a causal inverse.

Now we shall show that there is a class $S(\mathscr{H})$ of causal operators such that $F, K \in S(\mathscr{H})$ guarantees that $(I-FK)^{-1}$ and $(I-KF)^{-1}$ are causal operators. These operators will be called *strictly causal.*

5.1.4.1 Definition. A causal operator T is called strictly causal if for every $\varepsilon > 0$ there exists a partition of Ω such that for any finer partition,

$$t_1 < t_2 < \ldots < t_m$$

$$\|\Delta^i T \Delta^i\| < \varepsilon$$

where

$$\Delta^i := E^{t_i} - E^{t_{i-1}} \qquad i = 1, 2, \ldots, m.$$

Roughly speaking, if the transition operator of a system is strictly causal, then the value $y(t)$ of the output depends only on the 'strict' past of the input.

Example 1. Every Hilbert–Schmidt Volterra operator is strictly causal. In fact, let Ω be the real line and let T be a Hilbert–Schmidt Volterra operator of $L^2(\Omega)$; then

$$Ty := \int_{-\infty}^{t} k(t, \tau) y(\tau)\, d\tau$$

with

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |k(t, \tau)|^2\, dt\, d\tau < \infty. \tag{$*$}$$

In this case,

$$[\Delta^i f](t) = (E^{t_i} - E^{t_{i-1}}) f(t) = \begin{cases} f(t) & \text{if } t_{i-1} \leq t < t_i \\ 0 & \text{elsewhere} \end{cases}$$

$$[T \Delta^i f](t) = \begin{cases} 0 & \text{for } t \leq t_{i-1} \\ \int_{t_{i-1}}^{t} k(t, \tau) f(\tau)\, d\tau & \text{for } t_{i-1} < t < t_i \\ \int_{t_{i-1}}^{t_i} k(t, \tau) f(\tau)\, d\tau & \text{for } t_i \leq t \end{cases}$$

$$[\Delta^i T \Delta^i f](t) = \begin{cases} 0 & \text{for } t \leq t_{i-1} \\ \int_{t_{i-1}}^{t} k(t, \tau) f(\tau)\, d\tau & \text{for } t_i < t \leq t_{i-1} \\ 0 & \text{for } t_i < t. \end{cases}$$

Hence

$$\|\Delta^i T \Delta^i\| \leq \|\Delta^i T \Delta^i\|_2 = \int_{t_{i-1}}^{t_i} \int_{t_{i-1}}^{t_i} |k(t, \tau)|^2\, dt\, d\tau$$

(for the meaning of $\|\,.\,\|_2$ in this case, see § 4.8.1).

Now, considering $(*)$, the right-hand side of this inequality can be as small as we please if the partition is fine enough.

Example 2. If $\Omega=\{0, 1, 2, \dots, n\}$ and T is an operator represented by an $n\times n$ matrix, then T is strictly causal if and only if

$$a_{ik} = 0 \qquad \text{for} \quad i \leqslant k$$

i.e. the $n\times n$ matrix is lower triangular *with zero diagonal elements.*

In fact, in this case *there is no finer partition than* $t_i=i$ $(i=1, 2, \dots, n)$ and

$$\Delta^i \mathbf{x} = [E^i - E^{i-1}]\mathbf{x} = x_i \mathbf{e}_i$$

where x_i is the ith member of the infinite sequence $\mathbf{x}\in l^2$ and $\mathbf{e}_i\in l^2$ with 1 in the ith position and 0 in all other positions. Moreover,

$$T\Delta^i \mathbf{x} = x_i \mathbf{t}_i$$

where t_i is the ith column of the matrix $\mathbf{T}$ and, finally,

$$\Delta^i T \Delta^i x = x_i t_{ii} \mathbf{e}_i.$$

Hence

$$\|\Delta^i T \Delta^i x\| = |x_i|\,|t_{ii}|.$$

It follows that T is strictly causal if and only if $t_{kk}=0$ $(k=1, 2, \dots, n)$, i.e. the diagonal elements of the matrix $\mathbf{T}$ are 0.

Example 3. The multiplication operator on $L^2(-\infty, +\infty)$, which is both causal and anticausal, is not strictly causal. To show this, let

$$[Tf](t) = h(t) f(t) \qquad f\in L^2(-\infty, +\infty)$$

where h is a bounded continuous function:

$$\Delta^i T \Delta^i f = \begin{cases} h(t) f(t) & \text{if} \quad t\in[t_{i-1}, t_i] \\ 0 & \text{elsewhere} \end{cases}$$

by a computation similar to that in the previous example. Hence for every $\varepsilon>0$,

$$\|\Delta^i T \Delta^i f\|_2 = \Big(\int_{t_{i-1}}^{t_i} |h(t) f(t)|^2 \, \mathrm{d}t \Big)^{1/2} < \varepsilon$$

if the partition is fine enough. However, *this does not hold for the norm of the operator* $\Delta^i T \Delta^i$. In fact, let

$$t_i = i/n \qquad h(t_i) \neq 0$$

and

$$g_n(t) = \begin{cases} n^{1/2} & \text{if } t \in [(i-1)/n, i/n] \\ 0 & \text{elsewhere.} \end{cases}$$

Then $\|g_n\|_2 = 1$ and

$$\|\Delta^i T \Delta^i g_n\|_2^2 = n \int_{(i-1)/n}^{i/n} |h(t)|^2 \, \mathrm{d}t > \inf |h(t)|^2 > \frac{1}{2} |h(t_i)|^2$$

if n is large enough, since $h=h(t)$ is a continuous function. On the other hand, for every partition of the interval [0, 1] we can find a finer partition with $t_i = i/n$.

The main theories concerning strictly causal operators, including causal invertibility, will be discussed in § 5.3.

5.2 Causality in a Hilbert space

The material in the previous section has perhaps been presented in a more abstract form than is strictly necessary; however, in this abstract form there is an immediate extension of the theorems for causal operators to more general Hilbert spaces.

5.2.1. Let us consider a general Hilbert space $\mathscr{H}$ instead of $L^2(\Omega)$ and a one-parameter set $\{P^t;\ t \in \Lambda\}$ of projection operators of $\mathscr{H}$ totally ordered in the sense that for every pair P^s, P^t,

$$\text{either} \qquad P^s \leqslant P^t \qquad \text{or} \qquad P^t \leqslant P^s$$

instead of the truncation operators defined in 5.1.1 (∗).

5.2.1.1 Definition. The linear operator T of $\mathscr{H}$ is *causal with respect to* $\{P^t;\ t \in \Lambda\}$ if from

$$P^t x = P^t y \qquad t \in \Lambda, \quad x, y \in \mathscr{H}$$

it follows that

$$P^t T x = P^t T y.$$

It is easy to check that with the exception of 5.1.2.1 and 5.1.2.2, everything that we have said about causal operators in § 5.1 remains valid in this more general case.

5.2.2. Let us consider an RKHS $\mathscr{H}(R)$ with $\mathscr{D} = \Omega$. Then the causality, defined by 5.1.1.1, can be expressed by the kernel $R = R(s, t)$ as follows.

The operator T of $\mathscr{H}(R)$ is causal if from

$$(f(\,.\,)|R(\,.,t)) = (g(\,.\,)|R(\,.,t)) \qquad \text{for} \quad t \leqslant s$$

it follows that

$$([Tf](\,.\,)|R(\,.,t)) = ([Tg](\,.\,)|R(\,.,t)) \qquad \text{for} \quad t \leqslant s$$

where s is any fixed point in Ω.

5.2.2.1 Theorem. A linear operator T is causal if and only if

$$\{R(\,.,t);\ t \leqslant s\}^{\perp}$$

i.e. the orthogonal complement of each linear space generated by $\{R(\,.\,,t);$ $t \leqslant s\}$ is an invariant subspace of T.

Proof. Let the linear operator T be causal and let $f(t)=0$ for every $t \leqslant s$; then $f \in \{R(\,.\,,t);\ t \leqslant s\}^{\perp}$ and $Tf(t)=0$ for every $t \leqslant s$; hence we also have $Tf \in \{R(\,.\,,t),\ t \leqslant s\}^{\perp}$.

Conversely, if $\{R(\,.\,,t);\ t \leqslant s\}^{\perp}$ for every s is an invariant subspace of the linear operator T and $f(t)=g(t)$ for $t \leqslant s$, then $f-g \in \{R(\,.\,,t);\ t \leqslant s\}^{\perp}$ and hence $T(f-g)=Tf-Tg \in \{R(\,.\,,t);\ t \leqslant s\}^{\perp}$, i.e.

$$\begin{aligned} 0 &= (Tf(.)-Tg(.)|R(.,t)) = (Tf(.)|R(.,t))-(Tg(.)|R(.,t)) \\ &= (Tf)(t)-[Tg](t) \qquad t \leqslant s. \end{aligned}$$

Now, let P^s be the projection operator onto

$$\overline{\{R(.,t);\ t \leqslant s\}}$$

i.e. onto the closed subspace generated by $\{R(\,.\,,t);\ t \leqslant s\}$. Then it follows from 5.2.1.1 and 5.1.1.2, that *the linear operator T of an* RKHS *$\mathscr{H}(R)$ is causal* (in the sense of 5.1.1.1) *if and only if it is causal with respect to P^s; $s \in \Omega$.*

Remark. 5.2.2.1 can also be derived from 5.1.1.2.

5.2.2.2 Theorem. $\{P^s;\ s \in \Omega\}$ has the following properties:

(i) $[P^s f](t) = f(t)$ for $t \leqslant s$;

(ii) If $g \in \mathscr{H}(R)$ and $g(t) = f(t)$ for $t \leqslant s$, then $\|g\| \geqslant \|P_s f\|$;

(iii) $f(t) = g(t)$ for $t \leqslant s$ if and only if $P^s f = P^s g$.

Proof.

$$P_s f := f - P^s f \qquad \text{and} \qquad (f(.)-P^s f(.)|R(.,t)) = 0 \qquad \text{if} \quad s \geqslant t$$

by the definition of P^s; hence (i) is satisfied. By applying the projection principle (2.4.1.3 or 2.5.4.2), we obtain (ii). The proof of (iii) is left to the reader as a simple exercise.

Remark. If $f \in L^2(\Omega)$, then the truncated f also belongs to $L^2(\Omega)$, but this is not so in an RKHS. If the kernel is continuous, then for every $f \in \mathscr{H}(R)$ $(f \neq \theta)$, many t can be found such that $E^t f \notin \mathscr{H}(R)$. Hence *we cannot define the causality by the truncation operators in an* RKHS *with continuous kernel.* However, 5.2.2.2 (i)–(iii) are common properties for $L^2(\Omega)$ and $\mathscr{H}(R)$ with $\mathscr{D} = \Omega$.

Now we can say something about the form of a causal operator in an RKHS. If T is a linear operator of $\mathscr{H}(R)$, then

$$[Tf](t) = (Tf(.)|R(., t)) = (f(.)|T^*R(., t))$$

and hence T is represented by the scalar product with

$$\mathscr{T}(., t) := T^* R(., t)$$

called the *kernel of the operator T.* It follows that T is causal if and only if

$$f \in \{R(., t);\ t \leqslant s\}^{\perp} \Rightarrow Tf \in \{\mathscr{T}(., t);\ t \leqslant s\}^{\perp}.$$

Hence

$$\{R(., t);\ t \leqslant s\}^{\perp} \subseteq \{\mathscr{T}(., t);\ t \leqslant s\}^{\perp}$$

and, from the definition of the orthogonal complement,

$$\{\mathscr{T}(., t);\ t \leqslant s\} \subseteq \overline{\{R(., t);\ t \leqslant s\}}$$

for *every* $s \in \Omega$.

Example 1. Let us consider the RKHS $\mathscr{H}(S)$ in Example 3 of § 3.3.1, with kernel

$$R(s, t) = \int_0^1 (t-\tau)_+ (s-\tau)_+ \, d\tau.$$

We compute $P^s f$ for $f \in \mathscr{H}(S)$ and the condition for the kernel of a linear operator of $\mathscr{H}(S)$ to be causal.

Recall that $\mathscr{H}(S)$ consists of functions f on [0, 1] with

$$f'' \in L^2[0, 1] \qquad f(0) = f'(0) = 0$$

and

$$\|f\| = \Big(\int_0^1 |f''(t)|^2 \, dt\Big)^{1/2}.$$

Hence

$$\frac{d^2}{dt^2} P^s f \in L^2[0, 1]$$

and

$$\left[\frac{d^2}{dt^2} P^s f\right](t) = \left[\frac{d}{dt^2} f\right](t) \qquad \text{for} \quad t \leqslant s$$

since

$$[P^s f](t) = f(t) \qquad \text{for} \quad t \leqslant s$$

and

$$\left[\frac{d^2}{dt^2} P^s f\right](t) = 0 \qquad \text{for} \quad t > s \tag{$*$}$$

by the minimum property (ii) of $P^s f$ in Theorem 5.2.2.2. Now, applying the identity

$$\int_0^1 (t-\tau)_+ f''(\tau)\, d\tau = \int_0^t (t-\tau) f''(\tau)\, d\tau = \int_0^t f'(\tau)\, d\tau = f(t)$$

and

$$\frac{d}{dt} P^s f = \begin{cases} f'(t) & \text{if} \quad t \leqslant s \\ f'(s) & \text{if} \quad t > s \end{cases}$$

since $[P^s f]'$ is continuous and $[P^s f]''(t) = 0$ for $t > s$; we have

$$[P^s f](t) = \int_0^s f'(\tau)\, d\tau + \int_s^t f'(s)\, d\tau = f(s) + (t-s) f'(s)$$

if $s < t$ and $[P^s f](t) = f(t)$ if $s \geqslant t$ on the basis of property (i) in Theorem 5.2.2.2.

Remark. It is easy to see that the above result also derives from the fact that the operators $\{P^s;\ s \in [0, 1]\}$ are unitarily equivalent to the truncation operators $\{E^s;\ s \in [0, 1]\}$ of $L^2[0, 1]$.

If $\mathscr{T}(s, t)$ is the kernel of a linear operator T of $\mathscr{H}(S)$, then from the definition of scalar product in $\mathscr{H}(S)$,

$$(f(.)|\mathscr{T}(., t)) = \int_0^1 \frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t) \overline{f''(\tau)}\, d\tau.$$

We shall show that

$$\frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t) = 0 \qquad \text{for} \quad \tau > t. \tag{$**$}$$

In fact, it follows from $(*)$ and 5.1.1.2 that if T is causal then

$$g(t) = 0 \quad \text{for} \quad t < s \Rightarrow \int_0^1 \left(\frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t)\right) \overline{g(\tau)}\, d\tau = 0 \qquad \text{for} \quad t < s, \quad g \in L^2[0, 1]$$

i.e. for $t<s$,

$$\int_s^1 \left(\frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t)\right) \overline{g(\tau)}\, d\tau = 0 \qquad g \in L^2[0, 1], \quad t \in [0, 1]$$

and hence (**) holds.

Remark. It follows from the definition of $\mathscr{H}(S)$ and the kernel $\mathscr{T}(\tau, t)$ of T that *for every fixed* $t \in [0, 1]$,

$$\frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t) \in L^2[0, 1] \qquad \text{and} \qquad \mathscr{T}(\tau, t) \in \mathscr{H}(S).$$

Hence if

$$\int_s^1 \left(\frac{\partial^2}{\partial \tau^2} \mathscr{T}(\tau, t)\right) \overline{g(\tau)}\, d\tau = 0$$

for every $g \in L^2[0, 1]$, then this is true in particular for $g(\,.\,)=\mathscr{T}''(\,.\,, t)$ and hence $\mathscr{T}''(\,.\,, t)=0$ on $[s, 1]$.

If (**) holds then

$$\frac{\partial}{\partial \tau} \mathscr{T}(\tau, t) = \mathscr{T}'(t, t) \qquad \text{for} \quad \tau > t$$

since $\mathscr{T}(\,.\,, t)$ is a continuous function at $\tau=t$; moreover,

$$\begin{aligned} \mathscr{T}(s, t) &= \int_0^s \frac{\partial}{\partial \tau} \mathscr{T}(\tau, t)\, d\tau \\ &= \int_0^t \frac{\partial}{\partial \tau} \mathscr{T}(\tau, t)\, d\tau + \int_t^s \mathscr{T}'(t, t)\, d\tau \\ &= \mathscr{T}(t, t) + (s-t)\mathscr{T}'(t, t) \qquad \text{for} \quad s > t \end{aligned}$$

where (as before)

$$\mathscr{T}'(t, t) := \frac{\partial}{\partial \tau} \mathscr{T}(\tau, t)$$

at $\tau=t$.

Example 2. The RKHS with kernel

$$R(s, t) = \sup(s, t) \qquad s, t \in [0, 1]$$

plays an important role in the study of Wiener stochastic processes. By computations similar to those in the previous example we obtain

$$P^s f = \begin{cases} f(t) & \text{for} \quad t \leqslant s \\ f(s) & \text{for} \quad t > s \end{cases}$$

and for the kernel of a causal T,

$$\mathscr{T}(\tau, t) = \begin{cases} \mathscr{T}(\tau, t) & \text{for} \quad \tau \leqslant t \\ \mathscr{T}(t, t) & \text{for} \quad \tau > t \end{cases}$$

in the case of $\mathscr{H}(R)$ with kernel $R(s, t) = \sup(s, t)$ $(s, t \in [0, 1])$.

Example 3. If we define strictly causal operators in $\mathscr{H}(S)$ and in the RKHS of the Wiener processes as in 5.1.4.1 (with obvious modifications) then it turns out that a linear operator T of an RKHS is strictly causal if

$$\int_0^1 \int_0^1 \left| \frac{\partial^2}{\partial s^2} \mathscr{T}(s, t) \right|^2 \mathrm{d}s\,\mathrm{d}t < \infty$$

and

$$\int_0^1 \int_0^1 \left| \frac{\partial}{\partial s} \mathscr{T}(s, t) \right|^2 \mathrm{d}s\,\mathrm{d}t < \infty,$$

respectively.

∗5.3 Strictly causal operators

Certain Volterra integral operators and their 'discrete analogues' are the most important examples of strictly causal operators in L^2-spaces. Strictly causal operators in a general Hilbert space $\mathscr{H}$ have a representation similar to the Volterra integral operator. Moreover, $\lambda I - T$ has a causal inverse for every $\lambda \neq 0$ if T is strictly causal.

5.3.1. Let $\{P^t;\ t \in \Lambda\}$ be a one-parameter set of projection operators, as defined in § 5.2.1. Then a finite sequence

$$0 \leqslant P^1 \leqslant P^2 \leqslant \ldots \leqslant P^m \leqslant I \tag{$*$}$$

is called a *partition*, where $P^i \neq P^j$ for $i \neq j$ (0 and I are, respectively, the zero and identity operators and the abbreviation $P^i := P^{t_i}$ has been used).

$$0 \leqslant P'^1 \leqslant P'^2 \leqslant \ldots \leqslant P'^k \leqslant I$$

is called a *finer partition* than $(*)$ if

$$\{P^i;\quad i = 1, 2, \ldots, m\} \subset \{P'^i;\ i = 1, 2, \ldots, k\}.$$

Remark 1. 0 and I are not necessarily contained in $\{P^t;\ t\in\Lambda\}$; however, a partition always contains 0 and I as minimal and maximal elements.

Remark 2. The totally ordered $\{P^t;\ t\in\Lambda\}$ defines an ordering in Λ in the obvious manner

$$s<t \Leftrightarrow P^s \leqslant P^t \quad \text{and} \quad P^s \neq P^t.$$

Hence we shall sometimes write $s<t$ instead of $P^s \leqslant P^t$.

We are now ready to define strict causality.

5.3.1.1 Definition. Let T be a causal operator with respect to $\{P^t;\ t\in\Lambda\}$. Then T is called *strictly causal* if for every $\varepsilon>0$ there exists a partition such that for any finer partition,

$$\|\Delta^i T\Delta^i\| < \varepsilon$$

where

$$\Delta^1 = P^1,\quad \Delta^{n+1} = I-P^n \quad \text{and} \quad \Delta^i := P^i - P^{i-1};$$

$$i = 2, 3, \ldots, n.$$

An important norm estimate for the representations of strictly causal operators is the following.

5.3.1.2 Theorem.

$$\Big\|\sum_{k=1}^{n+1} \Delta^k T\Delta^k\Big\| = \max_k \|\Delta^k T\Delta^k\|.$$

Proof. Applying 4.10.3.3 and 4.10.3.2, we have

$$\|x\|^2 = \Big\|\sum_{k=1}^{n+1} \Delta^k x\Big\|^2 = \sum_{k=1}^{n+1} \|\Delta^k x\|^2 \tag{$*$}$$

and so

$$\Big\|\sum_{k=1}^{n+1} \Delta^k T\Delta^k x\Big\|^2 = \sum_{k=1}^{n+1} \|\Delta^k T\Delta^k x\|^2. \tag{$**$}$$

Moreover,

$$\sum_{k=1}^{n+1} \|\Delta^k T\Delta^k x\|^2 = \sum_{k=1}^{n+1} \|\Delta^k T\Delta^k (\Delta^k x)\|^2$$

$$\leqslant \sum_{k=1}^{n+1} \|\Delta^k T\Delta^k\|^2 \|\Delta^k x\|^2 \leqslant \max_k \|\Delta^k T\Delta^k\|^2 \sum_{k=1}^{n+1} \|\Delta^k x\|^2. \tag{$***$}$$

Comparing ($*$), ($**$) and ($***$), we have

$$\Big\|\sum_{k=1}^{n+1} \Delta^k T\Delta^k x\Big\|^2 \leqslant \max_k \|\Delta^k T\Delta^k\|^2 \|x\|^2$$

and hence

$$\Big\|\sum_{k=1}^{n+1} \Delta^k T \Delta^k\Big\| \leqslant \max\{\|\Delta^k T \Delta^k\|;\ k = 1, 2, \dots, n+1\}.$$

On the other hand, if the ith member on the left-hand side has the largest norm, i.e.

$$\|\Delta^i T \Delta^i\| = \max\{\|\Delta^k T \Delta^k\|;\ k=1,2,\dots,n+1\}$$

then from 4.10.3.3

$$\sum_{k=1}^{n+1} \Delta^k T \Delta^k (\Delta^i y) = \Delta^i T \Delta^i y$$

and hence, considering also (**),

$$\|\Delta^i T \Delta^i\| = \sup\{\|\Delta^i T \Delta^i\|\ \|y\|;\ \|y\| = 1\} \leqslant \Big\|\sum_{k=1}^{n+1} \Delta^k T \Delta^k\Big\|.$$

5.3.2. For the 'integral representation' of strictly causal operators we generalise the operator-valued Riemann integral introduced in § 4.11.2.

Let $L=L(t)$ $(t\in \Lambda)$ be an operator-valued function and let us consider the 'Riemann sum'

$$\sum_{k=1}^{n+1} L(t_k)\Delta^k \tag{$*$}$$

corresponding to the partition

$$0 \leqslant P^1 \leqslant P^2 \leqslant \dots \leqslant P^n \leqslant I$$

where Δ^k $(k=1, 2, \dots, n+1)$ are defined in 5.3.1.1.

5.3.2.1 Definition. The Riemann sums $(*)$ converge to the operator T if for every $\varepsilon>0$ there exists a partition such that for any finer partition, we have

$$\Big\|T - \sum_{k=1}^{m} L(t_k')\Delta'^k\Big\| < \varepsilon.$$

$(*)$ is called the *upper sum*. If we substitute $L(t_{k-1})$ in place of $L(t_k)$, then we obtain the corresponding *lower sum*. If the upper and lower sums converge to the same limit T, then we call this the *integral of* $L=L(t)$ and denote it by

$$T = \int_\Lambda L(t)\,\mathrm{d}P^t.$$

Remark. There is also a strong version of this integral, similar to 4.11.2.1.

5.3.2.2 Theorem. The causal linear operator T is strictly causal if and only if

$$T = \int_{\Lambda} P_t T \, \mathrm{d}P^t.$$

Proof. It is easy to check that every linear operator T can be partitioned as follows:

$$T = \sum_{i=1}^{n+1} (P^{i-1} T \Delta^i + \Delta^i T \Delta^i + P_i T \Delta^i).$$

If T is causal then

$$P^{i-1} T \Delta^i = P^{i-1} T P^{i-1}(P^i - P^{i-1}) = 0 \qquad (**)$$

since

$$P^{i-1}(P^i - P^{i-1}) = P^{i-1} - P^{i-1} = 0$$

and hence

$$\Big\| T - \sum_{i=1}^{n+1} P_i T \Delta^i \Big\| = \Big\| \sum_{i=1}^{n+1} \Delta^i T \Delta^i \Big\| = \max_i \|\Delta^i T \Delta^i\|.$$

It follows from 5.3.1.1 and 5.3.2.1 that T is strictly causal if and only if the Riemann sums

$$\sum_{i=1}^{n+1} (P_i T) \Delta^i$$

converge to the operator T.

We still have to show that the upper and lower integrals are the same. It follows from $(**)$ that for a causal operator T,

$$T = \sum_{i=1}^{n+1} \Delta^i T \Delta^i + P_i T \Delta^i = \sum_{i=1}^{n+1} (P^i - P^{i-1}) T \Delta^i + (I - P^i) T \Delta^i = \sum_{i=1}^{n+1} P_{i-1} T \Delta^i$$

since $P^k := I - P_k$; $k = 1, 2, \ldots$.

We conclude that *every causal operator T is equal to the lower integral of $P_t T$* and T is strictly causal if and only if it is also equal to the upper integral of $P_t T$, that is, if the theorem holds.

Remark. We have also shown in the proof that for a causal operator T,

$$T = \sum_{i=1}^{n+1} P_{i-1} T \Delta^i \qquad (***)$$

for every partition

$$0 \leqslant P^1 \leqslant P^2 \leqslant \ldots \leqslant P^n \leqslant I.$$

It is easy to show that for an *arbitrary* linear operator T, $P_{i-1} T \Delta^i$ is always causal and hence the right-hand side of $(***)$ is causal for every partition.

5.3.3. We now turn to the causal inverse problem. First let us consider a very simple case. Since the Neumann series is convergent for

$$|\lambda| > \limsup_n \|T^n\|^{1/n}$$

(see 4.2.3.1 and 4.3.2.2) and the limit of causal operators is again a causal operator, we have the following.

5.3.3.1 Theorem. For every causal operator T there exists a causal inverse $(\lambda I - T)^{-1}$ if

$$|\lambda| > \limsup_n \|T^n\|^{1/n}.$$

For the case of a strictly causal operator we begin with a partition of a linear operator T by the projection operators Δ^i; $i=1, 2, \ldots, n+1$. By straightforward calculations we can show that, for *every* linear T,

$$T = \sum_{j=1}^{n+1} \sum_{i=1}^{n+1} \Delta^i T \Delta^j.$$

In fact, for every $x \in \mathscr{H}$,

$$\sum_{i=1}^{n+1} \Delta^i T \Delta^j x = T \Delta^j x \qquad \text{and} \qquad \sum_{j=1}^{n+1} T \Delta^j x = T \sum_{j=1}^{n+1} \Delta^j x = Tx.$$

5.3.3.2 Theorem. If the linear operator T is causal, then

$$T = \sum_{i=1}^{n+1} \Delta^i T \Delta^i + \sum_{j=1}^{n+1} \sum_{i=j+1}^{n+1} \Delta^i T \Delta^j$$

and for the second term,

$$\Big(\sum_{j=1}^{n+1} \sum_{i=j+1}^{n+1} \Delta^i T \Delta^j\Big)^{n+1} = 0. \tag{$*$}$$

Proof. If $i<j$ and T is causal, then

$$\Delta^i T \Delta^j = (P^i - P^{i-1}) T \Delta^j = P^i T P^i \Delta^j - P^{i-1} T P^{i-1} \Delta^j$$

and

$$P^k \Delta^j = P^k (P^j - P^{j-1}) = P^k - P^k = 0$$

if $k \leqslant j-1$, by 4.10.2.1. For the second part of the theorem, observe that

$$(\Delta^i T \Delta^j)(\Delta^k T \Delta^m) = \begin{cases} 0 & \text{if } j \neq k \\ \Delta^i T^2 \Delta^m & \text{if } j = k. \end{cases}$$

Hence

$$\Big(\sum_{j=1}^{n+1} \sum_{i=j+1}^{n+1} \Delta^i T \Delta^j\Big)^2 = \sum_{j=2}^{n+1} \sum_{i=j+1}^{n+1} \Delta^i T^2 \Delta^j$$

i.e. the terms containing Δ^1 disappear in the square. Moreover, the terms containing Δ^2 disappear in the cube, and so on.

Remark. If T is a linear operator represented by an $n\times n$ matrix $\mathbf{A}$, then for the finest partition, $\Delta^i T\Delta^j$ is represented by the $n\times n$ matrix $\mathbf{A}_{ij}$ with

$$a_{nk} = \begin{cases} a_{ij} & \text{for} \quad n = i, \quad k = j \\ 0 & \text{elsewhere} \end{cases}$$

i.e. the jth element in the ith row is a_{ij} and all other elements are 0.

In this case the content of our theorem is the obvious partition of a triangular matrix into diagonal and 'strict' triangular matrices and the fact that a 'strict' triangular matrix $\mathbf{U}$ with n rowns and n columns has the property $\mathbf{U}^{n+1}=\mathbf{O}$.

From the foregoing it is clear that if T is causal and for the finest partition we have

$$\Delta^i T\Delta^i = 0 \qquad i = 1, 2, \ldots, n+1$$

then $(\lambda I-T)^{-1}\in B(\mathscr{H})$ and $(\lambda I-T)^{-1}$ is a causal operator for every $\lambda\neq 0$ since

$$\lambda I-T = \lambda I-\sum_{j=1}^{n+1}\sum_{i=j+1}^{n+1}(\Delta^i T\Delta^j)$$

and the Neumann series is finite in this case because of $(*)$. We shall show that this is also valid for every strictly causal T.

5.3.3.3 Theorem. If T is strictly causal, then $(\lambda I-T)^{-1}\in B(\mathscr{H})$ and $(\lambda I-T)^{-1}$ is a causal operator for every $\lambda\neq 0$.

Proof. If T is strictly causal then for every $\lambda\neq 0$ there exists a partition such that

$$\|\Delta^i T\Delta^i\| < |\lambda|$$

and hence, by 5.3.1.2, 5.3.3.1 and 4.2.3.1, if

$$B = \lambda I-\sum_{i=1}^{n+1}\Delta^i T\Delta^i$$

then $B^{-1}\in B(\mathscr{H})$ and causal.

B^{-1} is in the form of a Neumann series with members in the form $\gamma\Delta^i T^k\Delta^i$ and

$$\Delta^j(\Delta^i T^k\Delta^i) = (\Delta^i T^k\Delta^i)\Delta^j \qquad i, j = 1, 2, \ldots$$

by simple calculations; hence

$$\Delta^j B^{-1} = B^{-1}\Delta^j.$$

Now, if we apply the identity

$$B-\sum_{j=1}^{n+1}\sum_{i=j+1}^{n+1}\Delta^i T\Delta^j = B\Big(I-\sum_{j=1}^{n+1}\sum_{i=j+1}^{n+1}\Delta^i B^{-1}T\Delta^j\Big)$$

the theorem is clear since, by the considerations preceding 5.3.3.3, the second term on the right-hand side is also invertible with *causal* inverse.

$*$5.4 Automatic continuity of causal operators

We showed in 5.1.2.4 that every passive operator is causal. In this section we shall show that a passive operator is also continuous and that the same holds for time-invariant causal operators.

5.4.1. The following continuity principle is fundamental to the investigations of this section.

5.4.1.1 Theorem. If T is an everywhere-defined linear operator in a Banach space B and $\{L_\alpha;\ \alpha\in\Lambda\}$ is a set of bounded linear operators of B satisfying the conditions

(*a*) if $L_\alpha z=0$ for every $\alpha\in\Lambda$, then $z=\theta$;

(*b*) $L_\alpha T$ is a bounded linear operator of B for every $\alpha\in\Lambda$;

then T is also bounded.

Proof. We shall show that T is a closed operator and, by applying the closed-graph theorem, which says that an everywhere-defined closed operator of a Banach space is bounded, the proof is complete.

It follows from the definition of a closed operator (see 4.13.39) that if the linear operator T is closed, then

$$x_n\to\theta \quad\text{and}\quad Tx_n\to y\Rightarrow y=\theta \qquad (*)$$

since $T\theta=\theta$ in this case. First we shall show that if T is everywhere defined then the converse is also true. Let $(*)$ be satisfied for the linear operator T and $z_n\to z$, $Tz_n\to w$. Then $z_n-z\to\theta$ and $T(z_n-z)\to w-Tz$ and hence, by $(*)$,

$$w-Tz=\theta.$$

Now, if $x_n\to\theta$ and $Tx_n\to y$, then

$$L_\alpha Tx_n = L_\alpha(Tx_n)\to L_\alpha y$$

and also

$$L_\alpha Tx_n\to\theta$$

since $L_\alpha T$ is a bounded operator for every $\alpha \in \Lambda$. It follows that

$$L_\alpha y = \theta \qquad \alpha \in \Lambda$$

and hence $y=\theta$, i.e. $(*)$ is satisfied.

In a finite-dimensional $\mathscr{H}$ every linear operator is continuous, as we saw in § 1.7.3. *We shall therefore assume in the remainder of this section that* $\mathscr{H}$ *is infinite dimensional* and $\{P^t;\ t\in\Lambda\}$ is 'large enough' in the following sense:

(i) if $P^t z=\theta$ for every $t\in\Lambda$, then $z=\theta$;
(ii) for every $t\in\Lambda$ there exists $s\in\Lambda$ with $P^t\leqslant P^s$ $(P^t\neq P^s)$.

5.4.2. The proof that a passive operator is always causal was built on the simple observation that passivity, in Hilbert space language, means that the bilinear functional B_T defined in 5.1.2.4 is positive. A deeper investigation concerning B_T will prove that a passive operator is also continuous.

It follows from 5.1.2.4 $(*)$ and from

$$B_T(f, g) \geqslant |(Tf|P^t g)| - |(P^t f|Tg)|$$

that

$$\begin{aligned} |(P^t Tf|g)| &\leqslant B_T(f, f)^{1/2} B_T(g, g)^{1/2} + |(f|P^t Tg)| \\ &\leqslant B_T(f, f)^{1/2} B_T(g, g)^{1/2} + (f|f)^{1/2} (Tg|Tg)^{1/2}. \end{aligned}$$

Hence, introducing

$$\begin{aligned} R(f) &:= |(P^t Tf|f)|^{1/2} + (f|f)^{1/2} \\ S(g) &:= 2|(P^t Tg|g)|^{1/2} + (Tg|Tg)^{1/2} \end{aligned}$$

we obtain, by straightforward calculations,

$$|(P^t Tf|g)| \leqslant R(f) S(g).$$

We now define

$$F_f(g) := \frac{1}{R(f)} (P^t Tf|g) \qquad f\neq\theta.$$

It follows that $\{F_f(g);\ f\neq 0\}$ is bounded for every $g\in\mathscr{H}$ *with bound independent of* f. Hence, by the uniform boundedness principle (see the Appendix, A.2.2.1), $\{\|F_f\|;\ f\neq\theta\}$ is bounded, i.e. there exists $M>0$ such that

$$\|F_f\| < M \qquad f\neq\theta$$

and hence

$$\frac{1}{R(f)} |(P^t Tf|g)| < M\|g\|$$

which means that

$$\|P^t Tf\| = \sup\{|(P^t Tf|g)|;\ \|g\| = 1\} < MR(f). \tag{$*$}$$

Remark. Although it is not indicated, M depends on t.

We assert that for every operator $P^t T$ there exists $C=C_t$ such that

$$\frac{\|P^t T f\|}{\|f\|} < C \qquad f \neq \theta.$$

For the proof of this inequality, note that

$$0 < R(f) \leqslant \|P^t T f\|^{1/2} \|f\|^{1/2} + \|f\|$$

and hence

$$\frac{R(f)}{\|f\|} \leqslant \frac{\|P^t T f\|^{1/2}}{\|f\|^{1/2}} + 1.$$

From $(*)$ it follows that

$$\frac{\|P^t T f\|}{\|f\|} \leqslant M \frac{\|P^t T f\|^{1/2}}{\|f\|^{1/2}} + M$$

whence

$$\frac{\|P^t T f\|}{\|f\|} \leqslant (M+1)^2.$$

We have now proved that $P^t T$ is a bounded operator for every $t \in \Lambda$. If we apply 5.4.1.1, the theorem is also proved.

5.4.3. Every causal operator is continuous on at least one invariant subspace $P_t \mathscr{H}$. More precisely, the following theorem holds.

5.4.3.1 Theorem. If

(*a*) T is a linear operator of $\mathscr{H}$ (i.e. everywhere defined);
(*b*) T is causal with respect to $\{P^s;\ s \in \Lambda\}$ satisfying conditions 5.4.1 (i), (ii);

then there exists an invariant subspace $P_t \mathscr{H}$ such that T restricted to $P_t \mathscr{H}$ is bounded.

Proof. We shall prove that if for every $t \in \Lambda$ there exists $s \in \Lambda$ such that $P^s T$ is unbounded on $P_t \mathscr{H}$, then we can construct $x_0 \in \mathscr{H}$ so that $x_0 \notin \mathscr{D}(T)$ and hence T is not an everywhere-defined operator. By this contradiction it will be proved that there exists an invariant subspace $P_t \mathscr{H}$ of T such that every operator $P^s T$ $(s \in \Lambda)$ is bounded on $P_t \mathscr{H}$.

Applying 5.4.1.1, we conclude that T is also bounded on $P_t \mathscr{H}$.

The construction of $x_0 \in \mathscr{H}$, which constitutes the basis of the proof is the following. For $t_1 \in \Lambda$ there exists $s_1 \in \Lambda$ such that $P^{s_1} T$ is unbounded on $P_{t_1} \mathscr{H}$ and hence there exists $P_{t_1} x_1$ such that

$$\|P_{t_1} x_1\| = 1 \qquad \text{and} \qquad \|P^{s_1} T P_{t_1} x_1\| > 1.$$

For $t_2=s_1$ there exists $s_2 \in \Lambda$ such that $P^{s_2}T$ is unbounded on $P_{t_2}B$ and hence there exists $P_{t_2}x_2$ such that

$$\|P_{t_2}x_2\| - 1 \qquad \text{and} \qquad \|P^{s_2}TP_{t_2}x_2\| > 2^2(2+\frac{1}{2}\|TP_{t_1}x_1\|).$$

Moreover, it can be supposed that $t_2<s_2$ since

$$\|P^s z\| = \|P^s P^{s_2} z\| \leqslant \|P^{s_2} z\| \qquad \text{for} \quad s<s_2.$$

The following abbreviations will be used hereafter:

$$P^{s_k} := P^k \qquad P_{s_k} := P_k.$$

For any integer $k>0$ and $t_{k+1}=s_k$ there exists s_{k+1} such that $P^{k+1}T$ is unbounded on $P_k\mathscr{H}$ and hence there exists P_kx_k such that

$$\|P_kx_k\| = 1 \qquad s_{k+1} > s_k$$

and

$$\|P^{k+1}TP_kx_k\| > 2^k\left(k+\sum_{i=1}^{k-1}\frac{1}{2^i}\|TP_ix_i\|\right).$$

Now, it is obvious that

$$x_0 = \sum_{k=1}^{\infty}\frac{1}{2^k}P_kx_k \in \mathscr{H}$$

i.e. the infinite series on the right-hand side is convergent. We shall show that for any integer $N>0$,

$$\|Tx_0\| > N$$

and hence $x_0 \notin \mathscr{D}(T)$. In fact,

$$\|Tx_0\| \geqslant \|P^{N+1}Tx_0\|$$

$$= \left\|\sum_{k=1}^{N-1}\frac{1}{2^k}P^{N+1}TP_kx_k+\frac{1}{2^N}P^{N+1}TP_Nx_N+P^{N+1}T\sum_{k=N+1}^{\infty}\frac{1}{2^k}P_kx_k\right\|.$$

Consider the right-hand side of the equation; for the first term,

$$\left\|\sum_{k=1}^{N-1}\frac{1}{2^k}P^{N+1}TP_kx_k\right\| \leqslant \sum_{k=1}^{N-1}\frac{1}{2^k}\|TP_kx_k\|$$

and for the third term

$$P^{N+1}T\sum_{k=N+1}^{\infty}\frac{1}{2^k}P_kx_k = P^{N+1}T\sum_{k=N+1}^{\infty}\frac{1}{2^k}P^{N+1}P_kx_k = 0$$

since T is causal,

$$P^{N+1}P_k = P^{N+1}(I-P^k) = 0 \qquad \text{for} \quad k \geqslant N+1.$$

We conclude that

$$\|Tx_0\| \geqslant \left\| \frac{1}{2^N} P^{N+1} T P_N x_N + \sum_{k=1}^{N-1} \frac{1}{2^k} P^{N+1} T P_k x_k \right\|$$

$$\geqslant \frac{1}{2^N} \|P^{N+1} T P_N x_N\| - \sum_{k=1}^{N-1} \frac{1}{2^k} \|T P_k x_k\|$$

$$\geqslant \left(N + \sum_{k=1}^{N-1} \frac{1}{2^k} \|T P_k x_k\|\right) - \sum_{k=1}^{N-1} \frac{1}{2^k} \|T P_k x_k\| = N.$$

Remark. It is obvious that if $\mathscr{H} = \mathrm{L}^2(\Omega)$ or $\mathscr{H} = \mathscr{H}(R)$ with $\mathscr{D} = \Omega$, then 5.4.1 (i) and (ii) are satisfied. Moreover, $P_t\mathscr{H}$ *is* infinite dimensional for every $t \in \Lambda$ in these cases.

5.4.4. Now let T be a causal time-invariant operator of $\mathrm{L}^2(\Omega)$; we then have a much more powerful result.

5.4.4.1 Theorem. A causal time-invariant linear operator T of $\mathrm{L}^2(\Omega)$ is continuous.

Proof. Let $t_0 \in \Omega$; then for every $x \in \mathrm{L}^2(\Omega)$ and $t < t_0$,

$$\|T P_t x\| = \|U_{t_0 - t} T P_t x\| = \|T U_{t_0 - t} P_t x\| = \|T P_{t_0} U_{t_0 - t} x\|. \qquad (*)$$

Here we have applied 5.1.2.1 $(*)$ and the fact that U_t is an isometry (i.e. $\|U_t x\| = \|x\|$ for every $t \in \Omega$, $x \in \mathscr{H}$).

It follows from 5.4.3.1 that T is bounded on an invariant subspace $P_{t_0}\mathscr{H}$; hence

$$\|T P_{t_0} U_{t_0 - t} x\| \leqslant \|T\|_0 \|P_{t_0} U_{t_0 - t} x\| = \|T\|_0 \|U_{t_0 - t} P_t x\| = \|T\|_0 \|P_t x\| \quad (**)$$

where $\|T\|_0$ is the norm of T restricted to $P_{t_0}\mathscr{H}$. By comparing $(*)$ and $(**)$ we obtain the continuity of T on $P_t\mathscr{H}$ for every $t < t_0$.

If $t > t_0$, then $P_t\mathscr{H} \subset P_{t_0}\mathscr{H}$ and hence

$$\|T P_t x\| \leqslant \|T\|_0 \|P_{t_0} P_t x\| = \|T\|_0 \|P_t x\|.$$

We therefore conclude that T is bounded by $\|T\|_0$ on $\bigcup_{t \in \Omega} P_t\mathscr{H}$. Moreover, $\bigcup_{t \in \Omega} P_t\mathscr{H}$ is dense in $\mathscr{H}$ if $\mathscr{H} = \mathrm{L}^2(\Omega)$.

Remark. To extend our theorem from $\mathrm{L}^2(\Omega)$ to other Hilbert spaces we have the problem of defining $\{U_t;\ t \in \Lambda\}$ in such a way that 5.1.2.1 $(*)$ is satisfied if the truncation operators $\{E^t;\ t \in \Omega\}$ are replaced by $\{P^t;\ t \in \Lambda\}$.

5.4.5. One of the important features of the 'automatic' continuity theorems is the following. The simplest stability concept is that a system is stable if a bounded input implies a bounded output. If the bound is measured in L^2-norm, i.e. a signal $x=x(t)$ is bounded if

$$\int_{-\infty}^{+\infty} |x(t)|^2 \, dt < \infty$$

(for simplicity, Ω is the real line), then this means that the transition operator T of the system is an (everywhere-defined) operator of $L^2(\Omega)$. However, we can only apply Hilbert space theory in the case where T is a bounded operator.

Our theorem about time-invariant causal operators therefore says that the Hilbert space theory can be applied for every time-invariant input–output system that is stable in this simple sense.

A more sophisticated stability concept is connected with the time structure. $E^t Tx$, the part of the output before a certain time t, is always bounded; however, it 'blows up' after a time if the system is *not* stable. The mathematical expression of this property is that the L^2-norm of $E^t Tx$ is finite for every t and $x \in L^2(\Omega)$ and T is stable if $\|E^t Tx\| < M$, i.e. there is *a common bound for every t*. Investigations of this stability can be found, for example, in Desoer and Vidysagar 1975, pp 169–86 and Feintuch and Saeks 1982, pp 173–9.

Appendix

A.1 The adjoint T^* of an unbounded operator T

A.1.1. If T is a bounded operator but is not everywhere defined in the Hilbert space then we have two cases.

(*a*) The domain $\mathscr{D}(T)$ of T is dense in $\mathscr{H}$. There is then a unique extension T onto $\mathscr{H}$ via continuity as follows. For every $x \in \mathscr{H}$ there exists $\{x_n\}$; $x_n \in \mathscr{D}(T)$ such that $x_n \to x$ and hence $Tx_n \to y \in \mathscr{H}$ since T is continuous. We define $Tx = y$. (It is obvious that $y = \lim_n Tx_n$ is the same for every $\{x_n\}$ tending to x and hence there is only one extended T.)

(*b*) The domain $\mathscr{D}(T)$ is not dense in $\mathscr{H}$. There is then a unique extension onto $\overline{\mathscr{D}(T)}$, the closure of $\mathscr{D}(T)$, and $\overline{\mathscr{D}(T)}$ as a closed subspace of a Hilbert space $\mathscr{H}$ is also a Hilbert space in itself.

We conclude that a bounded linear operator can always be considered as an operator *on* a Hilbert space.

A.1.2. Let T be a linear operator with dense domain $\mathscr{D}(T)$. Then the adjoint T^* is defined by

$$(T^*y|x) = (y|Tx) \qquad y \in \mathscr{D}(T^*); \quad x \in \mathscr{D}(T)$$

as in the case of bounded T. More precisely, $\mathscr{D}(T^*)$ is the subspace

$$\{y\colon x \to (y|Tx) \text{ is a continuous functional on } \mathscr{D}(T)\}$$

and T^*y is the functional that associates the number $(y|Tx)$ to $x \in \mathscr{D}(T)$;

Remark 1. $T^*y \in \mathscr{H}$ exists by the Riesz–Fréchet Theorem.

Remark 2. The condition that $\mathscr{D}(T)$ is dense in $\mathscr{H}$ is not a serious restriction since it is always satisfied in the Hilbert space $\overline{\mathscr{D}(T)} \subseteq \mathscr{H}$.

Example 1. Let

$$Ty = \frac{\mathrm{d}}{\mathrm{d}t}\, y$$

with domain

$$\mathscr{D}(T) = \{y\colon\ y' \in L^2[0,1],\ y(0) = 0\}.$$

(Here and in what follows we use $\mathrm{d}y/\mathrm{d}t$ and y' interchangeably for the derivative of y.) Then, integrating by parts, we obtain

$$\int_0^1 y'(t)\overline{z(t)}\,\mathrm{d}t = -\int_0^1 y(t)\overline{z'(t)}\,\mathrm{d}t$$

if $z' \in L^2[0,1]$ and $z(1)=0$. Moreover,

$$\left|\int_0^1 y'(t)\overline{z(t)}\,\mathrm{d}t\right| \leq \left\|\frac{\mathrm{d}}{\mathrm{d}t}z\right\|_2 \|y\|_2.$$

It follows that

$$\{z\colon\ z' \in L^2[0,1],\ z(1) = 0\} \subset \mathscr{D}(T^*)$$

and

$$T^* z = -\frac{\mathrm{d}}{\mathrm{d}t}\, z \tag{$*$}$$

on this subset. We shall show that $(*)$ is the form of the adjoint operator T^* for every $z \in \mathscr{D}(T^*)$ and

$$\mathscr{D}(T^*) = \{z\colon\ z' \in L^2[0,1],\ z(1) = 0\}.$$

In fact, by the Riesz–Fréchet theorem, $z \in \mathscr{D}(T^*)$ if and only if there exists $z_* \in L^2[0,1]$ such that

$$\int_0^1 y'(t)\overline{z(t)}\,\mathrm{d}t = \int_0^1 y(t)\overline{z_*(t)}\,\mathrm{d}t \qquad y \in \mathscr{D}(T). \tag{$**$}$$

We can write

$$z_*(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\int_t^1 z_*(\tau)\,\mathrm{d}\tau$$

and, integrating by parts,

$$\int_0^1 y(t)\overline{\left(\frac{\mathrm{d}}{\mathrm{d}t}\int_t^1 z_*(\tau)\,\mathrm{d}\tau\right)}\,\mathrm{d}t = \int_0^1 y'(t)\overline{\int_t^1 z_*(\tau)\,\mathrm{d}\tau}\,\mathrm{d}t \tag{$***$}$$

for $y \in \mathscr{D}(T)$. Comparing $(**)$ and $(***)$, we have

$$z(t) = \int_t^1 z_*(\tau)\,\mathrm{d}\tau$$

since T maps $\mathscr{D}(T)$ *onto* $L^2[0, 1]$. It follows that

$$-\frac{d}{dt} z = z_* \qquad \text{and} \qquad z(1) = 0$$

if $z \in \mathscr{D}(T^*)$.

Remark 1. Here and in what follows, derivative means weak derivative in the sense of § 3.9.3. If we restrict ourselves to

$$\mathscr{D}(T) = \{y\colon\ y' \in L_0^2[0, 1], y(0) = 0\}$$

then weak derivatives can be avoided; however, the meaning of

$$z_*(t) = -\frac{d}{dt} \int_t^1 z_*(\tau)\, d\tau$$

remains problematic for certain $z_* \in L^2[0, 1]$.

Remark 2. We can also write

$$z_*(t) = \frac{d}{dt} \int_0^t z_*(\tau)\, d\tau$$

but this form does not yield the derived result.

To give $\mathscr{D}(T^*)$ precisely is the main difficulty in determining T^* for an unbounded T.

Example 2. If

$$Ty = i \frac{d}{dt} y$$

with domain

$$\mathscr{D}(T) = \{y\colon\ y' \in L^2[0, 1],\ y(0) = 0\}$$

and $i = \sqrt{-1}$, then we compute, as in the previous example, that

$$\int_0^1 i y'(t) \overline{z(t)}\, dt = \int_0^1 y(t) \left(\overline{i z'(t)}\right) dt$$

if $z' \in L^2[0, 1]$ and $z(1) = 0$ and

$$T^* z = i \frac{d}{dt} z \qquad \mathscr{D}(T^*) = \{z\colon\ z' \in L^2[0, 1], z(1) = 0\}.$$

However, $T \neq T^*$ in this case since $\mathscr{D}(T^*) \neq \mathscr{D}(T)$. (In other words, T is only identical to its *formal* adjoint (see § 3.9.3).)

Example 3. Let

$$Ty = \frac{d^2}{dt^2} y$$

with domain

$$\mathscr{D}(T) = \{y: y'' \in L^2[0, 1], y(0) = y(1) = 0\}.$$

Again, integrating by parts we obtain

$$\int_0^1 y''(t)\overline{z(t)}\,dt = \int_0^1 y(t)\overline{z''(t)}\,dt$$

if $z'' \in L^2[0, 1]$ and $z(1) = z(0) = 0$. Moreover,

$$\left|\int_0^1 y''(t)\overline{z(t)}\,dt\right| \leq \left\|\frac{d^2}{dt^2} z\right\|_2 \|y\|_2.$$

It follows that

$$\{z: z'' \in L^2[0, 1], z(0) = z(1) = 0\} \subseteq \mathscr{D}(T^*)$$

and

$$T^* z = \frac{d^2}{dt^2} z \qquad (*)$$

on this subset. We shall show that $(*)$ is the form of the adjoint operator T^* for every $z \in \mathscr{D}(T^*)$ and

$$\mathscr{D}(T^*) = \{z: z'' \in L^2[0, 1], z(0) = z(1) = 0\}.$$

In fact, by the Riesz–Fréchet theorem, $z \in \mathscr{D}(T^*)$ if and only if there exists $z_* \in L^2[0, 1]$ such that

$$\int_0^1 y''(t)\overline{z(t)}\,dt = \int_0^1 y(t)\overline{z_*(t)}\,dt \qquad y \in \mathscr{D}(T). \qquad (**)$$

Let

$$k(t, \tau) = \begin{cases} t(\tau - 1) & t < \tau \leq 1 \\ \tau(t-1) & 0 \leq \tau < t. \end{cases}$$

Then, by immediate calculation, we see that

$$z_*(t) = \frac{d^2}{dt^2} \int_0^1 k(t, \tau) z_*(\tau)\,d\tau$$

and, integrating by parts,

$$\int_0^1 y(t)\left(\frac{d^2}{dt^2} \int_0^1 k(t, \tau) z_*(\tau)\,d\tau\right) dt = \int_0^1 y''(t) \int_0^1 k(t, \tau) z_*(\tau)\,d\tau\,dt \qquad (***)$$

for $y \in \mathscr{D}(T)$. Comparing (**) and (***), we have

$$z(t) = \int_0^1 k(t, \tau) z_*(\tau) \, d\tau$$

since T maps $\mathscr{D}(T)$ *onto* $L^2[0, 1]$. It follows that

$$\frac{d^2}{dt^2} z = z_* \qquad \text{and} \qquad z(0) = z(1) = 0$$

if $z \in \mathscr{D}(T^*)$.

A.1.2.1 Theorem. T^* is a linear closed operator.
Proof. For $x, y \in \mathscr{D}(T^*)$, $z \in \mathscr{D}(T)$, $\lambda, \mu \in \Phi$, we have

$$\begin{aligned}(T^*(\lambda x + \mu y)|z) &= (\lambda x + \mu y|Tz) = \lambda(x|Tz) + \mu(y|Tz) \\ &= \lambda(T^* x|z) + \mu(T^* y|z) = (\lambda T^* x + \mu T^* y|z)\end{aligned}$$

and hence

$$T^*(\lambda x + \mu y) = \lambda T^* x + \mu T^* y$$

since $\mathscr{D}(T)$ is dense. For the second part of the theorem, let $x_n \to x$ and $T^* x_n \to h$; then

$$(h|z) = \lim_n (T^* x_n|z) = \lim (x_n|Tz) = (x|Tz)$$

and hence $x \in \mathscr{D}(T^*)$ and $T^* x = h$.

Remark. Observe that in the proof we did not assume that T is either linear or closed.

A.1.3. The set of pairs $[x, y]$; $x, y \in \mathscr{H}$ of a Hilbert space $\mathscr{H}$ forms a linear space when we define addition by

$$[x_1, y_1] + [x_2, y_2] := [x_1 + x_2, y_1 + y_2]$$

and product by λ as

$$\lambda[x, y] = [\lambda x, \lambda y] \qquad x_1, x_2, y_1, y_2, x, y \in \mathscr{H}, \ \lambda \in \Phi$$

and a Hilbert space when we define the scalar product

$$([x_1, y_1]|[x_2, y_2]) := (x_1|x_2) + (y_1|y_2).$$

The linear subspace of elements

$$[x, \theta] \qquad \text{and} \qquad [\theta, y] \qquad x, y \in \mathscr{H}$$

forms a closed subspace isomorphic with $\mathscr{H}$ and the Hilbert space of pairs

$\{[x, y];\ x, y \in \mathscr{H}\}$ is therefore called the orthogonal direct sum $\mathscr{H} \oplus \mathscr{H}$. Some properties of operators in $\mathscr{H}$ can be expressed in a simple way by $\mathscr{H} \oplus \mathscr{H}$.

Recall that the *graph* of an operator is the set of pairs $[x, Tx]$ and so the graph of a linear operator T, with domain and range in $\mathscr{H}$, is a linear subspace of $\mathscr{H} \oplus \mathscr{H}$. It is easy to show that the following theorem is true.

A.1.3.1 Theorem. The linear operator T is *closed* if and only if its graph

$$\mathscr{g}_T := [x, Tx]; \quad x \in \mathscr{D}(T)$$

is a closed linear subspace of $\mathscr{H} \oplus \mathscr{H}$.

The following connection exists between the graph of T and its adjoint T^*. Let us define

$$V[x, y] := [y, -x] \qquad x, y \in \mathscr{H}.$$

Then V is an isometry of $\mathscr{H} \oplus \mathscr{H}$ and $V^2 = -I$. It is easy to verify that

$$\mathscr{g}_T = [V\mathscr{g}_{T^*}]^\perp \tag{$*$}$$

if T is a closed operator.

A.1.3.2 Theorem. If T is closed, then $\mathscr{D}(T^*)$ is dense in $\mathscr{H}$.

Proof. It is obvious that the statement '$\mathscr{D}(T^*)$ is dense' is the same as '$\mathscr{D}(T^*)^\perp = \{\theta\}$'. So let $h \in \mathscr{D}(T^*)^\perp$, and we shall prove that $h = \theta$. In fact,

$$([\theta, h] \| [T^*g, -g]) = 0 \qquad g \in \mathscr{D}(T^*), \quad h \in \mathscr{D}(T^*).$$

But this means that

$$[\theta, h] \in [V\mathscr{g}_{T^*}]^\perp$$

and so, from the identity $(*)$,

$$[\theta, h] \in \mathscr{g}_T$$

i.e. $h = T\theta = \theta$.

Remark. Observe that in the proof we did not assume that T is linear, only that $T\theta = \theta$.

A.2 The uniform boundedness principle

A.2.1. A Banach space has the following property.

A.2.1.1 Theorem. Let B be a Banach space and let $\mathscr{Z}_k;\ k = 1, 2, \ldots$ be an infinite sequence of closed subsets, such that

$$B = \bigcup_{k=1}^{\infty} \mathscr{Z}_k.$$

Then there exists k such that $\mathscr{Z}_k$ contains an open sphere. This is the *Baire principle.*

Proof. Recall that an open sphere with centre x_0 and radius r is the subset

$$\mathscr{S}(x_0, r) := \{x\colon \|x-x_0\| < r\}$$

of B.

For the proof we shall suppose that none of the closed sets $\mathscr{Z}_k$, $k=1, 2, \dots$ contains an open sphere and we shall reach a contradiction.

If $\mathscr{Z}_1$ does not contain an open sphere then

$$\mathscr{S}(x_0, 1) \not\subset \mathscr{Z}_1$$

and hence for $x_1 \in \mathscr{S}(x_0, 1)$, $x_1 \notin \mathscr{Z}_1$, we have

$$\mathscr{S}(x_1, r_1) \quad \text{with} \quad r_1 < 1/2 \quad \text{such that} \quad \mathscr{S}(x_1, r_1) \subset \mathscr{Z}_1^c$$

where $\mathscr{Z}_1^c$ is the complement of the closed set $\mathscr{Z}_1$ and is therefore open (see 1.5.1.1 and 1.5.1.4). Moreover, we may suppose that

$$\mathscr{S}(x_1, r_1) \subset \mathscr{S}(x_0, 1).$$

Since $\mathscr{Z}$ also does not contain an open sphere,

$$\mathscr{S}(x_1, r_1) \not\subset \mathscr{Z}_2$$

and hence for $x_2 \in \mathscr{S}(x_1, r_1)$, $x_2 \notin \mathscr{Z}_2$, we have

$$\mathscr{S}(x_2, r_2) \quad \text{with} \quad r_2 < 1/2^2 \quad \text{such that} \quad \mathscr{S}(x_2, r_2) \subset \mathscr{Z}_2^c$$

where $\mathscr{Z}_2^c$ is the complement of the closed set $\mathscr{Z}_2$ and is therefore open. Moreover, we may suppose that

$$\mathscr{S}(x_2, r_2) \subset \mathscr{S}(x_1, r_1).$$

Continuing this process, we obtain a sequence

$$\mathscr{S}(x_0, 1) \supset \mathscr{S}(x_1, r_1) \supset \dots \supset \mathscr{S}(x_n, r_n) \supset \dots$$

and for $m > n$,

$$\|x_n - x_m\| < r_n < 1/2^n.$$

Hence the sequence $\{x_n\}$ of the centres is convergent. Let $\lim_n x_n = x$ $x \in B$. Since B is a Banach (i.e. complete) space and

$$x \in \mathscr{S}(x_n, r_n) \qquad n = 1, 2, \dots$$

but $\mathscr{S}(x_n, r_n) \cap \mathscr{Z}_k = \emptyset$ if $n \geq k$ and hence $x \notin \mathscr{Z}_k$; $k=1, 2, \dots$.

Thus we have the contradiction

$$x \in B \qquad \text{and} \qquad x \notin \bigcup_{k=1}^{\infty} \mathscr{Z}_k.$$

A.2.2. The uniform boundedness principle says that if we have a set $\{T_\alpha;\, \alpha \in \Lambda\}$ of linear operators and $\{T_\alpha x;\, \alpha \in \Lambda\}$ is bounded for every $x \in \mathscr{S}(\theta, 1)$ then it is *uniformly bounded* on the unit sphere $\mathscr{S}(\theta, 1)$, i.e. there is a common bound for

$$\{T_\alpha x;\ \alpha \in \Lambda, x \in \mathscr{S}(\theta, 1)\}.$$

In other words, we have the following theorem.

A.2.2.1 Theorem. If $\{T_\alpha;\, \alpha \in \Lambda\}$ is a (one-parameter) set of bounded linear operators and $\{\|T_\alpha x\|;\, \alpha \in \Lambda\}$ is bounded for each $x \in B$, then there is a common bound M and

$$\|T_\alpha\| < M.$$

Proof. Let

$$\mathscr{Z}_k = \{x\colon \|T_\alpha x\| \leqslant k\};\quad k = 1, 2, \ldots.$$

Then each $\mathscr{Z}_k$ is closed and

$$B = \bigcup_{k=1}^{\infty} \mathscr{Z}_k$$

since, by our condition, $\{\|T_\alpha x\|;\, \alpha \in \Lambda\}$ is bounded by a certain k for every $x \in B$.

It follows from the foregoing theorem that there exists $\mathscr{Z}_k$ containing an open sphere $\mathscr{S}(x_k, r_k)$. This means that for a certain k

$$\|T_\alpha x_k\| \leqslant k \quad \text{and} \quad \|x - x_k\| < r_k \quad \text{implies} \quad \|T_\alpha x\| < k \qquad (*)$$

for every $\alpha \in \Lambda$. It follows that

'if $\|x - x_k\| < r_k$ then $\|T_\alpha(x - x_k)\| < 2k$' or, equivalently,

$$\|T_\alpha\| < \frac{2k}{r_k}.$$

Remark 1. Observe that the bound k of $\{\|T_\alpha x\|;\, \alpha \in \Lambda\}$ depends on $x \in B$, i.e. $k = k(x)$.

Remark 2. The theorem remains valid if T is a linear operator from one Banach space B_1 into another Banach space B_2, with the same proof.

The uniform boundedness principle has important consequences. Some of them are as follows.

A.2.2.2 Theorem. If T is a linear operator defined by

$$\lim_n T_n x := Tx \qquad x \in B$$

where T_n; $n=1, 2, \ldots$ are bounded linear operators, then T is also bounded. This is the *Banach–Steinhaus Theorem.*

Proof. Now $\{\|T_n x\|;\ n=1, 2, \ldots\}$ is bounded for every $x \in B$ since a convergent sequence is bounded. Hence, by the previous theorem, $\|T_n\| < M$ and so

$$\|T_n x\| < M\|x\| \qquad n = 1, 2, \ldots.$$

It follows that

$$\|Tx\| = \lim_n \|T_n x\| \leqslant M\|x\|.$$

Again, let $\{T_\alpha;\ \alpha \in \Lambda\}$ be a set of bounded linear operators, but let the index set Λ be totally ordered. Then $\{T_\alpha x\}$ *is called convergent and*

$$\lim_\alpha T_\alpha x = y$$

if for every $\varepsilon > 0$ *there exists* $\alpha_0 \in \Lambda$ *such that*

$$\|y - T_\alpha x\| < \varepsilon \qquad \text{if} \quad \alpha > \alpha_0.$$

(Compare with 4.11.2.1 and 5.3.2.1.)

An immediate generalisation of Theorem A.2.2.2 is the following.

A.2.2.3 Theorem. Let

(*a*) $\{T_\alpha x\}$ $(\alpha \in \Lambda)$ be convergent for each $x \in \mathscr{D}$, where $\mathscr{D}$ is a dense subset of the Banach space B;

(*b*) $\{\|T_\alpha x\|\}$ $(\alpha \in \Lambda)$ be bounded for every $x \in B$.

Then the linear operator T defined by

$$Tx = \lim_\alpha T_\alpha x \qquad x \in \mathscr{D} \qquad (**)$$

is bounded and (extending all over B) $Tx = \lim_\alpha T_\alpha x$ for every $x \in B$.

Proof. By A.2.2.1, $\{\|T_\alpha\|\}$ is bounded, i.e. $\|T_\alpha\| < M$ and hence, as in A.2.2.2,

$$\|Tx\| = \lim_\alpha \|T_\alpha x\| \leqslant M\|x\| \qquad x \in \mathscr{D}$$

and so there is a unique extension of T onto B by A.1.1. It is easy to show that $(**)$ also holds for every $x \in B$.

* *A.2.2.4 Theorem.* If

(*a*) $T_\alpha x \to \theta$ for every $x \in \mathscr{D}$, where $\mathscr{D}$ is a dense subset of the Banach space B;

(*b*) $\{\|T_\alpha x\|\}$ $(\alpha \in \Lambda)$ is bounded for every $x \in B$

then $T_\alpha x \to \theta$ *uniformly*, i.e. $\sup_\alpha \|T_\alpha x\| \to 0$.

Proof. Applying the previous theorem for $T=0$, we conclude that $T_\alpha x \to \theta$ for every $x \in B$. As in the proof of A.2.2.1, we conclude that

$$\|T_\alpha x_k\| \leqslant k \quad \text{and} \quad \|x - x_k\| < r_k \quad \text{implies} \quad \|T_\alpha x\| < k$$

and hence, if $\|x-x_k\| < r_k$, then $\|T_\alpha(x-x_k)\| < 2k$. Thus for every $\varepsilon > 0$,

$$\|T_\alpha(x-x_k)\| < \varepsilon$$

if

$$\|x - x_k\| < \delta = \varepsilon r_k/2k$$

(independent of α).

A.3 The closed graph theorem

In our approach, the closed graph theorem will be derived from the uniform boundedness principle and the connections between T and T^*.

Another approach can be found in Gohberg and Goldberg (1981), where the closed graph theorem is derived immediately from the Baire principle and the uniform boundedness principle is proved afterwards, using the closed graph theorem.

Theorem. If T is a closed operator and $\mathscr{D}(T) = \mathscr{H}$, i.e. everywhere defined in the Hilbert space $\mathscr{H}$, then T is bounded.
Proof. First we shall show that if $\mathscr{D}(T) = \mathscr{H}$ then T^* is a bounded operator on $\mathscr{D}(T^*)$. In fact, if $\|x_\alpha\| \leqslant 1$, $x_\alpha \in \mathscr{D}(T^*)$, then

$$|(y|T^* x_\alpha)| = |(Ty|x_\alpha)| \leqslant \|Ty\| \qquad y \in \mathscr{H}$$

i.e. the functionals $\{T^* x_\alpha\}$ are bounded by $\|Ty\|$ for every $y \in \mathscr{H}$. Applying the uniform boundedness principle, it follows that

$$\|T^* x_\alpha\| < M.$$

T^* is also a closed operator, by A.1.2.1, and $\mathscr{D}(T^*)$ is also closed for a bounded and closed operator T^*. In fact, if $x_n \in \mathscr{D}(T^*)$ and $x_n \to x$, then $\{T^* x_n\}$ is also convergent since T^* is bounded and $x \in \mathscr{D}(T^*)$ since T^* is closed.

It follows from A.1.3.2 that $\mathscr{D}(T^*)$ is also dense in $\mathscr{H}$, and so $\mathscr{H} = \mathscr{D}(T^*)$.

To summarise, $T^* \in B(\mathscr{H})$ and consequently $T^{**} = (T^*)^* \in B(\mathscr{H})$; moreover, $T^{**} = T$.

Remark. Using the usual definition of the adjoint T^* of linear operators T in Banach spaces, the proof also works when T is a Banach space operator.

References and Further Reading

o Berberian S K 1961 *Introduction to Hilbert Space* (Oxford: Oxford University Press)

Brockett R W 1970 *Finite Dimensional Linear Systems* (New York: Wiley)

Desoer C A and Vidysagar M 1975 *Feedback Systems (Input–Output Properties)* (New York: Academic)

Dunford N and Schwartz J 1958, 1963 *Linear Operators* vols I and II (New York: Interscience)

Erdos J A and Longstaff W E 1973 The convergence of triangular integrals of operators on Hilbert space *Indiana Univ. J.* **22** 929–38

Feintuch A and Saeks R 1982 *Systems Theory (A Hilbert Space Approach)* (New York: Academic)

o Gelfand I M 1961 *Lectures in Linear Algebra* (New York: Interscience)

o Gohberg I and Goldberg S 1981 *Basic Operator Theory* (Basel: Birkhäuser)

Gohberg I and Krein M G 1970 Theory and applications of Volterra operators in Hilbert Space vol. 24 (Providence: Am. Math. Soc.)

o Groetsch C W 1980 *Elements of Applicable Functional Analysis* (New York: Dekker)

o Gulyás O 1967 Some remarks on the sampling theorem *Proc. Colloq. Inf. Theory* ed A Rényi (Budapest: Akadémiai Kiadó) 243–51

Hackenbroch W 1969 Integraldarstellung einer Klasse dissipativer linearer Operatoren *Math. Z.* **109** 273–87

Halmos P R 1967 *A Hilbert Space Problem Book* (New York: Van Nostrand)

Kailath T and Duttweiler D L 1972 An RKHS approach to detection and estimation problems *IEEE Trans.* **IT (Information Theory)–18** 730–45

o König H 1959 Zur Theorie der linearen dissipativen Transformationen *Arch. Math.* **10** 447–51

Lamperti J 1977 *Stochastic Processes* (Berlin: Springer)

o Loy R J 1974 Continuity of linear operators commuting with shift *J. Funct. Anal.* **17** 48–60

o Luenberger D G 1969 *Optimization by Vector Space Methods* (New York: Wiley)

o Lusternik L A and Sobolev V I 1961 *Elements of Functional Analysis* (New York: Ungar)

o Naylor A W and Sell G R 1982 *Linear Operator Theory in Engineering and Science* (Berlin: Springer)

o Parzen E 1967 *Stochastic Processes* (Cambridge: Holden Day)

o Riesz F and Sz Nagy B 1955 *Functional Analysis* (New York: Ungar)

Ringrose J R 1971 *Compact Non-Self-Adjoint Operators* (New York: Van Nostrand)

○ Shapiro H S 1971 Topics in approximation theory (Berlin: Springer) *Lect. Notes Math.* **187**

Showalter R E 1977 *Hilbert Space Methods for Partial Differential Equations* (New York: Pitman).

Subject Index

Notation Index

The aim of this book is to make Hilbert space theory accessible to applied mathematicians, engineers and scientists. Based on the author's courses on functional analysis for engineering students, all that is assumed is a knowledge of linear algebra and analysis.

Hilbert Space Methods in Science and Engineering covers the theoretical fundamentals, the geometry of Hilbert spaces, reproducing kernel Hilbert spaces, operator theory including causal operators. The construction of mathematical models using Hilbert space theory is emphasised, and the results which follow from these models are evaluated. In particular, mathematical models based on reproducing kernel Hilbert spaces and causal operators are presented here at an introductory level for the first time.

ISBN 963 05 4925 5