MATHEMATICS
IN SCIENCE
AND
ENGINEERING

Volume 146

Applications of Functional Analysis and Operator Theory

V Hutson
J S Pym

INTRODUCTION TO FUNCTIONAL ANALYSIS

INTRODUCTION TO FUNCTIONAL ANALYSIS

SECOND EDITION

Angus E. Taylor
University of California,
Berkeley and Los Angeles

David C. Lay
University of Maryland
College Park

ROBERT E. KRIEGER PUBLISHING COMPANY
MALABAR, FLORIDA

Second Edition 1980
Reprint Edition 1986

Printed and Published by
ROBERT E. KRIEGER PUBLISHING COMPANY, INC.
KRIEGER DRIVE
MALABAR, FLORIDA 32950

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Taylor, Angus Ellis, 1911-
Introduction to functional analysis.

Reprint. Originally published: New York : Wiley, c1980.
Bibliography: p.
Includes index.
1. Functional analysis. I. Lay, David C. II. Title.
(QA320.T3 1986) 515.7 86-7492
ISBN 0-89874-951-4

10 9 8 7 6 5 4 3

To our wives—
PATSY and LILLIAN

PREFACE

The central theme of this book is the theory of normed linear spaces and of linear mappings between such spaces. The text provides the necessary foundation for further study in many areas of analysis, and it strives to generate an appreciation for the unifying power of the abstract linear-space point of view in surveying the problems of linear algebra, classical analysis, and differential and integral equations. While the book is principally addressed to graduate students, it is also intended to be useful to mathematicians and users of mathematics who have need of a simple and direct presentation of the fundamentals of the theory of linear spaces and linear operators.

In many respects, this new edition is similar to the first edition written by Taylor. The prerequisites are the same—the reader should already be acquainted with the fundamentals of real and complex analysis and elementary point set topology. The manner and level of presentation are essentially unchanged, although the scope of the text has been broadened somewhat, and the emphasis on concrete examples and connections with classical mathematics has been retained.

The revision was made in order to incorporate recent developments in functional analysis and to make the selection of topics more appropriate for current courses in functional analysis. Significant additions to this new edition include a chapter on Banach algebras, and material on weak topologies and duality, equicontinuity, the Krein–Milman theorem, and the theory of Fredholm operators. Furthermore, there is greater emphasis on closed unbounded linear operators, with more illustrations drawn from ordinary differential equations. Two background chapters from the first edition (on topology and some topics in integration theory) have been omitted because the material has become part of the standard curriculum. A few facts from those chapters are now reviewed as needed.

The problems in the text, increased from 300 to nearly 500, have been carefully selected—they both illustrate and extend the theory, and they give the reader an opportunity to construct arguments similar to those in the text. The ℓ^p sequence spaces have been chosen for some of the concrete problems and examples, in order to minimize the technical difficulties. In addition, there are problems that relate to differential equations, integral equations, and the theory of analytic functions.

The book begins with a concise review of linear algebra and an introduction to linear problems in analysis, omitting topological considerations. An analytic version of the Hahn–Banach theorem is given in § I.10. A well-prepared reader may start with Chapter II, using Chapter I as a reference when necessary. The first part of Chapter II is on normed linear spaces. However, the basic theory of topological linear spaces is important because it is frequently used today in analysis and because it is relevant in the study of normed linear spaces. This basic theory is developed in parts of Chapters II and III. Certain optional topics are identified in the introductions to these chapters.

The subject of linear operators is begun in detail in Chapter IV, with some of the most important results in Chapters IV and V depending on completeness of the underlying spaces. However, unless needed for effective results, the hypothesis of completeness is not invoked. The exposition is not materially lengthened by this greater generality. The more specialized and distinctive theory of operators on Hilbert space is presented in §§ 3, 11, and 12 of Chapter IV.

Chapter V stresses the importance of complex contour integration and the calculus of residues in the spectral theory of linear operators. The methods apply to all closed linear operators, bounded or not. Although some familiarity with complex analysis is assumed in the text, the main theorems needed for Chapter V are reviewed in § V.1. This chapter also contains the famous Riesz theory of compact operators, as extended and perfected by later research workers, and its application to the classical "determinant-free" theorems for Fredholm integral equations of the second kind. The subject of invariant subspaces is treated briefly in this chapter as well.

Chapter VI presents the standard elementary theory of self-adjoint, normal, and unitary operators on Hilbert space. The discussion of the theory of compact symmetric operators and symmetric operators with compact resolvent is very important for applications to integral and differential equations. The completeness of the inner-product space under consideration is not required here. The spectral analysis of self-adjoint operators is performed with the aid of the Riesz representation theorem for linear functionals on a space of continuous functions. The treatment is deliberately kept as close as possible to classical analysis.

General Banach algebras and the Gelfand theory of commutative Banach algebras are discussed in Chapter VII. Most of the development assumes the existence of a unit, but some examples and problems show how the general theory would proceed for algebras such as $L^1(\boldsymbol{R})$ that lack a unit. The chapter is largely independent of Chapters IV to VI, except for the material in the early sections of these three chapters. The text returns to operator theory in the final section of Chapter VII, where two versions of the spectral theorem for normal operators on Hilbert space are derived from the Gelfand–Naimark theorem for commutative B^*-algebras.

The senior author (Taylor) wishes to include here a personal acknowledgment of thanks to his coauthor: When I left U.C.L.A. to become the academic vice-president of the University of California multicampus system, I soon realized that my busy schedule would not enable me to carry forward the task of revising the book and of tuning it anew to the needs of an oncoming generation of students. Fortunately, David Lay, to whom I had been close in his student days at U.C.L.A., was able and willing to coauthor this book. He has done most of the revision, but we have been in close touch throughout. I am both obligated by what he has done and immensely pleased by his accomplishment and judgment.

One of the pleasures in writing books comes from the opportunities afforded authors to improve the breadth and depth of their own understanding of a subject in relation to its origins and applications. We hope that some of the same pleasure will accrue to readers of this text.

We are especially grateful to Professor Steven Lay who worked closely with the junior author one summer and who made valuable contributions during the early stages of the revision. We are also indebted to Professor Denny Gulick whose suggestions and critical analysis of the manuscript led to substantial improvements in the text. Our appreciation is extended to Professors John Brace, Robert Ellis, and John Horváth, who class-tested portions of the text material and made helpful comments, and to Professor Seymour Goldberg for many stimulating discussions. We also wish to thank Professor Ronald Douglas for his review of the manuscript and useful remarks.

Finally, we wish to thank our many students and colleagues who have, through communications or conversations, influenced us in specific or perhaps even unrealized ways and thereby, we hope, helped make a better book.

Berkeley, California
College Park, Maryland
March 1979

Angus E. Taylor
David C. Lay

CONTENTS

INTRODUCTION

It is the purpose of this introduction to explain certain terminology used throughout the book and to list some inequalities for easy reference.

FUNCTIONS

Let X and Y be arbitrary nonempty sets and let $\mathscr{D}$ be a nonempty subset of X. A *function* f from $\mathscr{D}$ into Y is a rule that to each $x \in \mathscr{D}$ assigns a unique element $f(x)$ in Y. We sometimes denote the function by the expression $x \mapsto f(x)$. The *domain* of f is the set $\mathscr{D}$, often written as $\mathscr{D}(f)$, and the *range* of f is the set $\mathscr{R}(f) = \{f(x) : x \in \mathscr{D}\}$. The *graph* of f (sometimes identified with f itself) is the set of ordered pairs $\{(x, f(x)) : x \in \mathscr{D}\}$. This is a particular kind of nonempty subset of the Cartesian product $X \times Y$ of all ordered pairs (x, y), where $x \in X$, $y \in Y$. A function g is said to be a *restriction* of f, and f an *extension* of g, if $\mathscr{D}(g) \subset \mathscr{D}(f)$ and $g(x) = f(x)$ for $x \in \mathscr{D}(g)$.

We say that a function f is *injective*, or *one-to-one*, if for each y in the range $\mathscr{R}(f)$ there exists only one x in the domain $\mathscr{D}(f)$ such that $f(x) = y$, and we denote this unique x by $f^{-1}(y)$. When f is injective, the correspondence $y \mapsto f^{-1}(y)$ is a function f^{-1} called the *inverse* of f, whose domain is $\mathscr{R}(f)$ and range is $\mathscr{D}(f)$. We sometimes say that f^{-1} exists, in place of saying that f is injective.

If each y in Y is in the range of f, we say that f is *surjective*, or that f maps its domain *onto* Y. If f is both injective and surjective, we say that f is *bijective*. In this case f^{-1} maps Y onto the domain of f.

Given $A \subset X$, $B \subset Y$ and f as above, we use the following notation.

$$f(A) = \{f(x) : x \in A \cap \mathscr{D}(f)\},$$

$$f^{-1}(B) = \{x \in \mathscr{D}(f) : f(x) \in B\}.$$

We call $f(A)$ the image of A under f. Note that $f(A) = \varnothing$, where $\varnothing$ denotes the empty set, if $A \cap \mathscr{D}(f) = \varnothing$. We call $f^{-1}(B)$ the inverse image of B under f. We write $f^{-1}(B)$ even though f^{-1} may not exist as a function. [In fact, f^{-1} exists if and only if $f^{-1}(B)$ is a set consisting of a single element of $\mathscr{D}(f)$ whenever B is a set consisting of just one element of $\mathscr{R}(f)$.]

REAL AND COMPLEX NUMBERS

The real number system is denoted by $\boldsymbol{R}$, the complex numbers by $\boldsymbol{C}$. The *real* and *imaginary parts* of a complex number λ are written as $\operatorname{Re}\lambda$ and $\operatorname{Im}\lambda$, respectively. On occasion it will be necessary to work within the extended real number system $\boldsymbol{R} \cup \{+\infty\} \cup \{-\infty\}$. Algebraic operations in this system are discussed in Taylor [5, pages 178–180].* The *least upper bound* (*supremum*) and *greatest lower bound* (*infimum*) of a nonempty set S of real numbers always exist in the extended real number system; they are denoted by $\sup S$ and $\inf S$, respectively.

TOPOLOGY

We assume the reader is familiar with the basic definitions and theorems of topology, such as those presented in Taylor [5, pages 89–140]. The *closure* of a set A in a topological space X is denoted by $\bar{A}$, the *interior* by $\operatorname{int}(A)$. We say that a set V in X is a *neighborhood* of a point $x \in X$ if there exists an open set U such that $x \in U$ and $U \subset V$. (This usage follows the Bourbaki tradition and differs from Taylor [5], where a neighborhood of x means an open set containing x.) The space X is called a T_1 space if every set consisting of a single point is closed. The space is a *Hausdorff* space (or T_2 space) if for each pair of distinct points $x_1, x_2 \in X$ there exist disjoint neighborhoods of x_1 and x_2, respectively. If τ_1 and τ_2 are topologies for the same set X, then τ_1 is called *weaker* than τ_2 (equivalently, τ_2 is *stronger* than τ_1) if every τ_1-open set is τ_2-open.

THE KRONECKER DELTA

The symbol δ_{ij} denotes the number 1 if $i = j$ and the number 0 if $i \neq j$. Here i and j are positive integers.

INEQUALITIES

At a number of places in this book we use some of the standard inequalities concerning sums and integrals. We list the most commonly used ones here. The standard reference work on this subject is the book, *Inequalities*, by Hardy, Littlewood, and Pólya [1]. In what follows we refer to this book as H, L, and P, and cite by number the section in which the stated inequality is discussed. Most of the inequalities are given as exercises, with hints for

* References to the bibliography are made by listing the author's name and a number in brackets, identifying a book or article by that author.

solutions, in Taylor [5, pages 119–120, 278]. In all inequalities the quantities involved may be either real or complex. Sums are either all from 1 to n or from 1 to ∞, and in the latter case certain evident assumptions and implications of convergence are involved. For simplicity the inequalities for integrals are written for the case in which the functions are defined on a finite or infinite interval of the real axis. The inequalities are valid with more general interpretations of the set over which integration is extended.

Hölder's inequality for sums (H, L, and P, § 2.8): If $1<p<\infty$ and $p'=p/(p-1)$, then

$$\sum|a_ib_i| \leqslant \left(\sum|a_i|^p\right)^{1/p}\left(\sum|b_i|^{p'}\right)^{1/p'}.$$

The special case when $p=p'=2$ is called Cauchy's inequality (H, L, and P, § 2.4).

Minkowski's inequality for sums (H, L, and P, § 2.11): If $1\leqslant p<\infty$, then

$$\left(\sum|a_i+b_i|^p\right)^{1/p} \leqslant \left(\sum|a_i|^p\right)^{1/p}+\left(\sum|b_i|^p\right)^{1/p}.$$

Jensen's inequality (H, L, and P, § 2.10): If $0<p<q$, then

$$\left(\sum|a_i|^q\right)^{1/q} \leqslant \left(\sum|a_i|^p\right)^{1/p}.$$

Hölder's inequality for integrals (H, L, and P, § 6.9): If $1<p<\infty$ and $p'=p/(p-1)$, then

$$\int |f(x)g(x)|\,dx \leqslant \left(\int |f(x)|^p\,dx\right)^{1/p}\left(\int |g(x)|^{p'}\,dx\right)^{1/p'}.$$

The special case when $p=p'=2$ is called the Schwarz inequality (H, L, and P, § 6.5).

Minkowski's inequality for integrals (H, L, and P, § 6.13): If $1\leqslant p<\infty$, then

$$\left(\int |f(x)+g(x)|^p\,dx\right)^{1/p} \leqslant \left(\int |f(x)|^p\,dx\right)^{1/p}+\left(\int |g(x)|^p\,dx\right)^{1/p}.$$

I THE ABSTRACT APPROACH TO LINEAR PROBLEMS

The modern treatment of many topics in pure and applied mathematics is characterized by the effort that is made to strip away nonessential details and to show clearly the fundamental assumptions and the structure of the reasoning. This effort often leads to some degree of abstraction, with the concrete nature of the originally contemplated problem being temporarily put aside and the aspects of the problem that are of greatest significance being cast into axiomatic form. It is found that in this way there is a considerable gain in transparency and that diverse problems exhibit common characteristics that enable them all to be at least partially solved by the methods of a single general theory.

In this chapter we consider the algebraic aspects of such an abstract approach to linear problems. In essence, all linear problems are viewed in some measure as analogous to the linear problems exhibited in elementary algebra by the theory of systems of linear equations. The linear problems of analysis usually require topological as well as algebraic considerations. However, in this chapter, we exclude all concern with topology; the topological aspects of the abstract approach to linear problems will be taken up in later chapters.

The most profound results of the chapter are the extension theorems in § 10 (Theorems 10.1 and 10.4) and Theorem 11.2 on the existence of a complementary subspace. They all depend on Zorn's lemma (§ 9). Theorem 10.4 is one version of the important Hahn–Banach theorem. Other versions will be discussed in § III.2 and § III.3.

Chapter I culminates in § 13 with two theorems relating the range and null space of a linear operator to the null space and range of the transpose of the operator (Theorems 13.4 and 13.5). These theorems furnish information on existence and uniqueness theorems for certain kinds of linear problems. For the finite-dimensional case these theorems include the standard results concerning algebraic systems of linear equations. In the infinite-dimensional case the results are not as useful as results that can be obtained with the aid of metric or topological tools. Nevertheless, the material of § 13 points the way to more incisive results, some of which are given in § IV.8.

I.1 ABSTRACT LINEAR SPACES

We have as yet made no formal definition of what is meant by the adjective *linear* in the phrase "linear problems." We can cite various particular kinds of linear problems: the problems of homogeneous and inhomogeneous systems of linear equations in n "unknowns" in elementary algebra; the problems of the theory of linear ordinary differential equations (existence theorems, particular and general solutions, problems of finding solutions satisfying given conditions at one or two end points); boundary or initial-value problems in the theory of linear partial differential equations; problems in the theory of linear integral equations; linear "transform" problems, for example, problems related to Fourier and Laplace transforms. This is by no means an exhaustive list of the types of mathematical situations in which linear problems arise.

At the bottom of every linear problem is a mathematical structure called a *linear space*. We shall, therefore, begin with an axiomatic treatment of abstract linear spaces.

Definition. Let X be a set of elements, hereafter sometimes called *points*, and denoted by small italic letters: $x, y, \ldots$. We assume that each pair of elements x, y can be combined by a process called addition to yield another element z denoted by $z = x + y$. We also assume that each real number α and each element x can be combined by a process called multiplication to yield another element y denoted by $y = \alpha x$. The set X with these two processes is called a *linear space* if the following axioms are satisfied:

1. $x + y = y + x$.
2. $x + (y + z) = (x + y) + z$.
3. There is in X a unique element, denoted by 0 and called the zero element, such that $x + 0 = x$ for each x.
4. To each x in X corresponds a unique element, denoted by $-x$, such that $x + (-x) = 0$.
5. $\alpha(x + y) = \alpha x + \alpha y$.
6. $(\alpha + \beta)x = \alpha x + \beta x$.
7. $\alpha(\beta x) = (\alpha\beta)x$.
8. $1 \cdot x = x$.
9. $0 \cdot x = 0$.

Anyone who is familiar with the algebra of vectors in ordinary three-dimensional Euclidean space will see at once that the set of all such vectors forms a linear space. An abstract linear space embodies so many of the features of ordinary vector algebra that the word *vector* has been taken over into a more general context. A linear space is often called a *vector space*, and the elements of the space are called *vectors*.

In the foregoing list of axioms it was assumed that the multiplication operation was performed with *real* numbers, α, β. To emphasize this, if necessary, we call the space a *real* linear space, or a *real* vector space. An alternative notion of a linear space is obtained if it is assumed that any *complex* number α and any element x can be multiplied, yielding another element αx. The axioms are the same as before. The space is then called a *complex* linear space.

The notion of a vector space is defined even more generally in abstract algebra, by allowing the multipliers $\alpha, \beta, \ldots$ to be elements of an arbitrary commutative field. In this book, however, we confine ourselves to the two fields of real and complex numbers, respectively. The elements of the field are called *scalars*, to contrast with the *vector* elements of the linear space.

It is easy to see that $-1 \cdot x = -x$ and that $\alpha \cdot 0 = 0$. We write $x - y$ for convenience in place of $x + (-y)$. The following "cancellation" rules are also easily deduced from the axioms:

(1-1) $$x + y = x + z \text{ implies } y = z;$$

(1-2) $$\alpha x = \alpha y \text{ and } \alpha \neq 0 \text{ imply } x = y;$$

(1-3) $$\alpha x = \beta x \text{ and } x \neq 0 \text{ imply } \alpha = \beta.$$

Definition. A nonempty subset M of a linear space X is called a *linear manifold* in X if $x + y$ is in M whenever x and y are both in M and if also αx is in M whenever x is in M and α is any scalar.

In this definition and generally throughout the book, statements made about linear spaces, without qualification as to whether the space is real or complex, will be intended to apply equally to real spaces and complex spaces.

It will be seen at once that, if M is a linear manifold in X, it may be regarded as a linear space by itself. For, if x is in M, then $-1 \cdot x = -x$ is also in M, and $x - x = 0$ is also in M. The nine axioms for a linear space are now found to be satisfied in M. Another term for a linear manifold in X is *subspace* of X. A subspace of X is called *proper* if it is not all of X.

The set consisting of 0 alone is a subspace. We denote it by (0).

Suppose S is any nonempty subset of X. Consider the set M of all finite linear combinations of elements of S, that is, elements of the form $\alpha_1 x_1 + \cdots + \alpha_n x_n$, where n is any positive integer (not fixed), $x_1, \ldots, x_n$ are any elements of S, and $\alpha_1, \ldots, \alpha_n$ are any scalars. This set M is a linear manifold. It is called the linear manifold generated, or determined, by S. Sometimes we speak of M as the linear manifold *spanned* by S. It is easy to verify the truth of the following statements: (1) M consists of those vectors that belong to every linear manifold that contains S; that is, M is the intersection of all such manifolds. (2) M is the smallest linear manifold that contains S; that is, if N is a linear manifold that contains S, then M is contained in N.

One of the most important concepts in a vector space is that of linear dependence.

Definition. A finite set of vectors $x_1, \ldots, x_n$ in the space X is *linearly dependent* if there exists scalars $\alpha_1, \ldots, \alpha_n$, not all zero, such that $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$. If the finite set $x_1, \ldots, x_n$ is not linearly dependent, it is called *linearly independent*. In that case, a relation $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$ implies that $\alpha_1 = \cdots = \alpha_n = 0$. An infinite set S of vectors is called linearly independent if every finite subset of S is linearly independent; otherwise S is called linearly dependent.

Observe that if a set of vectors contains a linearly dependent subset, the whole set is linearly dependent. Also note that a linearly independent set cannot contain the vector 0.

We note the following simple but important theorem, of which use will be made in later arguments:

Theorem 1.1. *Suppose $x_1, \ldots, x_n$ is a set of vectors with $x_1 \neq 0$. The set is linearly dependent if and only if some one of the vectors $x_2, \ldots, x_n$, say x_k, is in the linear manifold generated by $x_1, \ldots, x_{k-1}$.*

Proof. Suppose the set is linearly dependent. There is a smallest integer k, with $2 \leq k \leq n$, such that the set $x_1, \ldots, x_k$ is linearly dependent. This dependence is expressed by an equation $\alpha_1 x_1 + \cdots + \alpha_k x_k = 0$, with not all the α's equal to zero. Necessarily, then, $\alpha_k \neq 0$, for otherwise $x_1, \ldots, x_{k-1}$ would form a linearly dependent set. Consequently, $x_k = \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$, where $\beta_i = -\alpha_i / \alpha_k$. This shows that x_k is in the manifold spanned by $x_1, \ldots, x_{k-1}$. On the other hand, if we assume that some x_k is in the linear manifold spanned by $x_1, \ldots, x_{k-1}$, then an equation of the form $x_k = \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}$ shows that the set $x_1, \ldots, x_k$ is linearly dependent, whence the same is true of the set $x_1, \ldots, x_n$. $\square$

It is convenient to say that x is a *linear combination* of $x_1, \ldots, x_n$ if it is in the linear manifold spanned by these vectors.

Using the notion of linear dependence, we can define the concept of a finite-dimensional vector space.

Definition. Let X be a vector space. Suppose there is some positive integer n such that X contains a set of n vectors that are linearly independent, while every set of $n+1$ vectors in X is linearly dependent. Then X is called *finite dimensional* and n is called the dimension of X. We sometimes write $\dim X = n$. A vector space with just one element (which must then be the zero element) is also called finite dimensional, of dimension zero. If X is not finite dimensional, it is called infinite dimensional.

As we shall see later, the spaces of greatest interest in analysis are infinite dimensional. Nevertheless, it will often be of use to consider finite-dimensional spaces. Such spaces are, moreover, the source of much of our intuitive perception about what to expect in dealing with linear spaces generally.

Definition. A finite set S in a space X is called a *basis* of X if S is linearly independent and if the linear manifold generated by S is all of X.

If $x_1, \ldots, x_n$ is a basis of X, the definition means that every x in X can be expressed in the form $x = \xi_1 x_1 + \cdots + \xi_n x_n$. Since the basis is a linearly independent set, the coefficients $\xi_1, \ldots, \xi_n$ are uniquely determined by x; that is, x cannot be expressed as a *different* linear combination of the basis elements.

It is readily seen that if X is n-dimensional, where $n \geq 1$, then X has a basis consisting of n elements. For, X certainly contains vectors $x_1, \ldots, x_n$ that form a linearly independent set. Now, for any x, the set of $n+1$ vectors $x_1, x_2, \ldots, x_n, x$ must be linearly dependent, by the definition of the dimensionality of X. Hence it is clear, from the proof of Theorem 1.1, that x is in the linear manifold spanned by $x_1, \ldots, x_n$. This shows that $x_1, \ldots, x_n$ form a basis of X. This same argument shows that every linearly independent set of n elements in an n-dimensional linear space X is a basis of X.

Next we wish to show that if X has a basis of n elements, then X is n-dimensional. First we prove a lemma.

Lemma 1.2. *If the finite set $x_1, \ldots, x_n$ generates X and if $y_1, \ldots, y_m$ are elements of X forming a linearly independent set, then $m \leq n$.*

Proof. If S is any linearly dependent finite ordered set of vectors $u_1, \ldots, u_p$ with $u_1 \neq 0$, let S' denote the ordered set that remains after the deletion of the first u_i that is a linear combination of its predecessors. Also, for any y let yS denote the ordered set $(y, u_1, \ldots, u_p)$. Now define S_1 to be the set $(y_m, x_1, \ldots, x_n)$, $S_2 = y_{m-1}S'_1$, $S_3 = y_{m-2}S'_2$, and so on. We make several observations: (1) S_1 spans X, is a linearly dependent set, and $y_m \neq 0$. Hence we can form S'_1 (see Theorem 1.1). (2) S'_1 and hence also S_2, spans X. We can continue in this way, constructing new sets S and S' as long as the y's last. Since the set of y's is linearly independent, the discarded element at each step must be an x. Since we can form $S'_1, \ldots, S'_m$, it follows that we discard m x's, and hence $n \geq m$, as asserted in the lemma. $\square$

Theorem 1.3. *If the linear space X has a basis of n elements, then X is n-dimensional, and conversely. Consequently, each basis of an n-dimensional linear space consists of n elements.*

Proof. If X has a basis of n elements, any linearly independent set in X has at most n elements, by Lemma 1.2. On the other hand, the basis is a set of n linearly independent elements. Hence X is n-dimensional, by definition. The proof of the converse was given just before Lemma 1.2. The second statement of the theorem follows immediately from the first. □

The following theorem will be useful later.

Theorem 1.4. *Let X be an n-dimensional linear space, and let the set $y_1, \dots, y_m$ be linearly independent, with $m < n$. Then there exists a basis of X composed of $y_1, \dots, y_m$ and $n-m$ other vectors.*

Proof. Let $x_1, \dots, x_n$ be a basis of X. Let S_1 be the ordered set $(y_1, \dots, y_m, x_1, \dots, x_n)$, and let $S_2 = S_1'$, $S_3 = S_2'$, and so on, where S' is related to S in the manner explained in the proof of Lemma 1.2. Observe that S_1 is certainly linearly dependent and that the deleted vector at each step is one of the x's (by Theorem 1.1), so that each S_k includes $y_1, \dots, y_m$. Since S_1 spans X, so does S_1', and so on. We can form S_k' if and only if S_k is linearly dependent. Ultimately we reach a stage where S_k is linearly independent and spans X. It is then a basis of X and includes $y_1, \dots, y_m$. Since the basis must have n elements, there are $n-m$ vectors in addition to $y_1, \dots, y_m$. □

It is natural to expect that if X is a linear space of dimension n, every subspace of X has dimension not exceeding n. This is indeed the case.

Theorem 1.5. *Suppose the vector space X is of dimension n, and let M be a proper subspace of X. Then M is of some finite dimension m, where $m < n$.*

Proof. If $M = (0)$, it is of dimension 0, by definition. Since X must contain nonzero elements, $n > 0$, so the assertion of the theorem is true in this case. We now assume that M contains nonzero elements. If x_0 is a nonzero element of M, it forms a linearly independent set in M. On the other hand, a linearly independent set in M cannot contain as many as n elements. For, a linearly independent set of n elements is a basis of X (as we saw following the definition of a basis), so that in such a case, X would be generated by M; this is impossible, since $M \neq X$, and the linear manifold generated by the manifold M is M itself. We now consider the nonempty class of all linearly independent sets in M. Each such set has a finite number of elements p, where $1 \leqslant p < n$. There is therefore a set for which p is largest, say $p = m$. It is immediate that $m < n$ and that M is of dimension m. □

Before going on to consider the general form of linear problems in linear spaces, we illustrate to some extent the great variety of possible examples of linear spaces.

I.2. EXAMPLES OF LINEAR SPACES

Example 1. The simplest important example of a real linear space is the set of all n-tuples of real numbers, $x = (\xi_1, \ldots, \xi_n)$. The definitions of addition and multiplication by scalars in this space are as follows:

If $x = (\xi_1, \ldots, \xi_n)$ and $y = (\eta_1, \ldots, \eta_n)$, then $z = x + y$, where $z = (\zeta_1, \ldots, \zeta_n)$, $\zeta_k = \xi_k + \eta_k$, $k = 1, \ldots, n$. The vector αx is $(\alpha\xi_1, \ldots, \alpha\xi_n)$. Here the ξ's, η's, and α are arbitrary real numbers, and n is an arbitrary, fixed positive integer. We define $0 = (0, \ldots, 0)$ and $-x = (-\xi_1, \ldots, -\xi_n)$. It is an easy matter to verify that the nine axioms for a linear space are satisfied.

The dimension of this space is n; we prove this by exhibiting a basis consisting of n elements. Let $e_1 = (1, 0, \ldots, 0)$, $e_2 = (0, 1, 0, \ldots, 0), \ldots, e_n = (0, 0, \ldots, 0, 1)$. If $x = (\xi_1, \ldots, \xi_n)$, observe that $x = \xi_1 e_1 + \cdots + \xi_n e_n$. Thus the set $e_1, \ldots, e_n$ generates the whole space. Moreover, the set is linearly independent, for $\alpha_1 e_1 + \cdots + \alpha_n e_n = (\alpha_1, \ldots, \alpha_n) = 0$ if and only if all the α's are zero. Thus $e_1, \ldots, e_n$ constitute a basis.

We call this space *n-dimensional real arithmetic space*, and denote it by $\boldsymbol{R}^n$.

The space $\boldsymbol{R}^n$ has a familiar geometrical interpretation, $\xi_1, \ldots, \xi_n$ being the coordinates of the point x in a system of Cartesian coordinates. Thus $\boldsymbol{R}^1$ is interpreted as a line, $\boldsymbol{R}^2$ as a plane, and so on. In the geometrical interpretation of $\boldsymbol{R}^n$ we may regard the element x either as a point or as the vector from 0 (the origin in $\boldsymbol{R}^n$) to that point. For geometrical interpretations this latter point of view is in many ways the most fruitful.

Example 2. The set of all n-tuples $x = (\xi_1, \ldots, \xi_n)$ of *complex* numbers forms a complex linear space, which we call *n-dimensional complex arithmetic space*, and denote by $\boldsymbol{C}^n$.

The n vectors $e_1 = (1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$ form a basis of $\boldsymbol{C}^n$. It will be observed that the elements of $\boldsymbol{R}^n$ belong to $\boldsymbol{C}^n$; however, $\boldsymbol{R}^n$ is not a subspace of $\boldsymbol{C}^n$, for, if α is complex and x is in $\boldsymbol{R}^n$, then αx is in $\boldsymbol{C}^n$ but not always in $\boldsymbol{R}^n$ (e.g., $i \cdot e_1 = (i, 0, \ldots, 0)$ is not in $\boldsymbol{R}^n$).

The set $\boldsymbol{C}^n$ can also be regarded as a *real* linear space, by using the real field for scalar multipliers. But then the space is not of dimension n, but of dimension $2n$, one possible basis consisting of the vectors $e_1, \ldots, e_n$ and $ie_1, \ldots, ie_n$. We shall hereafter always regard $\boldsymbol{C}^n$ as a *complex* space.

Example 3. Let $[a, b]$ (with $a<b$) be a finite closed interval of the real axis, and let x denote a continuous real-valued function whose value at the point s of $[a, b]$ is $x(s)$. Let $C[a, b]$ denote the set of all such functions, and define x_1+x_2, αx in the natural manner, that is, $z=x_1+x_2$, where the value of z at s is $z(s)=x_1(s)+x_2(s)$; $y=\alpha x$, where the value of y at s is $y(s)=\alpha x(s)$. It is clear that $C[a, b]$ is a real vector space. We might equally well have considered complex-valued continuous functions of s; in that case we would have obtained a complex vector space. We shall denote this space by $C[a, b]$ also; thus, in speaking of the space $C[a, b]$, we shall have to make clear by an explicit statement whether we are talking about real-valued or complex-valued functions. In either case, $C[a, b]$ is infinite dimensional. For, let $x_0(s)=1, x_n(s)=s^n, n=1, 2, \ldots$. Evidently, $x_0, x_1, \ldots, x_n$ all belong to $C[a, b]$. This set of elements is linearly independent, no matter how large n is; for, by well-known properties of polynomials, if $\alpha_0+\alpha_1 s+\cdots+\alpha_n s^n=0$ for every s such that $a \leqslant s \leqslant b$, then $\alpha_0=\alpha_1=\cdots=\alpha_n=0$. Therefore $C[a, b]$ cannot be finite dimensional.

The role of $[a, b]$ is inessential in the demonstration that $C[a, b]$ is a vector space and of infinite dimension. For example, we could equally well have considered continuous functions of a complex variable that ranges over some fixed infinite point set in the complex plane. An infinite set is essential to make the space infinite dimensional.

A great many of the linear spaces that are of interest in analysis are spaces whose elements are functions. We shall mention a few further examples.

Example 4. Let f be a function of the complex variable z, which is defined (single valued), analytic, and bounded in the open unit disc $|z|<1$. The class of all such functions becomes a complex vector space when $f+g$ and αf are defined in the natural way. This space is infinite dimensional. As a subspace we mention the class of all those functions f in the space for which $f(0)=0$. Another subspace is the class of those f's whose definitions can be extended to the boundary $|z|=1$ in such a way that each f is continuous in the closed disc $|z| \leqslant 1$. Both of these subspaces are infinite dimensional.

Example 5. Consider the complex-valued functions x of the real variable s such that the derivatives x' and x'' are defined and continuous on the closed interval $[0, \pi]$. The set of all such functions is a linear space. It is of interest in considering ordinary second-order linear differential equations with coefficients continuous on $[0, \pi]$. The set of all elements of the space that satisfy the equation $x''(s)+x(s)=0$ is a subspace of dimension 2. A basis of the subspace is furnished by the functions e^{is}, e^{-is}. Another basis is furnished by $\sin s$ and $\cos s$. Another subspace of interest is that consisting of all elements of the space such that $x(0)=x(\pi)=0$. This subspace is infinite dimensional, for it contains the infinite linearly independent set consisting of $\sin ns$, $n = 1, 2, \ldots$.

Example 6. As elements of a space consider infinite sequences $x = \{\xi_n\}(n = 1, 2, \ldots)$ such that $\Sigma_{n=1}^{\infty} |\xi_n|^2 < \infty$, the ξ's being complex numbers. Define $\alpha x = \{\alpha\xi_n\}$ and $\{\xi_n\} + \{\eta_n\} = \{\xi_n + \eta_n\}$. It is readily seen (by Minkowski's inequality with $p = 2$) that this is an infinite-dimensional complex linear space. We denote it by ℓ^2. This space was first extensively studied by D. Hilbert in his work on quadratic forms in infinitely many variables, with applications to the theory of integral equations. On this account, ℓ^2 is the classical prototype of what is known today as a *Hilbert* space. Hilbert spaces are linear spaces with a certain special kind of metrical structure. We discuss them extensively elsewhere in this book.

Example 7. Suppose $p \geq 1$ (p not necessarily an integer). Let $\mathscr{L}^p$ denote the class of all functions x of the real variable s such that $x(s)$ is defined for all s, with the possible exception of a set of measure zero, and is measurable and $|x(s)|^p$ is integrable (in the Lebesgue sense) over the range $(-\infty, \infty)$. Instead of $(-\infty, \infty)$ we could equally well consider $(0, \infty)$ or any finite interval (a, b). We write $\mathscr{L}^p(-\infty, \infty)$, $\mathscr{L}^p(0, \infty)$, $\mathscr{L}^p(a, b)$ to distinguish these various situations. Also, we usually omit the index p when $p = 1$, writing $\mathscr{L}$ for $\mathscr{L}^1$.

Let $\mathscr{D}_x$ denote the set on which x is defined. We define αx as the function $(\alpha x)(s) = \alpha x(s)$, with $\mathscr{D}_{\alpha x} = \mathscr{D}_x$, and $x + y$ as the function $(x + y)(s) = x(s) + y(s)$, with $\mathscr{D}_{x+y} = \mathscr{D}_x \cap \mathscr{D}_y$. Clearly $\alpha x \in \mathscr{L}^p$ if $x \in \mathscr{L}^p$; it is also true that $x + y \in \mathscr{L}^p$ if $x, y \in \mathscr{L}^p$. This latter fact follows from the inequality

$$|A + B|^p \leq 2^p[|A|^p + |B|^p], \tag{2-1}$$

where A and B are any real or complex numbers. For, if w is measurable, if z is integrable, and if $|w(s)| \leq |z(x)|$, then w is integrable. To see the truth of (2-1), observe that

$$\max\{|A|, |B|\} \leq |A| + |B| \leq 2 \max\{|A|, |B|\},$$

$$|A + B|^p \leq (|A| + |B|)^p \leq (2 \max\{|A|, |B|\})^p$$

$$= \max\{2^p|A|^p, 2^p|B|^p\} \leq 2^p|A|^p + 2^p|B|^p.$$

If we define $x = y$ to mean that $\mathscr{D}_x = \mathscr{D}_y$ and $x(s) = y(s)$ for every $s \in \mathscr{D}_x$, we may at first think that $\mathscr{L}^p$ is a linear space. It is not, however. For, if $\mathscr{L}^p$ were a linear space, the zero element z would necessarily be the function defined for *all* s, with $z(s) = 0$ for every s. But then $x + (-x) = z$ would not be true for all x, as we see by choosing an x for which $\mathscr{D}_x$ does not include every s. To get around this difficulty we proceed as follows: Define an equivalence relation $=^0$ in $\mathscr{L}^p$ by saying that $x =^0 y$ if $x(s) = y(s)$ a.e. (i.e., almost everywhere, which means except on a set of measure zero). The set of equivalence classes into which $\mathscr{L}^p$ is thus divided is denoted by L^p; here also, we write L for L^1. For the time being we shall denote an element of L^p that contains x by $[x]$. We define $[x]+[y]=[x+y]$. This definition of $[x]+[y]$ is unambiguous, for, if $x \overset{0}{=} x_1$ and $y \overset{0}{=} y_1$, it follows that $x_1+y_1 \overset{0}{=} x+y$. Likewise, we define $\alpha[x]=[\alpha x]$, noting that $\alpha x \overset{0}{=} \alpha x_1$ if $x \overset{0}{=} x_1$. With these definitions the class L^p becomes a linear space; the zero element of L^p is the equivalence class consisting of all $x \in \mathscr{L}^p$ such that $x(s)=0$ a.e.

In practice we usually ignore the notational distinction between $\mathscr{L}^p$ and L^p and write x instead of $[x]$. When we do this it must be remembered that x really denotes, not a single function, but a class of equivalent functions.

The space L^p is of interest in connection with Fourier transforms and various kinds of integral equation problems. The case $p=2$ is especially important.

As an example of a subclass of $\mathscr{L}^p$ that is useful in certain applications to first-order ordinary differential equation problems, we mention the following: the class of all x for which $x(s)$ is defined everywhere and is absolutely continuous on every finite interval and both $x(s)$ and $x'(s)$ define elements of $\mathscr{L}^p$.

Example 8. Let $[a, b]$ be a finite closed interval of the real axis. Let $BV[a, b]$ denote the class of real-valued functions of s that are defined and of bounded variation on $[a, b]$. This class is a linear space.

Example 9. Let x be a complex-valued function of the real variable s that is defined and has derivatives of all orders for every value of s. The class of all such functions x is evidently a linear space.

It would be easy to give many more examples of linear spaces whose elements are certain kinds of functions. We shall see later on in this book that the study of linear spaces leads us to introduce still other linear spaces composed of functions defined on the original linear spaces.

I.3 LINEAR OPERATORS

A linear operator is a certain kind of function whose domain is a linear space and whose range is contained in another linear space (possibly the same as the first one). For the terminology concerning functions we refer the reader to the Introduction. If A is a linear operator, it is customary to omit parentheses and to write Ax instead of $A(x)$ whenever it seems convenient. Also, if M is in the domain of A, we write AM or $A(M)$ for the set $\{Ax : x \in M\}$.

Definition. Let X and Y be linear spaces (both real or both complex). Let A be a function with domain $\mathscr{D}(A)$ in X and range $\mathscr{R}(A)$ in Y. Then A is called a ***linear operator*** if $\mathscr{D}(A)$ is a subspace of X and if

$$A(x_1+x_2)=Ax_1+Ax_2, \tag{3-1}$$

$$A(\alpha x)=\alpha Ax, \tag{3-2}$$

whenever α is a scalar and x_1, x_2, x are vectors in $\mathscr{D}(A)$. If $\mathscr{D}(A)=X$, we often say that A is a linear operator on (or from) X into Y.

It follows immediately by induction from (3-1) and (3-2) that

$$A(\alpha_1x_1+\cdots+\alpha_nx_n)=\alpha_1Ax_1+\cdots+\alpha_nAx_n, \tag{3-3}$$

for arbitrary n. Also, by taking $\alpha=0$ in (3-2), we see that

$$A(0)=0. \tag{3-4}$$

An important subset of the domain of A is the ***null space*** of A, $\mathscr{N}(A)$, where

$$\mathscr{N}(A)=\{x\in\mathscr{D}(A)\colon Ax=0\}.$$

It is readily verified that $\mathscr{N}(A)$ is a subspace of X. Likewise, $\mathscr{R}(A)$ is a subspace of Y.

We recall from the Introduction that the inverse function A^{-1} exists if and only if A is injective.

Theorem 3.1. *Let A be a linear operator with domain $\mathscr{D}(A)$ in X and range $\mathscr{R}(A)$ in Y, where X and Y are linear spaces. Then A^{-1} exists if and only if $\mathscr{N}(A)=(0)$. When A^{-1} exists, it is a linear operator.*

Proof. Suppose that A^{-1} exists. If $x\in\mathscr{D}(A)$ satisfies $Ax=0$, then $A(x)=A(0)$ [see (3-4)], and hence $x=0$ since A is injective. Thus $\mathscr{N}(A)=(0)$. Conversely, suppose that $\mathscr{N}(A)=(0)$ and suppose $Ax_1=Ax_2$. Then, by the linearity of A, we have $A(x_1-x_2)=0$. Hence $x_1-x_2\in\mathscr{N}(A)$, and so $x_1-x_2=0$ and $x_1=x_2$. This shows that A is injective. Hence A^{-1} exists. The proof that A^{-1} is linear, when it exists, is left to the reader. □

Theorem 3.2. *Let A be an injective linear operator with $\mathscr{D}(A)=X$ and $\mathscr{R}(A)\subset Y$. Then $\mathscr{R}(A)$ is of the same (possibly infinite) dimension as X.*

Proof. Suppose $x_1,\ldots,x_k$ are linearly independent vectors in X. Then $Ax_1,\ldots,Ax_k$ must be linearly independent in Y. For otherwise there would exist scalars $\alpha_1,\ldots,\alpha_k$, not all zero, such that

$$\begin{aligned}0&=\alpha_1Ax_1+\cdots+\alpha_kAx_k\\&=A(\alpha_1x_1+\cdots+\alpha_kx_k).\end{aligned}$$

Since A is injective, this would imply that $\alpha_1 x_1 + \cdots + \alpha_k x_k = 0$, contradicting the assumption that $x_1, \ldots, x_k$ are linearly independent. The fact that $Ax_1, \ldots, Ax_k$ are linearly independent whenever $x_1, \ldots, x_k$ are linearly independent implies that the dimension of $\mathscr{R}(A)$ is at least as great as the dimension of $\mathscr{D}(A) = X$. Furthermore, since A is injective, we can apply the same argument to the (injective) operator A^{-1} to conclude that the dimension of $\mathscr{R}(A^{-1}) = X$ is at least as great as the dimension of $\mathscr{D}(A^{-1}) = \mathscr{R}(A)$. This proves that $\dim X = \dim \mathscr{R}(A)$. $\square$

Definition. An injective linear operator whose domain is X and whose range is all of Y is said to be an *isomorphism of X onto Y*. Two linear spaces X and Y (with the same scalar field) are said to be *isomorphic* if there exists an isomorphism of X onto Y.

In less technical language, an isomorphism from X onto Y provides a one-to-one correspondence between the elements of X and the elements of Y such that the operations of vector addition and scalar multiplication are preserved under the correspondence. That is, if x_1 and x_2 have correspondents y_1 and y_2, respectively, then $y_1 + y_2$ corresponds to $x_1 + x_2$ and αy_1 corresponds to αx_1.

If X is isomorphic to Y and Y is isomorphic to Z, then X is isomorphic to Z. For, if A is an isomorphism of X onto Y and if B is an isomorphism of Y onto Z, then it is readily verified that the mapping BA, defined by $(BA)x = B(Ax)$, is an isomorphism of X onto Z.

The next theorem is an immediate consequence of Theorem 3.2 and the definition of isomorphic linear spaces.

Theorem 3.3. *If X and Y are isomorphic linear spaces, then X and Y have the same (possibly infinite) dimension.*

The linear problems of algebra and analysis are concerned with linear operators on various linear spaces. We mention two kinds of problems in very general terms: *existence* problems and *uniqueness* problems. Suppose A is a given linear operator, with domain a given space X and range in a given space Y. Then we can ask: "For which elements $y \in Y$ does there exist in X an element x such that $Ax = y$?" This is the same as asking: "Which elements of Y belong to $\mathscr{R}(A)$?" Existence problems are of this kind.

Example 1. Let $Y = C[0, 1]$ (real-valued functions). Let X be the subclass of Y consisting of those elements x that have first and second derivatives continuous on $[0, 1]$ and that are, moreover, such that $x(0) = x'(0) = 0$. Let p and q be members of Y, and define A on X by $Ax = y$ where

$$y(s) = x''(s) + p(s)x'(s) + q(s)x(s).$$

Then A is linear on X into Y. Is $\mathcal{R}(A)$ all of Y? This is the question as to whether the differential equation

$$x'' + px' + qx = y$$

has a solution such that $x \in X$, for each choice of $y \in Y$. Note that $x \in X$ implies that $x(0) = x'(0) = 0$. The "initial conditions" have been incorporated into the definition of the domain of A. There is a standard existence theorem in the theory of differential equations that does in fact assure us that $\mathcal{R}(A) = Y$. It also assures us that for a given y there is *only* one x in X such that $Ax = y$. Therefore A^{-1} exists.

Example 2. Let X be the space $BV[0, 1]$ defined in Example 8, § 2. Let Y be the space of all bounded sequences $y = \{\eta_i\}$ $(i = 1, 2, \ldots)$ where the definitions of the algebraic processes in Y are made as for the space ℓ^2 of Example 6, § 2. Define a linear operator A on X into Y by $Ax = y$, where

$$\eta_k = \int_0^1 s^k \, dx(s).$$

The integral here is a Stieltjes integral. The η_k's are called *moments*. A sequence $\{\eta_k\}$ that arises in this way is called a *moment sequence* arising from x. A moment sequence is certainly bounded, for $|\eta_k|$ cannot exceed the total variation of x. But which bounded sequences are moment sequences? That is, how can we recognize those bounded sequences that are in the range of A? This is a classical problem known as the *moment problem of Hausdorff*. In this case, $\mathcal{R}(A)$ is not all of Y. For a description of $\mathcal{R}(A)$ (i.e., of all moment sequences) we refer the reader to Chapter 3 of D. V. Widder, [1].

In addition to existence problems, there are *uniqueness* problems. In the case of any linear operator A we can ask: is the x such that $Ax = y$ unique in all cases in which a solution for x exists? This is the same as the question: does A^{-1} exist? By Theorem 3.1 this is equivalent to the question: is $\mathcal{N}(A) = (0)$, that is, does the equation $Ax = 0$ have a unique solution (namely, $x = 0$)? To answer this question for a particular operator, we must have a good deal of detailed knowledge about the operator.

Example 3. Let $Y = C[a, b]$, and let X be the subspace of $C[a, b]$ consisting of those functions x that have continuous first and second derivatives on $[a, b]$ and are such that $x(a) = x(b) = 0$. Define A on X into Y by $Ax = x''$, where x'' is the second derivative of x. It is easy to show that A^{-1} exists. For, $Ax = 0$ implies that $x(s) = c_1 s + c_2$, where c_1 and c_2 are constants, and the requirements $x(a) = x(b) = 0$ then lead to the conclusions $c_1 = c_2 = 0$.

It is also easy to show that $\mathcal{R}(A) = Y$. For a given $y \in C[a, b]$ the unique x in X such that $Ax = y$ is given by

$$x(s) = \int_a^s du \int_a^u y(t)\, dt - \frac{s-a}{b-a} \int_a^b du \int_a^u y(t)\, dt.$$

With a little manipulation this formula may be put in the form

$$x(s) = \int_a^b K(s, t) y(t)\, dt,$$

where

$$K(s, t) = \begin{cases} \dfrac{(s-b)(t-a)}{b-a} & \text{if } a \leqslant t \leqslant s, \\ \dfrac{(s-a)(t-b)}{b-a} & \text{if } s \leqslant t \leqslant b. \end{cases}$$

Note that $K(s, t) = K(t, s)$ and that K is continuous on the square where it is defined.

Sometimes there is an important connection between existence problems and uniqueness problems. This connection is simplest in the case of finite-dimensional spaces.

Theorem 3.4 *Let A be a linear operator on X into Y, where X and Y are both of the same finite dimension n. Then $\mathcal{R}(A) = Y$ if and only if A^{-1} exists.*

Proof. If A^{-1} exists, we know that $\dim \mathcal{R}(A) = \dim X = n$, by Theorem 3.2. It follows from Theorem 1.5 that $\mathcal{R}(A)$ cannot be a proper subset of Y and must, therefore, coincide with Y. Conversely, suppose that $\mathcal{R}(A) = Y$, and let $y_1, \ldots, y_n$ be a basis of Y. There exist $x_1, \ldots, x_n$ in X such that $Ax_k = y_k$, $k = 1, \ldots, n$, since $\mathcal{R}(A) = Y$. The set of x's is linearly independent. For, if $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$, it follows that $0 = A(\alpha_1 x_1 + \cdots + \alpha_n x_n) = \alpha_1 y_1 + \cdots + \alpha_n y_n$, whence $\alpha_1 = \cdots = \alpha_n = 0$, since the set of y's is linearly independent. It follows that the set $x_1, \ldots, x_n$ is a basis of X because X is n-dimensional. (See the remark preceding Lemma 1.2.) Now suppose $x \in \mathcal{N}(A)$. We can express x in the form $x = \alpha_1 x_1 + \cdots + \alpha_n x_n$, and so $0 = Ax = \alpha_1 y_1 + \cdots + \alpha_n y_n$. This implies that $\alpha_1 = \cdots = \alpha_n = 0$, and so $x = 0$. Thus $\mathcal{N}(A) = (0)$, and hence A^{-1} exists, by Theorem 3.1. □

Finite dimensionality is essential in Theorem 3.4. We can see that this is so by the following example. Let X be the linear space of Example 9, § 2, and let $Y = X$. Define A by setting $Ax = y$, where $y(s) = x'(s)$. In this case, $\mathcal{R}(A) = X$, for a solution of $Ax = y$ is given by $x(s) = \int_0^s y(t)\, dt$. The inverse A^{-1} does not exist, however, for $Ax = 0$ in the case of every constant function.

PROBLEMS

1. Let A be a linear operator from a linear space X into a linear space Y. Given a subspace M of X, show that AM is a subspace of Y.
2. Let A be a linear operator from X into Y, and let M be a subspace of X such that $M \cap \mathcal{N}(A) = (0)$.
   a. Show that the restriction A_M of A to the subspace M is an injective operator.
   b. Show that the spaces M and AM are isomorphic.
3. Fill in the details of the following alternate argument for the second part of the proof of Theorem 3.4: Assume that $\mathcal{R}(A) = Y$ but that A^{-1} does not exist so that there is an $x_1 \neq 0$ with $Ax_1 = 0$. Choose $x_2, \ldots, x_n$ so that $x_1, \ldots, x_n$ is a basis of X. Let $y_k = Ax_k$, $k = 2, \ldots, n$. Then Y is generated by $y_2, \ldots, y_n$.

I.4 LINEAR OPERATORS IN FINITE-DIMENSIONAL SPACES

Let X denote either the real space $\boldsymbol{R}^n$ or the complex space $\boldsymbol{C}^n$ (Examples 1 and 2, § 2). Let $Y = \boldsymbol{R}^m$ if $X = \boldsymbol{R}^n$, and $Y = \boldsymbol{C}^m$ if $X = \boldsymbol{C}^n$. Here m and n may be any positive integers. Let

$$(4\text{-}1) \qquad (\alpha_{ij}) \equiv \begin{Vmatrix} \alpha_{11} & \alpha_{12} & & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & & \alpha_{2n} \\ \cdot & \cdot & & \cdot \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{Vmatrix}$$

be any $m \times n$ matrix of scalars (real or complex according as X and Y are real or complex). Such a matrix defines a linear operator A on X into Y as follows: $Ax = y$, where $x = (\xi_1, \ldots, \xi_n)$, $y = (\eta_1, \ldots, \eta_m)$, and

$$(4\text{-}2) \qquad \sum_{j=1}^{n} \alpha_{ij}\xi_j = \eta_i \qquad i = 1, \ldots, m.$$

The linear problems connected with the operator A are, in this case, problems connected with the system of m linear equations (4-2). The special properties of the operator A must be found by an examination of the particular matrix (4-1).

We recall the notation

$$e_1 = (1, 0, \ldots, 0), \ldots, e_n = (0, \ldots, 0, 1)$$

introduced in Example 1, § 2. The vectors $e_1, \ldots, e_n$ form a basis of X. Let us write

$$f_1 = (1, 0, \ldots, 0), \ldots, f_m = (0, \ldots, 0, 1)$$

for the analogous basis in Y. Observe that $x = \xi_1 e_1 + \cdots + \xi_n e_n$ and $y =$

$\eta_1 f_1 + \cdots + \eta_m f_m$. Now let us define vectors $a_1, \ldots, a_n$ in Y by $a_k = Ae_k$. Since $\xi_j = \delta_{kj}$ (the Kronecker delta) if $x = e_k$, we see by (4-2) that

$$a_k = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} \delta_{kj} \right) f_i = \sum_{i=1}^{m} \alpha_{ik} f_i = (\alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{mk}).$$

In other words, the vector a_k appears as the kth column in the matrix (4-1).

In the foregoing situation we started with an arbitrary $m \times n$ matrix and used it to define a linear operator on X into Y. Let us now proceed the other way around. We shall start with an arbitrary linear operator A on X into Y and show that there is an $m \times n$ matrix of the form (4-1) such that it defines the operator A, in the sense that $Ax = y$ means precisely that the η's are defined in terms of the ξ's by equations (4-2).

If A is the given operator, we define $a_j = Ae_j$, $j = 1, \ldots, n$. The a's are vectors in Y and can, therefore, be expressed in terms of the basis $f_1, \ldots, f_m$. Let the expression of a_j be $a_j = \alpha_{1j} f_1 + \cdots + \alpha_{mj} f_m$. In this way we arrive at the set of scalars α_{ij}, with which we form the matrix (4-1). Now, if $x = (\xi_1, \ldots, \xi_n)$ and $y = Ax = (\eta_1, \ldots, \eta_m)$, we have to show that the equations (4-2) are valid. We have $x = \xi_1 e_1 + \cdots + \xi_n e_n$, $Ax = \xi_1 a_1 + \cdots + \xi_n a_n$, or

$$y = \sum_{j=1}^{n} \xi_j \left(\sum_{i=1}^{m} \alpha_{ij} f_i \right) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \alpha_{ij} \xi_j \right) f_i.$$

Since the expression of a vector in terms of the basis is unique, this last equation is equivalent to the equations (4-2).

The foregoing considerations show that the study of linear operators on $\boldsymbol{R}^n$ into $\boldsymbol{R}^m$ is closely related to the study of matrices of the form (4-1) with real elements α_{ij}. If we deal with $\boldsymbol{C}^n$ and $\boldsymbol{C}^m$, the only difference is that the matrix elements are from the complex number field.

If we now turn to a study of linear operators on X into Y, where X and Y are *arbitrary* finite-dimensional spaces, we shall find the same close connection between linear operators and matrices. The reason for this is that every n-dimensional real linear space is isomorphic to $\boldsymbol{R}^n$ and every n-dimensional complex linear space is isomorphic to $\boldsymbol{C}^n$, as we see from the following theorem.

Theorem 4.1. *If X and Y are both n-dimensional linear spaces with the same scalar field, they are isomorphic.*

Proof. Assume the scalar field is that of the real numbers, for definiteness. We shall show that X (and hence also Y) is isomorphic to $\boldsymbol{R}^n$, from which it will follow that X and Y are isomorphic. Let $x_1, \ldots, x_n$ be a basis of X. Every x in X has a unique representation $x = \xi_1 x_1 + \cdots + \xi_n x_n$, where $\xi_1, \ldots, \xi_n$ are real numbers. We define a linear operator A on X into $\boldsymbol{R}^n$ by writing $Ax = (\xi_1, \ldots, \xi_n)$. The facts that this is a linear operator, that it has all of $\boldsymbol{R}^n$ as its range, and that A^{-1} exists are all easily verified, and we omit the details. □

If X is n-dimensional, with a basis $x_1, \ldots, x_n$, the coefficients $\xi_1, \ldots, \xi_n$ in the representation $x = \xi_1 x_1 + \cdots + \xi_n x_n$ may be called the *coordinates* (or, also, the *components*) of x with respect to the basis $x_1, \ldots, x_n$. Thus the isomorphism of X and $\boldsymbol{R}^n$ is established by correlating x with the point in $\boldsymbol{R}^n$ whose coordinates with respect to the basis $e_1, \ldots, e_n$ of $\boldsymbol{R}^n$ are the same as the coordinates of x with respect to $x_1, \ldots, x_n$.

Suppose now that X and Y are any linear spaces of dimensions n and m, respectively, with the same scalar field. Let $x_1, \ldots, x_n$ be a basis in X, and $y_1, \ldots, y_m$ a basis in Y. Let A be a linear operator on X into Y. Since $Ax_j \in Y$, we can write

$$Ax_j = \sum_{i=1}^{m} \alpha_{ij} y_i, \qquad j = 1, \ldots, n. \tag{4-3}$$

The operator A, in conjunction with the two bases $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_m\}$ determines an $m \times n$ matrix (α_{ij}). With this matrix we can calculate the vector Ax for every $x \in X$. The calculations are essentially the same as those given earlier in the special case $X = \boldsymbol{R}^n$, $Y = \boldsymbol{R}^m$; the coordinates η_i of Ax are given in terms of the coordinates ξ_j of x by equations (4-2).

It must be kept in mind that, although the operator A is represented by the matrix (α_{ij}), *it is not the same thing as the matrix.* For, the operator A is not dependent on the particular bases that are chosen for X and Y, whereas the matrix *does* depend on these bases, and the same operator is represented by different matrices when different bases are chosen.

To conclude this section we give an example.

Example. Let X be the complex linear space consisting of all polynomials in the real variable s of degree not exceeding $n-1$, with complex coefficients. Let Y be the corresponding space of polynomials with degree not exceeding n. It is readily evident that X is of dimension n, one possible basis $x_1, \ldots, x_n$ being that defined by

$$x_1(s) = 1, x_2(s) = s, \ldots, x_n(s) = s^{n-1}.$$

Likewise, Y is of dimension $n+1$. X is a subspace of Y. For a basis in Y we take $y_1 = x_1, \ldots, y_n = x_n, y_{n+1}(s) = s^n$.

Now consider the operator A on X into Y defined by $Ax = y$, where

$$y(s) = \int_0^s x(t)\, dt.$$

We see that

$$Ax_j = \frac{1}{j} y_{j+1}, \qquad j = 1, \ldots, n.$$

The matrix (α_{ij}) in this case is therefore

$$\begin{Vmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{1}{3} & 0 \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ & & & & 0 \\ 0 & 0 & \cdots & \cdots & \frac{1}{n} \end{Vmatrix}$$

with $n+1$ rows and n columns.

We observe, incidentally, that the inverse A^{-1} exists. The range $\mathscr{R}(A)$ consists of all polynomials of degree $\leqslant n$ whose constant term is zero. That is, $\mathscr{R}(A)$ is the linear manifold spanned by $y_2, \ldots, y_{n+1}$. With $\mathscr{R}(A)$ as its domain, A^{-1} is defined by $A^{-1}y = x$, where $x(s) = y'(s)$.

PROBLEM

1. Let X be the complex linear space consisting of all polynomials in a real variable s of degree not exceeding four, with complex coefficients. Given $x \in X$, let Ax be the derivative of x; that is, $(Ax)(s) = x'(s)$. Determine the matrix that represents A (as an operator from X into X) with respect to the basis $x_1, \ldots, x_5$, where $x_1(s) = 1$ and $x_k(s) = s^{k-1}$, $k = 2, \ldots, 5$.

I.5 OTHER EXAMPLES OF LINEAR OPERATORS

In this section we give a number of illustrations of problems in analysis, as formulated in terms of linear operators.

Example 1. Let G be a nonempty, bounded, connected, open set in the Euclidean xy-plane. Let $\bar{G}$ be the closure of G; then the boundary $\bar{G}\backslash G$ is nonempty. Let H be the class of all real-valued functions f of x, y, which are defined and continuous on $\bar{G}$ and harmonic in G (i.e., each f must satisfy Laplace's equation in G). Let C denote the class of real-valued functions defined and continuous on $\bar{G}\backslash G$. Clearly H and C are both linear spaces.

Now, if $u \in H$, $u(x, y)$ is defined at every point $(x, y) \in \bar{G}$. If we think only of the values of u at the points of $\bar{G}\backslash G$, we arrive at an element f of C defined by $f(x, y) = u(x, y)$ when $(x, y) \in \bar{G}\backslash G$. We call f the restriction of u to the boundary of G. We define a linear operator A on H into C by writing $Au = f$; the fact that A is linear is evident.

Let us consider existence and uniqueness problems for A (see § 3 for a general discussion of these problems). In the present example these are problems of potential theory. The uniqueness problem is completely disposed of by the well-known fact that if $u \in H$ and $u(x, y)=0$ on $\bar{G}\backslash G$, then $u(x, y)=0$ at all points of G (since u must attain both its maximum and minimum values on the boundary of G). In our present notation this means that $u=0$ if $Au=0$; consequently A^{-1} exists. For the existence problem we want to know if $\mathscr{R}(A)$ is all of C. This is the famous Dirichlet problem: Is there for each $f \in C$ an element $u \in H$ such that f is the restriction of u to the boundary of G? The answer to this question is not always affirmative; $\mathscr{R}(A)$ may be a proper subspace of C. However, under certain conditions involving the nature of the set G and its boundary, it is known that $\mathscr{R}(A)=C$. In particular, under suitable conditions, $\mathscr{R}(A)=C$ and A^{-1} is defined by an integral, involving the Green's function of G, extended over the boundary of G.

Many boundary-value problems of both ordinary and partial differential equations are susceptible of formulation in terms of linear operators, by procedures analogous to those of Example 1. It is essential that the differential equations be linear and that either the differential equations or the boundary conditions be homogeneous.

Example 2. Let $X=Y=C[a, b]$ (see Example 3, § 2). Let $k(s, t)$ be defined for $a \leqslant s \leqslant b$, $a \leqslant t \leqslant b$ and such that for each $x \in X$ the Riemann integral

$$\int_a^b k(s, t)x(t)\,dt \tag{5-1}$$

exists and defines a continuous function of s on $[a, b]$. The values of k are to be either real or complex, depending on whether the values of the elements of $C[a, b]$ are taken as real or complex.

The integral (5-1) defines a linear operator K on X into X, if we take $Kx=y$ to mean

$$y(s)=\int_a^b k(s, t)x(t)\,dt. \tag{5-2}$$

The equation (5-2) is called an *integral equation*. It is of the particular sort known as an equation *of Fredholm type, of the first kind.*

Another operator T is obtained by defining $Tx=y$ to mean

$$y(s)=x(s)-\int_a^b k(s, t)x(t)\,dt. \tag{5-3}$$

In this case the integral equation is said to be *of Fredholm type, of the second kind.* Equations of this sort are of great importance. There is a well worked-out theory of the existence and uniqueness problems associated with such equations. The applications of this theory play a vital role in the theory of boundary-value problems in differential equations.

A special situation results if we assume that $k(s, t) = 0$ when $t > s$. The integral (5-1) then becomes

$$\int_a^s k(s, t)x(t)\, dt.$$

The equations (5-2) and (5-3), in this modified form, are said to be *of Volterra type*. The theory of Volterra equations of the second kind is particularly simple, much more so than the corresponding theory for equations of the Fredholm type. Later in this book we shall treat the theory of integral equations of the second kind, using abstract linear space methods. For this treatment it is necessary to introduce topological as well as algebraic considerations about linear spaces.

Example 3. Let $X = Y = C[0, \pi]$. In this case we assume specifically that we are dealing with complex-valued continuous functions. If $x \in X$ and if $x'(s)$ exists on $[0, \pi]$, we denote the derived function by x'. Likewise, x'' denotes the second derivative if $x''(s)$ exists on $[0, \pi]$. We now define a linear operator A as follows: Let $\mathscr{D}(A) = \{x : x', x'' \in C[0, \pi] \text{ and } x(0) = x(\pi) = 0\}$. For $x \in \mathscr{D}(A)$ define $Ax = y$ to mean

$$-x''(s) + \lambda x(s) = y(s), \tag{5-4}$$

where λ is a complex parameter. The operator A thus depends on λ. Clearly $\mathscr{R}(A) \subset X$.

We shall discuss the existence and uniqueness problem for A in some detail. This amounts to a discussion of finding solutions of the differential equation (5-4) that satisfy the end conditions $x(0) = x(\pi) = 0$.

Let one of the square roots of λ be denoted by μ. By the method of variation of parameters, the general solution of (5-4) (disregarding the end conditions) is found to be

$$x(s) = \frac{1}{\mu}\int_0^s y(t) \sinh \mu(t - s)\, dt + C_1 e^{\mu s} + C_2 e^{-\mu s}, \tag{5-5}$$

where C_1 and C_2 are arbitrary constants. This is in case $\lambda \neq 0$. If $\lambda = 0$, the solution may be found by direct integration. If we impose the end conditions $x(0) = x(\pi) = 0$, we find, first, that $C_1 + C_2 = 0$ and then that

$$0 = \frac{1}{\mu}\int_0^\pi y(t) \sinh \mu(t - \pi)\, dt + 2C_1 \sinh \pi\mu.$$

To solve this equation for C_1 we must assume that $\sinh \pi\mu \neq 0$. On substituting into (5-5) the values found for C_1, C_2, we have

$$x(s) = \frac{1}{\mu}\int_0^s y(t) \sinh \mu(t-s)\, dt - \frac{\sinh \mu s}{\mu \sinh \mu\pi}\int_0^\pi y(t) \sinh \mu(t-\pi)\, dt. \tag{5-6}$$

The corresponding formula when $\lambda = 0$ is

$$x(s) = \int_0^s y(t)(t-s)\, dt - \frac{s}{\pi}\int_0^\pi y(t)(t-\pi)\, dt. \tag{5-7}$$

It may now be seen that the equation $Ax = y$ admits a solution $x \in \mathcal{D}(A)$ for every $y \in X$, provided either that $\lambda = 0$ or that $\sinh \pi\mu \neq 0$ if $\lambda \neq 0$. Moreover, the solution is unique, for it may be argued from the standard theory of linear differential equations that (5-6) and (5-7) give, in their respective cases, the *only* solutions of (5-4) satisfying the given end conditions.

There remain to be considered the cases when $\lambda \neq 0$ but $\sinh \pi\mu = 0$. Now, since $\mu^2 = \lambda$ and since $\sinh \pi\mu = 0$ if and only if $\pi\mu = in\pi$, where n is an integer, we see that the exceptional cases are given by $\lambda = -n^2$, $n = 1, 2, 3, \ldots$. In these cases (5-4) cannot have a unique solution in $\mathcal{D}(A)$ (i.e., A^{-1} does not exist), for $Ax = 0$ in case $\lambda = -n^2$ and $x(s) = \sin ns$. If we reexamine our earlier work, assuming that $\lambda = -n^2 \neq 0$, we see that the equation $Ax = y$ has all its solutions $x \in \mathcal{D}(A)$ given by

$$x(s) = \frac{1}{n}\int_0^s y(t) \sin n(t-s)\, dt + C \sin ns, \tag{5-8}$$

where C is an arbitrary constant, if and only if the function y satisfies the condition

$$\int_0^\pi y(t) \sin nt\, dt = 0. \tag{5-9}$$

This condition serves as a description of $\mathcal{R}(A)$ in this case.

By writing $\int_0^\pi = \int_0^s + \int_s^\pi$ in (5-6) and by rearranging somewhat, we can write the formula as

$$x(s) = \int_0^s y(t) \frac{\sinh \mu t \sinh \mu(\pi - s)}{\mu \sinh \mu\pi}\, dt + \int_s^\pi y(t) \frac{\sinh \mu s \sinh \mu(\pi - t)}{\mu \sinh \mu\pi}\, dt.$$

If we define

$$k(s, t) = \frac{\sinh \mu t \sinh \mu(\pi - s)}{\mu \sinh \mu\pi} \qquad \text{when} \qquad 0 \leq t \leq s$$

and $k(s, t) = k(t, s)$ when $s \leq t \leq \pi$, (5-6) becomes

$$x(s) = \int_0^\pi k(s, t)y(t)\,dt.$$

We note that this formula is of the same type as (5-2). In our operator notation this formula is $x = A^{-1}y$. If we regard $x = A^{-1}y$ as an equation to be solved for y, we see that it is an integral equation of Fredholm type, of the first kind. It has a unique solution, for each $x \in \mathscr{D}(A)$, given by (5-4).

Example 4. Let $X = Y = L(0, \infty)$ (see Example 7, § 2). Let us define an operator A as follows: Consider those functions in $\mathscr{L}(0, \infty)$ that are absolutely continuous on $[0, a]$ for every finite a, with derivatives also belonging to $\mathscr{L}(0, \infty)$. Each such function determines an equivalence class in $\mathscr{L}(0, \infty)$, that is, an element of $L(0, \infty)$. The set of such elements of $L(0, \infty)$ is going to be $\mathscr{D}(A)$. For convenience we shall drop the notational distinction between an element of $L(0, \infty)$ and one of its representatives in $\mathscr{L}(0, \infty)$. Thus, if $x \in L(0, \infty)$, we shall write $x(s)$ to indicate a functional value at s of a representative of x. We then define $Ax = y$ to mean that

$$-x'(s) + \lambda x(s) = y(s) \tag{5-10}$$

almost everywhere on $(0, \infty)$. This defines $y \in L(0, \infty)$ uniquely if $x \in \mathscr{D}(A)$. As in Example 3, λ denotes a complex parameter.

To solve the equation $Ax = y$, we proceed formally to solve (5-10) by use of the integrating factor $e^{-\lambda s}$. If $x(s)$ is absolutely continuous and satisfies (5-10), it is necessarily of the form

$$x(s) = e^{\lambda s}\left\{C - \int_0^s y(t)e^{-\lambda t}\,dt\right\}, \tag{5-11}$$

where C is some constant and, vice versa, the function $x(s)$ so defined, for any C, is absolutely continuous and does satisfy (5-10). If $x(s)$ turns out to belong to $\mathscr{L}(0, \infty)$, it is clear from (5-10) that $x'(s)$ will also belong, because we assume that $y(s)$ belongs to $\mathscr{L}(0, \infty)$. The problem is then that of seeing whether the constant C can be chosen so that the integral $\int_0^\infty |x(s)|\,ds$ will be convergent. When we consider this problem, it is convenient to consider three cases, depending on whether the real part of λ is positive, negative, or zero. We write $\lambda = \alpha + i\beta$, α and β real. Note that

$$|x(s)| = e^{\alpha s}\left|C - \int_0^s y(t)e^{-\lambda t}\,dt\right|.$$

Case 1. $\alpha > 0$. Since

$$\int_0^s y(t)e^{-\lambda t}\,dt \to \int_0^\infty y(t)e^{-\lambda t}\,dt \qquad \text{as} \quad s \to \infty,$$

we see that $|x(s)|$ will be of the order of magnitude of $e^{\alpha s}$ when s is large unless

$$C = \int_0^\infty y(t)e^{-\lambda t}\,dt. \tag{5-12}$$

This is, therefore, a necessary condition for $x(s)$ to belong to $\mathscr{L}(0, \infty)$, when $\alpha > 0$. It is also sufficient; for, if it is satisfied, we have

$$x(s) = e^{\lambda s}\int_s^\infty y(t)e^{-\lambda t}\,dt = \int_0^\infty y(s+u)e^{-\lambda u}\,du, \tag{5-13}$$

from which it is easy to see that $x(s) \in \mathscr{L}(0, \infty)$ if $\alpha > 0$.

Case 2. $\alpha = 0$. The condition (5-12) is likewise necessary in this case. It is not sufficient, however, and whether or not $x(s) \in \mathscr{L}(0, \infty)$ will depend on the choice of y. For instance, if

$$y(t) = e^{i\beta t}t^{-2}$$

when t is large, then for large s

$$x(s) = e^{\lambda s}\int_s^\infty y(t)e^{-\lambda t}\,dt = e^{i\beta s}\int_s^\infty t^{-2}\,dt = \frac{e^{i\beta s}}{s},$$

and $x(s)$ is not in $\mathscr{L}(0, \infty)$.

Case 3. $\alpha < 0$. In this case the function $x(s)$ defined by (5-11) is in $\mathscr{L}(0, \infty)$, no matter how C and y are chosen. This is easily seen, for $e^{\lambda s}$ is in $\mathscr{L}(0, \infty)$ and

$$\int_0^s y(t)e^{\lambda(s-t)}\,dt$$

is also in $\mathscr{L}(0, \infty)$. In fact,

$$\begin{aligned}\int_0^\infty \left| \int_0^s y(t)e^{\lambda(s-t)}\,dt \right| ds &\leq \int_0^\infty ds \int_0^s |y(t)|e^{\alpha(s-t)}\,dt \\ &= \int_0^\infty e^{-\alpha t}|y(t)|\,dt \int_t^\infty e^{\alpha s}\,ds \\ &= -\frac{1}{\alpha}\int_0^\infty |y(t)|\,dt.\end{aligned}$$

We can sum up the discussion as follows: The inverse A^{-1} exists if $\alpha \geq 0$, and the solution $x = A^{-1}y$ is expressed by (5-13). If $\alpha > 0$, $\mathscr{R}(A)$ is all of X, but not so if $\alpha = 0$. When $\alpha < 0$, A^{-1} does not exist, but $\mathscr{R}(A)$ is all of X.

Example 5. Let $X = L(0, 2\pi)$. Let F be the class of functions f analytic in the unit disc $|z|<1$. For $x \in X$, with representative $x(s) \in \mathscr{L}(0, 2\pi)$, define $Ax = f$ to mean

$$f(z) = \frac{1}{2\pi}\int_0^{2\pi} \frac{x(s)}{1 - ze^{-is}}\, ds. \tag{5-14}$$

Then A is a linear operator on X into F. The inverse A^{-1} does not exist, for it is a simple matter to calculate that $Ax = 0$ if $x(s) = e^{-ims}$, $m = 1, 2, \ldots$ (use the geometric series for $(1 - ze^{-is})^{-1}$, and integrate term by term). The problem of characterizing the range $\mathscr{R}(A)$ has not yet been solved satisfactorily. One would like to know conditions on the analytic function f, necessary and sufficient in order that f can be represented by formula (5-14), with some $x \in L(0, 2\pi)$. A simple ***sufficient***, but not necessary, condition is that f be continuous when $|z| \leq 1$ as well as analytic when $|z|<1$. For then, by Cauchy's formula,

$$f(z) = \frac{1}{2\pi i}\int \frac{f(w)}{w - z}\, dw,$$

the complex line integral being extended counterclockwise around the unit circle $|w| = 1$. If we write $w = e^{is}$, $0 \leq s \leq 2\pi$, this becomes

$$f(z) = \frac{1}{2\pi}\int_0^{2\pi} \frac{f(e^{is})}{1 - ze^{-is}}\, ds.$$

Thus in this case $f = Ax$, where $x(s) = f(e^{is})$.

A less restrictive sufficient condition is known. Suppose $f \in F$, and suppose that there is a constant M such that

$$\int_0^{2\pi} |f(re^{i\theta})|\, d\theta \leq M \tag{5-15}$$

when $0 \leq r < 1$. Then it may be shown that $\lim_{r\to 1} f(re^{is}) = x(s)$ exists almost everywhere on $0 \leq s \leq 2\pi$, that $x(s) \in \mathscr{L}(0, 2\pi)$, and that $Ax = f$. See § 2.1 and § 3.3 of Duren [1].

Example 6. Let X be the class of all infinite arrays of the form $x = (\ldots, \xi_{-2}, \xi_{-1}, \xi_0, \xi_1, \xi_2, \ldots)$, where the ξ_n's are arbitrary complex numbers. We denote these more briefly as $x = \{\xi_n\}$, $n = 0, \pm1, \pm2, \ldots$. Define addition and scalar multiplication in the obvious way, that is,

$$\{\xi_n\} + \{\eta_n\} = \{\xi_n + \eta_n\}, \qquad \alpha\{\xi_n\} = \{\alpha\xi_n\}.$$

Then X is a complex linear space. Now consider the linear operator A on $L(0, 2\pi)$ into X defined as follows: If $f \in L(0, 2\pi)$ with representative $f(t) \in$

$\mathscr{L}(0, 2\pi)$, let $Af = x$, where $x = \{\xi_n\}$ is defined by

$$\xi_n = \frac{1}{2\pi}\int_0^{2\pi} f(t)e^{-int}\,dt, \qquad n = 0, \pm 1, \pm 2, \ldots.$$

The numbers ξ_n are the complex Fourier coefficients of f.

The inverse A^{-1} exists. This is simply the classical theorem that if $f(t) \in \mathscr{L}(0, 2\pi)$ and if all of its Fourier coefficients vanish, then $f(s) = 0$ almost everywhere [so that $f = 0$ as an element of $L(0, 2\pi)$]. The problem of characterizing $\mathscr{R}(A)$ has never been fully solved, though many things are known about $\mathscr{R}(A)$ from the theory of Fourier series. Many interesting and important results have been found by confining attention to subspaces of $L(0, 2\pi)$ and X. For instance, it is known that, if we consider the domain of A to be $L^2(0, 2\pi)$ instead of $L(0, 2\pi)$, the range of A is the subspace of X consisting of all $\{\xi_n\}$ for which the series

$$\sum_{n=-\infty}^{\infty} |\xi_n|^2$$

is convergent.

I.6 DIRECT SUMS AND QUOTIENT SPACES

Let X be a linear space. Given two linear manifolds M and N in X, we let $M + N$ denote the set of all vectors of the form $m + n$, where $m \in M$ and $n \in N$. The set $M + N$ is called the *sum* of M and N; it is the smallest linear manifold in X containing both M and N (see problem 1). When $M \cap N = (0)$, that is, when M and N have only the zero vector in common, we write $M \oplus N$ in place of $M + N$, and we call $M \oplus N$ the *direct sum* of M and N.

Theorem 6.1. *Let M and N be subspaces of a linear space X. Then $X = M \oplus N$ if and only if each x in X may be written in the form $x = m + n$, with $m \in M$, $n \in N$, in one and only one way.*

Proof. Suppose $X = M \oplus N$. Then $X = M + N$, and so each x in X may be written in the form $m + n$. Suppose that $x = m_1 + n_1 = m_2 + n_2$, with $m_1, m_2 \in M$ and $n_1, n_2 \in N$. Then $m_1 - m_2 = n_2 - n_1$, and so the element $m_1 - m_2$ belongs to both M and N. But $M \cap N = (0)$, by hypothesis, and so $m_1 - m_2 = 0$. Hence $m_1 = m_2$ and $n_1 = n_2$.

Conversely, suppose each x in X has a unique representation $x = m + n$, with $m \in M$, $n \in N$. Then surely $X = M + N$. If we take x in $M \cap N$, then we may represent x as the sum of an element in M and an element in N in two ways, namely, $x = x + 0$ and $x = 0 + x$. By hypothesis, these two representations must be the same. Hence $x = 0$. Thus $M \cap N = (0)$, and $X = M \oplus N$. $\square$

Theorem 6.2. *If $X = M \oplus N$, then*

$$\dim X = \dim M + \dim N. \tag{6-1}$$

Proof. If either M or N is infinite dimensional, then so is X, by Theorem 1.5. Since, by convention, $\infty + n = \infty$ and $\infty + \infty = \infty$, equality holds in (6-1). If both M and N are finite dimensional, let $m_1, \ldots, m_j$ be a basis for M and let $n_1, \ldots, n_k$ be a basis for N. Clearly the vectors $m_1, \ldots, m_j$, $n_1, \ldots, n_k$ span X, since $X = M + N$. In fact, since $M \cap N = (0)$, it is easy to see that these vectors are linearly independent and hence form a basis for X. Then $\dim X = j + k = \dim M + \dim N$, by Theorem 1.3. □

When $X = M \oplus N$ as in Theorems 6.1 and 6.2, we say that M and N are *complementary subspaces*, and we call N a *complement* of M (in X). It is a fact that, given any subspace M of X, there always exists at least one complement. This fact follows easily from Theorem 1.4 when X is finite dimensional. For the general case, see Theorem 11.2.

Theorem 6.3. *Let X and Y be linear spaces, and let A be a linear operator from X into Y. If W is a complement of $\mathcal{N}(A)$ in X, then the restriction of A to W is an isomorphism of W onto $\mathcal{R}(A)$.*

Proof. We have $X = \mathcal{N}(A) \oplus W$. The restriction A_W of A to W is easily seen to be a linear operator. If $A_W x = 0$ for some $x \in W$, then $Ax = 0$ and $x \in \mathcal{N}(A)$. This implies that $x = 0$, since $\mathcal{N}(A) \cap W = (0)$. Hence $\mathcal{N}(A_W) = (0)$, which shows that A_W is injective, by Theorem 3.1. Certainly A_W maps W into $\mathcal{R}(A)$. Furthermore, if $y \in \mathcal{R}(A)$, then $y = Ax$ where $x = u + w$ for some $u \in \mathcal{N}(A)$ and $w \in W$. Hence $y = Ax = Au + Aw = A_W w$, which shows that A_W maps W onto $\mathcal{R}(A)$. Thus A_W is bijective when considered as a linear operator from W onto $\mathcal{R}(A)$; that is, A_W is an isomorphism of W onto $\mathcal{R}(A)$. □

Theorem 6.4. *Let A be a linear operator from X into Y. Then*

$$\dim X = \dim \mathcal{N}(A) + \dim \mathcal{R}(A).$$

Proof. Let W be a complement of $\mathcal{N}(A)$ in X, so that $X = \mathcal{N}(A) \oplus W$. Then $\dim X = \dim \mathcal{N}(A) + \dim W$, by Theorem 6.2. Since $\dim W = \dim \mathcal{R}(A)$, by Theorems 6.3 and 3.3, this completes the proof. □

Quotient Spaces

Given a subspace M of a linear space X, let us say that any two elements x_1, x_2 in X are *equivalent modulo M* if their difference $x_1 - x_2$ belongs to M. Since M is a linear manifold, it is readily seen that "equivalence modulo M" has the usual properties of an equivalence relation—namely, reflexivity, symmetry, and transitivity. Hence X is divided into mutually disjoint equivalence classes, two elements being in the same equivalence class if and only if they are equivalent modulo M. Let $[x]$ denote the equivalence class that contains the element x. Then $[x]=[u]$ if and only if $x-u \in M$, and

$$[x]=\{u : x-u \in M\}.$$

The set of all these equivalence classes is denoted by X/M. We shall explain how to define addition in X/M and multiplication by scalars so that X/M becomes a linear space. First, we let $[x]+[y]$ be the element $[x+y]$ of X/M. To show that $[x]+[y]$ is unambiguously defined we have to show that $[u+v]=[x+y]$ if $[u]=[x]$ and $[v]=[y]$. This is at once evident, however, because $(x+y)-(u+v)=(x-u)+(y-v)$ and M is a linear manifold. Next, we define $\alpha[x]=[\alpha x]$, observing that $[\alpha u]=[\alpha x]$ if $[u]=[x]$. It is a routine matter to verify that X/M becomes a linear space as a result of these definitions. The zero element of X/M is $[0]$, which is the subset M. The linear space X/M is called the *quotient space of X modulo M.*

To get an intuitive geometric appreciation of the definition of X/M, consider the case in which X is three-dimensional and the points of X are represented in a three-dimensional rectangular coordinate system, the point $x=(\xi_1, \xi_2, \xi_3)$ having ξ_1, ξ_2, ξ_3 as its coordinates. If M is the linear manifold of points $(\xi_1, 0, 0)$ (i.e., the ξ_1-axis), the elements of X/M are the straight lines parallel to the ξ_1-axis. Each such line is uniquely determined by the point $(0, \xi_2, \xi_3)$ in which it intersects the plane $\xi_1=0$, and it is readily seen that X/M is isomorphic to the linear manifold N of points $(0, \xi_2, \xi_3)$. Note that N is a complement of M. Again, if M is the linear manifold of points $(\xi_1, \xi_2, 0)$, the elements of X/M are the planes parallel to the plane $\xi_3=0$. Each such plane is uniquely determined by the point $(0, 0, \xi_3)$ in which it is pierced by the ξ_3-axis, and X/M is isomorphic to the one-dimensional linear manifold N of points $(0, 0, \xi_3)$. As before, we observe that N is a complement of M. The isomorphism between X/M and a complement of M, which is apparent in these particular cases, persists in general, as we see in the next theorem.

Theorem 6.5. *Let M be a linear manifold in X, and let N be any complement of M, so that $X=M\oplus N$. Then X/M is isomorphic to N.*

Proof. Define a mapping ϕ from X into X/M by $\phi(x)=[x]$. It is easy to verify that ϕ is a linear mapping from X onto X/M whose null space is M. Hence the restriction of ϕ to N is an isomorphism of N onto X/M, by Theorem 6.3. □

The linear mapping ϕ of X onto X/M defined by $\phi(x)=[x]$ is called the *canonical mapping of X onto X/M.*

Theorem 6.6. *Let X and Y be linear spaces, and let A be a linear operator from X into Y. Then $\mathscr{R}(A)$ is isomorphic to $X/\mathscr{N}(A)$.*

Proof. Let W be a complement of $\mathscr{N}(A)$ in X. Then $X/\mathscr{N}(A)$ is isomorphic to W, by Theorem 6.5, and W is isomorphic to $\mathscr{R}(A)$, by Theorem 6.3. Hence $X/\mathscr{N}(A)$ is isomorphic to $\mathscr{R}(A)$. □

PROBLEMS

1. Let M and N be linear manifolds in a linear space X.
   a. The set $M+N$ is a linear manifold containing M and N.*
   b. If W is a linear manifold in X containing M and N, then W contains $M+N$. Thus $M+N$ is the smallest linear manifold containing M and N.
2. Let X and Y be linear spaces, and let A be a linear operator from X into Y. Suppose M and N are subspaces of X, and let AM and AN denote the images under A of M and N, respectively. If $M \cap N = (0)$ and $[M \oplus N] \cap \mathscr{N}(A) = (0)$, then
$$A(M \oplus N) = AM \oplus AN.$$
3. Deduce Theorem 3.4 from Theorem 6.4.
4. If M is a finite-dimensional subspace of X, then
$$\dim X/M = \dim X - \dim M.$$
5. Let M and N be linear manifolds in X. Then the quotient spaces $(M+N)/M$ and $N/(M \cap N)$ are isomorphic. [*Hint.* Let ϕ be the canonical mapping of the linear space $M+N$ onto $(M+N)/M$, and let A be the restriction of ϕ to the subspace N. Show that A is a linear mapping from N onto $(M+N)/M$ whose null space is $M \cap N$.]
6. (Alternate proof of Theorem 6.6) Given a linear operator A from X into Y, define $\hat{A}$ from $X/\mathscr{N}(A)$ into Y by $\hat{A}(\phi(x)) = Ax$, where ϕ is the canonical mapping of X onto $X/\mathscr{N}(A)$. Then $\hat{A}$ is a well-defined linear operator. Furthermore, $\hat{A}$ is a bijective mapping of $X/\mathscr{N}(A)$ onto $\mathscr{R}(A)$. Hence $\hat{A}$ is an isomorphism of $X/\mathscr{N}(A)$ onto $\mathscr{R}(A)$.
7. Suppose that M and N are linear manifolds in X such that $M \cap N = (0)$ and $\dim X/M \leqslant \dim N < \infty$. Then $X = M \oplus N$.

I.7 LINEAR FUNCTIONALS

The field of real numbers is a real linear space; it is one-dimensional. Likewise, the field of complex numbers is a one-dimensional complex linear space.

Let X be a real or complex linear space, and let Y be the linear space composed by the scalar field associated with X. A linear operator on X into Y

* For brevity, we shall often phrase problems in the form of assertions that must be proved.

is then called a *linear functional* on X. The set of all linear functionals on X will be denoted by X^f.

Since we shall have many occasions to deal with linear functionals, we shall make certain conventions of notation concerning them. If X is a linear space and x is a generic symbol for elements of X, we shall often use x' as a generic symbol for linear functionals on X. If $x_1 \in X$ and if $x_1' \in X^f$, there is no implication of a special relationship between x_1 and x_1'. Since x' is a function, we can speak of its value at the point x; by our previous usage this value would be denoted by $x'(x)$, or simply $x'x$. It will sometimes be convenient to have another notation. We shall write

$$\langle x, x' \rangle$$

for the value of x' at x. This notation has several advantages, as we shall see, especially when we replace x' by a more complicated symbol representing a linear functional on X.

If x_1' and x_2' are linear functionals on X and α is a scalar, we define $x_1' + x_2'$ and $\alpha x_1'$ by the formulas

$$\langle x, x_1' + x_2' \rangle = \langle x, x_1' \rangle + \langle x, x_2' \rangle, \tag{7-1}$$

$$\langle x, \alpha x_1' \rangle = \alpha \langle x, x_1' \rangle. \tag{7-2}$$

These are the usual definitions for adding functions and multiplying them by scalars. The functional x' such that $x'(x) = 0$ for every x is denoted by 0. Thus we use 0 to denote the zero scalar, the zero vector in X, and the zero functional; the meaning of 0 in a particular occurrence will be clear from the context.

In addition to the formulas (7-1) and (7-2), we note the formulas

$$\langle x_1 + x_2, x' \rangle = \langle x_1, x' \rangle + \langle x_2, x' \rangle, \tag{7-3}$$

$$\langle \alpha x, x' \rangle = \alpha \langle x, x' \rangle. \tag{7-4}$$

These are merely the properties (3-1), (3-2) of linear operators, expressed in the new notation for linear functionals.

Definition. Let X be a linear space. With the above definitions, the set X^f of all linear functionals on X becomes a linear space. We call it the *algebraic conjugate* of X.

We shall see later that the space X^f plays an important role in the study of linear operators with domain X.

Before going further we pause to give some illustrative examples of linear functionals. We refer to the examples of linear spaces in § 2.

Example 1. Let X be either $\boldsymbol{R}^n$ or $\boldsymbol{C}^n$ where $n \geq 1$ (see Examples 1, 2, of § 2). Since a linear functional on X is a linear operator with its range in the one-dimensional space $\boldsymbol{R}^1$ or $\boldsymbol{C}^1$ (depending on whether X is $\boldsymbol{R}^n$ or $\boldsymbol{C}^n$), we can see from § 4 that every linear functional $x' \in X^f$ has a representation

$$\langle x, x' \rangle = \alpha_1\xi_1 + \cdots + \alpha_n\xi_n, \qquad x = (\xi_1, \ldots, \xi_n), \tag{7-5}$$

where $\alpha_1, \ldots, \alpha_n$ are scalars and, conversely, that every n-tuple of scalars $(\alpha_1, \ldots, \alpha_n)$ can be used to define a functional x' by formula (7-5). From this we see that, if we define

$$Ax' = (\alpha_1, \ldots, \alpha_n),$$

then A is a linear operator on X^f into X. We also see that A^{-1} exists and that $\mathscr{R}(A) = X$. Hence, by the definition in § 3, X^f and X are isomorphic in case $X = \boldsymbol{R}^n$ or $X = \boldsymbol{C}^n$.

Example 2. Let $X = C[a, b]$ (see Example 3, § 2). Let s_0 be a fixed point in the interval $[a, b]$. The formula

$$\langle x, x' \rangle = x(s_0) \tag{7-6}$$

defines an element of X^f. Another example of a linear functional is furnished by

$$\langle x, x' \rangle = \int_a^b x(s)\, df(s), \tag{7-7}$$

where f is an arbitrary function of bounded variation defined on $[a, b]$ and the integral is a Stieltjes integral. In case f has a continuous derivative, (7-7) can be written

$$\langle x, x' \rangle = \int_a^b x(s) f'(s)\, ds,$$

the integral being a Riemann integral.

It may be observed that (7-6) is a special case of (7-7). If $s_0 = a$, (7-7) reduces to (7-6) when we define $f(a) = 0$ and $f(s) = 1$ if $a < s \leq b$. If $a < s_0 \leq b$, (7-7) reduces to (7-6) by choosing $f(s) = 0$ when $a \leq s < s_0$ and $f(s) = 1$ when $s_0 \leq s \leq b$.

Example 3. Let $X = \ell^2$ (see Example 6, § 2). Suppose that $a = \{\alpha_n\}$ is an element of ℓ^2, and define $x' \in X^f$ by

$$\langle x, x' \rangle = \sum_{i=1}^{\infty} \alpha_i \xi_i, \qquad \text{where} \qquad x = \{\xi_n\}. \tag{7-8}$$

The series does converge, for, by Cauchy's inequality,

$$\sum_{i=1}^{n} |\alpha_i \xi_i| \leq \left(\sum_{i=1}^{n} |\alpha_i|^2 \right)^{1/2} \left(\sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2}$$

for every value of n. Therefore, since $a \in \ell^2$ and $x \in \ell^2$, the partial sums of the series $\Sigma\, |\alpha_i \xi_i|$ are bounded, and the series in (7-8) is absolutely convergent. It is evident that x', so defined, is a member of X^f.

Example 4. Let $X = L(-\infty, \infty)$ (see Example 7, § 2). If f is a bounded measurable function defined almost everywhere on $(-\infty, \infty)$, we can use it to define a linear functional on X as follows:

$$\langle x, x' \rangle = \int_{-\infty}^{\infty} x(s) f(s)\, ds, \tag{7-9}$$

where $x(s)$ is a function in $\mathscr{L}(-\infty, \infty)$ representing the element $x \in L(-\infty, \infty)$, and the integral is a Lebesgue integral.

The reader will note that it was only in the case of Example 1 that we asserted anything about giving a general representation of *all* elements of X^f. Ordinarily, in dealing with the infinite-dimensional spaces arising in analysis, we find it most useful to limit attention to certain subspaces of X^f. These subspaces are arrived at by imposing a topological structure on the space X and then considering only those linear functionals on X that are continuous with respect to the topology in question. We shall see later that the functionals exhibited in Examples 2, 3, and 4 are typical for the respective spaces when suitable topologies are introduced.

Since X^f is a linear space, we may also consider its algebraic conjugate, which we denote by

$$(X^f)^f, \qquad \text{or simply } X^{ff}.$$

We shall denote elements of X^{ff} generically by x'', and we shall use the notations

$$x''(x') = \langle x', x'' \rangle$$

for the value of x'' at x'.

Corresponding to each x in X there is a unique x'' in X^{ff} defined by

$$\langle x', x'' \rangle = \langle x, x' \rangle \tag{7-10}$$

or, equivalently, $x''(x') = x'(x)$. That this formula does in fact define a linear functional on X^f is apparent from (7-1) and (7-2). The correspondence between x and x'' in (7-10) defines a linear operator J on X into $X^{ff}: Jx = x''$. The fact that J is linear is apparent from (7-3) and (7-4). The mapping of X in X^{ff} by the operator J is called the *canonical mapping* of X into X^{ff}.

It will be shown in Theorem 10.3 that J has an inverse; that is, $Jx=0$ implies $x=0$. If X is finite dimensional, this is shown in Theorem 8.2. It follows from this fact that the range of J is isomorphic to X. Hence we may identify X and the range of J, thus permitting ourselves to look upon X as a subspace of X^{ff}. When we adopt this point of view, we speak of the canonical imbedding of X in X^{ff}. If $\mathscr{R}(J)=X^{ff}$, we say that X is *algebraically reflexive*. It turns out, as we shall see later, that X is algebraically reflexive if and only if it is finite dimensional (Theorem 11.1).

I.8 LINEAR FUNCTIONALS IN FINITE-DIMENSIONAL SPACES

Let X be an n-dimensional linear space, and let $x_1, \ldots, x_n$ be a basis of X. Let $x=\xi_1x_1+\cdots+\xi_nx_n$ be the representation of an element $x \in X$. For any $x' \in X^f$ we have

$$x'(x)=\xi_1x'(x_1)+\cdots+\xi_nx'(x_n). \tag{8-1}$$

The coefficients $x'(x_1), \ldots, x'(x_n)$ are independent of x. This set of coefficients may be prescribed arbitrarily; that is, given any n scalars $\alpha_1, \ldots, \alpha_n$, there exists a unique $x' \in X^f$ such that $x'(x_i)=\alpha_i$, $i=1, \ldots, n$. This x' is given by

$$x'(x)=\sum_{i=1}^{n} \alpha_i\xi_i, \qquad \text{where} \qquad x=\sum_{i=1}^{n} \xi_ix_i. \tag{8-2}$$

The case in which one α_i is 1 and the others are 0 is of especial interest. If $\alpha_1=1$, $\alpha_2=\cdots=\alpha_n=0$, denote the corresponding x' by x_1'. Then $x_1'(x_1)=1$, $x_1'(x_2)=\cdots=x_1'(x_n)=0$. We define $x_2', \ldots, x_n'$ in similar fashion. Specifically, for $k=1, \ldots, n$, let $\alpha_i=\delta_{ik}$, and denote the corresponding x' in (8-2) by x_k'. The essential character of the functionals $x_1', \ldots, x_n'$ is then exhibited in the equations

$$\langle x_j, x_k'\rangle=\delta_{jk} \qquad j, k=1, \ldots, n. \tag{8-3}$$

To demonstrate this we observe that

$$x_j=\sum_{i=1}^{n} \delta_{ij}x_i, \qquad x_k'(x_j)=\sum_{i=1}^{n} \delta_{ik}\delta_{ij}=\delta_{jk}.$$

Theorem 8.1. *The set $x_1', \ldots, x_n'$ defined in the foregoing discussion is a basis of X^f. Hence the algebraic conjugate of an n-dimensional linear space is n-dimensional.*

Proof. The set $x_1', \ldots, x_n'$ is linearly independent. For, from $\beta_1x_1'+\cdots+\beta_nx_n'=0$ it follows with the aid of (8-3) that

$$0=\left\langle x_j, \sum_{i=1}^{n} \beta_ix_i'\right\rangle=\sum_{i=1}^{n} \beta_i\langle x_j, x_i'\rangle=\sum_{i=1}^{n} \beta_i\delta_{ji}=\beta_j,$$

or $\beta_1 = \cdots = \beta_n = 0$. Next, for a given $x' \in X^f$ we define $\alpha_i = x'(x_i)$ and prove that

$$x' = \alpha_1 x_1' + \cdots + \alpha_n x_n', \tag{8-4}$$

thereby completing the proof of the theorem. The meaning of (8-4) is that, for every $x \in X$,

$$x'(x) = \alpha_1 x_1'(x) + \cdots + \alpha_n x_n'(x). \tag{8-5}$$

The validity of this formula is readily verified with the aid of (8-2) and (8-3). □

Definition. The basis $x_1', \ldots, x_n'$ of X^f, as defined in the foregoing work, is said to be *dual* to the basis $x_1, \ldots, x_n$ of X.

Corresponding to each basis of X there is a uniquely defined dual basis of X^f.

Theorem 8.2. *If X is n-dimensional and if $x \in X$ is such that $x'(x) = 0$ for every $x' \in X^f$, it follows that $x = 0$. In other words, $Jx = 0$ implies $x = 0$, so that J^{-1} exists, where J is defined as in* § 7.

Proof. We use the basis $x_1, \ldots, x_n$ and its dual $x_1', \ldots, x_n'$. Suppose $Jx = 0$ and write $x = \xi_1 x_1 + \cdots + \xi_n x_n$ for suitable scalars $\xi_1, \ldots, \xi_n$. In the notation of (8-4) and (8-2), we have $\alpha_1 \xi_1 + \cdots + \alpha_n \xi_n = 0$ for all choices of $\alpha_1, \ldots, \alpha_n$. This implies $\xi_1 = \cdots = \xi_n = 0$, and so $x = 0$. Thus $Jx = 0$ implies that $x = 0$; that is, J^{-1} exists (Theorem 3.1). □

Theorem 8.3. *If X is an n-dimensional linear space, it is algebraically reflexive.*

Proof. We assume $n > 0$, since the case $n = 0$ is trivial. By Theorem 8.1, both X^f and X^{ff} are n-dimensional. Since J^{-1} exists (Theorem 8.2), $\mathcal{R}(J) = X^{ff}$, by Theorem 3.4. □

Theorem 8.4. *Let M be a proper subspace of an n-dimensional linear space X, and suppose $x_0 \in X \backslash M$. Then there exists an element $x' \in X^f$ such that $x'(x) = 0$ if $x \in M$ and $x'(x_0) = 1$.*

Proof. Suppose M is m-dimensional. The trivial case $m = 0$ is left to the reader, and we consider the case $0 < m < n$. Let $x_1, \ldots, x_m$ be a basis of M, and write $x_{m+1} = x_0$. The set $x_1, \ldots, x_{m+1}$ is linearly independent. By Theorem 1.4, we can find a basis of X whose first $m+1$ vectors are $x_1, \ldots, x_{m+1}$. Let $x_1', \ldots, x_n'$ be the dual basis of X^f, and let $x' = x_{m+1}'$. Then the conditions required in the theorem are satisfied. □

The proofs of Theorems 8.2 and 8.4 have been given by methods depending on finite dimensionality. But the theorems are true if the assumption of finite dimensionality of X is dropped. The proofs in this general case require more profound methods, however; see § 10.

PROBLEM

1. If X^f is n-dimensional, so is X.

I.9 ZORN'S LEMMA

What is commonly known as Zorn's lemma was originally formulated by Zorn as a proposition in the theory of sets, which he showed to be equivalent to the axiom of choice. Subsequently, this proposition has been given a formulation in terms of partially ordered sets, which is extremely convenient for many applications. In this book we state Zorn's lemma and use it as the need arises. It is immaterial whether the reader chooses to regard the lemma as an axiom of set theory or as a theorem derived from the axiom of choice.

Our statement of Zorn's lemma depends on the notion of a partially ordered set.

Definition. Let P be a set of elements. Suppose there is a binary relation defined between certain pairs of elements a, b of P, expressed symbolically by $a \prec b$, with the properties:

1. If $a \prec b$ and $b \prec c$, then $a \prec c$.
2. If $a \in P$ then $a \prec a$.
3. If $a \prec b$ and $b \prec a$, then $a = b$.

Then P is said to be *partially ordered* by the relation.

For example, if P is the set of all subsets of a given set X, the set inclusion relation ($A \subset B$) gives a partial ordering of P. The set of all complex numbers $z = x + iy$, $w = u + iv, \dots$ is partially ordered by defining $z \prec w$ to mean that $x \leqslant u$ and $y \leqslant v$, where $\leqslant$ has its usual meaning for real numbers.

If P is partially ordered and if, moreover, for every pair a, b in P either $a \prec b$ or $b \prec a$, then P is said to be *completely* ordered. (The adjectives *linearly*, *totally*, and *simply* are also used in place of *completely*.)

For example, the real numbers are completely ordered by the relation "a is less than or equal to b."

A subset of a partially ordered set P is itself partially ordered by the relation that partially orders P. Also, a subset of P may turn out to be completely ordered by this relation.

If P is a partially ordered set and S is a subset of P, an element $m \in P$ is called an *upper bound* of S if $a < m$ for every $a \in S$. An element $m \in P$ is said to be *maximal* if $a \in P$ and $m < a$ together imply $m = a$.

Zorn's Lemma. *Let P be a nonempty partially ordered set with the property that every completely ordered subset of P has an upper bound in P. Then P contains at least one maximal element.*

Some applications of this proposition will be found in §§10, 11.

I.10 EXTENSION THEOREMS FOR LINEAR OPERATORS

If X is a set, M a proper subset of X, and f a function defined on M, a function F defined on X is called an *extension* of f if $F(x) = f(x)$ when $x \in M$. In practice, we are usually concerned with extensions that conform to certain additional requirements. Often these requirements are of the sort that state that the extension F is to have certain of the properties possessed by f (e.g., be bounded, or continuous, or differentiable, or analytic, etc.). The question as to whether an extension of the required sort exists may be easy or difficult to answer, depending on the particular nature of the problem. In this section we are concerned with the existence of an extension when X is a linear space, M is a proper subspace, and f is a linear operator on M into a second linear space Y. We require the extension F to be linear on X into Y. There may also be further conditions on f and F.

Theorem 10.1. *Let X and Y be linear spaces and M a proper subspace of X. Let f be a linear operator defined on M into Y. Then there exists a linear operator F defined on X into Y such that F is an extension of f.*

Proof. Choose any element $x_0 \in X \backslash M$, and let $M_0 = M \oplus \{\alpha x_0 : \alpha \text{ any scalar}\}$. Then each element of M_0 has a unique representation in the form $x + \alpha x_0$, where $x \in M$ and α is a scalar.

We now define a linear operator F_0 on M_0 into Y as follows:

$$F_0(x + \alpha x_0) = f(x) + \alpha y_0,$$

where y_0 is some fixed vector in Y. For our present purpose it does not matter how y_0 is chosen. It is readily seen that F_0 is an extension of f and that it is a linear operator on M_0 into Y.

It now seems plausible to suppose that by a continuation of the procedure we can arrive at the required extension F defined on all of X. Indeed, if

X is finite dimensional, we shall arrive at the desired goal in a finite number of steps. For the general case the argument is made precise by use of Zorn's lemma.

Let g be a linear operator with domain $\mathscr{D}(g) \subset X$ and range $\mathscr{R}(g) \subset Y$; suppose M is a proper subspace of $\mathscr{D}(g)$ and that g is an extension of f. Let P be the class of all such operators g. If $g, h \in P$, let us define the relation $g < h$ to mean that $\mathscr{D}(g) \subset \mathscr{D}(h)$ and that h is an extension of g. This relation defines a partial ordering of P. Moreover, P is nonempty, for certainly $F_0 \in P$.

Now suppose that Q is a completely ordered subset of P. We shall define an element $G \in P$ that is an upper bound of Q. Let $\mathscr{D}(G)$ be the union of all the sets $\mathscr{D}(g)$ corresponding to elements $g \in Q$. This set $\mathscr{D}(G)$ is a subspace of X. For, suppose x_1, $x_2 \in \mathscr{D}(G)$. Then there exist elements $g_1, g_2 \in Q$ such that $x_k \in \mathscr{D}(g_k)$, $k = 1, 2$. We may suppose $g_1 < g_2$, since Q is completely ordered. Then $\mathscr{D}(g_1) \subset \mathscr{D}(g_2)$, and so $x_1 + x_2 \in \mathscr{D}(g_2) \subset \mathscr{D}(G)$. Closure of $\mathscr{D}(G)$ under multiplication by scalars is proved even more simply. Now suppose $x \in \mathscr{D}(G)$. Then $x \in \mathscr{D}(g)$ for some $g \in Q$. We shall define $G(x) = g(x)$. This definition is unambiguous, for, if $x \in \mathscr{D}(g_1)$ and $x \in \mathscr{D}(g_2)$, where $g_1, g_2 \in Q$, we have $g_1(x) = g_2(x)$ by the fact that Q is completely ordered. The proof that G is linear is like the proof that $\mathscr{D}(G)$ is a linear manifold. It is clear that $G \in P$ and that $g < G$ for every $g \in Q$.

We now know that P satisfies the conditions of Zorn's lemma and must, therefore, contain a maximal element, say F. The domain of F must be all of X, for otherwise we could regard $\mathscr{D}(F)$ as the M in the first part of the proof and thus obtain an element $g \in P$ with $g \neq F$, $F < g$, contrary to the maximality of F. The proof of the theorem is now complete, for F has the properties required in the theorem. □

Theorem 10.2. *Let M be a proper subspace of a linear space X, and suppose $x_0 \in X \setminus M$. Then there exists an element $x' \in X^f$ such that $\langle x, x' \rangle = 0$ if $x \in M$ and $\langle x_0, x' \rangle = 1$.*

Proof. We observe that this is the same as Theorem 8.4 with the hypothesis of finite dimensionality omitted. To prove the theorem, let $M_0 = M \oplus \{\alpha x_0 : \alpha \text{ any scalar}\}$ and define f on M_0 by $f(x + \alpha x_0) = \alpha$. Then f is a linear functional on M_0, and $f(x) = 0$ if $x \in M$, whereas $f(x_0) = 1$. By Theorem 10.1 (with Y the linear space of scalars), there exists an element of X^f that is an extension of f. If we denote this element by x', we see that it has the required properties. □

Theorem 10.3. *Let X be a linear space. If $x \in X$ and if $\langle x, x' \rangle = 0$ for every $x' \in X^f$, then $x = 0$. In other words, J^{-1} exists, where J is the canonical mapping defined in* § 7.

Proof. Applying Theorem 10.2 with $M=(0)$, we see that if $x_0 \neq 0$, there exists $x' \in X^f$ such that $\langle x_0, x'\rangle \neq 0$. This statement is merely the contrapositive form of Theorem 10.3, so the proof is complete. Note that this proposition is the same as Theorem 8.2 with the hypothesis of finite dimensionality omitted. □

Next, we consider an extension theorem for linear functionals that are subjected to a special kind of condition.

Definition. Let X be a real linear space, and let p be a real-valued function defined on X, with the properties:

1. $p(x+y) \leq p(x)+p(y)$.
2. $p(\alpha x)=\alpha p(x) \quad \text{if} \quad \alpha \geq 0$.

We call p a *sublinear functional* (or a *gauge*) on X.

As an example of sublinear functional, we cite the example of the space $X=\boldsymbol{R}^n$ with p defined by $p(x)=|\xi_1|+\cdots+|\xi_n|$, where $x=(\xi_1,\ldots,\xi_n)$. Another example is that of $X=C[a,b]$, with $p(x)=|x(a)|$ or $p(x)=\max|x(t)|$ for $a\leq t\leq b$.

Theorem 10.4 (Hahn–Banach). *Let X be a real linear space, and let M be a proper subspace of X. Let p be a sublinear functional defined on X, and let f be a linear functional defined on M such that $f(x)\leq p(x)$ for each $x\in M$. Then there exists a linear functional x' defined on X such that x' is an extension of f and $-p(-x)\leq x'(x)\leq p(x)$ for each $x\in X$.*

Proof. We begin with x_0 and M_0 as in the proof of Theorem 10.1. The space Y is now $\boldsymbol{R}$. The first step is to show that there exists a real number ξ_0 such that the linear functional F_0 defined on M_0 by

$$F_0(x+\alpha x_0)=f(x)+\alpha\xi_0 \tag{10-1}$$

satisfies the inequality

$$F_0(x+\alpha x_0)\leq p(x+\alpha x_0), \qquad x\in M, \alpha\in \boldsymbol{R}, \tag{10-2}$$

that is, $F_0(w)\leq p(w)$ for all $w\in M_0$. Now, if $x,y\in M$, then $f(x)+f(y)=f(x+y)\leq p(x+y)\leq p(x-x_0)+p(y+x_0)$, and so

$$f(x)-p(x-x_0)\leq p(y+x_0)-f(y). \tag{10-3}$$

Since the right side of (10-3) is independent of x, there exists $\xi_0\in\boldsymbol{R}$ such that

$$\xi_0=\sup_{x\in M}\{f(x)-p(x-x_0)\}.$$

Then for $x, y \in M$, we deduce that

$$f(x)-\xi_0 \leq p(x-x_0), \qquad f(y)+\xi_0 \leq p(y+x_0).$$

Taking $\alpha > 0$ and replacing x and y by $\alpha^{-1}x$, we have

$$f(\alpha^{-1}x)-\xi_0 \leq p(\alpha^{-1}x-x_0), \qquad f(\alpha^{-1}x)+\xi_0 \leq p(\alpha^{-1}x+x_0).$$

Multiplying these inequalities by α and using (10-1), we obtain (10-2) when α is either negative or positive; (10-2) is obviously valid if $\alpha = 0$. This completes the first phase of the proof.

The proof is completed by use of Zorn's lemma in essentially the same way as in the proof of Theorem 10.1. We consider the class P of all linear functionals g with $\mathscr{D}(g) \subset X$ such that M is a proper subspace of $\mathscr{D}(g)$, g is an extension of f, and $g(x) \leq p(x)$ when $x \in \mathscr{D}(g)$. The partial ordering of P is defined as before, and the rest of the proof is essentially as before. We obtain a maximal element x' of P whose domain is all of x such that x' is an extension of f and $x'(x) \leq p(x)$ for $x \in X$. Since x' is linear, we also have $x'(x) = -x'(-x) \geq -p(-x)$ for $x \in X$. $\square$

I.11 HAMEL BASES

The German mathematician Hamel conceived the notion of a "basis" for all real numbers as follows: Let H be a set of real numbers with the properties:

1. If $x_1, \ldots, x_n$ is any finite subset of H and if $r_1, \ldots, r_n$ are rational numbers for which $r_1x_1 + \cdots + r_nx_n = 0$, then $r_1 = \cdots = r_n = 0$.
2. Every real number x can be expressed as a finite linear combination of elements of H, with rational coefficients.

In terms of such a basis, Hamel [1] then discussed real functions f satisfying the equations $f(x+y) = f(x)+f(y)$ for all real x, y. If H is a basis in the foregoing sense, let a real function f be defined as follows: If $x \in H$, assign the values of $f(x)$ arbitrarily. Any real x has a unique representation $x = r_1x_1 + \cdots + r_nx_n$, where $x_1, \ldots, x_n$ are in H and $r_1, \ldots, r_n$ are rational (n may vary with x, of course). We then define $f(x) = r_1 f(x_1) + \cdots + r_n f(x_n)$, the values $f(x_1), \ldots, f(x_n)$ having already been assigned. With this definition f turns out to satisfy the condition $f(x+y) = f(x)+f(y)$ for every x and y. This procedure gives all possible functions f satisfying this condition.

To show the existence of a basis for all real numbers, Hamel used an argument based on the proposition that every set can be well ordered. This proposition is equivalent to the axiom of choice, and to Zorn's lemma.

Hamel's whole procedure can be adapted to the purpose of showing the nature of all possible linear functionals on a linear space X.

Definition. Let X be a linear space with some nonzero elements. A set $H \subset X$ is called a *Hamel basis* of X if:

1. H is a linearly independent subset of X.
2. The linear manifold spanned by H is all of X.

To show the existence of a Hamel basis, let P be the class whose members are the linearly independent subsets of X. Let P be partially ordered by the relation of set inclusion (i.e., $M < N$ if $M, N \in P$ and $M \subset N$).

It is easy to see that P satisfies the conditions of Zorn's lemma. For, if $x_1 \neq 0$, the set consisting of x_1 alone is in P and, if Q is a completely ordered subset of P, the subset of X obtained by taking the union of all the subsets of X comprised in Q is in P and is an upper bound for Q. Hence P must contain a maximal element, say H. Then the linear manifold spanned by H must be X. For otherwise some element x of X is not in the linear manifold spanned by H, and the set consisting of x and the elements of H is linearly independent and contains H as a proper subset, contrary to the maximal character of H. Thus H is a Hamel basis of X.

If $x' \in X^f$, we have $x'(x) = \alpha_1 x'(x_1) + \cdots + \alpha_n x'(x_n)$ whenever $x = \alpha_1 x_1 + \cdots + \alpha_n x_n$, the α's being any scalars. Conversely, suppose H is a Hamel basis of X. Define x' by assigning the value of $x'(x)$ arbitrarily when $x \in H$, and define $x'(x) = \alpha_1 x'(x_1) + \cdots + \alpha_n x'(x_n)$ if x is any element of X, uniquely represented in the form $x = \alpha_1 x_1 + \cdots + \alpha_n x_n$ where $x_1, \ldots, x_n$ are in H and $\alpha_1, \ldots, \alpha_n$ are scalars. Then $x' \in X^f$, as is easily seen.

By using the concept of a Hamel basis, we can show that no infinite-dimensional space is algebraically reflexive. First, we observe that if S is a linearly independent subset of X, there is a Hamel basis H of X such that S is a subset of H. This is shown by use of Zorn's lemma, taking P to be the class of all linearly independent subsets of X that contain S.

Theorem 11.1. *If X is infinite dimensional, it is not algebraically reflexive. Consequently (see Theorem 8.3), a linear space is algebraically reflexive if and only if it is finite dimensional.*

Proof. Let $H = \{x_i : i \in I\}$ be a Hamel basis of X. Then I is an infinite index set, and $x_i \neq x_j$ if $i \neq j$. Define $x_i' \in X^f$ by $x_i'(x_i) = 1, x_i'(x_j) = 0$ if $i \neq j$. Then the set $\{x_i' : i \in I\}$ is a linearly independent set in X^f. For, if we suppose that $\sum_{\nu=1}^n \alpha_\nu x'_{i(\nu)} = 0$, we have $0 = \sum_{\nu=1}^n \alpha_\nu x'_{i(\nu)}(x_{i(\mu)}) = \alpha_\mu$ when $\mu = 1, \ldots, n$. Now let H' be a Hamel basis of X^f that contains the set $\{x_i' : i \in I\}$.

Let $\{\beta_i : i \in I\}$ be a set of numbers such that $\beta_i \neq 0$ for infinitely many indices i. Define $x'' \in X^{ff}$ by setting $x''(x_i') = \beta_i$ and $x''(x') = 0$ if $x' \in H'$ but x' is not one of the elements x_i'. Consider an element $x_0'' \in X^{ff}$ of the form $x_0'' = Jx$, $x \in X$. We have $x_0''(x_i') = x_i'(x) = \alpha_i$, where α_i is the coefficient of x_i in the representation of x in terms of the Hamel base H. Since $\alpha_i = 0$ for all but a finite number of indices i, it follows that the set of α_i's cannot be the same as the set of β_i's, and therefore $x'' \neq x_0''$. This shows that X^{ff} contains elements not in the range of J, so that X is not algebraically reflexive. □

Theorem 11.2. *If M is a subspace of a linear space X, then there exists a complementary subspace N in X, so that $X = M \oplus N$.*

Proof. Since M is itself a linear space, it possesses a Hamel basis H_M. Let H be a Hamel basis of X that contains H_M (see the remark preceding Theorem 11.1), and let N be the linear manifold spanned by the elements in $H \backslash H_M$. It is easily verified that $X = M \oplus N$. In fact, each $x \in X$ is a linear combination of elements of H and so is the sum of a linear combination of elements of H_M and a linear combination of elements of $H \backslash H_M$. Thus $X = M + N$. Also, if $x \in M \cap N$, then x is a finite linear combination of elements of H in two different ways, namely, as a linear combination $\Sigma\, \alpha_i u_i$ of elements u_i in H_M and as a linear combination $\Sigma\, \beta_j v_j$ of elements v_j in $H \backslash H_M$. Then $0 = \Sigma\, \alpha_i u_i - \Sigma\, \beta_j v_j$. This implies that $\alpha_i = \beta_j = 0$ for all i,j, because the u_i and v_j belong to a linearly independent set H. Thus $x = 0$, which shows that $M \cap N = (0)$. □

PROBLEMS

1. If H_1 and H_2 are any two Hamel bases for X, they have the same cardinal number. For each $x \in H_1$ let $H_2(x)$ be the finite set of those elements of H_2 that are needed to represent x by using the basis H_2. Show that if $y \in H_2$, then $y \in H_2(x)$ for some $x \in H_1$, and hence that

$$H_2 = \bigcup_{x \in H_1} H_2(x).$$

   The cardinality argument from here on is like that used in another context later in this book; see the last part of the proof of Theorem II.6.12.
2. The cardinal number of a Hamel basis of X is called the dimension of X. In the case of a space X of infinite dimension, the cardinal of X itself can be shown to be the product $C \cdot \dim X$, where C is the cardinal of the set of all real numbers and $\dim X$ is the dimension of X. Since this product is just the larger of the two factors, the cardinal of X is C if $\dim X < C$ and $\dim X$ if $C \leq \dim X$. See Jacobson [1, Chapter 9] and Löwig [1, page 20].
3. The dimensions of X and X^f are related as follows, provided that $\dim X$ is infinite:

$$\dim X^f = C^{\dim X}.$$

   See Jacobson [1, Chapter 9]. It follows that X cannot be algebraically reflexive if $\dim X$ is infinite for, in that case, $\dim X < \dim X^f$. We can also write $\dim X^f = 2^{\dim X}$ in this case. For, we have $C = 2^{\aleph_0}$, and $\aleph_0 \aleph = \aleph$ for any infinite cardinal $\aleph$, whence $C^{\aleph} = 2^{\aleph}$.

1.12 THE TRANSPOSE OF A LINEAR OPERATOR

Suppose X and Y are linear spaces, and let A be a linear operator on X into Y. To each $y' \in Y^f$ let us make correspond the element $x' \in X^f$ defined by

$$\langle x, x'\rangle = \langle Ax, y'\rangle, \tag{12-1}$$

where x varies over all of X, and let us denote the function so defined by A^T; that is, $A^T y' = x'$. It is easy to check that A^T is a linear operator on Y^f into X^f. We leave the verifications to the reader. Equation (12-1) may now be written in the form

$$\langle x, A^T y'\rangle = \langle Ax, y'\rangle, \qquad x \in X,\ y' \in Y^f. \tag{12-2}$$

Definition. The operator A^T defined by (12-2) is called the *transpose* of A.

To see the motivation for this terminology, let us consider the case in which X and Y are finite dimensional. Let X be n-dimensional, with basis $x_1, \ldots, x_n$, and let Y be m-dimensional, with basis $y_1, \ldots, y_m$. Let the corresponding dual bases in X^f and Y^f be $x_1', \ldots, x_n'$ and $y_1', \ldots, y_m'$, respectively (see § 8). We know from § 4 that the operator A determines, and is determined by, a matrix

$$\left\| \begin{matrix} \alpha_{11} & \alpha_{12} & \cdots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \cdots & \alpha_{2n} \\ \cdot & \cdot & & \cdot \\ & & & \\ \alpha_{m1} & \alpha_{m2} & \cdots & \alpha_{mn} \end{matrix} \right\|. \tag{12-3}$$

The basic formulas are [see (4-3)]

$$Ax_j = \sum_{i=1}^{m} \alpha_{ij} y_i \qquad j = 1, \ldots, n. \tag{12-4}$$

Since A^T is a linear operator on Y^f into X^f, it follows that A^T is determined by a certain matrix

$$\left\| \begin{matrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \beta_{n1} & \beta_{n2} & \cdots & \beta_{nm} \end{matrix} \right\| \tag{12-5}$$

and the set of equations

$$A^T y_i' = \sum_{k=1}^{n} \beta_{ki} x_k', \qquad i = 1, \ldots, m. \tag{12-6}$$

If we use the relations between a basis and the dual basis, we shall be able to find the relation that the matrix (12-5) bears to the matrix (12-3). We know that $\langle x_j, x_k' \rangle = \delta_{jk}$ [see (8-3)] and likewise that $\langle y_k, y_i' \rangle = \delta_{ki}$. Consequently, by (12-4),

$$\langle Ax_j, y_i' \rangle = \left\langle \sum_{k=1}^{m} \alpha_{kj} y_k, y_i' \right\rangle = \sum_{k=1}^{m} \alpha_{kj} \delta_{ki} = \alpha_{ij}.$$

But also, by (12-6),

$$\langle Ax_j, y_i' \rangle = \langle x_j, A^T y_i' \rangle = \left\langle x_j, \sum_{k=1}^{n} \beta_{ki} x_k' \right\rangle = \sum_{k=1}^{n} \beta_{ki} \delta_{jk} = \beta_{ji}.$$

Therefore

$$\alpha_{ij} = \beta_{ji}. \tag{12-7}$$

In other words, the matrix (12-5) has for its kth row the kth column of the matrix (12-3). In the customary terminology of matrix algebra, the matrix (12-5) is the *transpose* of the matrix (12-3). It is on this account that the operator A^T is called the transpose of A.

The main reason for introducing the notion of the transpose of a linear operator is that the discussion of the existence of A^{-1} and the description of the range $\mathscr{R}(A)$ are facilitated by considering the transpose operator A^T. We shall see this in § 13.

I.13 ANNIHILATORS, RANGES, AND NULL SPACES

Our goal in this section is to discuss certain relations between a linear operator and its transpose. At first, however, we shall be concerned only with linear manifolds in a linear space X and in its algebraic conjugate X^f.

Definition. Suppose S is a subset of X. The *annihilator of S in X^f* is the set $S^\perp$ of all $x' \in X^f$ such that $\langle x, x' \rangle = 0$ if $x \in S$. If S is a subset of X^f, the *annihilator of S in X* is the set $S^\perp$ of all $x \in X$ such that $\langle x, x' \rangle = 0$ if $x' \in S$.

Clearly $X^\perp$ is the zero subspace of X^f. The fact that $(X^f)^\perp$ is the zero subspace of X follows from Theorem 10.3. The following properties of annihilators of subsets of X and X^f are easily verified.

(13-1) $S^\perp$ is a linear manifold, whether or not S is.

(13-2) $$S \subset S^{\perp\perp}, \text{ where } S^{\perp\perp} \text{ denotes the set } (S^{\perp})^{\perp}.$$

(13-3) $$S_1 \subset S_2 \text{ implies } S_2^{\perp} \subset S_1^{\perp}, \text{ and hence } S_1^{\perp\perp} \subset S_2^{\perp\perp}.$$

Replacing S by $S^{\perp}$ in (13-2), we have $S^{\perp} \subset S^{\perp\perp\perp}$. Applying (13-3) to the relation (13-2), we have $S^{\perp\perp\perp} \subset S^{\perp}$. Hence $S^{\perp} = S^{\perp\perp\perp}$, and the process of forming annihilators stops.

Theorem 13.1. *If S is a subset of X, then $S^{\perp\perp}$ is the linear manifold spanned by S. In particular, if M is a subspace of X, we have*

(13-4) $$M^{\perp\perp} = M.$$

Proof. Let M be the linear manifold spanned by S. Since $S^{\perp\perp}$ is a linear manifold containing S, we have $S^{\perp\perp} \supset M$. If we suppose $x_0 \in S^{\perp\perp} \backslash M$, we know by Theorem 10.2 that there exists $x' \in X^f$ such that $\langle x, x' \rangle = 0$ if $x \in M$, but $\langle x_0, x' \rangle \neq 0$. The first of these two conditions implies that $x' \in S^{\perp}$, because $S \subset M$. But then $\langle x_0, x' \rangle = 0$ because $x_0 \in S^{\perp\perp}$. We have now reached a contradiction. Therefore $S^{\perp\perp} \subset M$, and the proof is complete. □

Corollary 13.2. *If M_1 and M_2 are subspaces of X such that $M_1^{\perp} = M_2^{\perp}$, then $M_1 = M_2$.*

Proof. By (13-4), we have $M_1 = (M_1^{\perp})^{\perp} = (M_2^{\perp})^{\perp} = M_2$. □

Theorem 13.1 does not always hold for subsets of X^f. The linear manifold spanned by a subset S of X^f may be a proper subset of $S^{\perp\perp}$. Problem 1 gives an example of a linear manifold M in X^f such that $M \neq M^{\perp\perp}$.

Theorem 13.3. *Let M be a proper subset of X^f. Then $M = M^{\perp\perp}$ if and only if to each $x' \in X^f \backslash M$ corresponds some $x \in M^{\perp}$ such that $\langle x, x' \rangle \neq 0$.*

Proof. From (13-2) it follows that $M = M^{\perp\perp}$ if and only if $M^{\perp\perp} \subset M$. For a given $x' \in X^f$, the statement "to x' corresponds some $x \in M^{\perp}$ such that $\langle x, x' \rangle \neq 0$" is equivalent to the statement "x' is not in $M^{\perp\perp}$." Thus the theorem can be rephrased as follows: $M^{\perp\perp} \subset M$ if and only if $x' \in X^f \backslash M$ implies $x' \in X^f \backslash M^{\perp\perp}$. In this form the assertion is obviously true, since a set-theoretic inclusion relation is reversed when we pass from the sets to their complements in X^f. □

Definition. A subspace M of X^f is said to be *algebraically saturated* if $M = M^{\perp\perp}$.

Both (0) and X^f are algebraically saturated subspaces of X^f. This follows from the remark preceding (13-1). Also, if M is any subset of X, its annihilator $M^{\perp}$ is algebraically saturated, since $M^{\perp} = M^{\perp\perp\perp}$, as we observed earlier. It will follow easily from Theorem III.1.2 that every finite-dimensional subspace of X^f is algebraically saturated. In particular, if X is finite dimensional, then every subspace of X^f is algebraically saturated. (Cf. problem 2.)

Ranges and Null Spaces

We now consider an arbitrary linear operator A on X into Y, where X and Y are linear spaces with the same field of scalars. In § 3 we discussed the existence and uniqueness of solutions to equations of the form $Ax = y$, where y is a given element of Y. The existence problem leads us to seek various ways of characterizing the range $\mathscr{R}(A)$. One method is to use the transpose operator A^T.

Theorem 13.4.
(*a*) $\{\mathscr{R}(A)\}^\perp = \mathcal{N}(A^T)$.
(*b*) $\mathscr{R}(A) = \{\mathcal{N}(A^T)\}^\perp$.
(*c*) $\mathscr{R}(A) = Y$ *if and only if* $(A^T)^{-1}$ *exists.*

Proof. By definition of A^T, we have $\langle Ax, y'\rangle = \langle x, A^T y'\rangle$ for $x \in X$, $y' \in Y^f$. If $y' \in \{\mathscr{R}(A)\}^\perp$, we see that $\langle Ax, y'\rangle = 0$ for each $x \in X$, and so $A^T y' = 0$, or $y' \in \mathcal{N}(A^T)$. The reasoning is reversible, and so (*a*) is proved. Taking annihilators in (*a*) and using Theorem 13.1, we obtain (*b*). To prove (*c*) note that if $\mathscr{R}(A) = Y$, then $\mathcal{N}(A^T) = (0)$, by (*a*). This implies that $(A^T)^{-1}$ exists. Conversely, if $(A^T)^{-1}$ exists, then $\mathcal{N}(A^T) = (0)$ and $\mathscr{R}(A) = \{\mathcal{N}(A^T)\}^\perp = (0)^\perp = Y$, by (*b*). □

Recall from § 3 that solutions to equations of the form $Ax = y$ are unique if and only if A^{-1} exists, that is, if and only if $\mathcal{N}(A) = (0)$. The transpose operator can be used to describe this situation.

Theorem 13.5.
(*a*) $\mathcal{N}(A) = \{\mathscr{R}(A^T)\}^\perp$.
(*b*) $\{\mathcal{N}(A)\}^\perp = \mathscr{R}(A^T)$.
(*c*) A^{-1} *exists if and only if* $\mathscr{R}(A^T) = X^f$.

Proof. The formula $\langle x, A^T y'\rangle = \langle Ax, y'\rangle$ shows at once that $\mathcal{N}(A) \subset \{\mathscr{R}(A^T)\}^\perp$. The same formula shows that if $x \in \{\mathscr{R}(A^T)\}^\perp$, then $\langle Ax, y'\rangle = 0$ for every $y' \in Y^f$. But then $x \in \mathcal{N}(A)$, by Theorem 10.3. This proves (*a*). From (*a*) we obtain

$$\mathscr{R}(A^T) \subset \{\mathscr{R}(A^T)\}^{\perp\perp} \subset \{\mathcal{N}(A)\}^\perp.$$

Now take $x' \in \{\mathcal{N}(A)\}^\perp$ so that $\langle x, x'\rangle = 0$ whenever $Ax = 0$. If $y = Ax_1$ and

$y = Ax_2$, then $0 = A(x_1 - x_2)$ and therefore $\langle x_1 - x_2, x'\rangle = 0$, or $\langle x_1, x'\rangle = \langle x_2, x'\rangle$. From this observation it follows that if $y = Ax$ and if we define a mapping of $\mathscr{R}(A)$ into the scalar field by $f(y) = \langle x, x'\rangle$, then f is well defined. (The value of $f(y)$ depends on y and not on the particular x such that $Ax = y$.) It is readily verified that f is linear.

Now let $y' \in Y^f$ be an extension of f. Then $\langle x, A^T y'\rangle = \langle Ax, y'\rangle = f(Ax) = \langle x, x'\rangle$. Since this is true for all x, it follows that $A^T y' = x'$, so that $x' \in \mathscr{R}(A^T)$. Hence $\{\mathscr{N}(A)\}^\perp \subset \mathscr{R}(A^T)$, which completes the proof of (b). Part (c) follows immediately from (a) and (b) and the fact (noted earlier) that $\{X^f\}^\perp = (0)$. □

The proof of (b) above is due to Howard Wicke. It improves the original form of the theorem. Note from (b) that $\mathscr{R}(A^T)$ is algebraically saturated, since it is the annihilator of a set in X.

PROBLEMS

1. Let X be the space (c) of all convergent sequences $x = \{\xi_n\}$, with addition and scalar multiplication defined in the natural way. Let M be the subspace of X^f consisting of all linear functionals x' expressible in the form

$$\langle x, x'\rangle = \sum_{i=1}^{\infty} \xi_i \xi_i' \qquad \text{where} \qquad \sum_{i=1}^{\infty} |\xi_i'| < \infty.$$

Let x_0' be the particular element of X^f defined by

$$\langle x, x_0'\rangle = \lim_{n\to\infty} \xi_n.$$

   a. Show that $M^\perp = (0)$, and hence $M^{\perp\perp} = X^f$.
   b. Show that $x_0' \in M^{\perp\perp} \backslash M$. This will prove that M is not algebraically saturated.

2. Prove that if X is finite dimensional, then every subspace of X^f is algebraically saturated. (Do not use Theorem III.1.2 but, rather, use the fact that X is algebraically reflexive.)
3. Let X be an n-dimensional linear space, and let M be an m-dimensional subspace of X. Then the annihilator M has dimension $n - m$.
4. Let M be a subspace of a linear space X. Show that $(X/M)^f$ is isomorphic to $M^\perp$. [*Hint.* If u' is a linear functional on X/M, define Tu' for $x \in X$ by $(Tu')(x) = u'([x])$. Show that T is an injective linear mapping of $(X/M)^f$ into X^f with range $M^\perp$.] Then show that problem 3 follows easily from this result.
5. If M is a subspace of a linear space X, then the quotient space $X^f/M^\perp$ is isomorphic to the algebraic conjugate M^f. [*Hint.* For $x' \in X^f$, let Tx' be the restriction of x' to elements of M. Show that T is a linear mapping of X^f onto M^f whose null space is $M^\perp$.]

I.14 CONCLUSIONS

If we suppose that spaces X, Y and an operator A come to us as known things, the theorems in §13 will be effectively useful to us in furnishing information about $\mathcal{R}(A)$ and $\mathcal{N}(A)$ only to the extent that we know or can find out the relevant information about $\mathcal{R}(A^T)$ and $\mathcal{N}(A^T)$. For this we must, at the very least, know some things about the spaces X^f, Y^f and the operator A^T.

In the case in which X and Y are finite dimensional, the spaces X^f and Y^f are also finite dimensional. In this case the operator A is fully represented by a matrix, and the transpose A^T is represented by the transposed matrix. For the finite-dimensional case all the results of § 13 are well known in algebra (though perhaps in a different terminology) as part of the theory of systems of linear equations or of the theory of matrices.

For infinite-dimensional spaces the results of § 13 are not as useful as could be wished, because the algebraic conjugate of an infinite-dimensional space is not as amenable to study as the algebraic conjugate of a finite-dimensional space. The theory for finite-dimensional spaces is much simplified by the fact that we have simple "concrete representations" of such spaces. If X is an n-dimensional linear space with real scalars, it is isomorphic to the arithmetic space $\boldsymbol{R}^n$ (see the proof of Theorem 4.1), and we can regard $\boldsymbol{R}^n$ as a concrete representation of X. Likewise, if A is a linear operator on X into Y (spaces of dimensions n and m respectively), we have a concrete representation of A by a certain $m \times n$ matrix. The purpose of such a representation is to replace one object of study by another object whose properties are more readily apparent because the structure of the object is more familiar or more susceptible to intuitive perception.

In order to obtain results of the type presented in § 13, but more useful, it seems essential to abandon the purely algebraic approach to linear problems for infinite-dimensional spaces and to bring in topological considerations. If we impose a specified topological structure on a linear space X, it is natural to consider those linear functionals on X that are continuous with respect to the topology. These form a subspace of the algebraic conjugate space X^f; this subspace may be called the *topological conjugate* of X. In a good many interesting special cases the topological conjugate of a given topological linear space X can be studied more effectively than the algebraic conjugate of X. This may be the case, for one thing, because the topological conjugate is significantly smaller than the algebraic conjugate (i.e., has fewer elements). It may also turn out that there is a useful concrete representation of the topological conjugate space. Thus, for example, if X is the space of real-valued continuous functions defined on $0 \leqslant t \leqslant 1$, with a certain natural topology for X, the topological conjugate of X is isomorphic to a certain space whose elements form a subclass of the class of all real-valued functions of bounded variation defined on $0 \leqslant t \leqslant 1$; see Theorem III.5.6.

This necessarily somewhat vague and general discussion is intended to suggest in part why the abstract approach to linear problems, as far as infinite-dimensional spaces are concerned, is more fruitful if suitable topological as well as algebraic methods are employed. There are still other reasons why it is useful to take account of a topology for a given linear space. A given linear operator may have important properties that find their appropriate expression in terms of the topological or metric notions of "nearness," "small variations," and other related ideas.

The greatest successes of the abstract approach to linear problems have been achieved in the study of Hilbert space. A Hilbert space is a very special type of linear space with a topology. It has many of the properties of Euclidean space and is, therefore, amenable to the use of geometric reasoning guided by intuition. But perhaps the decisive reason for the importance of Hilbert space, and for the greater success that has been achieved with Hilbert space than with other spaces, is the fact that the topological conjugate of a Hilbert space is isomorphic with the given Hilbert space.

II TOPOLOGICAL LINEAR SPACES

In this chapter we study vector spaces that possess a topological structure such that the operations of vector addition and scalar multiplication are continuous with respect to the topology. For our purposes, the most useful spaces of this type are *normed linear spaces*. In such a vector space the topology is introduced, by means of a metric, in a simple and natural way by assuming that each vector in the space has a length and that the rules governing the lengths of vectors conform to a few simple and natural geometric principles. Most of the latter part of the book concerns the use of normed linear spaces in functional analysis. Actually, many of the deepest results involve *Banach* spaces—normed linear spaces that are complete as metric spaces. We begin the study of these spaces in § 4.

Of special interest among normed linear spaces are the inner-product spaces, in which the length, or *norm*, of a vector is defined in terms of a certain function of two vectors, called the *inner product*. It is formally analogous to the dot product of ordinary vector analysis. The classical example of an inner-product space is the space of all continuous real-valued functions on a closed real interval, say $[-\pi, \pi]$, with the inner product of two such functions f, g given by a Riemann integral

$$(f, g) = \int_{-\pi}^{\pi} f(t)g(t)\,dt.$$

Important problems in physics and engineering can be formulated in terms of a space such as this. However, this space is not complete as a metric space. Fortunately, most of the theory we develop will apply to such a space. The exposition in § 6 on inner-product spaces would not be shortened significantly by assuming completeness everywhere; moreover, a few results such as Parseval's formula (Theorem 6.10) gain in clarity when the inessential hypothesis of completeness of the space is removed. Those results depending on completeness are considered separately in the section on Hilbert spaces (§ 7).

In the second part of the chapter, beginning with § 9, we examine more general topological linear spaces. Some knowledge of these spaces is advantageous, even if one restricts his or her attention mainly to normed linear spaces. In particular, the study of *convex sets* and *locally convex spaces* in § 10 and § 11 will be needed as background for the discussions in Chapter III. Sections 9 to 13 contain basic definitions and some elementary theorems for topological linear spaces. The reader who is interested mainly in normed linear spaces should look at § 9 as far as Theorem 9.2, study § 10, read § 11 as far as Theorem 11.3, and then skip to Chapter III.

II.1 NORMED LINEAR SPACES

A *norm* on a linear space X is a real-valued function, whose value at x we denote by $\|x\|$, with the properties:

1. $\|x_1+x_2\| \leq \|x_1\| + \|x_2\|$.
2. $\|\alpha x\| = |\alpha| \|x\|$.
3. $\|x\| \geq 0$.
4. $\|x\| \neq 0$ if $x \neq 0$.

Actually, property (3) is a consequence of properties (1) and (2). For $\|0\|=0$ by (2) (put $\alpha=0$), and from (1) and (2), $0=\|x-x\| \leq \|x\|+\|-x\|=2\|x\|$, whence $0 \leq \|x\|$.

A linear space on which a norm is defined becomes a metric space if we define $d(x_1, x_2)=\|x_1-x_2\|$, as is easily verified. A linear space that is a metric space in this way is called a *normed linear space*, or a *normed vector space*, and the topology associated with the metric is called the *norm topology*.

The set $\{x : \|x-x_0\|<r\}$, where $r>0$, is called the *open ball* (or sphere) of radius r with center x_0; the closure of this set is $\{x: \|x-x_0\| \leq r\}$, which we call the *closed ball* of radius r with center x_0. By the *surface* of this ball we mean the set $\{x: \|x-x_0\|=r\}$.

A set S in a normed linear space is bounded if and only if it is contained in some ball; an equivalent condition is that S be contained in some ball with center at 0, which means that $\|x\|$ is bounded for $x \in S$.

If f is a function with domain $\mathscr{D}(f) \subset X$ and range $\mathscr{R}(f) \subset Y$, where X and Y are normed linear spaces, continuity of f at $x_0 \in \mathscr{D}(f)$ is expressed by the condition: to each $\varepsilon>0$ corresponds some $\delta>0$ such that $\|f(x)-f(x_0)\|<\varepsilon$ if $x \in \mathscr{D}(f)$ and $\|x-x_0\|<\delta$.

To show that a normed linear space X is a topological linear space, we must verify the continuity of addition and scalar multiplication. The inequality

$$\|(x_1+x_2)-(y_1+y_2)\| \leq \|x_1-y_1\|+\|x_2-y_2\|$$

shows that vector addition is a continuous function from $X \times X$ into X. Let $\boldsymbol{K}$

denote the scalar field ($\boldsymbol{R}$ or $\boldsymbol{C}$) associated with X. The relations

$$\begin{aligned}\|\alpha x - \alpha_0 x_0\| &= \|\alpha(x - x_0) + (\alpha - \alpha_0)x_0\| \\ &\leqslant |\alpha|\,\|x - x_0\| + |\alpha - \alpha_0|\,\|x_0\|,\end{aligned}$$

where α, $\alpha_0 \in \boldsymbol{K}$, show that scalar multiplication is a continuous function from $\boldsymbol{K} \times X$ into X.

It is also to be observed that the norm itself is a uniformly continuous function from X into the real numbers. This is evident from the inequality

$$|\|x_1\| - \|x_2\|| \leqslant \|x_1 - x_2\|,$$

which follows easily from property (1) of the norm.

We make free use of the definitions and terminology concerning linear spaces as given in Chapter I. A subspace of a normed linear space is itself a normed linear space. However, a subspace of a normed linear space may or may not be closed, and the distinction between closed and nonclosed subspaces is often important. The terms *subspace* and *linear manifold* are used interchangeably. As a result of the continuity of addition and scalar multiplication, it is easy to see that, if M is a linear manifold in X, the closure $\bar{M}$ is also a linear manifold. If S is any set in X, the closure in X of the linear manifold generated by S (as defined in § I.1) is called the closed linear manifold generated by S. The word "generated" in this definition is sometimes replaced by "determined" or "spanned."

Definition. Two normed linear spaces X and Y are said to be *isometrically isomorphic* or, more briefly, *congruent*, if there is a one-to-one correspondence between the elements of X and Y that makes the two spaces isomorphic in the sense defined in § I.3 and isometric as metric spaces.

In order that X and Y be congruent, it is necessary and sufficient that there exist a linear operator T with domain X and range Y such that $\|Tx\| = \|x\|$ for every $x \in X$. Here we use the same symbol for the norm in X as for the norm in Y. We shall meet examples of congruent spaces when we deal with concrete representations of various linear spaces whose elements are linear operators.

Definition. Two normed linear spaces X and Y are said to be *topologically isomorphic* if there is a linear operator T (with inverse T^{-1}) that establishes the isomorphism of X and Y and that furthermore has the property that T and T^{-1} are continuous on their respective domains. In other words, X and Y are topologically isomorphic provided there is a homeomorphism T of X onto Y that is also a linear operator. For this reason X and Y may be called *linearly homeomorphic*.

As we shall see later, there exist examples of pairs of spaces that are topologically isomorphic but not congruent.

It is convenient at this point to consider a few facts about continuous linear operators in normed linear spaces. Linear operators, without regard to continuity, were defined in § I.3.

Theorem 1.1. *Let X and Y be normed linear spaces and T a linear operator on X into Y. If T is continuous at some point $x_0 \in X$, then T is uniformly continuous on all of X. Furthermore, T is (uniformly) continuous on X if and only if there is a constant M such that $\|Tx\| \leq M\|x\|$ for every x in X.*

Proof. If T is continuous at x_0, then to each $\varepsilon > 0$ corresponds $\delta > 0$ such that

$$\|Tx - Tx_0\| < \varepsilon \qquad \text{whenever} \qquad \|x - x_0\| < \delta. \tag{1-1}$$

Choose $z \in X$ and suppose y satisfies $\|y - z\| < \delta$. Let $x = y - z + x_0$. Then $x - x_0 = y - z$, so x satisfies (1-1). It follows from (1-1) and the linearity of T that $\|Ty - Tz\| = \|Tx - Tx_0\| < \varepsilon$. This demonstrates the uniform continuity of T.

If $\|Tx\| \leq M\|x\|$ for all x, it is clear that T is continuous at 0 [recall that $T(0) = 0$]. Conversely, if T is continuous at 0, there exists $\delta > 0$ such that $\|Tu\| \leq 1$ if $\|u\| \leq \delta$. Given any x, we may write $x = cu$ where $\|u\| = \delta$ and $c = \delta^{-1}\|x\|$. Then

$$\|Tx\| = c\|Tu\| \leq c = \delta^{-1}\|x\|.$$

Thus we may take M to be δ^{-1}. □

If T is a continuous linear operator on X into Y, we define the *norm* $\|T\|$ of T by

$$\|T\| = \sup_{\|x\| \leq 1} \|Tx\|. \tag{1-2}$$

The reason for this terminology and notation will appear later. The quantity in (1-2) is clearly finite, by Theorem 1.1. Conversely, if the supremum in (1-2) is finite, then T must be continuous. To prove this, observe that

$$\|Tu\| \leq \sup_{\|x\| \leq 1} \|Tx\| = \|T\|, \qquad \|u\| \leq 1.$$

Hence, if $x \neq 0$ and $u = (1/\|x\|)x$, then $\|Tx\| = \|Tu\| \cdot \|x\| \leq \|T\|\|x\|$. Since the inequality

$$\|Tx\| \leq \|T\|\|x\| \tag{1-3}$$

also holds for $x = 0$, we see from Theorem 1.1 that T is continuous. The following formulas for the norm of T are useful:

$$\|T\| = \sup_{\|x\|=1} \|Tx\| = \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|}. \tag{1-4}$$

(Here we assume that X contains some nonzero elements.)

When T is continuous, (1-3) shows that a set K lying in a ball $\|x\| \leq r$ is carried into the set $T(K)$ lying in the ball $\|y\| \leq \|T\| \cdot r$ in Y. Hence $T(K)$ is bounded if K is bounded. Conversely, if T is linear and has the property that $T(K)$ is bounded whenever K is bounded, then T is continuous. We leave the proof of this assertion to the reader.

Theorem 1.2. *Suppose T is a linear operator on X into Y, where X and Y are normed linear spaces. Then the inverse T^{-1} exists and is continuous on its domain of definition if and only if there exists a constant $m > 0$ such that*

$$m\|x\| \leq \|Tx\| \tag{1-5}$$

for every x in X.

Proof. If (1-5) holds and $Tx = 0$, it follows that $x = 0$. Then T^{-1} exists, by Theorem I.3.1. Now $y = Tx$ is equivalent to $x = T^{-1}y$. Hence (1-5) is equivalent to $m\|T^{-1}y\| \leq \|y\|$, or $\|T^{-1}y\| \leq (1/m)\|y\|$, for all y in the range of T, which is the domain of T^{-1}. This implies that T^{-1} is continuous, by Theorem 1.1. We leave the converse proof to the reader. □

Theorem 1.3. *If X and Y are normed linear spaces, they are topologically isomorphic if and only if there exist a linear operator T with domain X and range Y, and positive constants m, M such that*

$$m\|x\| \leq \|Tx\| \leq M\|x\| \tag{1-6}$$

for every x in X.

Proof. This theorem is a direct corollary of Theorems 1.1 and 1.2. □

Theorem 1.4. *Let X be a linear space, and suppose two norms $\|x\|_1$ and $\|x\|_2$ are defined on X. These norms define the same topology on X if and only if there exist positive constants m, M such that*

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1 \tag{1-7}$$

for every x in X.

Proof. Let X_i be the normed linear space that X becomes with the norm $\|x\|_i$, $i = 1, 2$. Let $Tx = x$, and consider T as an operator with domain X_1 and range X_2. Condition (1-7) is precisely the condition that both T and T^{-1} are continuous. They are both continuous if and only if the open sets in X_1 are the same as the open sets in X_2. The conclusion now follows. □

II.2 EXAMPLES OF NORMED LINEAR SPACES

In this section we define and establish standard notations for a number of spaces that will be referred to at various places throughout the book. In defining the norms in these various spaces, we frequently leave it for the reader to verify that the norm, as defined, actually has the properties required of a norm.

In many of our examples, the elements of the spaces are functions defined on some set T. The values of the functions may be either real or complex; we get a real or complex space according to whether the values of the functions are real or complex. It is always to be understood that, if x_1, x_2 and x are functions defined on T, the functions x_1+x_2 and αx are defined by

$$(x_1+x_2)(t)=x_1(t)+x_2(t), \qquad (\alpha x)(t)=\alpha x(t), \qquad t \in T.$$

Example 1. The spaces $\boldsymbol{R}^n$ and $\boldsymbol{C}^n$ (defined in Examples 1 and 2, § I.2) can be made into normed linear spaces in a variety of ways. Let us consider $\boldsymbol{C}^n$. If $x=(\xi_1, \ldots, \xi_n) \in \boldsymbol{C}^n$ and $p \geqslant 1$, let us define

(2-1) $$\|x\|=(|\xi_1|^p+\cdots+|\xi_n|^p)^{1/p}.$$

The triangularity property $\|x_1+x_2\| \leqslant \|x_1\|+\|x_2\|$ is the same as Minkowski's inequality for finite sums, namely,

$$\left(\sum_{i=1}^n |\xi_i+\eta_i|^p\right)^{1/p} \leqslant \left(\sum_{i=1}^n |\xi_i|^p\right)^{1/p}+\left(\sum_{i=1}^n |\eta_i|^p\right)^{1/p}.$$

For reference to Minkowski's inequality see the Introduction. When $\boldsymbol{C}^n$ is considered as a normed space with the norm (2-1), we denote the space by $\ell^p(n)$. We can make the real space $\boldsymbol{R}^n$ into a normed space in the same way. We shall use $\ell^p(n)$ for both the real and the complex space with norm (2-1); whether the real or complex space is under discussion at a given time will usually be clear from the context, or we shall make a specific statement if necessary.

We can also define a norm in $\boldsymbol{R}^n$ and $\boldsymbol{C}^n$ by the formula

$$\|x\|=\max\{|\xi_1|, \ldots, |\xi_n|\}.$$

The notation for the space with this norm is $\ell^\infty(n)$. This notation is natural, since

(2-2) $$\max_{1 \leqslant k \leqslant n} |\xi_k|=\lim_{p \to \infty}(|\xi_1|^p+\cdots+|\xi_n|^p)^{1/p}.$$

Example 2. We also have spaces analogous to $\ell^p(n)$ and $\ell^\infty(n)$ for *infinite* sequences. The space ℓ^p, where $p \geq 1$, is defined to consist of all sequences $x = \{\xi_n\}$ such that $\Sigma_{n=1}^{\infty} |\xi_n|^p < \infty$. The norm in ℓ^p is defined by

$$\|x\| = \left(\sum_{n=1}^{\infty} |\xi_n|^p \right)^{1/p}. \tag{2-3}$$

Minkowski's inequality for infinite sums is used in showing that the requisite properties of a norm are fulfilled.

The space ℓ^∞ is defined to consist of the *bounded* sequences $x = \{\xi_n\}$, with the norm

$$\|x\| = \sup_n |\xi_n|.$$

For any $x \in \ell^\infty$ we have

$$\sup_n |\xi_n| = \lim_{n\to\infty} [\lim_{p\to\infty} (|\xi_1|^p + \cdots + |\xi_n|^p)^{1/p}]. \tag{2-4}$$

Some interest attaches to various subspaces of ℓ^∞. We mention in particular the space (c) of all *convergent* sequences $x = \{\xi_n\}$, and the space (c_0) of all sequences converging to zero.

Sometimes it is convenient to denote the norm in ℓ^p by $\|x\|_p$ and that in ℓ^∞ by $\|x\|_\infty$. It is worth noticing that the elements of ℓ^p form a subclass of the elements of ℓ^q if $1 \leq p < q \leq \infty$ and that, then, $\|x\|_q \leq \|x\|_p$ if $x \in \ell^p$ (see Jensen's inequality in the Introduction).

The space ℓ^p is separable if $1 \leq p < \infty$. To show that this is so, let us first introduce some terminology. A point $x \in \ell^p$ will be called rational if $x = \{\xi_n\}$ and each ξ_n is rational. In the case of the complex field, a number ξ_n is called rational if its real and imaginary parts are rational. A point $x = \{\xi_n\}$ will be called of *finite type* if the set of n for which $\xi_n \neq 0$ is finite. The set of all rational points of finite type is readily seen to be countable. It is also everywhere dense in ℓ^p. For, suppose $\varepsilon > 0$ and $x \in \ell^p$. We choose N so that $\Sigma_{n=N+1}^{\infty} |\xi_n|^p < \varepsilon^p/2$, and then we choose a rational point of finite type, say, $y = \{\eta_n\}$ such that $\eta_n = 0$ if $n > N$ and $|\xi_k - \eta_k| < (2N)^{-1/p}\varepsilon$ if $k = 1, \ldots, N$. Then

$$\|x - y\|^p = \sum_{n=1}^{N} |\xi_n - \eta_n|^p + \sum_{n=N+1}^{\infty} |\xi_n|^p < N\frac{\varepsilon^p}{2N} + \frac{\varepsilon^p}{2} = \varepsilon^p,$$

so that $\|x - y\| < \varepsilon$. Thus ℓ^p is separable.

The space ℓ^∞ is not separable, however. For, if $\{x_n\}$ is any countable set in ℓ^∞, with $x_n = (\xi_1^{(n)}, \xi_2^{(n)}, \ldots)$, let $x = \{\xi_n\}$ be the element of ℓ^∞, defined by $\xi_k = \xi_k^{(k)} + 1$ if $|\xi_k^{(k)}| \leq 1$ and $\xi_k = 0$ if $|\xi_k^{(k)}| > 1$. Then the kth component of $x - x_k$ is $\xi_k - \xi_k^{(k)}$, and $|\xi_k - \xi_k^{(k)}| \geq 1$, so that $\|x - x_k\| \geq 1$. Thus the set $\{x_k\}$ cannot be dense in ℓ^∞, and ℓ^∞ is not separable.

Example 3. Let T be any nonempty set, and let $B(T)$ be the class of all bounded functions x defined on T, with either real or complex values. Then $B(T)$ becomes a normed linear space if we define

$$\|x\| = \sup_{t \in T} |x(t)|.$$

The scalar field for $B(T)$ will be the same as the field in which the function values are required to lie.

Observe that ℓ^∞ is the special case of $B(T)$ in which T is the set of positive integers. In the special case where T is an interval $[a, b]$ of the real axis, we shall denote $B(T)$ by $B[a, b]$.

Example 4. Suppose T is a topological space. Let $C(T)$ denote the class of bounded and continuous functions x defined on T. Then $C(T)$ is a linear manifold in $B(T)$, and we consider $C(T)$ as a normed linear space on its own account, with the norm as defined in $B(T)$. If T is compact, the requirement that the values $x(t)$ be bounded is fulfilled automatically as a consequence of the assumption that x is continuous on T. This is because the continuous image of a compact set is compact, and a compact set of real or complex numbers is bounded. Moreover, if $x \in C(T)$, there is some point $t_m \in T$ such that

$$|x(t_m)| = \max_{t \in T} |x(t)| = \|x\|.$$

This is because $|x(t)|$ is continuous on T and a compact set of real numbers is closed as well as bounded.

If T is a finite closed interval on the real axis, we shall usually denote $C(T)$ by $C[a, b]$ (see Example 3, § I.2).

Example 5. The class $\mathscr{L}^p = \mathscr{L}^p(-\infty, \infty)$ and the corresponding linear space L^p were defined in Example 7, § I.2. Here we assume $1 \leqslant p < \infty$. If $x \in \mathscr{L}^p$, we write

$$\|x\|_p = \left(\int_{-\infty}^{\infty} |x(t)|^p \, dt \right)^{1/p}. \tag{2-5}$$

Evidently $\|x\|_p = 0$ is equivalent to $x(t) = 0$ a.e. and $\|x\|_p = \|y\|_p$ if $x \overset{0}{=} y$. Minkowski's inequality for integrals (see the Introduction) states that

$$\|x + y\|_p \leqslant \|x\|_p + \|y\|_p$$

if $x, y \in \mathscr{L}^p$. Hence, if $[x] \in L^p$ and we define $\|[x]\| = \|x\|_p$, L^p becomes a normed linear space. As stated in § I.2, we shall usually write x in place of $[x]$. Also, we shall write $\|x\|$ instead of $\|x\|_p$ if the situation is such that the use of the index is not essential.

Instead of functions defined on $(-\infty, \infty)$, we can consider any measurable set E in n-dimensional Euclidean space. The classes $\mathscr{L}^p$ and L^p are then defined with reference to Lebesgue integrals over E, and we indicate the dependence on E, if necessary, by writing $\mathscr{L}^p(E)$ and $L^p(E)$. It is also possible to consider the case in which E is replaced by a more general measure space and the integrals are defined in terms of a more general measure.

Example 6. Let (a, b) be an interval, finite or infinite, of the real axis. A measurable function x defined on (a, b), with real or complex values, is called *essentially* bounded if there is some $A \geqslant 0$ such that the set $\{t : |x(t)| > A\}$ has measure 0 [i.e., such that $|x(t)| \leqslant A$ a.e. on (a, b)]. If such a constant exists, there is a least one; we call this smallest possible A the *essential least upper bound* of x and denote it by $\sup^0 |x(t)|$. The notation $\sup^0$ distinguishes $\sup^0 |x(t)|$ from the ordinary least upper bound $\sup |x(t)|$. Of course, it can happen that $\sup^0 |x(t)| < \sup |x(t)|$, and it can even happen that $\sup^0 |x(t)| < \infty$ but $\sup |x(t)| = \infty$. Another characterization of $\sup^0 |x(t)|$ is the following. It is the largest number B such that if $\varepsilon > 0$, the set $\{t : |x(t)| > B - \varepsilon\}$ has positive measure.

Let $\mathscr{L}^\infty$ denote the class of all measurable and essentially bounded functions x defined on (a, b). If $x, y \in \mathscr{L}^\infty$, we write $x =^0 y$ if $x(t) = y(t)$ a.e. We then define the linear space L^∞ in relation to $\mathscr{L}^\infty$ just as we defined L^p in relation to $\mathscr{L}^p$ (see Example 7, § I.2). If $x =^0 y$, it is clear that $\sup^0 |x(t)| = \sup^0 |y(t)|$. Hence if we define

$$\|[x]\| = \sup{}^0 |x(t)|,$$

L^∞ becomes a normed linear space.

If (a, b) is a finite interval and $x \in \mathscr{L}^\infty$, then $x \in \mathscr{L}^p$ for every $p > 0$, and it can be proved that

$$\lim_{p \to \infty} \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} = \sup{}^0 |x(t)|. \tag{2-6}$$

In fact, if $A < \sup^0 |x(t)|$ and E is the set where $|x(t)| > A$, we have $m(E) > 0$, where $m(E)$ is the Lebesgue measure of E, and

$$A[m(E)]^{1/p} \leqslant \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} \leqslant (b - a)^{1/p} \sup{}^0 |x(t)|.$$

Letting $p \to \infty$, we see that

$$A \leqslant \underline{\lim} \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} \leqslant \overline{\lim} \left(\int_a^b |x(t)|^p \, dt \right)^{1/p} \leqslant \sup{}^0 |x(t)|.$$

Since A can be as near $\sup^0 |x(t)|$ as we please, this justifies (2-6). If the interval (a, b) is infinite, (2-6) is replaced by

$$\lim_{n\to\infty}\lim_{p\to\infty}\left(\int_{E_n}|x(t)|^p\,dt\right)^{1/p}=\sup{}^0|x(t)|, \tag{2-7}$$

where E_n is a sequence of finite intervals such that $E_1 \subset E_2 \subset \cdots$, each E_n lies in (a, b), and $\bigcup_n E_n = (a, b)$.

Example 7. The linear space $BV[a, b]$ of functions of bounded variation on $[a, b]$ was defined in Example 8, § I.2. If $x \in BV[a, b]$ and $V(x)$ denotes the total variation of $x(t)$ for $a \leq t \leq b$, we can define a norm by

$$\|x\| = |x(a)| + V(x). \tag{2-8}$$

Many interesting spaces can be formed from classes of analytic functions of a complex variable. Let $\mathfrak{A}$ denote the class of all functions $f(z)$ that are defined and analytic in the unit disk $|z|<1$ of the complex plane. This class is a complex linear space. We shall in the next two examples describe some subspaces of $\mathfrak{A}$ that become normed linear spaces with appropriately defined norms.

Example 8. Suppose $0<p<\infty$. For any $f \in \mathfrak{A}$ and $0 \leq r < 1$ let

$$\mathfrak{M}_p[f; r] = \left(\frac{1}{2\pi}\int_0^{2\pi}|f(re^{i\theta})|^p\,d\theta\right)^{1/p}. \tag{2-9}$$

The class H^p is by definition composed of those $f \in \mathfrak{A}$ such that $\sup_{0\leq r<1} \mathfrak{M}_p[f; r] < \infty$. If $1 \leq p$, it is clear by Minkowski's inequality that $f+g$ is in H^p if f and g are in H^p. If $0<p<1$, this same conclusion follows from the following inequality {Hardy, Littlewood, and Pólya, [1], formula (6.13-6)}:

$$\int_0^{2\pi}|f(re^{i\theta})+g(re^{i\theta})|^p\,d\theta \leq \int_0^{2\pi}|f(re^{i\theta})|^p\,d\theta + \int_0^{2\pi}|g(re^{i\theta})|^p\,d\theta. \tag{2-10}$$

It is thus seen that H^p is a linear space. If $1 \leq p$, we define

$$\|f\| = \sup_{0\leq r<1} \mathfrak{M}_p[f; r]. \tag{2-11}$$

This has the properties of a norm; thus H^p is a normed linear space if $1 \leq p < \infty$. If $0<p<1$, (2-11) does not define a norm, for the triangularity condition $\|f+g\| \leq \|f\| + \|g\|$ is not always satisfied.

When $1 \leq p < \infty$, it can be shown that H^p is in isometric correspondence with a subspace of $L^p(0, 2\pi)$. This follows from the following facts: If $f \in H^p$, then $\lim_{r\to 1} f(re^{i\theta})$ exists for almost all values of θ, thus defining a function which we denote by $f(e^{i\theta})$. Moreover, this latter function belongs to $\mathscr{L}^p(0, 2\pi)$, and

$$\|f\|_{H^p} = \left(\frac{1}{2\pi}\int_0^{2\pi} |f(e^{i\theta})|^p \, d\theta\right)^{1/p}.$$

For these facts about H^p, we refer the reader to P. Duren [1] and F. Riesz [4].

The space H^∞ consists of those $f \in \mathfrak{A}$ such that the values of $f(z)$ are bounded. We define a norm on H^∞ by

$$\|f\| = \sup_{|z|<1} |f(z)|.$$

That H^∞ is a natural notation for this space is seen from the fact that, if $f \in \mathfrak{A}$ and $0 \leq r < 1$, then

$$\max_{|z|=r} |f(z)| = \lim_{p\to\infty} \mathfrak{M}_p[f; r],$$

while, if $f \in H^\infty$, then

$$\|f\| = \sup_{0\leq r<1} \{\max_{|z|=r} |f(z)|\}.$$

Example 9. Let Δ denote the closed unit disc $|z| \leq 1$, and let $A(\Delta)$ denote the class of functions that are defined (single valued) and continuous on Δ and analytic on the interior of Δ. Evidently $A(\Delta)$ is a linear subclass of H^∞ and is a normed linear space with the norm defined as in H^∞. Because of the maximum modulus theorem and the fact that f is continuous, it is clear that for $f \in A(\Delta)$

$$\|f\| = \max_{|z|=1} |f(z)|. \tag{2-12}$$

The space $A(\Delta)$ is separable. In fact, the polynomial functions in z having complex rational coefficients form a set everywhere dense in $A(\Delta)$. To see this, observe that if $f \in A(\Delta)$, then f can be approximated uniformly on Δ by the functions $f_r(z) = f(z/r)$, $r > 1$. But f_r is analytic on the disc $|z| < r$ and hence has a power series that converges uniformly on Δ; the partial sums are polynomials that can be approximated uniformly by polynomials whose coefficients have rational real and imaginary parts.

The spaces H^p with $1 \leq p \leq \infty$ and the space $A(\Delta)$ have been investigated in connection with more general studies of Banach spaces composed of functions that are analytic on the interior of Δ. See Hoffman [1, Chapter 3] and Taylor [4].

Example 10. Let X be a normed linear space, and let K be the associated scalar field. This field is itself a normed linear space, with the absolute value $|\alpha|$ as the norm of α (see Example 1). Each linear functional $x' \in X^f$ may thus be viewed as a linear operator from X into K. Referring to Theorem 1.1, we see that a linear functional $x' \in X^f$ is continuous if and only if there is a constant M such that $|x'(x)| \leq M\|x\|$, for all $x \in X$. When x' is continuous, the smallest such M is denoted by $\|x'\|$. Equivalently,

$$\|x'\| = \sup_{\|x\| \leq 1} |x'(x)|. \tag{2-13}$$

The set of continuous linear functionals on X is a linear manifold in X^f and is denoted by X'. A norm on X' is given by (2-13). The vector space X' normed in this way is called the *normed conjugate* of X.

Some authors call X' the *dual* of X, others call it the *adjoint* of X. Sometimes it is denoted by X^*.

The normed conjugate of a normed linear space will be studied in detail in Chapter III.

PROBLEMS

1. The real space $C[a, b]$ is separable if $[a, b]$ is a finite closed interval of the real axis. A proof may be constructed using the Stone–Weierstrass theorem.
2. The space $L^\infty(a, b)$ of Example 6 is not separable. To prove this, consider the collection of characteristic functions of subintervals (a, s) where $a < s < b$.
3. The space $L^p(a, b)$ is separable for $1 \leq p < \infty$. (A countable dense set is the collection of all simple functions of the form $\varphi = \sum_{i=1}^n c_i f_i$, where each f_i is the characteristic function of a subinterval of (a, b) with rational endpoints and each c_i is a rational number or a complex number whose real and imaginary parts are rational.) Use this fact to deduce that H^p is separable, for $1 \leq p < \infty$.

II.3 FINITE-DIMENSIONAL NORMED LINEAR SPACES

We saw in Theorem I.4.1 that two linear spaces of the same finite dimension over the same scalar field are isomorphic. Now we shall see that if each of the two spaces has a norm, the spaces are topologically isomorphic according to the definition of § 1.

Theorem 3.1. *Let X_1 and X_2 be two normed linear spaces of the same finite dimension n, with the same scalar field. Then X_1 and X_2 are topologically isomorphic.*

Proof. The case $n = 0$ is trivial, and we assume $n \geq 1$. It will suffice to prove that if X is an n-dimensional normed linear space, it is topologically isomorphic to $\ell^2(n)$, for the relation of topological isomorphism is transitive (as well as being reflexive and symmetric). Let $x_1, \ldots, x_n$ be a basis of X and define $T: \ell^2(n) \to X$ by $T(\xi_1, \ldots, \xi_n) = \xi_1 x_1 + \cdots + \xi_n x_n$. We know from § I.4 that T is an isomorphism of $\ell^2(n)$ onto X. Furthermore, T is continuous, since scalar multiplication and vector addition are continuous operations on X. It remains to show that T^{-1} is continuous. We first observe that the surface S of the open unit ball B in $\ell^2(n)$ is compact, by the Heine–Borel theorem. (The proof for *complex* Euclidean n-space is the same as the usual proof of the Heine–Borel theorem for real Euclidean n-space.) Then $T(S)$ must be compact (since T is continuous) and hence is closed in X. Since T is an isomorphism, $0 \notin T(S)$. Thus there is an open set U containing 0, which is disjoint from $T(S)$. Choose $\delta > 0$ such that, if $V = \{x : \|x\| < \delta\}$, then $V \subset U$. We claim that $V \subset T(B)$. If $x \notin T(B)$, then $x = Tz$ for some $z \in \ell^2(n)$ with $\|z\| \geq 1$. If x also belonged to V, then $\|z\|^{-1}x$ would belong to V. This is impossible, since $\|z\|^{-1}x = T(\|z\|^{-1}z) \in T(S)$ and $V \cap T(S) = \varnothing$. Thus $V \subset T(B)$, so that $T^{-1}(V) \subset B$ and $T^{-1}(\delta^{-1}V) \subset \delta^{-1}B$. But $\delta^{-1}V = \{x : \|x\| < 1\}$, and we see that T^{-1} maps the open unit ball in X into a bounded set in $\ell^2(n)$. Thus T^{-1} is continuous. □

It is clear from Theorem 1.3 that if X and Y are topologically isomorphic normed linear spaces and if one of them is complete (as a metric space), the other is also complete. Now $\ell^1(n)$ is evidently complete (as a consequence of the completeness of $\ell^1(1)$, the real or complex number field). Thus we have:

Theorem 3.2. *A finite-dimensional normed linear space is complete.*

As a corollary, we have:

Theorem 3.3. *If X is a normed linear space, any finite-dimensional subspace of X is necessarily closed.*

Another important result is the following:

Theorem 3.4. *If X is a finite-dimensional normed linear space, each closed and bounded set in X is compact.*

Proof. This proposition is true (by classical analysis) for the particular finite-dimensional space $\ell^2(n)$. It then follows, by virtue of Theorem 3.1, that the theorem is true for any finite-dimensional space X, for the properties of being bounded and closed are transferred from a set S to its image S_1 in $\ell^2(n)$ by the topological isomorphism, and the compactness is then carried back from S_1 to S. □

The converse of Theorem 3.4 is also true. Before proving the converse, we consider a general theorem due to F. Riesz, which is useful in many arguments.

Theorem 3.5 (Riesz's Lemma). *Suppose X is a normed linear space. Let X_0 be a subspace of X such that X_0 is closed and a proper subset of X. Then for each θ such that $0<\theta<1$ there exists a vector $x_\theta \in X$ such that $\|x_\theta\|=1$ and $\|x-x_\theta\| \geq \theta$ if $x \in X_0$.*

Proof. Select any $x_1 \in X \backslash X_0$, and let

$$d=\inf_{x \in X_0}\|x-x_1\|.$$

Since X_0 is closed, it follows that $d>0$. There exists $x_0 \in X_0$ such that $\|x_0-x_1\| \leq \theta^{-1} d$ (because $\theta^{-1} d>d$). Let $x_\theta=h(x_1-x_0)$, where $h=\|x_1-x_0\|^{-1}$. Then $\|x_\theta\|=1$. If $x \in X_0$, then $h^{-1} x+x_0 \in X_0$ also, and so

$$\|x-x_\theta\|=\|x-h x_1+h x_0\|=h\|(h^{-1} x+x_0)-x_1\| \geq h d.$$

But $hd=\|x_1-x_0\|^{-1} d \geq \theta$, by the way in which x_0 was chosen. Thus $\|x-x_\theta\| \geq \theta$ if $x \in X_0$. □

We can restate Riesz's lemma as follows: *If X_0 is a closed and proper subspace of X, there exist on the surface of the unit sphere in X points whose distance from X_0 is as near* 1 *as we please.* This is the most that can be said in general, however. It need not be true that there are points on the unit sphere whose distance from X_0 is *equal* to 1.

Example. Let X be that subspace of the real space $C[0,1]$ (see § 2, Example 4) consisting of all continuous functions x on $[0,1]$ such that $x(0)=0$. For X_0 we take the subspace of all $x \in X$ such that $\int_0^1 x(t)\,dt=0$. Now suppose that $x_1 \in X$, $\|x_1\|=1$, and $\|x_1-x\| \geq 1$ if $x \in X_0$. Corresponding to each $y \in X \backslash X_0$, we let

$$c=\frac{\int_0^1 x_1(t)\,dt}{\int_0^1 y(t)\,dt}.$$

Then $x_1-cy \in X_0$, and so $1 \leq\|x_1-(x_1-cy)\|=|c|\|y\|$, or

$$\left|\int_0^1 y(t)\,dt\right| \leq\left|\int_0^1 x_1(t)\,dt\right|\|y\|$$

for each $y \in X$. Now, we can make $|\int_0^1 y(t)\,dt|$ as close to 1 as we please while maintaining $\|y\|=1$ (e.g., let $y_n(t)=t^{1/n}$, and let $n \rightarrow \infty$). Thus we see that

$$1 \leq\left|\int_0^1 x_1(t)\,dt\right|.$$

But since $\|x_1\| = \max_{0\leq t\leq 1} |x_1(t)| = 1$ and $x_1(0) = 0$, the continuity of x_1 shows that we must have

$$\left|\int_0^1 x_1(t)\,dt\right| < 1,$$

and thus we have a contradiction. Therefore, with X_0 and X as given here, there is no point on the surface of the unit sphere in X at unit distance from X_0.

We now come to the converse of Theorem 3.4.

Theorem 3.6. *Let X be a normed linear space, and suppose the surface S of the unit sphere in X is compact. Then X is finite dimensional.*

Proof. Suppose that X is not finite dimensional. Choose $x_1 \in S$, and let X_1 be the subspace generated by x_1. Then X_1 is a proper subspace of X, and it is closed (by Theorem 3.3). Hence, by Riesz's lemma, there exists $x_2 \in S$ such that $\|x_2 - x_1\| \geq \frac{1}{2}$. Let X_2 be the (closed and proper) subspace of X generated by x_1, x_2; then there must exist $x_3 \in S$ such that $\|x_3 - x\| \geq \frac{1}{2}$ if $x \in X_2$. Proceeding by induction, we obtain an infinite sequence $\{x_n\}$ of elements of S such that $\|x_n - x_m\| \geq \frac{1}{2}$ if $m \neq n$. This sequence can have no convergent subsequence. This is impossible since S is compact. Thus X must be finite dimensional.

Another proof, using the definition of compactness directly, may be given as follows: The family of all open spheres of radius $\frac{1}{2}$ with centers on S is an open covering of S. Since S is compact, there must exist a finite number of points $x_1, \ldots, x_n$ on S such that S is covered by the set of open spheres of radius $\frac{1}{2}$ with centers at $x_1, \ldots, x_n$. Let M be the finite-dimensional, and therefore closed, subspace of X generated by $x_1, \ldots, x_n$. Then M must be all of X. For, if not, by Riesz's lemma there exists a point $x_0 \in S$ whose distance from M is greater than $\frac{1}{2}$, and this point x_0 cannot be in any of the spheres that cover S. Since $M = X$, X is finite dimensional. □

PROBLEMS

1. If T is a linear operator from X into Y, where X and Y are normed linear spaces and X is finite dimensional, then T is necessarily continuous.
2. Let X be a normed linear space, and let X_0 be a proper subspace of X of finite dimension. Then there exists an x_1 in X such that the distance from x_1 to X_0 is exactly 1.
3. a. Sketch the unit balls in the real spaces $\ell^1(2)$, $\ell^2(2)$, and $\ell^\infty(2)$.
   b. Let T be the identity mapping on the set $\boldsymbol{R}^2$, considered as a linear

operator from $\ell^2(2)$ onto $\ell^1(2)$. Find the largest m and the smallest M such that

$$m\|(x, y)\|_2 \leq \|T(x, y)\|_1 \leq M\|(x, y)\|_2$$

for all $(x, y) \in \ell^2(2)$. (The norms in $\ell^1(2)$ and $\ell^2(2)$ are indicated by subscripts.)

II.4 BANACH SPACES

If a normed linear space is complete, it is called a *Banach space*. Stefan Banach gave axioms for a complete normed linear space in his doctoral thesis (1920). In the years after the publication (in 1932) of Banach's path-breaking book about methods in analysis based on such spaces, the name "Banach space" came into standard usage. While many propositions about normed linear do not require the hypothesis of completeness, a number of theorems of critical importance do depend on completeness. In particular, there are some important theorems in the theory of linear operators that make use of the category theorem of Baire and this is made possible by the assumption of completeness for the normed linear spaces under consideration. In some work it is necessary to construct elements of a normed linear space by means of infinite series or integrals, and completeness is then needed to ensure the existence of limits.

All of the spaces described in § 2 are Banach spaces. We shall discuss this later in this section. For an example of an incomplete normed linear space, consider the subspace X of ℓ^p $(1 \leq p < \infty)$ consisting of the sequences $x = \{\xi_n\}$ having only finitely many nonzero terms. This space is not complete under the ℓ^p norm. For, if $x = \{\xi_n\}$ is an element of ℓ^p that is not in X and if $x_n = (\xi_1, \xi_2, \ldots, \xi_n, 0, 0, \ldots)$, then it is clear that $x_n \to x$. Since the limit of a convergent sequence is unique and since $x \notin X$, we see that $\{x_n\}$ is a Cauchy sequence in X that has no limit in X.

An incomplete normed linear space X may be enlarged to form a Banach space $\hat{X}$ in which X is dense. The space $\hat{X}$, called the Banach space completion of X, is essentially unique, in the sense that any other Banach space containing X as a dense subspace must be isometrically isomorphic to $\hat{X}$ (problem 4). One way to construct $\hat{X}$ is first to obtain the completion $\hat{X}$ of X as a metric space. (This well-known step is described in § 3-8 of Taylor [5].) It is then possible in a straightforward way to extend from X to $\hat{X}$ the definitions of vector addition, scalar multiplication, and the norm of a vector in such a way that $\hat{X}$ becomes a Banach space. In § III.4 we shall describe another very neat and natural method for obtaining $\hat{X}$.

There is a particular kind of extension theorem for linear operators that we shall need later but which is convenient to discuss here.

Theorem 4.1. *Let X and Y be normed linear spaces, and let T be a continuous linear operator on X into Y. Then there is a uniquely determined continuous linear operator $\hat{T}$ on $\hat{X}$ into $\hat{Y}$ such that $\hat{T}x = Tx$ if $x \in X$. The relation $\|\hat{T}\| = \|T\|$ is valid.*

Proof. To define $\hat{T}$ we suppose $\hat{x} \in \hat{X}$ and select a sequence $\{x_n\}$ from X such that $x_n \to \hat{x}$. Then $\{x_n\}$ is a Cauchy sequence and

$$\|Tx_n - Tx_m\| = \|T(x_n - x_m)\| \leq \|T\| \|x_n - x_m\|,$$

so that $\{Tx_n\}$ is a Cauchy sequence in Y. Consequently, $Tx_n \to \hat{y}$, where $\hat{y}$ is some element of $\hat{Y}$. It is readily proved that $\hat{y}$ depends only on $\hat{x}$ and T, not on the particular sequence $\{x_n\}$. We define $\hat{T}\hat{x} = \hat{y}$. It is a simple matter to verify that $\hat{T}x = Tx$ if $x \in X$ and that $\hat{T}$ is linear. We see that $\|x_n\| \to \|\hat{x}\|$, $\|Tx_n\| \leq \|T\| \|x_n\|$, and hence $\|\hat{T}\hat{x}\| \leq \|T\| \|\hat{x}\|$. Thus $\hat{T}$ is continuous and $\|\hat{T}\| \leq \|T\|$ (see § 1). On the other hand, if $x \in X$, we have $\|Tx\| = \|\hat{T}x\| \leq \|\hat{T}\| \|x\|$, so that $\|T\| \leq \|\hat{T}\|$. Hence $\|\hat{T}\| = \|T\|$. The uniqueness assertion about $\hat{T}$ is easily justified by using the fact that X is dense in $\hat{X}$. □

Theorem 4.2. *If X is a Banach space and X_0 is a linear manifold in X, then X_0, considered as a normed linear space by itself, is a Banach space if and only if X_0 is closed in X.*

Proof. A subset of a complete metric space is itself complete if and only if it is closed. □

It is useful to note that a Cauchy sequence $\{x_n\}$ in a metric space is bounded, irrespective of whether or not the space is complete. Hence, if $\{x_n\}$ is a Cauchy sequence in a normed linear space, the sequence of norms $\|x_n\|$ is a bounded set of real numbers.

Sequences in a normed linear space X can be used to discuss infinite series. Let $\Sigma_1^\infty x_n$ be a formal series of elements of X. The "partial sums," $s_n = x_1 + \cdots + x_n$, are elements of X. We say the series $\Sigma_1^\infty x_n$ is *convergent* if there is an $x \in X$ such that the sequence $\{s_n\}$ of partial sums converges to x, and we write $x = \Sigma_1^\infty x_n$. We say the series $\Sigma_1^\infty x_n$ is *absolutely convergent* if $\Sigma_1^\infty \|x_n\|$ is convergent.

Theorem 4.3. *A normed linear space X is a Banach space if and only if every absolutely convergent series in X is convergent.*

Proof. Suppose X is complete, and let $\Sigma_1^\infty x_n$ be absolutely convergent. Let $s_n = x_1 + \cdots + x_n$. Then $\{s_n\}$ is a Cauchy sequence, for, if $m < n$, $\|s_n - s_m\| \leq \|x_{m+1}\| + \cdots + \|x_n\|$, and we can employ the Cauchy criterion on the series $\Sigma \|x_n\|$. Since X is complete, the sequence $\{s_n\}$ is convergent.

Now suppose that every absolutely convergent series in X is convergent. Let $\{x_n\}$ be a Cauchy sequence in X. By induction, it is possible to select a subsequence $\{u_k\}$ of $\{x_n\}$ such that $\|u_{k+1}-u_k\|<2^{-k}$, $k=1,2,\ldots$. Then $\Sigma_1^\infty \|u_{k+1}-u_k\|<\infty$. By hypothesis, $\Sigma_1^\infty (u_{k+1}-u_k)$ must converge. This is a telescoping series whose mth partial sum is $u_{m+1}-u_1$; hence $\{u_k\}$ must converge to some $x\in X$. Since the original sequence $\{x_n\}$ is a Cauchy sequence, it follows that $\{x_n\}$ also converges to x. Thus X is complete. □

We turn to a discussion of completeness for examples of normed linear spaces given in § 2. We already know (Theorem 3.2) that every finite-dimensional normed linear space is complete.

The space ℓ^p, where $p\geq 1$, is complete. Let $\{x_n\}$ be a Cauchy sequence in ℓ^p, with $x_n=(\xi_1^{(n)},\xi_2^{(n)},\ldots)$. For each fixed k, $\{\xi_k^{(n)}\}$ is a Cauchy sequence, because

$$|\xi_k^{(n)}-\xi_k^{(m)}|\leq\left(\sum_{i=1}^{\infty}|\xi_i^{(n)}-\xi_i^{(m)}|^p\right)^{1/p}=\|x_n-x_m\|.$$

Let $\xi_k=\lim\limits_{n\to\infty}\xi_k^{(n)}$. We shall first prove that the sequence $\{\xi_k\}$ is an element of ℓ^p. We know that $\|x_n\|$ is bounded, say $\|x_n\|\leq M$. Now, for any k,

$$\left(\sum_{i=1}^{k}|\xi_i^{(n)}|^p\right)^{1/p}\leq\|x_n\|\leq M.$$

Letting $n\to\infty$, we obtain

$$\left(\sum_{i=1}^{k}|\xi_i|^p\right)^{1/p}\leq M.$$

Since k is arbitrary, this shows that $\{\xi_k\}\in\ell^p$ and that its norm does not exceed M. Let $x=\{\xi_k\}$. It remains to prove that $\|x_n-x\|\to 0$. Suppose $\varepsilon>0$. Then there exists an integer N such that $\|x_n-x_m\|<\varepsilon$ if $N\leq m$ and $N\leq n$. Therefore, for any k,

$$\left(\sum_{i=1}^{k}|\xi_i^{(n)}-\xi_i^{(m)}|^p\right)^{1/p}\leq\|x_n-x_m\|<\varepsilon$$

if $N\leq m$ and $N\leq n$. Keeping k and n fixed, we let $m\to\infty$. This gives

$$\left(\sum_{i=1}^{k}|\xi_i^{(n)}-\xi_i|^p\right)^{1/p}\leq\varepsilon$$

if $N\leq n$. Since this is true for all k, we can let $k\to\infty$, and we obtain the result that $\|x_n-x\|\leq\varepsilon$ if $N\leq n$. This finishes the proof that ℓ^p is complete.

The form of the foregoing argument is such that it can be adapted to proving the completeness of a number of spaces. To avoid much repetition of essentially the same argument, we shall formulate a general theorem that embodies the principle of the argument.

Theorem 4.4. *Let Y be a linear space, and let $\mathscr{F}$ be a certain family of real-valued functions defined on Y. Suppose that the class X of $x \in Y$ such that $\sup_{f \in \mathscr{F}} f(x) < \infty$ is a linear manifold in Y and the function $\|x\|$ on X defined by*

$$\|x\| = \sup_{f \in \mathscr{F}} f(x)$$

is a norm on X. Suppose that to each Cauchy sequence $\{x_n\}$ (relative to this norm) in X is associated a certain element $y \in Y$ such that

(*a*) $\lim_{n \to \infty} f(x_n) = f(y)$, *for each $f \in \mathscr{F}$, and*

(*b*) $\lim_{m \to \infty} f(x_m - x_n) = f(y - x_n)$, *for each n and each $f \in \mathscr{F}$.*

Then X is a Banach space (i.e., X is complete).

Before proving the theorem it may be helpful to see how the hypotheses of the theorem can be satisfied in the case when X is the space ℓ^p. Let Y be the linear space of *all* sequences $y = \{\eta_i\}$, where $\eta_1, \eta_2, \ldots$ are scalars. Let $\mathscr{F}$ be the countable family of functions $f_1, f_2, \ldots,$ with f_k defined by

$$f_k(y) = \left(\sum_{i=1}^{k} |\eta_i|^p \right)^{1/p}, \qquad y \in Y.$$

Then if $\{x_n\}$ is a Cauchy sequence in ℓ^p, where $x_n = \{\eta_i^{(n)}\}$, let $\eta_i = \lim_{n \to \infty} \eta_i^{(n)}$ (we observed earlier that these limits exist), and let the associated $y \in Y$ be $\{\eta_i\}$. Then Theorem 4.4 could be applied with $X = \ell^p$. Now observe how the proof of Theorem 4.4 follows exactly the same lines as the proof that ℓ^p is complete.

Proof of Theorem 4.4. Suppose that $\{x_n\}$ is a Cauchy sequence in X, and let y be the element of Y associated with $\{x_n\}$. Then there is some M such that $\|x_n\| \leqslant M$ for all n. Also, $f(x_n) \leqslant \|x_n\| \leqslant M$ for each $f \in \mathscr{F}$, by definition of the norm. Hence $f(y) \leqslant M$, by property (a). Since this is true for all $f \in \mathscr{F}$, we must have $y \in X$. Now, if $\varepsilon > 0$, there exists an integer N such that $\|x_m - x_n\| < \varepsilon$ if $m, n \geqslant N$. Therefore, for each $f \in \mathscr{F}$, $f(x_m - x_n) < \varepsilon$ if $m, n \geqslant N$. By property (b), $f(y - x_n) \leqslant \varepsilon$. This is true for each $f \in \mathscr{F}$ and each $n \geqslant N$. Therefore $\|y - x_n\| \leqslant \varepsilon$ if $n \geqslant N$, proving that $\{x_n\}$ converges to y. □

Hints for demonstrating the completeness of some of the spaces described in § 2 are given in the problems below. Proofs of the completeness of the spaces L^p and L^∞ of Examples 5 and 6 are usually given in a course on Lebesgue integration. (See Taylor [5, pages 270–278].) The space in Example 10 is considered in the following theorem.

Theorem 4.5. *Let X be a normed linear space. Then the normed conjugate X' of X is complete.*

Proof. Let $\{x'_n\}$ be a Cauchy sequence in X'. Then, given $\varepsilon > 0$, there exists a positive integer N such that $\|x'_m - x'_n\| < \varepsilon$ whenever $m, n \geqslant N$. Consequently, for each $x \in X$,

$$|x'_m(x) - x'_n(x)| \leqslant \|x'_m - x'_n\| \|x\| < \varepsilon \|x\|, \tag{4-1}$$

when $m, n \geqslant N$. This shows that $\{x'_n(x)\}$ is a Cauchy sequence for each $x \in X$. Since the scalar field ($\boldsymbol{R}$ or $\boldsymbol{C}$) is complete, $\{x'_n(x)\}$ converges to a limit depending on x, which we denote by $x'(x)$, thus defining a function x' on X. It is easily verified that x' is a linear function; we omit the details. Letting m become infinite in (4-1), we see that

$$|x'(x) - x'_n(x)| \leqslant \varepsilon \|x\|, \qquad n \geqslant N. \tag{4-2}$$

From this we have

$$|x'(x)| \leqslant |x'(x) - x_N(x)| + |x'_N(x)| \leqslant (\varepsilon + \|x'_N\|)\|x\|,$$

implying that $x' \in X'$. It also follows from (4-2) that $\|x' - x'_n\| \leqslant \varepsilon$ when $n \geqslant N$; therefore $\{x'_n\}$ converges to x' in the norm topology of X'. Consequently, X' is complete. □

PROBLEMS

1. Theorem 4.4 may be used to prove that a number of spaces are complete.
   a. *$B(T)$ is complete* (see Example 3, § 2). Let Y be the linear space of *all* scalar-valued functions y defined on T. Let $\mathscr{F}$ be the family of functions f_t, $t \in T$, defined by $f_t(y) = |y(t)|$.
   b. *$C(T)$ is complete* (see Example 4, § 2). Use Theorem 4.2.
   c. *$BV[a, b]$ is complete* (see Example 7, § 2). Take Y as in (a), with $T = [a, b]$. For $\mathscr{F}$ take the family of all functions f_Δ, defined by

   $$f_\Delta(y) = |y(a)| + \sum_{i=1}^{n} |y(t_i) - y(t_{i-1})|,$$

   where Δ is a partition of $[a, b]$ by the points $t_0, t_1, \ldots, t_n$, $(a = t_0 < t_1 < \cdots < t_n = b)$.
   d. *H^p is complete,* $1 \leqslant p \leqslant \infty$ (see Example 8, § 2). Let Y be the class $\mathfrak{A}$ (defined previously) of all functions y of the complex variable t, defined and analytic when $|t| < 1$. When $1 \leqslant p < \infty$, take $\mathscr{F}$ to be the family of functions f_r defined by

   $$f_r(y) = \mathfrak{M}_p[y; r], \qquad 0 \leqslant r < 1.$$

(In § 2 we used f instead of y in $\mathfrak{M}_p[y; r]$.) In order to apply Theorem 4.4, it helps to know that

$$|x(t)| \leq \frac{\|x\|}{1-|t|}$$

if $x \in H^p$ and $|t| < 1$. The proof of this inequality runs as follows: Let

$$x(t) = \sum_{n=0}^{\infty} \xi_n t^n$$

$$\xi_n = \frac{1}{2\pi r^n} \int_0^{2\pi} x(re^{i\theta}) e^{-in\theta}\, d\theta, \qquad 0 < r < 1.$$

Then

$$r^n |\xi_n| \leq \mathfrak{M}_p[x; r] \leq \|x\|,$$

whence $|\xi_n| \leq \|x\|$ and

$$|x(t)| \leq \sum_{n=0}^{\infty} \|x\| |t|^n = \frac{\|x\|}{1-|t|}.$$

When $p = \infty$, take $\mathscr{F}$ to be the family of functions f_r defined by

$$f_r(y) = \max_{|t|=r} |y(t)|, \qquad 0 \leq r < 1.$$

e. $A(\Delta)$ *is complete* (see Example 9, § 2). Use Theorem 4.2.

2. Give an example of a series in a normed linear space that is absolutely convergent but not convergent.
3. Let X and Y be normed linear spaces. If T is a continuous linear operator from X into Y and if Σx_n is a convergent series in X, then $\Sigma T x_n$ converges in Y.
4. Let X be a normed linear space, and let $\hat{X}$ and Y be Banach spaces, each containing X as a dense subspace. Show that Y is isometrically isomorphic to $\hat{X}$.

II.5 QUOTIENT SPACES

Quotient spaces were introduced in § I.6 in a purely algebraic setting. When X is a *normed* linear space and M is a *closed* subspace of X, it is possible to define a norm on the quotient space X/M in a natural way. Recall that, for $x \in X$, the element $[x]$ of X/M is the set $\{u : x - u = m,$ some $m \in M\}$. Equivalently, $[x] = \{x - m : m \in M\}$. We define the *quotient norm* on X/M by

(5-1) $$\|[x]\| = \inf_{u \in [x]} \|u\| = \inf_{m \in M} \|x - m\|.$$

Theorem 5.1. *If M is a closed subspace of a normed linear space X, then* (5-1) *defines a norm on X/M. If X is a Banach space, then so is X/M.*

Proof. Note that $\|[x]\|$ is the distance from x to M. Since M is closed, $\|[x]\|=0$ if and only if $x \in M$, that is, if and only if $[x]$ is the zero element of X/M. Now for $\alpha \neq 0$, we have

$$\|\alpha[x]\| = \inf_{u \in [x]} \|\alpha u\| = |\alpha| \inf_{u \in [x]} \|u\| = |\alpha| \, \|[x]\|.$$

Also,

$$\begin{aligned}\|[x]+[y]\| &= \inf\{\|u+v\| : u \in [x],\ v \in [y]\}\\ &\leqslant \inf_{u \in [x]} \|u\| + \inf_{v \in [y]} \|v\| = \|[x]\| + \|[y]\|.\end{aligned}$$

Thus (5-1) defines a norm on X/M.

Now suppose X is a Banach space. Let $\Sigma_1^\infty U_k$ be an absolutely convergent series of elements of X/M; that is, $\Sigma_1^\infty \|U_k\| < \infty$. Because of the way in which the norm is defined in X/M, there exists $x_k \in U_k$ such that $\|x_k\| < \|U_k\| + 2^{-k}$, for $k = 1, 2, \ldots$. Consequently, $\Sigma_1^\infty \|x_k\| < \infty$. Since X is complete, the series $\Sigma_1^\infty x_k$ converges to some $x \in X$. Let $s_n = x_1 + \cdots + x_n$, $n = 1, 2, \ldots$. Then $[x - s_n] = [x] - \Sigma_1^n [x_k] = [x] - \Sigma_1^n U_k$. Since $\|[x - s_n]\| \leqslant \|x - s_n\|$, it follows that $\Sigma_1^\infty U_k$ converges to $[x]$. By Theorem 4.3, X/M is a Banach space. □

Closed Sets in X/M

Let ϕ be the canonical mapping of X onto X/M; that is, let $\phi(x) = [x]$ for $x \in X$. From (5-1), we have

$$\|\phi(x)\| = \inf_{m \in M} \|x - m\| \leqslant \|x\|, \qquad x \in X.$$

Thus the linear mapping ϕ is continuous, by Theorem 1.1. It can be shown that, for each open set U in X, the image set $\phi(U)$ is open in X/M (see problem 8 of § 9). However, it is *not* true that $\phi(F)$ is closed in X/M whenever F is closed in X. This follows from the next theorem and the fact that in general there exist closed sets F and M in X such that $F+M$ is not closed in X, where $F+M = \{x+m : x \in F, m \in M\}$. (See problem 4.) Observe that $\phi^{-1}(\phi(F)) = F + M$ (problem 1).

Theorem 5.2. *Let M be a closed linear manifold in a normed linear space X. Let F be a (not necessarily closed) subset of X. Then $\phi(F)$ is closed in X/M if and only if $F+M$ is closed in X.*

Proof. Since ϕ is continuous and the inverse image under ϕ of $\phi(F)$ is $F+M$, we know that $F+M$ must be closed whenever $\phi(F)$ is closed. Now suppose $F+M$ is closed. An arbitrary sequence in $\phi(F)$ is of the form $\{\phi(x_n)\}$, where each x_n is in F. Suppose that such a sequence converges (with respect to the norm in X/M) to an element of X/M, say $\phi(x)$. Then $\phi(x_n - x) =$

$\phi(x_n)-\phi(x)\to 0$, and hence the distance from $x_n - x$ to M goes to zero as $n\to\infty$. So there exists a sequence $\{y_n\}$ in M such that $x_n - x - y_n \to 0$. For each n, $x_n - y_n$ lies in the closed set $F+M$. Thus $x\in F+M$ and $\phi(x)\in\phi(F)$, proving that $\phi(F)$ is closed. □

Theorem 5.2 will be needed in Chapter VII. As an immediate application, we have the following result.

Theorem 5.3. *Let M be a closed subspace of a normed linear space X, and let F be a finite-dimensional subspace of X. Then $F+M$ is closed in X.*

Proof. Let ϕ be the canonical mapping of X onto X/M. Then $\phi(F)$ is a finite-dimensional subspace of X/M, and so $\phi(F)$ is closed, by Theorem 3.3. Hence $F+M$ is closed, by Theorem 5.2. □

PROBLEMS

1. Let ϕ be the canonical mapping of a normed linear space X onto a quotient space X/M, where M is a closed subspace of X. If $F\subset X$, then $\phi^{-1}(\phi(F))=F+M$. Also, $\|\phi\|=1$.
2. Let X and Y be normed linear spaces, and let T be a linear operator on X into Y such that $\mathcal{N}(T)$ is closed in X. For $[x]$ in $X/\mathcal{N}(T)$, define $\hat{T}([x])=Tx$. Then $\hat{T}$ is an isomorphism of $X/\mathcal{N}(T)$ onto $\mathcal{R}(T)$. (See problem 6 of § I.6.) Show that $\hat{T}$ is continuous if and only if T is continuous, in which case $\|\hat{T}\|=\|T\|$.
3. Suppose M is a closed subspace of a normed linear space X such that $\dim X/M<\infty$. Then $M+N$ is closed for every subspace N.
4. Let $e_1, e_2, \ldots,$ be the canonical unit vectors in ℓ^2, (where e_k denotes the sequence with 1 in the kth place and zeros elsewhere). Let M be the closed linear manifold generated by the set $\{e_{2k}: k=1,2,\ldots\}$, and let F be the closed linear manifold generated by the set $\{e_{2k}+(1/k)e_{2k-1}: k=1,2,\ldots\}$. Show that $F+M$ is dense but not closed in ℓ^2. [*Hint.* The sequence $x=(1,0,\frac{1}{2},0,\frac{1}{3},\ldots)$ is in $\ell^2\backslash(F+M)$.]

II.6 INNER-PRODUCT SPACES

Definition. A complex linear space X is called an inner-product space if to each pair of elements x, y of X there is associated a complex number (x, y) (called the *inner product* of x and y) with the following properties:

1. $(x+y, z)=(x, z)+(y, z)$, for $x, y, z\in X$.
2. $(x, y)=\overline{(y, x)}$ [the bar denoting complex conjugate].
3. $(\alpha x, y)=\alpha(x, y)$, for all scalars α.
4. $(x, x)\geq 0$ [it must be real by (2)], and $(x, x)\neq 0$ if $x\neq 0$.

A *real* linear space X is called an inner-product space if there is defined on $X \times X$ a real-valued function (x, y) with the properties (1) to (4). To simplify the exposition, we shall treat the cases of real X and complex X together, writing bars for complex conjugates where they are needed for the complex case. They can be ignored in the real case, since $\bar{\alpha} = \alpha$ if α is real.

Observe that as a consequence of the properties listed, (x, y) has the further properties:

$$(x, y+z) = (x, y) + (x, z)$$

$$(x, \alpha y) = \bar{\alpha}(x, y).$$

The first important fact to be noted about inner-product spaces is that we can use the inner product to define a norm. We must first prove the following theorem.

Theorem 6.1 (Cauchy–Schwarz Inequality). *If X is an inner-product space, then*

(6-1) $$|(x, y)| \leqslant (x, x)^{1/2}(y, y)^{1/2}.$$

The equality holds in (6-1) *if and only if x and y are linearly dependent.*

Proof. For arbitrary complex (or real) λ,

(6-2) $$0 \leqslant (x - \lambda y, x - \lambda y) = (x, x) - \bar{\lambda}(x, y) - \lambda\overline{(x, y)} + \lambda\bar{\lambda}(y, y).$$

The inequality (6-1) is obvious if $y = 0$, since in that case $(x, y) = 0$, so assume $y \neq 0$ and let $\lambda = (x, y)/(y, y)$. Substituting this in (6-2), we obtain

(6-3) $$0 \leqslant (x, x) - \frac{|(x, y)|^2}{(y, y)},$$

which is equivalent to (6-1). Equality holds in (6-3) if and only if equality holds in (6-2). But equality in (6-2) may be written as $0 = \|x - \lambda y\|^2$, which is true if and only if $x = \lambda y$. □

Theorem 6.2. *If X is an inner-product space, $(x, x)^{1/2}$ has the properties of a norm.*

Proof. We write $\|x\| = (x, x)^{1/2}$. The only requisite property of $\|x\|$ that is not immediately apparent is the triangular inequality property. Now

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

For complex A we write $\operatorname{Re} A = \frac{1}{2}(A + \bar{A}) =$ real part of A. Since $|\operatorname{Re} A| \leqslant |A|$, we see that

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

With (6-1) this gives

$$\|x+y\|^2 \leq \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2.$$

Consequently $\|x+y\| \leq \|x\| + \|y\|$. □

When X is an inner-product space we shall henceforth consider X also as a normed linear space, with the norm $\|x\| = (x, x)^{1/2}$. It is an interesting fact that the inner product is itself expressible in terms of this norm. If X is a *real* inner-product space,

$$(x, y) = \tfrac{1}{4}[\|x+y\|^2 - \|x-y\|^2]. \tag{6-4}$$

If X is a complex inner-product space, this *polarization identity* becomes

$$(x, y) = \tfrac{1}{4}\|x+y\|^2 - \tfrac{1}{4}\|x-y\|^2 + \frac{i}{4}\|x+iy\|^2 - \frac{i}{4}\|x-iy\|^2. \tag{6-5}$$

The validity of these formulas is readily verified by direct simplification of the expressions on the right, using the properties of the inner product.

Suppose for a moment that X is an arbitrary normed linear space, with norm $\|\cdot\|$. Then the polarization identity can be used to define a scalar-valued function f on $X \times X$. (Note that f will satisfy $f(x, x) = \|x\|^2$.) It can be shown, however, that this function will be an inner product on X if and only if the norm satisfies the *parallelogram law*:

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2. \tag{6-6}$$

If X is an inner-product space, the verification of (6-6) is similar to the verification of the polarization identity. The reader is referred to Day [2, page 153], for a proof that any norm that satisfies (6-6) must come from an inner product.

Throughout the rest of this section the discussion and theorems are concerned with inner-product spaces, and we shall not repeat this fact whenever stating theorems.

Theorem 6.3. *The inner product (x, y) is a continuous function on $X \times X$.*

Proof. It is readily verified that

$$(x_1, y_1) - (x_2, y_2) = (x_1 - x_2, y_1 - y_2) + (x_2, y_1 - y_2) + (x_1 - x_2, y_2).$$

Therefore, using (6-1),

$$|(x_1, y_1) - (x_2, y_2)| \leq \|x_1 - x_2\|\,\|y_1 - y_2\| + \|x_2\|\,\|y_1 - y_2\| + \|x_1 - x_2\|\,\|y_2\|.$$

It follows that the inner product is continuous at $(x_2, y_2) \in X \times X$. □

Examples of Inner-Product Spaces

The space ℓ^2 is an inner-product space, with inner product

$$(x, y) = \sum_{k=1}^{\infty} \xi_k \bar{\eta}_k,$$

where $x = \{\xi_k\}$, $y = \{\eta_k\}$.

The n-dimensional space $\ell^2(n)$ is an inner-product space, with

$$(x, y) = \sum_{k=1}^{n} \xi_k \bar{\eta}_k.$$

The real space $\ell^2(3)$ is ordinary three-dimensional Euclidean space, where the inner product is the "scalar product" of classical vector analysis.

The space $L^2(a, b)$ is an inner-product space with inner product

$$(x, y) = \int_a^b x(t)\overline{y(t)}\, dt.$$

Suppose ρ is an arbitrary positive measurable function on the interval $[a, b]$. Let $L^2_\rho(a, b)$ be the space (of equivalence classes) of all Lebesgue measurable functions x defined on $[a, b]$ such that

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

Then $L^2_\rho(a, b)$ is an inner-product space, with inner product

$$(x, y) = \int_a^b x(t)\overline{y(t)}\rho(t)\, dt.$$

The spaces just mentioned are complete. We can get many examples of incomplete inner-product spaces by selecting nonclosed linear manifolds in ℓ^2 or $L^2_\rho(a, b)$. For instance, the subspace of $L^2(a, b)$ consisting of elements corresponding to functions continuous on $[a, b]$ is an incomplete inner-product space.

Orthogonality

One of the most important notions in an inner-product space is that of orthogonality. We say that x and y are *orthogonal* if $(x, y) = 0$, and we write $x \perp y$. (Observe that $x \perp y$ is equivalent to $y \perp x$; also, $x \perp 0$ for every x.)

If x is orthogonal to each element of a set S, we say that x is orthogonal to S, and write $x \perp S$. If $x \perp y$ and $x \perp z$, then $x \perp (y+z)$ and $x \perp (\alpha y)$ for every scalar α. Also, since the inner product is continuous, it follows that if $x \perp y_n$ and $y_n \to y$, then also $x \perp y$. Hence, if $x \perp S$, then x is also orthogonal to the linear manifold generated by S and to the closure of this manifold.

A set S of vectors is called an *orthogonal set* if $x \perp y$ for every pair x, y in S such that $x \neq y$. If, in addition, $\|x\| = 1$ for every x in S, the set is called an *orthonormal set.*

Theorem 6.4. *Let $x_1, \ldots, x_n$ be an orthogonal set. Then*

$$\|x_1 + \cdots + x_n\|^2 = \|x_1\|^2 + \cdots + \|x_n\|^2.$$

Proof. When $n = 2$, this "Pythagorean Theorem" follows immediately from the definition of the norm and the fact that $(x_1, x_2) = 0$. The very simple details of an induction argument are left to the reader. $\square$

Examples of orthonormal sets. In the space ℓ^2, let e_k, $k = 1, 2, \ldots$, be that element $x = \{\xi_n\}$ for which $\xi_n = 0$ if $n \neq k$ and $\xi_k = 1$. The set of all the e_k's, or any subset of this set, is an orthonormal set.

In the complex space $L^2(0, 2\pi)$ let u_n be the element corresponding to the function $(1/\sqrt{2\pi})e^{int}$, $n = 0, \pm 1, \pm 2, \ldots$. Since

$$\frac{1}{2\pi}\int_0^{2\pi} e^{imt} e^{-int}\,dt = \begin{cases} 1 & \text{if} \quad m = n \\ 0 & \text{if} \quad m \neq n \end{cases}$$

we see that the collection of the elements u_n is an orthonormal set.

If we consider the *real* space $L^2(0, 2\pi)$, the elements determined by the functions

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\cos t, \frac{1}{\sqrt{\pi}}\cos 2t, \ldots$$

$$\frac{1}{\sqrt{\pi}}\sin t, \frac{1}{\sqrt{\pi}}\sin 2t, \ldots$$

form an orthonormal set.

Other examples of orthonormal sets are discussed after Theorem 6.5 and in § 8.

The Gram–Schmidt Process

Orthonormal sets play a central role in the theory of inner-product spaces. In fact, the rest of the theorems in § 6 (as well as four of the seven theorems of § 7) involve orthonormal sets in one way or another. When a linear manifold M is generated by a sequence of vectors, it is frequently useful to know that we may replace this sequence by an orthonormal sequence that also generates M. In concrete applications it is sometimes necessary to have an explicit method of constructing such an orthonormal sequence. The proof of the next theorem provides this. It is called the Gram–Schmidt orthogonalization process, in honor of J. P. Gram (1850–1916) and Erhard Schmidt (1876–1959).

Theorem 6.5. *Suppose $\{x_n\}$ is a finite or countable linearly independent set of vectors. Then there exists an orthonormal set having the same cardinal number that generates the same linear manifold as the given set $\{x_n\}$.*

Proof. Certainly $x_1 \neq 0$. We define $y_1, y_2, \ldots$ and $u_1, u_2, \ldots$ recursively, as follows:

$$y_1 = x_1 \qquad\qquad u_1 = \frac{y_1}{\|y_1\|},$$

$$y_2 = x_2 - (x_2, u_1)u_1 \qquad\qquad u_2 = \frac{y_2}{\|y_2\|},$$

$$y_{n+1} = x_{n+1} - \sum_{k=1}^{n} (x_{n+1}, u_k)u_k, \qquad u_{n+1} = \frac{y_{n+1}}{\|y_{n+1}\|},$$

The process terminates if the set of x's is finite. Otherwise it continues indefinitely. It is clear by induction that u_n is a linear combination of $x_1, \ldots, x_n$, and vice versa. Thus the linear manifold generated by the u's is the same as that generated by the x's. Observe that $y_n \neq 0$ because the set $x_1, \ldots, x_n$ is linearly independent. By direct calculation, we see that $(y_{n+1}, u_i) = 0$ if $i = 1, \ldots, n$. From this it follows that the u's form an orthonormal set. □

As an application of the Gram–Schmidt process, we consider the real space $L^2(-1, 1)$ and the sequence of elements corresponding to the polynomial functions

$$1, t, t^2, \ldots.$$

It is easy to see that this set is linearly independent. (Use the fact that a polynomial function of degree n has at most n zeros.) The "orthogonalization" of this set by the Gram–Schmidt process yields the set of polynomial functions

$$\sqrt{n+\tfrac{1}{2}}P_n(t), \qquad n = 0, 1, 2, \ldots,$$

where the $P_n(t)$ are the classical *Legendre polynomials*. For example,

$$P_0(t) \equiv 1, \qquad P_1(t) = t, \qquad P_2(t) = \tfrac{3}{2}t^2 - \tfrac{1}{2}, \qquad P_3(t) = \tfrac{5}{2}t^3 - \tfrac{3}{2}t.$$

In general,

$$P_n(t) = \frac{1}{2^n n!} \frac{d^n}{dt^n}(t^2-1)^n.$$

Subspaces of Finite Dimension

Although we are primarily interested in infinite orthonormal sets, it is convenient, first, to examine the situation for finite orthonormal sets in a way that will lead to the more general theorems we desire.

Theorem 6.6. *Let $S = \{u_1, \ldots, u_n\}$ be a finite orthonormal set in X, and let M be the subspace of X generated by S. Then S is a basis for M and each element $x \in M$ has the unique representation $x = \sum_{i=1}^n \xi_i u_i$, where $\xi_i = (x, u_i)$, $i = 1, \ldots, n$.*

Proof. If $x = \xi_1 u_1 + \cdots + \xi_n u_n$, then

$$(x, u_i) = \xi_1(u_1, u_i) + \cdots + \xi_n(u_n, u_i) = \xi_i,$$

by the orthonormality relations. If $x = 0$, it follows that $\xi_1 = \cdots = \xi_n = 0$. Hence we see that the u_i's form a linearly independent set and, therefore, form a basis for the subspace that they generate. □

If M is a subspace of X of finite dimension (and hence closed by Theorem 3.3) and if $x \in X$ but $x \notin M$, then x lies a positive distance from M. It then makes sense to try to find a vector x_M in M such that

$$\|x - x_M\| = \operatorname{dist}(x, M) \equiv \inf_{y \in M} \|x - y\|.$$

This is an abstract setting for the classical "problem of best approximation." This problem arises, for example, when one wants to approximate a function by a linear combination of a given finite set of linearly independent functions. In § 7 we shall see that in a complete inner-product space the problem of best approximation always has a unique solution. (This is not the case, for example, in the Banach space $C[0, 1]$.) And this unique solution will exist even if M is a closed infinite-dimensional subspace. The case when M is finite dimensional is considered in the next theorem.

Theorem 6.7. *Let $S = \{u_1, \ldots, u_n\}$ be an orthonormal set, and let M be the subspace generated by S. Then, given $x \in X$, there is exactly one element x_M in M such that $\|x - x_M\| = \operatorname{dist}(x, M)$, namely,*

$$x_M = \sum_{i=1}^{n} (x, u_i) u_i. \tag{6-7}$$

Furthermore, $(x - x_M) \perp M$, *and*

$$\|x - x_M\|^2 = \|x\|^2 - \|x_M\|^2 = \|x\|^2 - \sum_{i=1}^{n} |(x, u_i)|^2. \tag{6-8}$$

Proof. Define x_M by (6-7). For $u_j \in S$, we have

$$(x - x_M, u_j) = (x, u_j) - \sum_{i=1}^{n} (x, u_i)(u_i, u_j) = (x, u_j) - (x, u_j) = 0.$$

It follows that $(x - x_M) \perp S$, and hence $(x - x_M) \perp M$. It is clear that $x_M \in M$. If $m \in M$, then Theorem 6.4 implies that

$$\|x - m\|^2 = \|x - x_M\|^2 + \|x_M - m\|^2. \tag{6-9}$$

It follows that the infimum of $\|x - m\|$, for $m \in M$, is attained if and only if $m = x_M$. Also, from (6-9), if $x \in M$, then x must equal x_M. Finally, using Theorem 6.4 and the fact that S is an orthonormal set, we have

$$\begin{aligned} \|x\|^2 &= \|x - x_M\|^2 + \|x_M\|^2 = \|x - x_M\|^2 + (x_M, x_M) \\ &= \|x - x_M\|^2 + \sum_{i=1}^{n} |(x, u_i)|^2. \end{aligned}$$

This proves (6-8). $\square$

Bessel's Inequality

When M is an infinite-dimensional linear manifold the series analogous to those in (6-7) and (6-8) have infinitely many terms. We shall investigate the convergence of (6-7) and the existence of x_M in § 7, where X will be a complete inner-product space. However, completeness of X is not needed in the following important generalizations of (6-8). (Compare (6-8) with (6-13), below.)

Theorem 6.8. *Let S be an orthonormal set. For a fixed $x \in X$, the set of those $u \in S$ such that $(x, u) \neq 0$ is either finite or countably infinite. If $x, y \in X$, then*

$$\sum_{u \in S} |(x, u)\overline{(y, u)}| \leq \|x\| \, \|y\|, \tag{6-10}$$

it being understood that the sum on the left includes all $u \in S$ for which $(x, u)\overline{(y, u)} \neq 0$ and is, therefore, either a finite series or an absolutely convergent series with a countable infinity of terms.

Proof. Let $u_1, \ldots, u_n$ be any finite collection of distinct elements of S. Then (6-8) implies that

$$\sum_{i=1}^{n} |(x, u_i)|^2 \leq \|x\|^2. \tag{6-11}$$

Using (6-11) together with Cauchy's inequality (see page 3), we obtain

$$\begin{aligned} \sum_{i=1}^{n} |(x, u_i)\overline{(y, u_i)}| &\leq \left(\sum_{i=1}^{n} |(x, u_i)|^2\right)^{1/2} \left(\sum_{i=1}^{n} |(y, u_i)|^2\right)^{1/2} \\ &\leq \|x\| \, \|y\|. \end{aligned} \tag{6-12}$$

It follows from (6-11) that if $x \in X$ and n is a positive integer, the number of elements u of S such that $|(x, u)| \geq 1/n$ cannot exceed $n^2\|x\|^2$. Since $|(x, u)| \geq 1/n$ for some n if $(x, u) \neq 0$, the set of those $u \in S$ such that $(x, u) \neq 0$ is a countable union of finite sets and is, therefore, either finite or countably infinite. The inequality (6-10) now follows from (6-12). □

The special case of (6-10) where $x = y$ is called *Bessel's inequality*:

$$\sum_{u \in S} |(x, u)|^2 \leq \|x\|^2. \tag{6-13}$$

Parseval's Formula

It is often important to know when Bessel's inequality is actually an equality. This problem is considered in Theorem 6.10. It is convenient, however, first to prove a useful representation theorem, a generalization of Theorem 6.6.

Theorem 6.9. *Let S be an orthonormal set, and let M be the closed linear manifold generated by S. For each $x \in M$,*

$$x = \sum_{u \in S} (x, u)u,$$

where the summation is taken over any indexing of the (countable) set $\{u \in S : (x, u) \neq 0\}$.

Proof. Because x is in M, x is the limit of a sequence of finite linear combinations of elements of S. Given $\varepsilon > 0$, there exist a finite set $\{u_1, \ldots, u_n\} \subset S$ and a set of scalars $\{a_1, \ldots, a_n\}$ such that

$$\left\| x - \sum_{i=1}^{n} a_i u_i \right\|^2 < \varepsilon.$$

Applying Theorem 6.7 to the subspace generated by $\{u_1, \ldots, u_n\}$, we conclude that

(6-14)
$$\left\| x - \sum_{i=1}^{n} (x, u_i)u_i \right\|^2 \leq \left\| x - \sum_{i=1}^{n} a_i u_i \right\|^2 < \varepsilon.$$

If we let M_0 be the closed linear manifold spanned by the set $\{u \in S : (x, u) \neq 0\}$, it is clear that $\sum_{i=1}^{n} (x, u_i)u_i \in M_0$. Since $\varepsilon > 0$ was arbitrary, it follows from (6-14) that $x \in M_0$. The set $\{u \in S : (x, u) \neq 0\}$ is countable by Theorem 6.8, so let $v_1, v_2, \ldots,$ be an enumeration of this set. Then x is the limit of a sequence of finite linear combinations of the v_i. Such a finite linear combination can always be written in the form $\sum_{i=1}^{n} a_i v_i$, where n is sufficiently large and where some of the coefficients may be zero. For each positive integer n and for any choice of scalars, $a_1, \ldots, a_n$, we have

$$\left\| x - \sum_{i=1}^{n} (x, v_i)v_i \right\|^2 \leq \left\| x - \sum_{i=1}^{n} a_i v_i \right\|^2,$$

by Theorem 6.7. Since $x \in M_0$, the right side can be made arbitrarily small. This proves that

(6-15)
$$x = \sum_{i=1}^{\infty} (x, v_i)v_i,$$

and since the ordering of the v_i was arbitrary, this proves the theorem. □

Theorem 6.10 (Parseval's Formula). *Let S be an orthonormal set, and let M be the closed linear manifold generated by S. If $x \in X$, the Parseval formula*

(6-16)
$$\|x\|^2 = \sum_{u \in S} |(x, u)|^2$$

is valid if and only if $x \in M$.

Proof. For $x \in M$ we can use (6-15) and the notation of the preceding theorem. Then

(6-17)
$$(x, x) = \left(x, \lim_{n \to \infty} \sum_{i=1}^{n} (x, v_i)v_i\right) = \lim_{n \to \infty} \left(x, \sum_{i=1}^{n} (x, v_i)v_i\right)$$
$$= \lim_{n \to \infty} \sum_{i=1}^{n} \overline{(x, v_i)}(x, v_i) = \sum_{i=1}^{\infty} |(x, v_i)|^2,$$

since the inner product is continuous (Theorem 6.3). As in Theorem 6.8, we agree to write

$$\sum_{u \in S} |(x, u)|^2 \qquad \text{in place of} \qquad \sum_{i=1}^{\infty} |(x, v_i)|^2.$$

This proves (6-16).

Now suppose that (6-16) holds for an element $x \in X$. Let $u_1, u_2, \ldots$ be an enumeration of the set $\{u \in S : (x, u) \neq 0\}$, so that we may write

$$\|x\|^2 = \sum_{i=1}^{\infty} |(x, u_i)|^2. \tag{6-18}$$

Let $s_n = \sum_{i=1}^{n} (x, u_i)u_i$, $n = 1, 2, \ldots$. For $n = 1, 2, \ldots$, we may apply (6-8) of Theorem 6.7 to the subspace generated by $\{u_1, \ldots, u_n\}$:

$$\|x - s_n\|^2 = \|x\|^2 - \sum_{i=1}^{n} |(x, u_i)|^2.$$

From (6-18) it follows that $s_n \to x$, and hence that $x \in M$. □

Complete Orthonormal Sets

Definition. An orthonormal set S in the inner-product space X is said to be *complete* if there exists no orthonormal set of which S is a proper subset. In other words, S is complete if it is maximal with respect to the property of being orthonormal.

Theorem 6.11. *Every inner-product space X having a nonzero element contains a complete orthonormal set. Moreover, if S is any orthonormal set in X, there is a complete orthonormal set containing S as a subset.*

Proof. Let S be an orthonormal set in X. Such sets certainly exist; for instance, if $x \neq 0$, the set consisting merely of $x/\|x\|$ is orthonormal. Let P be the class of all orthonormal sets having S as a subset. It is easily verified that P satisfies the conditions of Zorn's lemma (§ I.9). Thus P contains a maximal member. □

Theorem 6.12. *Let S_1 and S_2 be two complete orthonormal sets in an inner-product space X. Then S_1 and S_2 have the same cardinal number (i.e., there is a one-to-one correspondence between the elements of S_1 and S_2).*

Proof. Suppose that one of the sets S_1, S_2, say S_1, is finite, and let M be the closed linear manifold generated by S_1. If $M \neq X$, take $x \in X \setminus M$, and let $x_M = \sum_{u \in S_1} (x, u)u$. By Theorem 6.7, $x - x_M \neq 0$. But clearly $(x - x_M) \perp S_1$. This contradicts the maximality of S_1, so we conclude $M = X$ and X is finite dimensional. An orthonormal set is linearly independent, as we saw in the proof of Theorem 6.6, so S_1 is a basis for X. Also, S_2 is a basis since it is a linearly independent and maximal set and, consequently, S_1 and S_2 have the same finite number of elements. For the remainder of the proof we assume, therefore, that both S_1 and S_2 are infinite sets. For each $x \in S_1$ let $S_2(x) = \{y \in S_2 : (x, y) \neq 0\}$. By Theorem 6.8, $S_2(x)$ is an at most countably infinite set. If

$y \in S_2$, then $\|y\| = 1$, and it follows from the completeness of S_1 that there is some $x \in S_1$ such that $y \in S_2(x)$ (since otherwise S_1 together with y would be an orthonormal set). Hence we see that

$$S_2 = \bigcup_{x \in S_1} S_2(x). \tag{6-19}$$

Now let A_1, A_2 be the cardinal numbers of the sets S_1, S_2, respectively, and let A_0 be the cardinal number of the set of positive integers. The relation (6-19) shows that $A_2 \leqslant A_0 A_1$. But it is known that if A is any infinite cardinal number, then $A_0 A = A$ (see, for example, Sierpinski [1, § 103]). Thus we see that $A_2 \leqslant A_1$. So far we have used only the completeness of S_1. If S_2 is also complete, we can exchange the roles of S_1 and S_2 and so conclude $A_1 \leqslant A_2$. But then $A_1 = A_2$. □

Theorem 6.13. *Let S be an orthonormal set in X. Suppose that Parseval's formula holds for each $x \in X$. Then S is complete.*

Proof. If S is not complete, there exists some $x \neq 0$ such that $x \perp S$, and hence also $x \perp M$, where M is the closed linear manifold generated by S. But the hypothesis of this theorem is equivalent to supposing that $M = X$, by Theorem 6.10. Then $x \perp x$, which implies $x = 0$. This contradiction implies S is complete. □

If the linear manifold generated by an orthonormal set S is dense in X, then S must be complete, by Theorems 6.10 and 6.13. The converse is not true. In fact, there exists an inner-product space X such that if S is *any* complete orthonormal set in X, then the closed linear manifold generated by S is a proper subset of X (see Dixmier [1]). This situation cannot occur when X is complete (see Theorem 7.4).

Separable Inner-Product Spaces

The theorems proved above have not required that X be a separable space. However, the most common inner-product spaces are separable. Therefore it is useful to know that in such spaces it is not necessary to consider uncountable orthonormal sets.

Theorem 6.14. *Suppose X is separable, and let S be an orthonormal set in X. Then S is either a finite or a countably infinite set.*

Proof. Each pair u_1, u_2 of distinct elements of an orthonormal set S are a distance $\sqrt{2}$ apart, for

$$\|u_1 - u_2\|^2 = (u_1 - u_2, u_1 - u_2) = (u_1, u_1) + (u_2, u_2) = 2.$$

Let $\{y_n\}$ be a countable set everywhere dense in X, and let $B_n = \{x \in X : \|x - y_n\| < \sqrt{2}/2\}$, $n = 1, 2, \ldots$. Since $\{y_n\}$ is dense, $S = \bigcup_{n=1}^{\infty} (S \cap B_n)$. It thus suffices to prove that, for each n, $S \cap B_n$ cannot have more than one element. Indeed, if $u_1, u_2 \in S \cap B_n$,

$$\|u_1 - u_2\| \leq \|u_1 - y_n\| + \|y_n - u_2\| < \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{2} = \sqrt{2},$$

which is impossible, unless $u_1 = u_2$. □

Theorem 6.15. *Let X be an infinite-dimensional inner-product space. Then X is separable if and only if there exists a countable orthonormal set S such that the closed linear manifold generated by S is X.*

Proof. Since X is separable, there exists a countable set $\{x_n\}$ that is everywhere dense in X. Let y_1 be the first nonzero element in the sequence $\{x_n\}$, y_2 the first x_n that is not in the linear manifold generated by y_1, and y_{k+1} the first x_n that is not in the linear manifold generated by $y_1, \ldots, y_k$. It is clear that the x's and the y's generate the same linear manifold and hence the same closed linear manifold. This manifold is the whole space X because the set of x's is dense in X. Applying the Gram–Schmidt process to the y's, we obtain an orthonormal set that generates the closed linear manifold X and which must therefore be complete, by Theorems 6.10 and 6.13. Conversely, if $\{u_n\}$ is a countable orthonormal set whose linear span is dense in X, let L be the set of all finite sums $\Sigma\, \alpha_k u_k$, where $\alpha_k = a_k + ib_k$ and a_k, b_k are rational numbers. (If X is a *real* space, each α_k must be real.) Clearly L is countable. Routine calculations show that L is dense. Thus X is separable. □

PROBLEMS

1. Let x and y be nonzero vectors in a complex inner-product space X.
   a. $\|x + y\| = \|x\| + \|y\|$ if and only if y is a positive multiple of x.
   b. $\|x - y\| = |\,\|x\| - \|y\|\,|$ if and only if y is a positive multiple of x.
   c. Given $z \in X$, $\|x - y\| = \|x - z\| + \|z - y\|$ if and only if $z = \alpha x + (1 - \alpha)y$ for some real number α with $0 \leq \alpha \leq 1$.
2. Let $\{x_n\}$ and $\{y_n\}$ be sequences in an inner-product space.
   a. If $x_n \to x$ and $y_n \to y$, then $(x_n, y_n) \to (x, y)$.
   b. If $(x_n, x) \to (x, x)$ and $\|x_n\| \to \|x\|$, then $x_n \to x$.
3. If $\{u_n\}$ is an orthonormal sequence in an inner-product space X, then $(u_n, x) \to 0$ for all $x \in X$.

II.7 HILBERT SPACE

A complete inner-product space is called a *Hilbert space* (real or complex according to whether the scalar field is real or complex). A finite-dimensional inner-product space is complete (Theorem 3.2) and is, therefore, a Hilbert space. The name Hilbert space is in honor of David Hilbert, who made use of what we now designate as the unit ball in the space ℓ^2 in his work on the theory of integral equations at the beginning of the twentieth century. The first abstract axiomatic study of Hilbert space was carried out by John von Neumann in 1928 in conjunction with his pioneering work on the mathematical foundations of quantum mechanics. Early writers on abstract Hilbert space theory required their Hilbert spaces to be infinite dimensional and separable. For a time it seemed that separability was indispensable for the general theory. However, ways were found to develop the theory without separability, and the current practice is to invoke this hypothesis only when it is needed.

A finite-dimensional Hilbert space is often called a *Euclidean space.* The term *unitary space* is sometimes applied to a finite-dimensional *complex* Hilbert space.

An incomplete inner-product space X, being a normed linear space, can be completed in the manner described in § 4. Let $\hat{X}$ be the completion of X. The function f defined by

$$f(x, y) = \|x+y\|^2 + \|x-y\|^2 - 2\|x\|^2 - 2\|y\|^2$$

is continuous on $\hat{X} \times \hat{X}$, since vector addition, scalar multiplication and the norm are all continuous functions. Since X is an inner-product space, the norm satisfies the parallelogram law (6-6) and $f(x, y) = 0$ on the set $X \times X$ that is dense in $\hat{X} \times \hat{X}$. It follows that (6-6) holds in $\hat{X} \times \hat{X}$, and so (6-4) or (6-5) defines an inner product on $\hat{X} \times \hat{X}$, with $\|x\|^2 = (x, x)$ for $x \in \hat{X}$. We see then that any incomplete inner-product space X can be extended to form a Hilbert space $\hat{X}$ in which X is everywhere dense. Because of this fact, an incomplete inner-product space may be descriptively called a *pre-Hilbert space.* This terminology has been used by a number of writers.

When X is a Hilbert space, all Cauchy sequences in X will converge and we can generalize Theorem 6.6.

Theorem 7.1. *Let X be a Hilbert space, and let $\{u_n\}$ be a countably infinite orthonormal set in X. Then a series of the form $\sum_{n=1}^{\infty} \xi_n u_n$ is convergent if and only if $\sum_{n=1}^{\infty} |\xi_n|^2 < \infty$. In this case the series $\sum_{n=1}^{\infty} \xi_n u_n$ converges to the same element, no matter how the terms are rearranged, and we have the relations*

$$\xi_n = (x, u_n), \qquad x = \sum_{n=1}^{\infty} \xi_n u_n$$

between the coefficients ξ_n and the element x defined by the series.

Proof. Let $s_n = \xi_1 u_1 + \cdots + \xi_n u_n$. Then if $m < n$, the orthonormality relations and the relation between the norm and inner product lead to the formula

$$\|s_n - s_m\|^2 = \left\| \sum_{i=m+1}^{n} \xi_i u_i \right\|^2 = \sum_{i=m+1}^{n} |\xi_i|^2.$$

Since X is complete, it is now clear that the sequence $\{s_n\}$ is convergent in X if and only if $\sum_{i=1}^{\infty} |\xi_i|^2 < \infty$. If this latter condition is satisfied and if $x = \lim_{n\to\infty} s_n$, we prove that $\xi_i = (x, u_i)$ as follows: By Theorem 6.6, we know that $\xi_i = (s_n, u_i)$ if $1 \leq i \leq n$. But $s_n \to x$ and hence $(s_n, u_i) \to (x, u_i)$ when $n \to \infty$, by the continuity of the inner product. Therefore $\xi_i = (x, u_i)$. Clearly x is in the closed linear manifold generated by $u_1, u_2, \ldots$, so the series converging to x does not depend on the order of its terms, by Theorem 6.9. ☐

The Riesz–Fischer theorem in the theory of Fourier series is a concrete instance of Theorem 7.1. If x is a function in the real space $\mathscr{L}^2(0, 2\pi)$, its Fourier series is

$$\frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos nt + \sum_{n=1}^{\infty} b_n \sin nt,$$

where

$$a_n = \frac{1}{\pi} \int_0^{2\pi} x(s) \cos ns \, ds, \qquad b_n = \frac{1}{\pi} \int_0^{2\pi} x(s) \sin ns \, ds. \tag{7-1}$$

The constants $a_0, a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are called the Fourier coefficients of x. The Riesz–Fischer theorem asserts that if $a_0, a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ are sequences of real constants such that

$$\sum_{n=1}^{\infty} (a_n^2 + b_n^2) < \infty,$$

there exists a function x of class $\mathscr{L}^2(0, 2\pi)$ having a_n and b_n as its Fourier coefficients. We take X to be $L^2(0, 2\pi)$ and let $u_1, u_2, u_3, u_4, u_5, \ldots$ be the elements of X corresponding to

$$\frac{1}{\sqrt{2\pi}} \qquad \frac{\cos t}{\sqrt{\pi}}, \qquad \frac{\sin t}{\sqrt{\pi}}, \qquad \frac{\cos 2t}{\sqrt{\pi}}, \qquad \frac{\sin 2t}{\sqrt{\pi}}, \tag{7-2}$$

The coefficients $\xi_1, \xi_2, \ldots$ in Theorem 7.1 are related to the standard Fourier coefficients by the formulas

$$\xi_1 = a_0\sqrt{\pi/2}, \qquad \xi_2 = a_1\sqrt{\pi}, \qquad \xi_3 = b_1\sqrt{\pi}, \qquad \xi_4 = a_2\sqrt{\pi}, \qquad \ldots.$$

The Riesz–Fischer theorem was proved independently by F. Riesz and E. Fischer in 1907. The classical theory of Fourier series and the Riesz–Fischer theorem originated before the axiomatic development of Hilbert space theory. It is perhaps for this reason that the (orthogonal) functions $\{1, \cos t, \sin t, \cos 2t, \sin 2t, \ldots\}$ seemed "more natural" to use than the (complete orthonormal) set of functions in (7-2).

When $u_1, u_2, \ldots$ is a complete orthonormal set in Hilbert space, the numbers $\xi_n = (x, u_n)$, $n = 1, 2, \ldots$, are often called the *Fourier coefficients of x relative to* $\{u_n\}$, even though this does not quite coincide with the classical situation in (7-1). Likewise, the series $\Sigma_{n=1}^{\infty} (x, u_n)u_n$ is called the Fourier series of x relative to $\{u_n\}$.

An important property of a Hilbert space is that a Fourier series of the form $\Sigma (x, u_n)u_n$ must converge, if the u_n belong to an orthonormal set. This fact leads to several important theorems.

Theorem 7.2. *Let S be an orthonormal set in a Hilbert space X, and let M be the closed linear manifold generated by S. For each $x \in X$ there is exactly one element x_M in M such that $\|x - x_M\| = \operatorname{dist}(x, M)$. This element is defined unambiguously by*

$$x_M = \sum_{u \in S} (x, u)u. \tag{7-3}$$

Furthermore, $(x - x_M) \perp M$, *and*

$$\|x - x_M\|^2 = \|x\|^2 - \|x_M\|^2. \tag{7-4}$$

Proof. Let $u_1, u_2, \ldots$ be an enumeration of those u's for which $(x, u) \neq 0$. By Bessel's inequality and Theorem 7.1, the series $\Sigma_n (x, u_n)u_n$, if infinite, is convergent to some element x_M. Evidently $x_M \in M$, so by Theorem 6.9 the series defining x_M does not depend on the order of its terms. The rest of the proof follows the same lines as the proof of Theorem 6.7. □

The vector x_M in Theorem 7.2 is called the *orthogonal projection of x on the subspace M.*

Theorem 7.3. *Let S be an orthonormal set in a Hilbert space X, and let M be the closed linear manifold generated by S. The following statements are equivalent*:

(*a*) *S is complete (i.e., maximal).*
(*b*) $M = X$.
(*c*) *If $x \perp S$, then $x = 0$.*
(*d*) $x = \Sigma_{u \in S} (x, u)u$ *for all* $x \in X$.
(*e*) $\|x\|^2 = \Sigma_{u \in S} |(x, u)|^2$ *for all* $x \in X$.

Proof. If $M \neq X$, take $x \in X \setminus M$ and let x_M be the element in M given by Theorem 7.2. If we set $z = (x - x_M)/\|x - x_M\|$, then the set consisting of S and z is orthonormal, and yet $z \notin S$. Thus S cannot be complete. Hence (*a*) implies

(b). By Theorem 6.9, (b) implies (d). Trivially, (d) implies (c). Given $x \in X$, we have $(x - x_M) \perp M$, by Theorem 7.2. If (c) is true, then $x - x_M = 0$; that is, $x = x_M \in M$. Thus (c) implies (b). Next, (b) and (e) are equivalent by Theorem 6.10. Finally, (e) implies (a) by Theorem 6.13. □

From Theorem 7.3 it follows that each closed linear manifold M in a Hilbert space X is generated by an orthonormal set S. For, M contains a maximal orthonormal set S, by Theorem 6.11, and since M is itself a Hilbert space, the closed linear manifold generated by S must be all of M. Consequently, Theorem 7.2 may be applied to any closed linear manifold M to conclude that to each $x \in X$ there corresponds a unique $x_M \in M$ such that $\|x - x_M\| = \text{dist}(x, M)$ and $(x - x_M) \perp M$.

The set of all elements orthogonal to a subspace M is denoted by $M^\perp$ and is called the *orthogonal complement* of M. It is readily verified that $M^\perp$ is a closed linear manifold and $M \cap M^\perp = (0)$. If M is closed, then we may write each $x \in X$ as the sum of an element in M and an element orthogonal to M, namely, $x = x_M + (x - x_M)$. Hence $X = M + M^\perp$. This proves the following theorem (cf. Theorem I.6.1).

Theorem 7.4. *If M is a closed subspace of a Hilbert space X, then every $x \in X$ may be uniquely represented in the form $x = y + z$ where $y \in M$ and $z \in M^\perp$; that is,*

$$X \doteq M \oplus M^\perp.$$

Representations of Hilbert Space

Theorem 7.5. *Every Euclidean space X of dimension n $(n \geqslant 1)$ is congruent to $\ell^2(n)$.*

Proof. For the definition of congruence see § 1. We can choose for X an orthonormal basis $u_1, \dots, u_n$. Then if $x = \xi_1 u_1 + \cdots + \xi_n u_n$, the correspondence $x \mapsto (\xi_1, \dots, \xi_n)$ establishes the congruence of X and $\ell^2(n)$. The isometric character of the correspondence follows by direct calculation, for, if $y = \eta_1 u_1 + \cdots + \eta_n u_n$, we have

$$(x, y) = \xi_1 \bar{\eta}_1 + \cdots + \xi_n \bar{\eta}_n.$$

Hence

$$\|x\| = (|\xi_1|^2 + \cdots + |\xi_n|^2)^{1/2},$$

and the right member of this equation is the norm of $(\xi_1, \dots, \xi_n)$ in $\ell^2(n)$. □

It is interesting to observe that if two inner-product spaces X and Y are congruent and if T is the operator that maps X isomorphically and isometrically onto Y, we not only have $\|Tx\| = \|x\|$, but also $(Tx_1, Tx_2) = (x_1, x_2)$, for every pair x_1, x_2 in X. That this is so follows directly from formulas (6-4) and (6-5).

Theorem 7.6 *Let X be an infinite-dimensional separable Hilbert space. Then X is congruent to ℓ^2.*

Proof. We know (Theorem 6.15) that X contains a countable complete orthornormal set $\{u_n\}$. If $x \in X$, let $\xi_n = (x, u_n)$. Then the sequence $\{\xi_n\}$ belongs to ℓ^2 and

$$\|x\|^2 = \sum_{n=1}^{\infty} |\xi_n|^2,$$

by Parseval's formula. Moreover, every $\{\xi_n\} \in \ell^2$ arises from some $x \in X$ in this way, by Theorem 7.1. The correspondence $x \mapsto \{\xi_n\}$ clearly establishes an isometric and isomorphic correspondence between the elements of X and ℓ^2, so that these spaces are congruent. □

In order to get a result like that of Theorem 7.6 for nonseparable Hilbert spaces, we must first construct a space somewhat like ℓ^2, but of such a character that it need not be separable.

Definition of the Space $\ell^2[Q]$. Let Q be any nonempty set of elements. Let $\ell^2[Q]$ be the class of all complex-valued functions x defined on Q such that the set of $q \in Q$ for which $x(q) \neq 0$ is either finite or countable and, moreover,

$$\sum_{q \in Q} |x(q)|^2 < \infty.$$

This class becomes a complex inner-product space if we define $x + y$ and αx as usual with functions and define the inner product by

$$(x, y) = \sum_{q \in Q} x(q)\overline{y(q)}.$$

The space $\ell^2[Q]$ is complete. This is clear if Q is a finite set with n elements or if Q is countable, for in these cases $\ell^2[Q]$ is congruent to $\ell^2(n)$ and to ℓ^2, respectively. (In fact, $\ell^2[Q]$ is $\ell^2(n)$ if Q is the set of integers $1, \ldots, n$, and $\ell^2[Q]$ is ℓ^2 if Q is the set of *all* positive integers.) If Q is uncountable, the fact that $\ell^2[Q]$ is complete follows readily from the fact that ℓ^2 is complete (see § 4).

We have defined the *complex* space $\ell^2[Q]$; it is clear that if we require all the functions x to be real valued, we get a real inner-product space.

Now, for each $p \in Q$ let x_p be the function defined by $x_p(q) = 0$ if $p \neq q$, $x_p(p) = 1$. The set of all the x_p's is an orthonormal set; clearly its cardinal number is the same as that of the set Q. Hence, by Theorem 6.14, $\ell^2[Q]$ is not separable if Q is uncountable. The orthonormal set formed by the x_p's is complete. For, if $x \in \ell^2[Q]$, the definition of the inner product shows that $(x, x_p) = x(p)$, and thus, by the definition of the norm,

$$\|x\|^2 = \sum_{p \in Q} |(x, x_p)|^2.$$

Therefore the orthonormal set is complete, by Theorem 6.13.

This example shows that there exist Hilbert spaces having complete orthonormal sets whose cardinal number is any specified infinite cardinal.

Theorem 7.7 *Let X be a Hilbert space for which the cardinal number of any one (and hence every) complete orthonormal set is A. Let Q be any set the cardinal number of whose elements is A. Then X is congruent to $\ell^2[Q]$.*

The proof is just like that of Theorem 7.6.

PROBLEM

1. a. Let f be a holomorphic function on the disc $|z| < R$, where $R > 0$, say

(7-5)
$$f(z) = \sum_0^\infty c_n z^n.$$

If $0 < r < R$, prove that

(7-6)
$$\frac{1}{2\pi} \int_0^{2\pi} |f(re^{i\theta})|^2 \, d\theta = \sum_0^\infty |c_n|^2 r^{2n}.$$

[*Hint.* The function $\theta \mapsto f(re^{i\theta})$ is in $L^2(0, 2\pi)$.]

b. (See Example 8, § 2.) Let f be a holomorphic function on the unit disc, with $f(z)$ given by (7-5). Then $f \in H^2$ if and only if $\Sigma |c_n|^2 < \infty$, in which case

$$\|f\|_2 = \left(\sum_0^\infty |c_n|^2 \right)^{1/2}.$$

c. Use (7-6) to prove Liouville's theorem: Suppose the series in (7-5) converges and defines $f(z)$ for all values of z (not merely $|z| < R$), and suppose there exists $M > 0$ such that $|f(z)| < M$ for all z, then $f(z) \equiv c_0$.

d. Use (7-6) to prove the Maximum Modulus Theorem: If $f(z)$ is given by (7-5) for $|z| < 1$ and if $|f(z)| \leq |f(a)|$ for some $|a| < 1$ and all $|z| < 1$, then $f(z) \equiv f(a)$ for $|z| < 1$.

II.8 EXAMPLES OF COMPLETE ORTHONORMAL SETS

In this section we give examples of orthonormal sets of functions that are useful in several branches of analysis, probability theory, mathematical physics, and electrical engineering. The functions in Examples 1 to 4 can arise, for instance, when one wants to solve certain boundary value problems for differential equations. A classical method is to look for a solution that is represented as a "Fourier series" of orthonormal functions.

Example 1. It was mentioned in § 6 that the functions $(1/\sqrt{2\pi})e^{int}$ $n = 0, \pm 1, \pm 2, \ldots$, determine an orthonormal set in the complex space $L^2(0, 2\pi)$. It is a fact of paramount importance in the theory of Fourier series that this orthonormal set is complete; that is, if

$$\int_0^{2\pi} x(t)e^{int}\,dt = 0$$

for all integers n, then $x(t) = 0$ almost everywhere. For a proof see Akhiezer and Glazman [1, § 11]. By Theorem 7.3, the completeness also finds its expression in the Parseval relation

$$\frac{1}{2\pi}\int_0^{2\pi} |x(t)|^2\,dt = \sum_{n=-\infty}^{\infty} |c_n|^2$$

where

$$c_n = \frac{1}{2\pi}\int_0^{2\pi} x(t)e^{-int}\,dt.$$

Example 2. In the study of certain differential equations, it is appropriate to look for solutions in a "weighted" $L_\rho(a, b)$ space rather than in $L^2(a, b)$. Consider the interval $[-1, 1]$ and define a weight function ρ by

$$\rho(t) = (1-t)^\alpha(1+t)^\beta,$$

where $\alpha > -1$ and $\beta > -1$. If the Gram–Schmidt process is applied to the polynomial functions $1, t, t^2, \ldots$, in the space $L^2_\rho(-1, 1)$, we obtain the *Jacobi polynomials*, up to scalar multiples. Completeness of the orthonormal system so obtained is proved in Szegö [1, pages 39–40].

The particular form of the Jacobi polynomials depends on α and β. For example, if $\alpha = \beta = 0$, then $\rho(t) = 1$ and we have the *Legendre polynomials*. If $\alpha = \beta = -\frac{1}{2}$, then $\rho(t) = (1-t^2)^{-1/2}$ and the corresponding Jacobi polynomials are called the *Chebysheff polynomials*.

Example 3. In the space $L^2(-\infty, \infty)$ one example of a complete orthonormal system is furnished by the elements corresponding to the Hermite functions, which are defined in terms of the Hermite polynomials. These polynomials are

$$H_n(t) = (-1)^n e^{t^2}\frac{d^n}{dt^n}e^{-t^2}, \qquad n = 0, 1, 2, \ldots,$$

and the Hermite function ψ_n is defined as

$$\psi_n(t) = (2^n n!\sqrt{\pi})^{-1/2} H_n(t) e^{-t^2/2}.$$

The fact that the ψ_n's are pairwise orthogonal is a consequence of the fact that

$$\psi_n''(t) - t^2\psi_n(t) = -(2n+1)\psi_n(t).$$

For a proof that the ψ_n's form a complete orthonormal set see Akhiezer and Glazman [1, § 11].

The Hermite polynomials also arise by applying the Gram–Schmidt process to the polynomials $1, t, t^2, \ldots$, in the space $L^2_\rho(-\infty, \infty)$, where

$$\rho(t) = e^{-t^2}.$$

Example 4. Another interesting example is that of the Laguerre functions

$$\phi_n(t) = \frac{1}{n!} e^{-t/2} L_n(t), \qquad n = 0, 1, 2, \ldots,$$

where

$$L_n(t) = e^t \frac{d^n}{dt^n}(t^n e^{-t})$$

is the Laguerre polynomial of degree n. The functions $\{\phi_n\}$ determine a complete orthonormal system in the space $L^2(0, \infty)$. See Courant and Hilbert [1, volume I, pages 95–96]; this proof shows that the closed linear manifold determined by the functions $\{\phi_n\}$ contains all the elements of $L^2(0, \infty)$ corresponding to piecewise continuous functions in $\mathscr{L}^2(0, \infty)$. Since elements of $L^2(0, \infty)$ corresponding to functions of this latter type are dense in $L^2(0, \infty)$, it follows by Theorem 7.3 that the Laguerre functions determine a complete orthonormal system.

Example 5. An orthonormal set in $L^2(0, 1)$ that is currently receiving much attention is the *Walsh system*. Introduced in 1922 by J. Walsh, the Walsh functions received little attention until the late 1960s. Now they are a fundamental tool for research and practical applications in fields such as coding theory, signal processing, and pattern recognition. The Walsh functions have properties similar to those of the trigonometric functions but are better adapted for computer applications because they are all step functions that assume only the values ± 1.

To construct the Walsh functions $\{W_n\}_0^\infty$, first define

$$r_k(t) = \text{sign}\,[\sin(2^k \cdot 2\pi t)], \qquad k = 0, 1, 2, \ldots,$$

where the value of $r_k(t)$ at a discontinuity is taken as the right-hand limit rather than zero. Now set $W_0(t) \equiv 1$. For $n \geqslant 1$, consider the binary representation of n,

$$n = \sum_{k=0}^{m} n_k \cdot 2^k, \qquad n_k \in \{0, 1\},$$

and let $W_n(t)$ be the product of the functions $r_k(t)$ whose indices k correspond to those n_k that are 1. Thus

$$W_0(t) \equiv 1, \qquad W_n(t) = \prod_{k=0}^{m} [r_k(t)]^{n_k}.$$

For completeness of $\{W_n\}$ see in Alexits [1, pages 60–61]. For historical background and applications of Walsh functions see Harmuth [1].

The constant function 1 and the functions r_k, $k = 0, 1, 2, \ldots$, form a subset of $\{W_n\}$ called the *Rademacher* functions. They were discovered by H. Rademacher in 1922, independently of Walsh. Although not a complete set, these orthonormal functions are often used in probability theory.

We shall see later on (cf. Theorem VI.4.4 and the latter part of § V.7) that certain types of integral equations lead to the determination of complete orthonormal sets. This situation occurs, in particular, when certain types of boundary value problems for differential equations are recast as problems of integral equations (see § VI.5).

II.9 TOPOLOGICAL LINEAR SPACES

A topological linear space is a linear space X that is also a topological space such that addition and scalar multiplication are continuous operations; that is, if K is the underlying scalar field, then the mapping $(x, y) \mapsto x + y$ is continuous on $X \times X$ into X and the mapping $(\alpha, x) \mapsto \alpha x$ is continuous on $K \times X$ into X. The space K has its usual topology (that of the real or complex field), and the product spaces are given the usual topologies for Cartesian products. Throughout this section X will denote such a topological linear space.

As an immediate consequence of this definition, we see that for each $a \in X$, the translation mapping $f_a : X \to X$ defined by $f_a(x) = x + a$ is a homeomorphism of X onto X. Likewise, for each nonzero scalar α, the mapping $g_\alpha : X \to X$ defined by $g_\alpha(x) = \alpha x$ is a homeomorphism (for the inverse of g_α is $g_{1/\alpha}$).

For use in this section and elsewhere we introduce the following notations, where S, T are sets in X, a is a fixed vector, and α is a fixed scalar:

$$S + T = \{x + y : x \in S, y \in T\},$$

$$a + S = \{a + x : x \in S\},$$

$$\alpha S = \{\alpha x : x \in S\},$$
$$-S = (-1) \cdot S.$$

Lemma 9.1. *If U is an open (resp., closed) set in X, then $a + U$ and αU are open (resp., closed) for each $a \in X$ and each nonzero scalar α. If U is open and if F is an arbitrary subset of X, then $F + U$ is open.*

Proof. The first two statements follow from the remarks above. Given F and U, we have $F + U = \bigcup_{a \in F} (a + U)$. If U is open, then $F + U$ is a union of open sets and hence is open. □

Definition. A *neighborhood* of a point $x \in X$ is any set $V \subset X$ such that there is an open set $U \subset X$ with $x \in U \subset V$. The collection of all neighborhoods of x will be denoted by $\mathcal{N}(x)$. A subcollection $\mathcal{U}$ of $\mathcal{N}(x)$ is called a *base of neighborhoods of* x, or simply, a *base at* x, if given $V \in \mathcal{N}(x)$ there exists $U \in \mathcal{U}$ such that $U \subset V$.

If $\mathcal{U}$ is a base of neighborhoods of 0, then $\mathcal{N}(0)$ consists of all sets $V \subset X$ with the property that V contains some $U \in \mathcal{U}$. Since translation is a homeomorphism, we have $\mathcal{N}(x) = \{x + V : V \in \mathcal{N}(0)\}$. Thus from $\mathcal{U}$ we can recover $\mathcal{N}(x)$ for each x in X.

Definition. A set S in X is *balanced* if $\alpha s \in S$ whenever $s \in S$ and α is a scalar such that $|\alpha| \leq 1$. (Some authors use *circled*; the Bourbaki books use *équilibré*.) If S is any set in X, the *balanced hull* of S is the set $\{\alpha s : s \in S, |\alpha| \leq 1\}$.

The intersection of balanced sets is balanced. The balanced hull of S is the intersection of all balanced sets that contain S. In particular, a balanced set coincides with its balanced hull.

Definition. A set S in X is *absorbing* if to each $x \in X$ there corresponds some $\varepsilon > 0$ such that $\alpha x \in S$ if $0 < |\alpha| \leq \varepsilon$. An equivalent formulation is: S is absorbing if to each $x \in X$ there corresponds some $r > 0$ such that $x \in \alpha S$ if $|\alpha| \geq r$.

It is clear at once from the continuity of products and the fact that $0 \cdot x = 0$ for every x that each neighborhood of 0 is absorbing. A balanced set S is absorbing if and only if to each $x \in X$ there corresponds some $\alpha \neq 0$ such that $\alpha x \in S$. The concepts of balanced sets and absorbing sets do not require a topology for their definition; they make sense in any linear space.

We shall say that a topology τ for a linear space X is a *linear* topology if τ is compatible with the linear structure of X, that is, if τ makes X into a topological linear space. The next theorem shows how such a topology may be constructed for a linear space. The theorem will be used later, in the proof of Theorem 11.3.

Theorem 9.2. *A topological linear space X has a base $\mathscr{U}$ at 0 with the following properties:*

(*a*) *Each member of $\mathscr{U}$ is balanced and absorbing.*

(*b*) *If $U \in \mathscr{U}$, there exists $V \in \mathscr{U}$ such that $V + V \subset U$.*

(*c*) *If U_1 and U_2 are in $\mathscr{U}$, there exists $U_3 \in \mathscr{U}$ such that $U_3 \subset U_1 \cap U_2$.*

Conversely, given a nonempty collection $\mathscr{U}$ of nonempty subsets of a linear space X such that $\mathscr{U}$ satisfies (a) to (c), there is a unique linear topology for X having $\mathscr{U}$ as a base at 0.

Proof. If X is a topological linear space, let $\mathscr{U}$ be the family of all balanced neighborhoods of 0, which is the same as the family of balanced hulls of *all* neighborhoods of 0. We leave the verification of properties (*a*) to (*c*) to the reader.

Now suppose $\mathscr{U}$ is a collection with properties (*a*) to (*c*). For each $x \in X$, let $\mathscr{N}(x)$ be the family of all subsets of X that contain a set of the form $x + U$ for some $U \in \mathscr{U}$. We define a set S to be open if $S \in \mathscr{N}(x)$ for each $x \in S$. Clearly the empty set and the whole space X are open by this definition. It is easy to see that arbitrary unions of open sets are open. From property (*c*) one proves that finite intersections of open sets are open. Hence the open sets form a topology for X. One can show that for each x, $\mathscr{N}(x)$ is precisely the family of neighborhoods of x for this topology (see problem 2 at the end of this section). Thus it follows from the construction of the $\mathscr{N}(x)$ that $\mathscr{U}$ is a base of neighborhoods of 0.

Now we must prove that this topology is compatible with the linear structure of X. Continuity of addition in X follows from (*b*). Before proving continuity of scalar multiplication, we observe the following.

(9-1) If $U \in \mathscr{U}$ and $\alpha \neq 0$, there exists $V \in \mathscr{U}$ such that $\alpha V \subset U$.

To prove this we observe by (*b*) and induction that if $U \in \mathscr{U}$ and n is any positive integer, there exists $V \in \mathscr{U}$ such that $2^n V \subset U$. Now consider U and α in (9-1). Choose n so that $|\alpha| \leq 2^n$, and $V \in \mathscr{U}$ so that $2^n V \subset U$. Since V is balanced, $\alpha 2^{-n} V \subset V$; that is, $\alpha V \subset 2^n V \subset U$.

Now suppose α_0 and x_0 are fixed. We shall show that given $U \in \mathscr{U}$ there is a $V \in \mathscr{U}$ and an $\varepsilon > 0$ such that $\alpha x \in \alpha_0 x_0 + U$ (i.e., $\alpha x - \alpha_0 x_0 \in U$) whenever $|\alpha - \alpha_0| < \varepsilon$ and $x \in x_0 + V$. Now, for arbitrary α and x,

(9-2) $$\alpha x - \alpha_0 x_0 = \alpha(x - x_0) + (\alpha - \alpha_0) x_0.$$

We know from (*b*) that there exists $W \in \mathcal{U}$ such that $W + W \subset U$. Since W is absorbing, we may choose $\varepsilon > 0$ such that $\beta x_0 \in W$ if $|\beta| < \varepsilon$. Having chosen ε, we choose $V \in \mathcal{U}$ such that $(\varepsilon + |\alpha_0|)V \subset W$; this is possible, by (9-1). Now suppose that α and x are such that $|\alpha - \alpha_0| < \varepsilon$ and $x - x_0 \in V$, but otherwise arbitrary. Then $(\alpha - \alpha_0)x_0 \in W$. Clearly $|\alpha| \leq |\alpha - \alpha_0| + |\alpha_0| < \varepsilon + |\alpha_0|$. Since V is balanced, it follows that

$$\frac{\alpha}{\varepsilon + |\alpha_0|}(x - x_0) \in V, \qquad \text{or} \qquad \alpha(x - x_0) \in (\varepsilon + |\alpha_0|)V \subset W.$$

But then, by (9-2), $\alpha x - \alpha_0 x_0 \in W + W \subset U$. This completes the proof that multiplication is continuous.

The uniqueness assertion of the theorem follows from the fact that the neighborhood system at each point is completely determined by a base of neighborhoods of 0. □

The definition of a topological linear space makes no provisions for the separation axioms of topology (cf. Taylor [5, pages 102–104]). Thus the space need not be a T_1-space. Consider, for instance, a linear space X having more than one element, with $\varnothing$ and X as the only open sets. The next theorem shows, however, that a topological linear T_1-space must necessarily be a T_3-space, and so, in particular, a Hausdorff space.

Theorem 9.3. *A topological linear space is regular. Hence it is a Hausdorff space if and only if the set* $\{0\}$ *is closed.*

Proof. Let U be an open neighborhood of 0. It follows from Theorem 9.2 that there is a neighborhood V of 0 such that $V + V \subset U$. The closure $\bar{V}$ of V is a closed neighborhood of 0, and we shall show that $\bar{V} \subset U$. If $x \in \bar{V}$, then $x - V$ is a neighborhood of x, and so $[x - V] \cap V \neq \varnothing$. Take y in $[x - V] \cap V$, and write $y = x - z$, $z \in V$. Then $x = y + z \in V + V \subset U$. This proves that U contains $\bar{V}$. Since translations are homeomorphisms, every open neighborhood of a point a must contain a closed neighborhood of a; that is, the space is regular. Consequently, the space is a Hausdorff space if and only if singleton points are closed. But this will be true if and only if the set $\{0\}$ is closed. □

It is also true that a topological linear T_1-space is completely regular (cf. Taylor [5, page 104]). For a proof of a somewhat more general result that is easily rephrased for the present situation, see Weil [1, page 13].

Theorem 9.4. *All* n*-dimensional topological linear* T_1*-spaces with the same scalar field are topologically isomorphic.*

This is a generalization of Theorem 3.1. A proof is outlined in problem 3. The T_1-space requirement is important, for the theorem is false without it (see problem 4).

It is not the purpose of this book to develop the general theory of topological linear spaces extensively. The principal interest is in normed linear spaces. However, it is essential for some purposes to make excursions into more general types of topological linear spaces, particularly those of the type known as *locally convex* spaces (see § 11). In Chapter III consideration is given to what are known as "weak topologies" on normed linear spaces. The use of such topologies is important for obtaining insights into the properties of Banach spaces and their conjugate spaces. Topological linear spaces with topologies not definable by means of a norm are of importance in the theory of distributions, with applications to partial differential equations. For accounts of the general theory of topological linear spaces, see Horváth [1], Köthe [1], and Schaefer [1]. For applications to distributions and partial differential equations, see Gelfand and Shilov [2, (especially volume 2)], Horváth [2], and Schwartz [1].

PROBLEMS

1. Let M be a linear manifold in a topological linear space. Show that the closure $\bar{M}$ of M is also a linear manifold.
2. Let $\mathscr{U}$ be a nonempty collection of nonempty subsets of a linear space X such that $\mathscr{U}$ satisfies (a) to (c) of Theorem 9.2. Define the sets $\mathscr{N}(x)$ and the topology on X as in the proof of Theorem 9.2. Given $x \in X$, show that $\mathscr{N}(x)$ coincides with the collection of all neighborhoods of x for that topology. [*Suggestion.* Given $x \in X$ and $V \in \mathscr{N}(x)$, let $V_1 = \{y : V \in \mathscr{N}(y)\}$. Show that V_1 is an open neighborhood of x contained in V. To show that V_1 is open, take $y \in V_1$ and use property (b) to produce $W \in \mathscr{U}$ such that $V \in \mathscr{N}(y+w)$ for all $w \in W$. It will follow that $y + W \subset V_1$. Hence $V_1 \in \mathscr{N}(y)$ for all $y \in V_1$.]
3. To prove Theorem 9.4 it suffices to show that an n-dimensional topological linear T_1-space X is topologically isomorphic to $\ell^2(n)$. If we write $x = \xi_1 x_1 + \cdots + \xi_n x_n$, where $x_1, \ldots, x_n$ is a basis for X, the main difficulty lies in proving that the point $(\xi_1, \ldots, \xi_n)$ of $\ell^2(n)$ depends continuously on x. [*Suggestions.* If $\varepsilon > 0$ is given, let K be the set of points in $\ell^2(n)$ for which $|\xi_1|^2 + \cdots + |\xi_n|^2 = \varepsilon^2$, and let S be the corresponding set in X. The fact that X is a T_1-space is used in proving that S is closed. It is possible to choose a balanced neighborhood U of 0 in X such that $U \cap S = \varnothing$. One can then show that $x \in U$ implies $|\xi_1|^2 + \cdots + |\xi_n|^2 < \varepsilon^2$.]
4. Give an example of a two-dimensional topological linear space that is not topologically isomorphic to $\ell^2(2)$.
5. If X and Y are topological linear spaces, a linear operator T on X into Y is continuous at all points of X if T is continuous at some point of X.

6. If X and Y are topological linear spaces and X is a finite-dimensional T_1-space, then every linear operator on X into Y is continuous.

7. Let f be a function from a real topological linear space X into $\boldsymbol{R}$ such that $f(x+y)=f(x)+f(y)$ for $x, y \in X$. If f is bounded on some neighborhood of 0 in X, then f is a continuous linear functional on X.

8. Let M be a closed subspace of a topological linear space X, and let ϕ be the canonical mapping of X onto the quotient space X/M (see § I.6). Given a set S in X/M, let $\phi^{-1}(S)=\{x \in X : \phi(x) \in S\}$, and define S to be open in X/M if $\phi^{-1}(S)$ is open in X. This defines the *quotient topology* on X/M.
   a. The quotient topology is the strongest topology on X/M relative to which ϕ is continuous.
   b. If U is an open set in X, then $\phi(U)$ is open in X/M; that is, ϕ is an open mapping. In fact, the quotient topology is the weakest topology on X/M relative to which ϕ is an open mapping.
   c. X/M is a topological linear T_1-space under the quotient topology.
   d. If the topology of X arises from a norm, the quotient topology on X/M is the same as that generated by the quotient norm (see § 5). [*Suggestion.* Given $\varepsilon>0$, let $B(x, \varepsilon)=\{y \in X : \|y-x\|<\varepsilon\}$ and let $B([x], \varepsilon)$ be defined similarly. Show that $B([x], \varepsilon) \subset \phi(B(x, \varepsilon))$.]

9. There is a generalization of the concept of completeness for a topological linear space. Since the topology may not be that of a metric space, it turns out to be desirable to use Moore–Smith convergence instead of sequential convergence in defining completeness. We rely on the terminology concerning *directed sets* and *nets* used by Kelley [1, Chapter 2], in connection with Moore–Smith convergence. Suppose X is a topological linear space. A net f in X (i.e., a function with range in X and domain some directed set D) is called a *Cauchy net* if for each neighborhood U of 0 in X there is some $d_0 \in D$ such that $f(d_1)-f(d_2) \in U$ if $d_0 < d_1$ and $d_0 < d_2$. The net is said to converge to x if for each neighborhood U of 0 there is some $d_0 \in D$ such that $f(d)-x \in U$ when $d_0 < d$. If a net converges to a point, it is a Cauchy net. If every Cauchy net in X converges to a point in X, we say that X is complete. An alternative approach to completeness is through the concept of *Cauchy filters* (see Bourbaki [1, Chapter II] and Horváth [1, pages 75–79, 124–135].
   a. If X and Y are topological linear spaces that are topologically isomorphic, they are both complete if one is.
   b. For a normed linear space the present definition of completeness is equivalent to the notion of completeness for metric spaces.
   c. Let M be a closed subspace of the topological linear space X. If X is complete, so is M. If X is a T_1-space and if M is complete, it is closed in X, regardless of whether X is complete. For the proof it is necessary to know that if $S \subset X$, a point x is in $\bar{S}$ if and only if there is a net in S converging to x. Also, when X is T_1, it is also T_2, and then a net in X cannot converge to two different points.
   d. A finite-dimensional subspace of a topological linear T_1-space is closed.
   e. A compact topological linear T_1-space consists of only the zero vector.
   f. If X is a topological linear T_1-space that is locally compact, it is finite

dimensional. We sketch the argument. Using the local compactness, we obtain a balanced compact neighborhood V of 0. Let U be the interior of V. Then there exist $a_1, \ldots, a_n \in V$ such that V is covered by the union of the sets $a_i + \frac{1}{2}U$ $(1 \leq i \leq n)$. Let M be the linear manifold generated by $a_1, \ldots, a_n$. Then M is closed in X and X/M is a T_1-space. Let ϕ be the canonical mapping of X onto X/M. Since $M + \frac{1}{2}V \supset V$, it follows that $\phi(V) = \frac{1}{2}\phi(V)$. From this, one can show that $\phi(V) = X/M$. Finally, we observe that $\phi(V) = X/M$ is a compact topological linear T_1-space, so that X/M is zero dimensional; that is, $X = M$. Thus X is finite dimensional.

II.10 CONVEX SETS

Let X be a linear space. If x_1 and x_2 are distinct points of X and if α is a real number with $0 \leq \alpha \leq 1$, the set of all points $(1-\alpha)x_1 + \alpha x_2$ is called the *line segment* joining x_1 and x_2, or simply, the line segment x_1x_2. The line segment x_1x_2 with the two endpoints omitted is called the *open line segment* x_1x_2.

Definition. A set S in X is *convex* if S contains the line segment x_1x_2 whenever x_1 and x_2 are distinct points of S. This is equivalent to the requirement that if $x_1, x_2 \in S$, then also $\alpha_1x_1 + \alpha_2x_2 \in S$ whenever α_1 and α_2 are positive numbers such that $\alpha_1 + \alpha_2 = 1$. A set S in X is *absolutely convex* if S is both convex and balanced. This is equivalent (cf. Theorem 10.3) to the requirement that if $x_1, x_2 \in S$, then also $\lambda x_1 + \mu x_2 \in S$ whenever λ and μ are scalars such that $|\lambda| + |\mu| \leq 1$.

Note that if X is a complex linear space, the condition that a set S be balanced involves complex scalars, but the convexity of S still involves only (positive) real scalars.

The empty set and a set consisting of one point are convex, since the conditions of the definition are vacuously satisfied by such sets. Similarly, the empty set and the set whose only element is the zero vector are absolutely convex. A linear manifold is absolutely convex. The intersection of a family of convex (resp. absolutely convex) sets is convex (resp. absolutely convex).

In a normed linear space, open balls $\{x : \|x - x_0\| < r\}$ and closed balls $\{x : \|x - x_0\| \leq r\}$ are convex. Open balls and closed balls are absolutely convex if and only if they are centered at the origin.

Theorem 10.1. *Let α and β be scalars. Let S and T be convex (absolutely convex) sets, then $\alpha S + \beta T$ is convex (absolutely convex). If α and β are real and positive and if S is convex, then $\alpha S + \beta S = (\alpha + \beta)S$.*

Proof. The first two statements are immediate consequences of the definitions of convexity and absolute convexity. The third statement is proved by noting that, if S is convex,

$$\alpha S+\beta S=(\alpha+\beta)\left[\frac{\alpha}{\alpha+\beta}S+\frac{\beta}{\alpha+\beta}S\right]=(\alpha+\beta)S. \quad \square$$

Definition. If S is any set in X, the *convex hull* of S is defined to be the intersection of all convex sets that contain S. We sometimes denote it by S_c.

Theorem 10.2. *The convex hull of S consists of all points that are expressible in the form $\alpha_1x_1+\cdots+\alpha_nx_n$, where $x_1,\ldots,x_n$ are any points of S, $\alpha_k>0$ for each k and $\Sigma_k\,\alpha_k=1$. The index n is not fixed.*

Proof. Let T be the set of points expressible in the manner described; that is, let T be the set of all finite *convex linear combinations* of elements of S. It is readily seen by the definition of convexity that T is convex. Since T clearly contains S, T must contain the convex hull of S. On the other hand, any convex set W containing S must contain all finite convex linear combinations of elements of S; that is, $W\supset T$. (Use finite induction on the number of elements used in the convex linear combinations.) It follows that T is contained in the convex hull of S. $\square$

Definition. If S is a set in X, let S_b denote the balanced hull of S, and let S_{bc} denote the convex hull of S_b. The set S_{bc} is called the *balanced and convex hull of S*, or the *absolutely convex hull of S*.

This name for S_{bc} is justified by the final assertion in the following theorem.

Theorem 10.3. *If $S\subset X$, the set S_{bc} consists of all finite sums $\Sigma\,\alpha_kx_k$, where the x_k's are elements of S and $\Sigma\,|\alpha_k|\leqslant 1$. S_{bc} is the intersection of all absolutely convex sets that contain S.*

Proof. Let T be the set described in the theorem. We recall that $S_b=\{\alpha x:|\alpha|\leqslant 1, x\in S\}$. If $x\in S_{bc}$, we know from Theorem 10.2 that x is a finite sum, say

$$(10\text{-}1)\qquad x=\sum_1^n \beta_k\alpha_kx_k,\qquad\text{with}\qquad |\alpha_k|\leqslant 1,\qquad \beta_k>0,\qquad \sum_1^n\beta_k=1,$$

and $x_k\in S$, $k=1,\ldots,n$. It is thus evident that $S_{bc}\subset T$. Now suppose $x\in T$, say $x=\Sigma\,\lambda_kx_k$, where $x_k\in S$ and $\lambda=\Sigma\,|\lambda_k|\leqslant 1$. If $\lambda=0$, then $x=0\in S_b$, so $x\in S_{bc}$. If $\lambda\neq 0$, we must show that x may be written in the form (10-1). This is accomplished by letting $\beta_k=|\lambda_k|/\lambda$ and $\alpha_k=(\lambda\lambda_k)/|\lambda_k|$. Thus $x\in S_{bc}$, which proves that $T=S_{bc}$. It is clear from the form of the elements of T that S_{bc} is balanced. It is of course convex. Verification of the last assertion in the theorem is left to the reader. $\square$

The reader can easily construct a set in the plane ($\boldsymbol{R}^2$) that is convex but whose balanced hull (using real scalars) is not convex. This will show that the balanced hull of the convex hull of S is not always the same as S_{bc}.

Next, we consider convex sets in a topological linear space X. We denote the interior and closure of a set S in X by $\operatorname{int}(S)$ and $\bar{S}$, respectively.

Theorem 10.4. *Let S be a subset of a topological linear space X. (a) If S is convex, then so are* $\operatorname{int}(S)$ *and $\bar{S}$. (b) If S is balanced, then so is $\bar{S}$. (c) If S is absolutely convex, then so are* $\operatorname{int}(S)$ *and $\bar{S}$.*

Proof. Let $U = \operatorname{int}(S)$. Then for $0 < \alpha < 1$,

$$(1-\alpha)U + \alpha U \subset S$$

because S is convex. Now $(1-\alpha)U + \alpha U$ is open, by Lemma 9.1, and hence must be contained in the interior U of S. This implies that $\operatorname{int}(S)$ is convex. Next, take $x, y \in \bar{S}$ and $0 < \alpha < 1$. Let W be any neighborhood of $(1-\alpha)x + \alpha y$. If we show that W contains a point in S, it will follow that $(1-\alpha)x + \alpha y \in \bar{S}$ and hence that $\bar{S}$ is convex. Now the function $f(u, v) = (1-\alpha)u + \alpha v$ is continuous, and so there exist neighborhoods U and V of x and y, respectively, such that $f(u, v) \in W$ if $u \in U$ and $v \in V$. Since $x, y \in \bar{S}$, there exist $x_1 \in U \cap S$ and $y_1 \in V \cap S$. Then $f(x_1, y_1) \in W$. Also, $f(x_1, y_1) \in S$ because S is convex. This completes the proof that $\bar{S}$ is convex. The proofs of (b) and (c) are left as exercises for the reader. $\square$

If S is any set in X, the closure of the convex hull of S is called the *closed convex hull* of S. It is the intersection of all closed and convex sets that contain S. The sets S and $\bar{S}$ have the same closed convex hull. Similar statements hold for the *closed absolutely convex hull* of S.

The next theorem is used in the proof of Theorem 12.5 (and in problem 3 of § 11). The result would be obvious and a formal statement unnecessary if we confined ourselves to normed linear spaces.

Theorem 10.5. *If X is a topological linear space and U is a convex neighborhood of* 0, *then U contains an absolutely convex open neighborhood of* 0.

Proof. By Theorem 9.2, U contains a balanced neighborhood V of 0. The convex hull W of V is also a neighborhood of 0. Since V is its own balanced hull, W is the absolutely convex hull of V. Since U is convex, it follows that $W \subset U$. Finally, $\operatorname{int}(W)$ is an open neighborhood of 0 (since W is a neighborhood) which is absolutely convex. $\square$

Dirichlet's Principle

The following discussion of Dirichlet's principle illustrates how the concepts of convexity and orthogonality in an inner-product space are useful in presenting a famous problem in classical analysis from a linear space point of view.

Let R be a bounded open set with boundary $B(R)$ in Euclidean space of three dimensions. We suppose, furthermore, that $B(R)$ is smooth enough to enable us to apply the divergence theorem to vector fields that are continuously differentiable in $\bar{R}$. In particular, then, with adequate continuity and differentiability we have the identity of Green

$$\int_R g\nabla^2 h\,dV + \int_R \nabla g\cdot\nabla h\,dV = \int_{B(R)} g\frac{\partial h}{\partial n}\,dS, \tag{10-2}$$

where g and h are real-valued functions defined in R, ∇^2 denotes the Laplacian, ∇ the gradient, and n the outward normal to $B(R)$.

We define a real inner-product space $\mathscr{F}$ as follows: The elements f of $\mathscr{F}$ are the real-valued functions that are twice continuously differentiable in $\bar{R}$. The inner product is

$$(f_1, f_2) = \int_{B(R)} f_1 f_2\,dS + \int_R \nabla f_1\cdot\nabla f_2\,dV. \tag{10-3}$$

The space $\mathscr{F}$ is not complete. Let $\mathscr{G}$ be the subspace of those elements $g\in\mathscr{F}$ that vanish on $B(R)$, and let $\mathscr{H}$ be the subspace of those elements $h\in\mathscr{F}$ that are harmonic (i.e., $\nabla^2 h=0$) in R. From (10-2) and (10-3) we see that $(g, h)=0$ if $g\in\mathscr{G}$ and $h\in\mathscr{H}$; that is, $\mathscr{G}$ and $\mathscr{H}$ are orthogonal.

Now consider any $h\in\mathscr{H}$ and any $f\in\mathscr{F}$ such that $f=h$ on $B(R)$. Let $g=f-h$. Then $g\in\mathscr{G}$. Since $f=g+h$, the orthogonality shows that

$$\|f\|^2 = \|g\|^2 + \|h\|^2,$$

and hence $\|h\|\le\|f\|$. Thus among all elements of $\mathscr{F}$ that are equal to the given h on $B(R)$, h has the smallest norm. This characterization of h by a minimal property is what is known as *Dirichlet's principle* (though originally the principle was not stated in the language of norms).

In the early work in the calculus of variations, attempts were made to use Dirichlet's principle to prove the existence of a solution of the Dirichlet problem. Let us see how we might attempt to give an existence proof by linear-space methods. Suppose f_0 is a given element of $\mathscr{F}$. Denote by K the set of elements of $\mathscr{F}$ that coincide with f_0 on $B(R)$. This set is precisely the set $f_0+\mathscr{G}$. Since $\mathscr{G}$ is a closed linear manifold in $\mathscr{F}$, K is a closed set, and it is easy to see that K is convex. If f_0 does not vanish identically on $B(R)$, the distance d from 0 to K is positive, and there exists a sequence $\{f_n\}$ in K such that $\|f_n\|\to d$. It turns out that, because K is convex, $\{f_n\}$ is necessarily a Cauchy sequence (see problem 6). But since $\mathscr{F}$ is not complete, we cannot be certain that the sequence has a limit in $\mathscr{F}$. (It is precisely here that this method fails in its attempt to provide an existence proof for the Dirichlet problem.) If a limit h *does* exist in $\mathscr{F}$, then $h \in K$. In this case, the fact that h is the element of K with minimal norm allows us to conclude, by the standard reasoning of the calculus of variations, that h satisfies Laplace's equation, which is the Euler equation for this situation. Thus we can say: If $f_0 \in \mathscr{F}$, the Dirichlet problem for the region R, with the values of f_0 on $B(R)$ as assigned boundary values, is solvable if and only if there is a point of $f_0 + \mathscr{G}$ at minimal distance from 0. This "point" is then the solution of the Dirichlet problem.

It is interesting to observe that if the Dirichlet problem is solvable, in the sense just described, for each $f \in \mathscr{F}$, we can write

$$\mathscr{F} = \mathscr{G} \oplus \mathscr{H}.$$

Hence, if $f = g + h$, where $g \in \mathscr{G}$ and $h \in \mathscr{H}$, then h is the solution of the Dirichlet problem with boundary values of f. We can describe h as the orthogonal projection of $\mathscr{F}$ on $\mathscr{H}$.

This concept of orthogonal projections in connection with the Dirichlet problem became very well known in the 1930s, and was utilized in various ways by research workers in the calculus of variations. It led naturally to various schemes for overcoming the difficulties arising from an incomplete space. One line of development was that in which the space was taken to be complete and the basic Dirichlet problem was modified by generalizing the sense in which a function takes on prescribed values at the boundary of a region. See the paper of Weyl [1, especially pages 411–414] for such an approach.

PROBLEMS

1. Let S be a convex set in a topological linear space X. Suppose $x_0 \in \operatorname{int}(S)$ and $y_0 \in \bar{S}$. Then every point y expressible in the form $y = \alpha x_0 + (1-\alpha)y_0$, $0 < \alpha < 1$, is an interior point of S.
2. Under the conditions on S and X in problem 1, if $\operatorname{int}(S) \neq \varnothing$, we have $\overline{\operatorname{int}(S)} = \bar{S}$ and $\operatorname{int}(S) = \operatorname{int}(\bar{S})$.
3. Let $A_1, \ldots, A_n$ be compact convex sets in a Hausdorff topological linear space X. Let $A = A_1 \cup \cdots \cup A_n$. Then the convex hull A_c is compact. Begin by showing that A_c is composed of all the points x that are expressible in the form $x = \alpha_1 x_1 + \cdots + \alpha_n x_n$, with $x_k \in A_k$, $0 \leq \alpha_k$, and $\alpha_1 + \cdots + \alpha_n = 1$. Then let B be the compact set in $\ell^2(n)$ composed of all points $(\alpha_1, \ldots, \alpha_n)$ with $\alpha_k \geq 0$ and $\alpha_1 + \cdots + \alpha_n = 1$. Observe that A_c is the image in X of the product set $B \times A_1 \times \cdots \times A_n$ by the mapping $(\alpha_1, \ldots, \alpha_n, x_1, \ldots, x_n) \mapsto \alpha_1 x_1 + \cdots + \alpha_n x_n$.
4. In a normed linear space X the convex hull of a precompact set S is precompact. (*Note.* S is precompact if to each $\varepsilon > 0$ corresponds some finite

set $\{x_1, \ldots, x_n\}$ in X such that S is contained in $\bigcup_{i=1}^{n} B(x_i, \varepsilon)$, where $B(x_i, \varepsilon) = \{y \in X : \|y - x_i\| < \varepsilon\}$.) *Suggestion.* Use problem 3. Given $\varepsilon > 0$, choose $A = \{x_1, \ldots, x_n\}$ so that S is contained in $B = \bigcup_{1}^{n} B(x_i, \varepsilon/2)$. Show that B_c is contained in the union of all open balls of radii $\varepsilon/2$ with centers in A_c. Then observe that A_c is compact and so deduce that S_c is precompact.

5. In a Banach space the closed convex hull of a compact set is compact. Use problem 4.
6. Let K be a nonempty convex set in an inner-product space X, and let $d = \inf\{\|x\| : x \in K\}$.
   a. If $\{x_n\}$ is a sequence in K such that $\|x_n\| \to d$, then $\{x_n\}$ is a Cauchy sequence. [*Hint.* Show that $\|x_n + x_m\| \geqslant 2d$ and use the parallelogram law, page 75.]
   b. If X is complete, then there exists a unique $x \in K$ such that $\|x\| = d$.

II.11 LOCALLY CONVEX SPACES

A topological linear space is said to be a *locally convex space* if every neighborhood of 0 contains a convex neighborhood of 0. In this section we discuss a few basic facts about such spaces and describe several important examples. The theory of locally convex spaces will be developed more fully in parts of Chapter III.

If X is a locally convex space and $\mathcal{U}$ is the family of all absolutely convex open neighborhoods of 0, then $\mathcal{U}$ is a base at 0 (by Theorem 10.5). In practice, the description of a base at 0 is often given in terms of seminorms.

Seminorms

Let X be a linear space (real or complex). A *seminorm* on X is a real-valued function p defined on X such that

1. $p(x+y) \leqslant p(x) + p(y)$ if $x, y \in X$.
2. $p(\alpha x) = |\alpha| p(x)$ if $x \in X$ and α is any scalar.

If p has the further property that $p(x) \neq 0$ if $x \neq 0$, then p is a *norm*. As with a norm, the properties of p imply the further properties $p(x) \geqslant 0$, $p(0) = 0$, and

$$|p(x) - p(y)| \leqslant p(x - y). \tag{11-1}$$

An important example of a seminorm is given by $p(x) = |x'(x)|$, where x' is a linear functional on X.

Lemma 11.1. *Let p be a seminorm on a linear space X. Then the sets*

$$V_1 = \{x : p(x) < 1\} \qquad \text{and} \qquad V_2 = \{x : p(x) \leqslant 1\}$$

are absorbing and absolutely convex.

Proof. If $x, y \in V_1$ and if λ and μ are scalars such that $|\lambda|+|\mu|\leqslant 1$, then

$$\begin{aligned} p(\lambda x+\mu y) &\leqslant |\lambda| p(x)+|\mu| p(y) \\ &\leqslant |\lambda|+|\mu| \leqslant 1. \end{aligned}$$

Thus $\lambda x+\mu y \in V_1$, and so V_1 is absolutely convex. Similarly, V_2 is absolutely convex. The fact that V_1 and V_2 are absorbing follows from the homogeneity property (2) of the seminorm p. □

Lemma 11.2. *Let p be a seminorm on a topological linear space. Then p is continuous from X into $\boldsymbol{R}$ if and only if $V_2=\{x : p(x)\leqslant 1\}$ is a neighborhood of 0 in X.*

Proof. If p is continuous, then the set V_1 of Lemma 11.1 is clearly open. Since $0 \in V_1 \subset V_2$, V_2 is a neighborhood of 0. Conversely, if V_2 is a neighborhood of 0, so is εV_2 for each $\varepsilon>0$. It is readily seen that $\varepsilon V_2=\{x : p(x)\leqslant \varepsilon\}$. From (11-1) it follows that for each $a \in X$ and each x in $a+\varepsilon V_2$, we have $|p(x)-p(a)|\leqslant p(x-a)\leqslant \varepsilon$ because $x-a$ is in εV_2. This implies that p is continuous at a. □

It follows from this lemma that if p and q are seminorms such that q is continuous and $p(x)\leqslant q(x)$ for all $x \in X$, then p is continuous. This is because $\{x : q(x)\leqslant 1\}\subset\{x : p(x)\leqslant 1\}$.

Theorem 11.3. *Let $\mathscr{P}$ be a nonempty family of seminorms on a linear space X. For each $p \in \mathscr{P}$ let $V(p)$ be the set $\{x : p(x)<1\}$. Let $\mathscr{U}$ be the family of all finite intersections*

$$r_1 V(p_1) \cap r_2 V(p_2) \cap \cdots \cap r_n V(p_n), \qquad r_k>0, \qquad p_k \in \mathscr{P}. \tag{11-2}$$

Then $\mathscr{U}$ is a base at 0 for a topology that makes X into a locally convex space. Furthermore, this topology is the weakest linear topology for X with respect to which all the seminorms in $\mathscr{P}$ are continuous.

Proof. It follows easily from Lemma 11.1 that the members of $\mathscr{U}$ are absolutely convex and absorbing. Furthermore, if $U \in \mathscr{U}$, it is easily seen that $V=\frac{1}{2}U \in \mathscr{U}$ and $V+V \subset U$ (use Theorem 10.1). Thus $\mathscr{U}$ satisfies conditions (a) and (b) of Theorem 9.2. Condition (c) is obviously satisfied, because $U_1 \cap U_2 \in \mathscr{U}$ whenever $U_1, U_2 \in \mathscr{U}$. Since the members of $\mathscr{U}$ are convex, it follows from Theorem 9.2 that $\mathscr{U}$ determines a topology for X that makes X into a locally convex space having $\mathscr{U}$ as a base at 0. By Lemma 11.2, each $p \in \mathscr{P}$ is continuous with respect to this topology. Now let τ be a topology making X into a topological linear space such that each $p \in \mathscr{P}$ is continuous. Then, for $r>0$, the sets $rV(p)$ are (open) τ-neighborhoods of 0, and hence each $U \in \mathscr{U}$ is a τ-neighborhood of 0. Thus the topology on X determined by $\mathscr{U}$ is weaker than τ. □

Definition. The topology for X described in Theorem 11.3 is called the topology generated by the family $\mathscr{P}$ of seminorms.

Observe that if $\mathscr{P}$ contains just one element p and if p is a norm, then $\mathscr{U}$ is the family of all spheres $\{x : p(x) < r\}$, and the topology for X generated by $\mathscr{P}$ is the topology of X as a normed linear space with norm p.

Let $\mathscr{P}$ be a family of seminorms on a linear space X, and let τ be the topology generated by $\mathscr{P}$. For technical reasons it is sometimes convenient to describe τ in terms of a base of *closed* τ-neighborhoods of 0. Such a base is the collection $\mathscr{V}$ of finite intersections of the form (11-2), where $V(p_k)$ is now interpreted to be the set $\{x : p_k(x) \leqslant 1\}$. (To prove this, observe that the sets in $\mathscr{V}$ are τ-neighborhoods of 0 by Lemma 11.2, and each set $\{x : p(x) < r\}$ contains a set of the form $\{x : p(x) \leqslant r/2\}$.)

Theorem 11.4. *Let X be a locally convex space whose topology is generated by a family $\mathscr{P}$ of seminorms. Then X is a Hausdorff space if and only if to each $x \neq 0$ there corresponds $p \in \mathscr{P}$ such that $p(x) \neq 0$.*

The proof is left to the reader.

Continuous Linear Operators

If T is a linear operator from a topological linear space X into a locally convex space Y, it is sometimes useful to discuss the continuity of T in terms of seminorms. As with normed linear spaces, it is enough to know about the continuity of T at $0 \in X$. For, suppose that T is continuous at 0, and let x be an element of X. Every neighborhood of Tx is of the form $Tx + V$, where V is a neighborhood of 0 in Y. Let U be a neighborhood of $0 \in X$ such that $T(U) \subset V$; then $T(x + U) = Tx + T(U) \subset Tx + V$, which shows that T is continuous at x.

Theorem 11.5. *Let X be a topological linear space, let Y be a locally convex space, and let $\mathscr{P}$ be a family of seminorms that generates the topology of Y. A linear operator T on X into Y is continuous if and only if the composite mapping $p \circ T$ is continuous for each $p \in \mathscr{P}$.*

Proof. If T is continuous and if $p \in \mathscr{P}$, then $p \circ T$ is certainly continuous. (It is also a seminorm.) Conversely, suppose that $p \circ T$ is continuous for each $p \in \mathscr{P}$, and let V be a neighborhood of 0 in Y. Then V contains a neighborhood W of the form (11-2). Using the notation of (11-2), we define a seminorm p by

$$p(y) = \max_k \frac{1}{r_k} p_k(y), \qquad y \in Y.$$

It is clear that $p \circ T$ is continuous, since each $p_k \circ T$ is continuous, by hypothesis. Hence the set $U = \{x : p(Tx) < 1\}$ is an open neighborhood of 0 in X. Observe that

$$T(U) \subset \{y : p(y) < 1\} \subset W \subset V.$$

This proves that T is continuous at 0. □

Examples of Locally Convex Spaces

Example 1. Let T be any nonempty set. Let X be a linear space composed of real-valued (or complex-valued) functions x defined on T. X need not comprise *all* functions defined on T. For example, X might consist of all bounded functions or of all functions meeting some prescribed condition, provided that $x+y$ and αx meet the condition whenever x and y do. As a family of seminorms, take

$$p_t(x) = |x(t)|$$

for each $t \in T$. This family generates a locally convex topology for X. With this topology, X is a Hausdorff space, by Theorem 11.4. We note that $x_n \to x$ (convergence in the sense of this topology) means that $x_n(t) \to x(t)$ (convergence in the usual sense for scalars) for each $t \in T$. For this reason the topology is called the topology of *pointwise convergence*.

There is an aspect of this topology for X that is interesting and significant. For each $t \in T$ the function $x \mapsto x(t)$ is continuous on X for the topology just defined. For if $x_0 \in X$ and $\varepsilon > 0$ and U is the neighborhood of 0 defined by $U = \{x : p_t(x) < \varepsilon\}$, we see that $|x(t) - x_0(t)| < \varepsilon$ means the same thing as $x \in x_0 + U$. On the other hand, suppose that τ is any topology for X that has the property that the mappings $x \mapsto x(t)$ are all continuous. Then, given $x_0 \in X$, a basic open neighborhood of x_0 for the topology of pointwise convergence has the form

$$V = x_0 + \bigcap_{k=1}^{n} \{x : p_{t_k}(x) < r_k\} = \bigcap_{k=1}^{n} \{x : p_{t_k}(x - x_0) < r_k\}$$

$$= \bigcap_{k=1}^{n} \{x : |x(t_k) - x_0(t_k)| < r_k\},$$

for $t_1, \dots, t_n \in T$ and $r_1, \dots, r_n > 0$. For each k, the set $\{x : |x(t_k) - x_0(t_k)| < r_k\}$ is τ-open, because the function $x \mapsto x(t_k)$ is continuous. Hence V is τ-open. This shows that the topology of pointwise convergence on X is the weakest topology for X with respect to which $x(t)$ is a continuous function of x for each t. Accordingly, this topology is also called the *weak topology of X as a space of functions.*

One specific example of T and X is that in which T is the set of positive integers and X is the space of *all* sequences $x=(\xi_1, \xi_2, \ldots)$. The weak topology of X arises from the demand that each "coordinate" ξ_n be a continuous function of x.

Example 2. Let T be a topological space. Let X be the linear space of all continuous real-valued (or complex-valued) functions on T. For each compact set $K \subset T$, define

$$q_K(x) = \sup_{t\in K} |x(t)|, \qquad x \in X.$$

The (locally convex) topology on X generated by this family of seminorms is the topology of *uniform convergence on compact sets.* When X is given this topology, it is frequently denoted by $\mathscr{C}(T)$. In applications to partial differential equations, T is frequently taken to be an open subset Ω of $\boldsymbol{R}^n$.

If T is not compact and if X is given the topology of *uniform convergence* (on T), then the resulting space is not in general a topological linear space. For example, if $T=\boldsymbol{R}$, then the topology of uniform convergence is determined by the metric

$$d(x, y) = \min\{1, \sup_{t\in \boldsymbol{R}} |x(t)-y(t)|\}.$$

Now let $x_n \in X$ be defined by $x_n(t)=t/n$, $n=1, 2, \ldots$. Then $d(x_n, 0)=1$, so the sequence $\{x_n\}$ cannot converge to 0. If X were a topological linear space, scalar multiplication would be continuous and, if $x(t)=t$, then the sequence $\{(1/n)x\}$ would converge to 0. This contradiction shows the topology of uniform convergence is incompatible with the linear structure of X.

Example 3. If f is a continuous function defined on an open subset Ω of $\boldsymbol{R}^n$, the *support* of f is the closure in $\boldsymbol{R}^n$ of the set $\{x : f(x)\neq 0\}$. Let X be the linear space of all continuous functions on Ω whose support is a compact subset of Ω (depending on the function in question). A norm topology may be put on X by defining

$$\|f\|_\Omega = \sup_{x\in\Omega} |f(x)|.$$

The finiteness of $\|f\|_\Omega$ stems from the fact that f has compact support. Each neighborhood of 0 for this topology contains a ball $\{f\in X : \|f\|_\Omega < r\}$ for some $r>0$. (This is the topology of uniform convergence on Ω.) A stronger and more useful topology on X may be described as follows. Let $\mathscr{U}$ consist of all absolutely convex absorbing sets U in X that have the property that for each compact set K in Ω, there exists an $r>0$ such that U contains the set

$$\{f\in X : \text{the support of } f \text{ is in } K, \|f\|_\Omega < r\}.$$

Then $\mathscr{U}$ satisfies the conditions of Theorem 9.2 and so is a base at 0 for a (locally convex) topology on X. The space of all continuous functions whose support is a compact set in Ω, together with the topology determined by $\mathscr{U}$, is often denoted by $\mathscr{K}(\Omega)$.

Example 4. In order to construct this example, it is necessary to introduce some notation. The set of all n-tuples $\ell=(\ell_1,\ldots,\ell_n)$, where each ℓ_i is a natural number, will be denoted by N^n, and each element of N^n will be called a *multi-index*. The *order* $|\ell|$ of a multi-index $\ell=(\ell_1,\ldots,\ell_n)$ is the number $|\ell|=\ell_1+\cdots+\ell_n$. If f is a function of the n variables $x_1,\ldots,x_n$ and if f has at least $|\ell|$ continuous partial derivatives, then we let

$$\partial^\ell f=\frac{\partial^{|\ell|}f}{\partial x_1^{\ell_1}\partial x_2^{\ell_2}\ldots\partial x_n^{\ell_n}}.$$

Now let Ω be an open set in R^n, let m be a natural number, and consider the linear space X of all real-valued (or complex-valued) functions defined on Ω such that the partial derivatives $\partial^\ell f$ exist and are continuous on Ω for all $\ell\in N^n$ with $|\ell|\leq m$. For each compact subset K of Ω and each multi-index $\ell\in N^n$, $|\ell|\leq m$, we define a seminorm on X by

$$p_{K,\ell}(f)=\sup_{x\in K}|\partial^\ell f(x)|. \tag{11-3}$$

This space of functions with the locally convex topology generated by these seminorms is often denoted by $\mathscr{E}^m(\Omega)$. When $m=0$, this is just the space $\mathscr{C}(\Omega)$ of Example 2.

Example 5. A function f such that $\partial^\ell f$ is defined and continuous on Ω for all $\ell\in N^n$ is said to be *infinitely differentiable on* Ω. The space $\mathscr{E}(\Omega)$ is the collection of all such functions, together with the topology defined by the family of all seminorms (11-3), where K varies over all compact subsets of Ω and ℓ varies over all of N^n.

Example 6. The space $\mathscr{D}(\Omega)$ consists of all infinitely differentiable functions on Ω whose support is a compact subset of Ω, together with the locally convex topology determined by a base $\mathscr{U}$ at 0, which we shall now describe. Let $\mathscr{U}$ consist of all absolutely convex absorbing sets U in $\mathscr{D}(\Omega)$ that have the property that for each compact set K in Ω there exists $r>0$ and a natural number m such that U contains the set of all $f\in\mathscr{D}(\Omega)$ whose support is in K and $p_{K,\ell}(f)<r$, for $|\ell|\leq m$, where $p_{K,\ell}$ is given by (11-3).

The spaces in Examples 3 to 6 arise in the theory of distributions and have important applications to partial differential equations and quantum mechanics. We shall refer to these spaces again in § III.5. Additional information may be found in Horváth [1] and R. E. Edwards [1].

PROBLEMS

1. Suppose that p and q are seminorms on a linear space X such that for some $\alpha > 0$, $p(x) \leqslant \alpha$ whenever $q(x) \leqslant 1$. Show that $p(x) \leqslant \alpha q(x)$ for all x. [*Suggestion.* Fix x and take β such that $q(x) < \beta$. Deduce that $p(x) \leqslant \alpha\beta$, and from this conclude that $p(x) \leqslant \alpha q(x)$.]
2. Let $\mathscr{P}$ be a nonempty family of seminorms on a linear space X. For each nonempty finite subset $\{p_1, \ldots, p_n\}$ of $\mathscr{P}$, define

$$p(x) = \max\{p_1(x), \ldots, p_n(x)\}, \qquad x \in X.$$

Show that p is a seminorm. Let $\mathscr{P}_1$ be the family of all seminorms formed in this manner. Show that $\mathscr{P}$ and $\mathscr{P}_1$ generate the same topology for X, and show that a base at 0 for this topology is given by all sets of the form $\{x : p(x) < r\}$ where $r > 0$ and $p \in \mathscr{P}_1$.
3. Let X be a locally convex space. An absorbing, closed and absolutely convex subset of X is called a *barrel* (*tonneau* in French). Show that X has a base at 0 composed of barrels.
4. A topological linear space X is *barreled* (*tonnelé* in French) if every barrel in X is a neighborhood of 0. Let X be a locally convex space that is also a Baire space (i.e., every nonempty open subset of X is of second category in X). Prove that X is barreled. [If B is a barrel, then $X = \bigcup_{n=1}^{\infty} nB$.]
5. Let X be the space of all continuous functions on $[0, 1]$, endowed with the $\mathscr{L}^1(0, 1)$ norm. Let $B = \{x \in X : \sup_{0 \leqslant t \leqslant 1} |x(t)| \leqslant 1\}$. Then B is a barrel in X, but it is not a neighborhood of 0. [If B were a neighborhood of 0, the $\mathscr{L}^1(0, 1)$ norm and the sup norm would determine the same topology on X.]
6. Show that the topology on $\mathscr{K}(\Omega)$ defined in Example 3 is strictly stronger than the topology of uniform convergence on Ω. [Let g be a real-valued function that is continuous and strictly positive on Ω and such that $\inf\{g(x) : x \in \Omega\} = 0$. Let U be the set of all $f \in \mathscr{K}(\Omega)$ such that $|f(x)| \leqslant g(x)$ for $x \in \Omega$. Then U is an absolutely convex neighborhood of 0 in $\mathscr{K}(\Omega)$, but U is not a neighborhood of 0 for the topology of uniform convergence on Ω.]

II.12 MINKOWSKI FUNCTIONALS

In this section we shall show that the topology of any locally convex space is generated by a suitable family of seminorms. We shall also characterize normable linear spaces in the larger class of topological linear spaces.

Definition. Let K be a convex, absorbing set in a linear space X. For each $x \in X$ let A_x be the set of those real α such that $\alpha > 0$ and $x \in \alpha K$. Since K is absorbing, A_x is not empty. The *Minkowski functional* (or *gauge*) of K is the functional p defined by

$$p(x) = \inf A_x. \tag{12-1}$$

If X is a normed linear space and $K = \{x : \|x\| < r\}$, where r is fixed and positive, it is easy to see that $p(x) = \|x\|/r$. If K is a closed and bounded convex set in Euclidean space of n dimensions, if 0 is an interior point of K and $x \neq 0$, we can think of $p(x)$ as follows: Let y be that unique positive multiple of x in which the ray from 0 through x intersects the boundary of K. Then $x = p(x)y$, and $p(x)$ is the distance from 0 to x divided by the distance from 0 to y.

Theorem 12.1. *Let K be a convex, absorbing set. The Minkowski functional of K has the properties:*

(*a*) $p(0) = 0, \qquad p(\lambda x) = \lambda p(x)$ *if* $\lambda > 0$, *and*

(*b*) $p(x+y) \leq p(x) + p(y)$.

If K is also balanced, then p is a seminorm, that is, p also satisfies

(*c*) $p(\lambda x) = |\lambda| p(x)$.

Proof. If $\lambda > 0$, the conditions that $x \in \alpha K$ and $\lambda x \in \lambda \alpha K$ are clearly equivalent, so $p(\lambda x) = \lambda p(x)$. Since K is absorbing, $0 \in K$ and so $p(0) = 0$. Property (*b*) follows from the convexity of K and Theorem 10.1. For, if $x \in \alpha K$ and $y \in \beta K$ with $\alpha, \beta > 0$, then $x + y \in \alpha K + \beta K = (\alpha + \beta)K$, and so $p(x+y) \leq \alpha + \beta$. Taking the infimum of such α and such β, we obtain (*b*). If K is balanced, (*c*) will follow from (*a*) if we verify (*c*) in the case when $|\lambda| = 1$. But when $|\lambda| = 1$ and K is balanced, the conditions that $\lambda x \in \alpha K$ and $x \in \alpha K$ are equivalent, so (*c*) is true. □

Theorem 12.2. *Let K be a convex absorbing set. If p is the Minkowski functional of K, then $K_1 \subset K \subset K_2$, where*

$$K_1 = \{x : p(x) < 1\} \qquad \textit{and} \qquad K_2 = \{x : p(x) \leq 1\}.$$

Proof. If $x \in K$, then $1 \in A_x$, and so $p(x) \leq 1$; that is, $x \in K_2$. If $x \in K_1$, then there exists $\alpha \in A_x$ such that $0 < \alpha < 1$ and $x \in \alpha K$. Then $\alpha^{-1}x \in K$. Also, $0 \in K$ since K is absorbing, and so the convexity of K implies that $\alpha \cdot \alpha^{-1}x + (1-\alpha) \cdot 0 = x \in K$. □

Next, we add the assumption that X is a topological linear space.

Theorem 12.3. *An absolutely convex absorbing set K in a topological linear space is a neighborhood of 0 if and only if its Minkowski functional p is* *continuous. In this case the interior of K is the set* $K_1 = \{x : p(x) < 1\}$ *and the closure of K is the set* $K_2 = \{x : p(x) \leq 1\}$.

Proof. Suppose K is a neighborhood of 0. Then, since $K \subset K_2$, we see from Lemma 11.2 that p is continuous. Conversely, if p is continuous, then K_1 is open. Hence K is a neighborhood of 0 because $K \supset K_1$. To prove the second statement of the theorem, it suffices to prove that

$$\operatorname{int}(K) \subset K_1 \subset K \subset K_2 \subset \bar{K},$$

since K_1 is open and K_2 is closed when p is continuous. Suppose $x \in \operatorname{int}(K)$. Since scalar multiplication is continuous, there exists $\alpha > 1$ such that $\alpha x \in K$. Then $\alpha^{-1} \in A_x$ and $p(x) \leq \alpha^{-1} < 1$, so $x \in K_1$. We know that $K_1 \subset K \subset K_2$ from Theorem 12.2. To prove $K_2 \subset \bar{K}$, it suffices to consider an x for which $p(x) = 1$. Then $\alpha x \in K_1 \subset K$ if $0 < \alpha < 1$. Letting $\alpha \to 1$, we see that $\alpha x \to x$, whence $x \in \bar{K}$. □

Theorem 12.4. *Let X be a topological linear space. Then X is locally convex if and only if there exists a family* $\mathscr{P}$ *of seminorms that generate the topology on X.*

Proof. The "if" part of the theorem is Theorem 11.3. Now suppose X is a locally convex space, and denote its topology by τ. Let $\mathscr{U}$ be the base at 0 consisting of all absolutely convex neighborhoods of 0, and let $\mathscr{P}$ be the family of Minkowski functionals of the sets in $\mathscr{U}$. From Theorem 12.1 it follows that each p in $\mathscr{P}$ is a seminorm, and so $\mathscr{P}$ determines a topology $\tau_{\mathscr{P}}$ for X. By Theorem 12.3, each p in $\mathscr{P}$ is τ-continuous. Consequently, $\tau_{\mathscr{P}}$ is weaker than τ, by Theorem 11.3. On the other hand, if $U \in \mathscr{U}$, then the Minkowski functional p of U is $\tau_{\mathscr{P}}$-continuous. By Theorem 12.3, U must be a $\tau_{\mathscr{P}}$-neighborhood of 0. Thus τ is weaker than $\tau_{\mathscr{P}}$, which proves that $\tau = \tau_{\mathscr{P}}$. □

If X is a locally convex space with a given topology τ, the largest family of seminorms that generates τ is the family of all τ-continuous seminorms.

Normable Spaces

A set S in a topological linear space X is said to be *bounded* if for each neighborhood U of 0 there is a positive scalar α such that $S \subset \alpha U$. When X is a normed linear space, this definition is easily seen to be equivalent to the usual definition in terms of the norm; that is, S is bounded if and only if there is some positive β such that $\|x\| < \beta$ for each x in S.

A topological linear space X is said to be *normable* if there exists (i.e., if it is possible to define) a norm on X such that the norm topology is the same as the given topology for X.

Theorem 12.5 (Kolmogorov's Normability Criterion) *A topological linear space X is normable if and only if (a) X is a T_1-space and (b) there exists in X a convex and bounded neighborhood of* 0.

Proof. The conditions are evidently necessary. For the converse, let U be a convex and bounded neighborhood of 0. Then there exists an absolutely convex open neighborhood V of 0 contained in U (Theorem 10.5). By Theorem 12.1, the Minkowski functional of V is a seminorm p. Given a neighborhood W of 0, there exists $\alpha > 0$ such that $V \subset U \subset \alpha W$, since U is bounded, and hence $\alpha^{-1} V \subset W$. It follows that $\{rV : r > 0\}$ is a base of neighborhoods of 0; that is, the topology of X is generated by the single seminorm p. Now condition (a) implies that X is a Hausdorff space (Theorem 9.4), and so $p(x) \neq 0$ whenever $x \neq 0$ (Theorem 11.4). Thus p is actually a norm, and X is normable. □

PROBLEMS

1. Let $\mathscr{P}$ be the family of all seminorms on a linear space X, where $X \neq (0)$. This family is nonempty, and the topology it generates is the strongest locally convex topology on X. Moreover, it is a T_1-topology. *Suggestion.* Given $x_0 \neq 0$, choose a Hamel basis for X of which x_0 is a member. For any $x \in X$ let $p(x)$ be the absolute value of the coefficient of x_0 in the representation of x in terms of the Hamel basis.
2. A set T in a topological linear space X is said to *absorb* a set S if there exists $r > 0$ such that $S \subset \alpha T$ when $|\alpha| \geqslant r$. (Thus T is "absorbing", as defined in § 9, if and only if T absorbs each point of X.) A set S is bounded if and only if each balanced neighborhood of 0 absorbs S.
3. A set S is bounded if and only if it has the following property: If $\{\alpha_n\}$ is a sequence of scalars such that $\alpha_n \to 0$ and if $\{x_n\}$ is a sequence of elements of S, the sequence $\{\alpha_n x_n\}$ is convergent to 0.
4. If S is a set in a locally convex space X, S is bounded if and only if each continuous seminorm is bounded on S. Instead of *all* continuous seminorms, it is sufficient to consider those in a family that generates the topology on X.
5. Let S be a bounded set in a topological linear space. The closure of S is bounded, and the closed balanced hull of S is bounded. Is the closed absolutely convex hull of S bounded? *Note.* Finding the answer to the last question may present some difficulties. Consider this question: Is it possible for a space to contain a bounded neighborhood of 0 whose convex hull is the whole space? What about $L^p(a, b)$ when $0 < p < 1$? See problem 2 in § 13, following.
6. If X, Y are topological linear spaces and T is a continuous linear operator from X into Y, then T maps bounded sets into bounded sets.

II.13 METRIZABLE TOPOLOGICAL LINEAR SPACES

A topological linear space X is *metrizable* if there exists a metric on X such that the metric topology coincides with the given topology of X. Normed linear spaces are obviously metrizable, but there are other important metrizable topological linear spaces as well.

In general, there exist many metrics that determine the topology of a given metrizable topological linear space X. If a metric d on X has the property that $d(x+z, y+z) = d(x, y)$ for all $x, y, z \in X$, we say that d is *invariant* (or *translation-invariant*). We call X a *metric linear space* if it is a topological linear space with topology derived from a specified invariant metric. It will follow from Theorem 13.1 below that for each metrizable topological linear space such an invariant metric always exists.

If X is metrizable, then its topology is certainly T_1. Also, there is a countable base of neighborhoods at each point x in X (e.g., the balls of radii $1/n$ with centers at $x, n = 1, 2, \ldots$). These necessary conditions are also sufficient, not merely to guarantee that the space is metrizable but that it is metrizable with an invariant metric. We state this formally:

Theorem 13.1 *Let X be a topological linear T_1-space in which there exists a countable base at 0. Then X is metrizable with an invariant metric.*

We omit the rather lengthy proof of this theorem. See Horváth [1, pages 111–113]. The proof when X is locally convex is outlined in problem 4.

Now suppose that X is a metric linear space with an invariant metric d, and write $|x|$ in place of $d(x, 0)$. Then it is easy to see that

(13-1) $$|x+y| \leq |x| + |y|,$$

(13-2) $$|-x| = |x|,$$

(13-3) $$|x| = 0 \text{ if and only if } x = 0.$$

There is nothing to guarantee that $|\alpha x| = |\alpha||x|$, however, and so $|x|$ may not be a norm. An example will be given below. Of course, $d(x, 0)$ is a continuous function of x and αx is a continuous function of α and x. It follows that:

(13-4) If $\{x_n\}$ is a sequence in X with $|x_n - x| \to 0$ and if $\{\alpha_n\}$ is a sequence of scalars with $\alpha_n \to \alpha$, then $|\alpha_n x_n - \alpha x| \to 0$.

The following theorem shows how a metric linear space may be characterized by specifying properties of $|x|$ rather than properties of a metric.

Theorem 13.2. *Let X be a linear space, and suppose $x \mapsto |x|$ is a real-valued function on X with properties* (13-1) *to* (13-4). *Set* $d(x, y) = |x - y|$.

Then d is an invariant metric on X and, with this metric, X is a metric linear space.

The proof is left to the reader.

Examples

Some simple examples of nonnormed, but metric, linear spaces are based on the inequality

$$\frac{|A+B|}{1+|A+B|} \leqslant \frac{|A|}{1+|A|} + \frac{|B|}{1+|B|}, \tag{13-5}$$

which is valid if A and B are arbitrary real or complex numbers. A proof of (13-5) follows readily from the observation that $t(1+t)^{-1}$ increases as the real variable t increases if $t > -1$ (check by computing the derivative).

Example 1. *The space* (s). The class of all sequences $x = \{\xi_n\}$ becomes a metric linear space if we define

$$|x| = \sum_{n=1}^{\infty} \mu_n \frac{|\xi_n|}{1+|\xi_n|}, \tag{13-6}$$

where $\{\mu_n\}$ is any fixed sequence of positive numbers such that $\Sigma\, \mu_n$ is convergent. We denote this space by (s).

Convergence in (s) is merely componentwise convergence. That is, if $x^{(i)} = \{\xi_n^{(i)}\}$, then $|x^{(i)} - x| \to 0$ as $i \to \infty$ if and only if $\lim_{i\to\infty} \xi_n^{(i)} = \xi_n$ for each n. Proof is left to the reader. It follows that (s) is a complete metric space. Furthermore, (s) is a locally convex space, by Example 1, § 11.

Example 2. *The space S.* Let $[a, b]$ be a finite closed interval of the real axis. Consider the family of all real-valued functions defined on $[a, b]$ that are measurable in the Lebesgue sense, but not necessarily bounded. By using almost-everywhere equality as an equivalence relation, we find that this family gives rise to a family of equivalence classes, which we make into a linear space (see, e.g., the discussion of $\mathscr{L}^p$ and L^p in Example 7, § I.2). If x is one of the equivalence classes, we define

$$|x| = \int_a^b \frac{|x(t)|}{1+|x(t)|}\, dt, \tag{13-7}$$

where $x(t)$ denotes a representative function for x. We then obtain a metric linear space, which we denote by S. We can also consider the case in which the functions are complex-valued.

Convergence in S is interpretable in terms of convergence *in measure.* We recall the definition: $x_n(t)$ converges in measure to $x(t)$ if for each $\varepsilon > 0$ the measure of the set $\{t : |x_n(t) - x(t)| \geq \varepsilon\}$ converges to zero as $n \to \infty$. Let $E_n(\varepsilon)$ be the set in question. By decomposing $[a, b]$ into the union of $E_n(\varepsilon)$ and its complement, we see from (13-7) that

$$|x_n - x| \leq m(E_n(\varepsilon)) + \varepsilon(b-a),$$

where m refers to Lebesgue measure. Also, since $\alpha(1+\alpha)^{-1}$ is an increasing function of α when $\alpha > -1$, we see that

$$|x_n - x| \geq \int_{E_n(\varepsilon)} \frac{|x_n(t) - x(t)|}{1 + |x_n(t) - x(t)|}\, dt \geq \frac{\varepsilon}{1+\varepsilon} m(E_n(\varepsilon)).$$

From the foregoing considerations it is clear that $|x_n - x| \to 0$ as $n \to \infty$ is equivalent to convergence in measure of $x_n(t)$ to $x(t)$. From a well-known fact about convergence in measure, it then follows that S is a complete metric space (see, e.g., Taylor [5, page 267]). However, it can be shown that S is *not* a locally convex space (see problem 7 in § III.2).

Using Theorem 13.1, we can show that many of the examples in § 11 of locally convex spaces are metrizable. If $\Omega \subset \boldsymbol{R}^n$ is an open set, the spaces $\mathscr{C}(\Omega)$, $\mathscr{E}^m(\Omega)$ $(1 \leq m \leq \infty)$ (see Example 4, § 11) are metrizable. We shall show how their topologies, which are evidently T_2, can be generated by countable families of seminorms. Let $\{K_j\}$ be a sequence of compact sets in $\boldsymbol{R}^n$ such that, $K_j \subset \operatorname{int}(K_{j+1})$, $j = 1, 2, \ldots$, and

$$\bigcup_{j=1}^{\infty} \operatorname{int}(K_j) = \Omega.$$

The reader may check that every compact subset of Ω is contained in some K_j. It follows that the topologies on $\mathscr{C}(\Omega)$ and $\mathscr{E}^m(\Omega)$ $(1 \leq m \leq \infty)$ generated by the families $\{q_{K_j} : j = 1, 2, \ldots\}$ and $\{q_{K_j,p} : j = 1, 2, \ldots, |p| \leq m\}$ are the same as those generated by $\{q_K : \text{compact } K \subset \Omega\}$ and $\{q_{K,p} : \text{compact } K \subset \Omega, |p| \leq m\}$. A similar argument shows that $\mathscr{C}(T)$ (Example 2, § 11) is metrizable if T is a σ-compact, locally compact Hausdorff space.

If X is the linear space of all real-valued functions on an uncountable set T (cf. Example 1, § 11), then the topology of pointwise convergence is not metrizable. For, if $d(x_1, x_2)$ is a metric that generates the topology of X, then the sets

$$U_n = \left\{x : d(x, 0) < \frac{1}{n}\right\}, \qquad n = 1, 2, \ldots,$$

form a base at 0 with the property that $\bigcap_{n=1}^{\infty} U_n = \{0\}$. Now contained in each U_n

is a neighborhood V of the form (11-3):

$$V = \{x : |x(t_i)| \leq r_i, \ i = 1, \ldots, m\},$$

where $r_i > 0$ and $t_i \in T$, $i = 1, \ldots, m$. Any function x that is zero at $t_1, \ldots, t_m$ belongs to V and hence to U_n. But we can find such a finite set of points for each n, so there is an at most countable set of points $\{t_i\} \subset T$ such that if $x(t_i) = 0$, $i = 1, 2, \ldots$, then $x \in \bigcap_{n=1}^{\infty} U_n$. Since T is uncountable, there exists a function that is zero on $\{t_n\}$ but not identically zero, and this function must be in $\bigcap_{n=1}^{\infty} U_n$. [In fact, there is an uncountable set of linearly independent functions of this kind, and so $\bigcap_{n=1}^{\infty} U_n$ contains an infinite-dimensional linear manifold.] This contradicts $\bigcap_{n=1}^{\infty} U_n = \{0\}$, and so X is not metrizable.

Fréchet Spaces

A locally convex complete metric linear space is usually called a *Fréchet space*. (Some authors omit the condition of local convexity.) The space (s), for example, is a Fréchet space. Much of the theory of linear operators that we shall present later for Banach spaces is actually valid for Fréchet spaces (and often for complete metric linear spaces that are not locally convex). However, the proofs in the more general setting usually involve only technical modifications of Banach space proofs, and we prefer to confine the discussion in later chapters to Banach spaces. Thus it will suffice here simply to mention some of the more important Fréchet spaces that arise in applications.

The space $\mathscr{C}(\Omega)$ is a Fréchet space. Only the completeness remains to be verified. Let $\{x_k\}$ be a Cauchy sequence in $\mathscr{C}(\Omega)$. For each compact subset K of Ω and each $\varepsilon > 0$ there exists an integer N, depending on K and ε, such that

$$|x_j(t) - x_k(t)| < \varepsilon$$

for $t \in K$ and $j, k \geq N$. We see that for each fixed $t \in \Omega$, $\{x_k(t)\}$ is a Cauchy sequence of scalars that converges to some number $x(t)$. If $K \subset \Omega$ is a compact neighborhood of $t_0 \in \Omega$, then $|x(t) - x_k(t)| \leq \varepsilon$ for $t \in K$ and $k \geq N$ (some N); that is, $x_k \to x$ uniformly on K. It follows that x is continuous at t_0; consequently, $x \in \mathscr{C}(\Omega)$. But then $\{x_k\}$ converges to x in the topology for $\mathscr{C}(\Omega)$ of uniform convergence on compact sets.

The spaces $\mathscr{E}^m(\Omega)$ $(1 \leq m \leq \infty)$ are also Fréchet spaces. The verification of their completeness is straightforward and will be omitted. The reader may consult Horváth [1, page 136].

Boundedness

In a metric linear space the concept of boundedness, as defined in § 12 for a topological linear space, does not need to coincide with the usual concept of boundedness for a metric space. A set may be metrically bounded but not bounded in the sense of § 12. This is easily shown to be possible in the space (s). In this case, the space as a whole is metrically bounded but is not bounded as a topological linear space. In a normed space, however, a set is bounded in norm if and only if it is bounded in the sense of § 12.

PROBLEMS

1. A topological linear space is said to be *locally bounded* if there exists in it a bounded neighborhood of 0. A locally bounded topological linear T_1-space is metrizable. (The converse is not true.)
2. Consider $L^p(a, b)$, where (a, b) is any interval of the real axis and $0<p<1$. For $x \in L^p(a, b)$, let

$$|x| = \int_a^b |x(s)|^p \, ds.$$

This function of x satisfies (13-1) to (13-4), and so determines an invariant metric on $L^p(a, b)$. [For (13-1), show that $(a+b)^p \leqslant a^p + b^p$ if $0<p<1$ and $a \geqslant 0$, $b \geqslant 0$. To do this, consider $(1+t)^p$, where $0 \leqslant t \leqslant 1$, and note the comparative sizes of c^p and c in the two cases $c \geqslant 1, 0 \leqslant c \leqslant 1$.] With this metric, $L^p(a, b)$ is a complete locally bounded metric linear space. A base of neighborhoods of 0 is given by the sets $\{x : |x| < 1/n\}$, $n = 1, 2, \ldots$. It can be shown that the space is not locally convex; in fact, the only nonempty open convex set in $L^p(a, b)$ is the whole space (see problem 6, § III.2).
3. Consider ℓ^p, where $0<p<1$ and $|x|$ is defined by

$$|x| = \sum_{i=1}^{\infty} |\xi_i|^p, \qquad x = (\xi_1, \xi_2, \ldots).$$

   a. The function $x \mapsto |x|$ determines an invariant metric on ℓ^p and makes ℓ^p into a locally bounded complete metric linear space.
   b. The space ℓ^p is not locally convex. [*Suggestion.* Let $B = \{x : |x| \leqslant 1\}$ and, for $m \geqslant 1$, let x_m be the element of ℓ^p with $1/m$ in the first m coordinates and zeros elsewhere. Then for each $\alpha > 0$, the elements αx_m, $m = 1, 2, \ldots,$ are all in the convex hull of αB. Show that this would be impossible if ℓ^p were locally convex.]
4. Let X be a locally convex T_1-space whose topology is generated by a countable family of seminorms, $p_0, p_1, p_2, \ldots$.
   a. For $n = 0, 1, \ldots,$ define $q_n(x) = \max\{p_0(x), \ldots, p_n(x)\}$. Then $\{q_n\}$ is an "increasing" sequence of seminorms that generates the same topology for X.

b. For $x \in X$, let

$$|x| = \sum_{n=0}^{\infty} \frac{1}{2^n} \frac{q_n(x)}{1+q_n(x)}.$$

Show that this function satisfies (13-1) to (13-3).

c. Show that X is a metrizable locally convex space. [First show that the neighborhood $\{x : q_m(x) \leq 2^{-k}\}$ contains the set $\{x : |x| \leq 2^{-m-k-1}\}$. Then show that the set $\{x : |x| \leq 2^{-k}\}$ contains the neighborhood $\{x : q_{k+1}(x) \leq 2^{-k-2}\}$.]

5. If x' is a continuous linear functional on the space (s), there exists a finite set $\alpha_1, \ldots, \alpha_N$ such that $x'(x) = \sum_{i=1}^{N} \alpha_i \xi_i$ for each $x \in (s)$.

III LINEAR FUNCTIONALS AND WEAK TOPOLOGIES

This chapter is devoted to the study of the relations between a topological linear space X and the conjugate or dual space X' of all continuous linear functionals on X. As we shall see, many geometric and topological properties of X may be described in terms of continuous linear functionals. When X is a Hausdorff locally convex space, the Hahn–Banach theorem, discussed in § 2, ensures the existence of enough elements in X' to make possible a rich theory of the duality between X and X'. The most important case for our purposes is when X is a normed linear space. The special features of this case are pointed out throughout the chapter. Sections 3 and 10 are devoted entirely to normed linear spaces. As an application, we show at the end of § 3 how the Hahn–Banach theorem can be used to give an existence proof in connection with the Dirichlet problem.

Concrete representations of the duals of many important spaces are given in § 5. Among the spaces considered are Hilbert space, ℓ^p, $L^p(a, b)$, $C[a, b]$, and $C(T)$ where T is a compact Hausdorff space. Radon measures and distributions are treated briefly at the end of the section.

The action of X and X' on one another leads naturally to the construction of "weak" topologies on X and X'. We describe these topologies and their main properties in § 6 to § 8. Readers interested mainly in normed spaces may omit § 8. The important and useful principle of uniform boundedness is discussed in § 9.

Recall from Chapter II that when X is a normed linear space, its conjugate X' has a natural Banach space structure. The interplay between the norm topologies and the weak topologies on X and X' forms the main subject of § 10. One of the main results in this section is Alaoglu's theorem about the weak*-compactness of the unit ball in X' (Theorem 10.2). Section 11 concerns the well-known Krein–Milman theorem. Among the applications in § 11 are a proof of the Stone–Weierstrass theorem and a discussion of the problem of best approximation in normed linear spaces.

III.1 LINEAR VARIETIES AND HYPERPLANES

Throughout this section, X denotes a linear space and $\boldsymbol{K}$ its field of scalars. $\boldsymbol{K}$ may be either the real or the complex field.

Definition. A set $M \subset X$ is called a *linear variety* (or affine manifold) if $M = x_0 + M_0$, where x_0 is a fixed vector and M_0 is a subspace of X.

We call M a translation of M_0. Note that $x_0 \in M$, because $0 \in M_0$. It is easy to see that if $M = x_0 + M_0$ and $x_1 \in M$, then also $M = x_1 + M_0$. This follows at once from the fact that if x_1 and x_2 are in M, then $x_1 - x_2 \in M_0$.

A linear variety is a convex set. A set consisting of a single point is a linear variety, for $\{0\}$ is a subspace.

A subspace X_0 of X will be called *maximal* if it is not all of X and if there exists no subspace X_1 in X such that $X_0 \neq X_1$, $X_1 \neq X$ and $X_0 \subset X_1$. A linear variety that results from the translation of a maximal subspace is called a *hyperplane*. In particular, a hyperplane containing 0 is the same thing as a maximal subspace. (Some authors reserve the term "hyperplane" for a maximal subspace and use "affine hyperplane" for translates of hyperplanes.) Hyperplanes obtained by translating a hyperplane H are said to be parallel to H.

Let H be a subspace. It is easily verified that H is a hyperplane containing 0 if and only if

$$X = H \oplus \{\lambda a : \lambda \in \boldsymbol{K}\}$$

for each $a \in X \backslash H$. From Theorem I.6.5, it follows that H is a hyperplane containing 0 if and only if the quotient space X/H is one-dimensional.

Theorem 1.1. (*a*) *If $x' \in X^f$ and $x' \neq 0$, then the set $H = \{x : x'(x) = 0\}$ is a hyperplane containing* 0. *A subset M of X is a hyperplane parallel to H if and only if $M = \{x : x'(x) = \alpha\}$ for some scalar α.*

(*b*) *To each hyperplane H containing* 0 *corresponds an $x' \in X^f$ such that $x' \neq 0$ and $H = \{x : x'(x) = 0\}$, and any other $y' \in X^f$ corresponding to H in this manner is a* (*nonzero*) *scalar multiple of x'.*

(*c*) *To each hyperplane M not containing* 0 *corresponds exactly one $x' \in X^f$ such that $M = \{x : x'(x) = 1\}$.*

Proof. (*a*) Since $x' \neq 0$, its null space $H = \{x : x'(x) = 0\}$ is evidently a proper subspace of X. If $a \in X \backslash H$, then

$$x - \frac{x'(x)}{x'(a)} a \in H$$

for each $x \in X$. It follows that $X = H \oplus \{\lambda a : \lambda \in \boldsymbol{K}\}$. Thus H is a hyperplane containing 0. Now suppose that $M = a + H$ for some $a \notin H$. It is easy to see that

$M=\{x: x'(x)=x'(a)\}$. On the other hand, if $M=\{x: x'(x)=\alpha\}$ and if $a \in M$, then one readily verifies that $M=a+H$.

(*b*) Let H be a hyperplane containing 0. Let $a \in X \backslash H$, so that $X=H \oplus \{\lambda a: \lambda \in \boldsymbol{K}\}$. Since each $x \in X$ may be expressed uniquely as $x=h+\lambda a$ (where $h \in H$), the linear functional x', given by $x'(x)=\lambda$, is well defined. Clearly H is the null space of x'. Furthermore, if H is the null space of some other $y' \in X^f$, then

$$y'(x)=y'(h+\lambda a)=\lambda y'(a)=y'(a) \cdot x'(x)$$

for each $x \in X$, and so y' is a multiple of x'. The scalar $y'(a)$ is nonzero since $a \notin H$.

(*c*) If M is a hyperplane not containing 0, then $M=a+H$ for some hyperplane H containing 0 and some $a \notin H$. By part (*b*), there exists $y' \in X^f$ such that $H=\{x: y'(x)=0\}$. Note that $\lambda=y'(a) \neq 0$, and let $x'=\lambda^{-1} y'$. As in the proof of (*a*), we have $M=\{x: x'(x)=x'(a)=1\}$. Now suppose that we also have $M=\{x: w'(x)=1\}$ for some $w' \in X^f$. Then by part (*a*), M is parallel to the hyperplane $H_1=\{x: w'(x)=0\}$. Hence H is parallel to H_1; and, since they both contain 0, we have $H=H_1$. Then $w'=\alpha x'$ for some α, by (*b*). Since $w'(a)=1=x'(a)$, we conclude that $\alpha=1$ and $w'=x'$. □

This is a convenient place at which to insert the following useful theorem.

Theorem 1.2. *Suppose that $x_1', \ldots, x_n'$ are linearly independent elements of X^f. If $y' \in X^f$, then either y' is a linear combination of $x_1', \ldots, x_n'$, or else there exists $x_0 \in X$ such that $x_1'(x_0)=\cdots=x_n'(x_0)=0$ and $y'(x_0)=1$.*

Proof. For the case $n=1$, suppose that $y'(x)=0$ whenever $x_1'=0$. Then the subspace $M=\{x: y'(x)=0\}$ contains the hyperplane $N=\{x: x_1'(x)=0\}$. Hence either $M=N$ or $M=X$. In either case y' must be a multiple of x_1', by Theorem 1.1. We now proceed inductively, assuming the theorem true for every set of $n-1$ linearly independent elements of X^f. Given a linearly independent set $\{x_1', \ldots, x_n'\}$, each subset of $n-1$ elements is linearly independent, and so for each k, $k=1, \ldots, n$, there exists an x_k such that $x_k'(x_k)=1$ and $x_j'(x_k)=0$ if $j \neq k$. For any $x \in X$, let

$$y=x-\sum_{k=1}^{n} x_k'(x) x_k. \tag{1-1}$$

We then see that $x_j'(y)=0$ for each j. Given $y' \in X^f$, suppose there is no $x_0 \in X$ such that $y'(x_0)=1$ while $x_j'(x_0)=0$, $j=1, \ldots, n$. It follows that $y'(y)=0$ for every y of the form (1-1). But then

$$y'(x)=\sum_{k=1}^{n} x_k'(x) y'(x_k), \qquad x \in X.$$

This is equivalent to

$$y' = \sum_{k=1}^{n} y'(x_k)x'_k.$$

Thus the theorem is true for every linearly independent set of n elements of X^f. By induction, this proves the theorem. □

We now consider the case when X is a topological linear space.

Theorem 1.3. *A hyperplane M in a topological linear space X is either closed or dense in X.*

Proof. Since translation is a homeomorphism, we may assume that M is a hyperplane containing 0. Then $\bar{M}$ is a subspace containing M (problem 1 in § II.9). Since M is maximal, either $\bar{M} = M$ or $\bar{M} = X$. □

Theorem 1.4. *Let x' be a nonzero linear functional on a topological linear space X. Then the following assertions are equivalent*:

(*a*) *x' is continuous*;
(*b*) *there exists a continuous seminorm p such that $|x'(x)| \leqslant p(x)$, $x \in X$*;
(*c*) *x' is bounded on some neighborhood of 0 in X*;
(*d*) *the hyperplane $H = \{x : x'(x) = 0\}$ is closed.*

Proof. Since the mapping $x \mapsto |x'(x)|$ is itself a seminorm, it is clear that (a) implies (b). If p is a continuous seminorm, then the set $V = \{x : p(x) \leqslant 1\}$ is a neighborhood of 0, by Lemma II.11.2. It follows that (b) implies (c). Now suppose that $|x'(x)| < B$ for some constant B and all x in some neighborhood U of 0. If $M = \{x : x'(x) = B + 1\}$, then $M \cap U = \varnothing$. Thus M is not dense. By Theorem 1.3, M must be closed. Since M is a translate of $H = \{x : x'(x) = 0\}$, the hyperplane H is closed. Thus (c) implies (d). Finally, suppose that (d) is true and take $\varepsilon > 0$. Then the translated hyperplane $L = \{x : x'(x) = \varepsilon\}$ is closed, and $X \backslash L$ is an open neighborhood of 0. By Theorem II.9.2, there exists a balanced neighborhood U of 0 such that $U \subset X \backslash L$. Then $x'(U)$ is a balanced set of scalars. Since $U \cap L = \varnothing$, we must have $|x'(x)| < \varepsilon$ for $x \in U$. This shows that x' is continuous at 0. Since x' is linear, it must be continuous everywhere. Thus (d) implies (a). □

PROBLEMS

1. Let X be a topological linear space, and let x' be a nonzero linear functional on X such that the hyperplane $H = \{x : x'(x) = 0\}$ is closed. Use problems 6 and 8 of § II.9 to show that x' is continuous. Prove also that x' is an open mapping, that is, $x'(U)$ is an open set of scalars whenever U is open in X.

2. Let X be a topological linear space (not necessarily locally convex), and let M be a closed hyperplane in X not containing 0. Then $X\backslash M$ contains an open absolutely convex neighborhood of 0.
3. Let X be a linear space, and let M be a finite-dimensional subspace of X^f. Then M is algebraically saturated (see page 46).

III.2 THE HAHN–BANACH THEOREM

The term "Hahn–Banach theorem" is applied today to several closely related results. The most general of these was presented in Theorem I.10.4; a similar result for normed linear spaces will appear in Theorem 3.1. The theorems in this section lie somewhere in between. Geometric forms of the Hahn–Banach theorem are given first, in Theorems 2.4 and 2.5. Several important corollaries follow. Finally, analytic formulations of the Hahn–Banach theorem are presented in Theorems 2.11 and 2.12.

Several results below are often referred to as separation theorems. Theorems 2.4 and 2.5 involve the separation of an open convex set and a (nonintersecting) linear variety. The separation of a point from a disjoint closed subspace and, more generally, from a disjoint closed convex set is described in Theorems 2.7 and 2.9. Other separation results are considered in the problems at the end of the section.

First, suppose that X is a real linear space. If x' is a nonzero linear functional and if $\alpha \in \boldsymbol{R}$, then the hyperplane $\{x : x'(x) = \alpha\}$ determines four convex sets called *half spaces*:

$$\{x : x'(x) < \alpha\}, \qquad \{x : x'(x) > \alpha\}, \tag{2-1}$$

$$\{x : x'(x) \leqslant \alpha\}, \qquad \{x : x'(x) \geqslant \alpha\}. \tag{2-2}$$

If X is a topological linear space and x' is continuous, the half spaces of (2-1) are open and those of (2-2) are closed. Conversely, it is readily verified that if a half space of the form (2-1) is open for some $\alpha \in \boldsymbol{R}$ or if a half space of the form (2-2) is closed, then x' must be continuous (see problem 1).

We say that a set S *lies on one side* of M if S lies entirely in one of the four half spaces determined by $M = \{x : x'(x) = \alpha\}$. If S lies on one side of M and does not intersect M, we say that S is *strictly* on one side of M.

Lemma 2.1. *Let X be a real linear space. In order that a convex subset S of X lie strictly on one side of a hyperplane M, it is necessary and sufficient that the intersection $S \cap M$ be empty.*

Proof. The condition is necessary, by definition. Suppose that $M = \{x : x'(x) = \alpha\}$ and $S \cap M = \varnothing$. Since $x'(S)$ is evidently a convex set in $\boldsymbol{R}$, it must lie either to the right or to the left of the number α. Thus S is contained in a half space (2-1). □

Lemma 2.2. *If L is a subspace of a topological linear space X such that* $\dim X/L \geq 2$, *then the complement of L, $X\backslash L$, is a connected set.*

Proof. We shall prove that $X\backslash L$ is pathwise connected, hence connected. Suppose $a, b \in X\backslash L$, and let $[a]$, $[b]$ denote their images in X/L under the canonical mapping. If $[a]$ and $[b]$ are linearly independent, then the line segment joining a to b does not pass through L. Indeed, if $(1-\lambda)a+\lambda b \in L$ for some $\lambda \in (0, 1)$, then $(1-\lambda)[a]+\lambda[b]=[0]$, and this would imply that $[a]$ and $[b]$ are linearly dependent, contrary to our assumption. Now suppose that $[a]$ and $[b]$ are linearly dependent. Since $\dim X/L \geq 2$, there exists $[c] \in X/L$ such that $[a]$ and $[c]$ are linearly independent. Then $[b]$ and $[c]$ must also be linearly independent. From the discussion above we know that the line segment from a to c and the line segment from c to b both lie in $X\backslash L$. Thus a and b can be joined by a path contained in $X\backslash L$. □

In the next lemma, $L \oplus \boldsymbol{R}a$ denotes the set $\{x+\lambda a : x \in L, \lambda \in \boldsymbol{R}\}$.

Lemma 2.3. *Let K be a nonempty open convex set in a real topological linear space X, and let L be a subspace that does not intersect K. Then either L is a hyperplane or there is a point $a \in X\backslash L$* such *that the subspace $L \oplus \boldsymbol{R}a$ does not intersect K.*

Proof. The set $\bigcup_{\lambda>0} \lambda K$ (called a cone without vertex) is open, and so is the set

$$U = L + \bigcup_{\lambda>0} \lambda K$$

(Lemma II.9.1). We note first that $U \cap L = \varnothing$ for, if $x \in L$ could be written in the form $x = y+\lambda k$ (where $y \in L$, $k \in K$, and $\lambda>0$), then $k=\lambda^{-1}(x-y)$ would be in $K \cap L$, which is empty by hypothesis. Since L is a subspace, it follows that $(-U)\cap L=\varnothing$. We now claim that $U \cap(-U)=\varnothing$. If $x \in U \cap(-U)$, then $x = y_1+\lambda_1 k_1 = y_2-\lambda_2 k_2$, where $y_1, y_2 \in L$, $k_1, k_2 \in K$ and $\lambda_1, \lambda_2>0$. Then $\lambda_1 k_1+\lambda_2 k_2 = y_2-y_1 \in L$, and it follows that

$$\frac{\lambda_1}{\lambda_1+\lambda_2} k_1+\frac{\lambda_2}{\lambda_1+\lambda_2} k_2 \in K \cap L,$$

since K is convex. Since $K \cap L=\varnothing$, this contradiction implies that

$$U \cap(-U)=\varnothing.$$

Let us suppose that L is not a hyperplane. Then the dimension of X/L is at least 2 and $X\backslash L$ is connected (Lemma 2.2). Since U and $-U$ are disjoint open sets that are contained in $X\backslash L$, it is impossible to have $X\backslash L = U \cup(-U)$. Thus there exists an $a \in X$ with $a \notin L$ and $a \notin U \cup(-U)$. We claim that

$(L\oplus \boldsymbol{R}a)\cap K=\varnothing$. For, if $k=y+\lambda a$ is in K for some $y\in L$, $\lambda\in\boldsymbol{R}$, then $\lambda\neq 0$ since $K\cap L=\varnothing$, and $a=\lambda^{-1}k-\lambda^{-1}y\in\lambda^{-1}K+L$. Thus $a\in U$ if $\lambda^{-1}>0$ or $a\in -U$ if $\lambda^{-1}<0$, which contradicts the fact that $a\notin U\cup(-U)$. □

The proof of this lemma may appear more intuitive after we consider a simple example. Let X be three-dimensional Euclidean space, and let K be a quarter of the unit ball in X:

$$K=\{(x,y,z): y>0,\ x^2+y^2<1,\ 0<z<\sqrt{1-x^2-y^2}\}.$$

Let L be the y-axis, that is, $L=\{(x,y,z): x=z=0\}$. Then

$$\bigcup_{\lambda>0}\lambda K=\{(x,y,z): y>0,\ z>0\},$$

and

$$U=L+\bigcup_{\lambda>0}\lambda K=\{(x,y,z): z>0\}.$$

Clearly $U\cup(-U)$ is all of X except the xy-plane. Any element of the form $(x,y,0)$ with $x\neq 0$ will do for the vector a. As a matter of fact, it should be clear that if $L\oplus\boldsymbol{R}a$ is a two-dimensional subspace not intersecting K, then it must be the xy-plane and a must have the form $(x,y,0)$ with $x\neq 0$.

The Geometric Form of the Hahn–Banach Theorem

Theorem 2.4. *Let K be a nonempty open convex set in a real topological linear space X, and let L be a linear variety that does not intersect K. Then there exists a closed hyperplane M that contains L and is such that K lies strictly on one side of M.*

Proof. We may assume that L is a subspace since translation is a homeomorphism and it preserves convexity of sets. Let

$$\mathcal{M}=\{N: N \text{ is a subspace such that } L\subset N\subset X \text{ and } N\cap K=\varnothing\}.$$

We let $\mathcal{M}$ be partially ordered by the relation of set inclusion (i.e., $M<N$ if $M,N\in\mathcal{M}$ and $M\subset N$). Suppose $\{N_\alpha\}$ is a completely ordered subset of $\mathcal{M}$. Letting $N=\bigcup_\alpha N_\alpha$, we readily see that $N\in\mathcal{M}$ and N is an upper bound for $\{N_\alpha\}$. By Zorn's lemma (§ I.9), there is a maximal element M in $\mathcal{M}$. Thus there can be no element $a\in X\backslash M$ such that the subspace $M\oplus\boldsymbol{R}a$ does not intersect K. It follows from Lemma 2.3 that M is a hyperplane. Since $M\cap K=\varnothing$, M cannot be dense and hence is closed (Theorem 1.3). By Lemma 2.1, K lies strictly on one side of M. □

As we shall see in Theorem 2.5, a slightly modified version of Theorem 2.4 is valid in either real or complex topological linear spaces. Suppose that X

is a complex linear space. We can also regard X as a real linear space simply by restricting the operation of scalar multiplication to only real scalars. Let us denote by X_r the real linear space obtained from X by adopting this point of view. If L is a subspace of X, then it is *a fortiori* a subspace of X_r. However, a subspace L of X_r is a (complex) subspace of X if and only if $iL = L$. When X is a topological linear space, the mapping $(\alpha, x) \mapsto \alpha x$ is a continuous function on $C \times X$ and *a fortiori* on $R \times X$. If we give X_r the same topology as X, then X_r becomes a real topological linear space.

Now consider a linear functional $x' \in X^f$, where X is a complex linear space. Given $x \in X$, the imaginary part of the scalar $x'(x)$ is the real part of $-ix'(x)$, that is, the real part of $-x'(ix)$. It follows that

$$x'(x) = \operatorname{Re} x'(x) - i \operatorname{Re} x'(ix). \tag{2-3}$$

Thus the functional x' is determined by its real part, $\operatorname{Re} x'$. It is readily verified that $\operatorname{Re} x' \in (X_r)^f$; that is, $\operatorname{Re} x'$ is a (real) linear functional on the *real* linear space X_r. Conversely, if we start with *any* element $x_1' \in (X_r)^f$ and define x' by

$$x'(x) = x_1'(x) - ix_1'(ix),$$

it is easy to verify that $x'(\alpha x) = \alpha x'(x)$ for all *complex* scalars and that $x' \in X^f$. If X is a topological linear space, it is clear from (2-3) that $x' \in X^f$ is continuous on X (i.e., $x' \in X'$) if and only if $\operatorname{Re} x'$ is continuous on X_r.

Theorem 2.5. *Let K be a nonempty open convex set in a topological linear space X, and let L be a linear variety that does not intersect K. Then there exists a closed hyperplane M that contains L and does not intersect K.*

Proof. We may assume the scalars are complex for, if they are real, the theorem reduces to Theorem 2.4. As in the proof of that theorem, we assume that L is a subspace of X. Then L is a subspace of X_r and, by Theorem 2.4, there is a closed real hyperplane M_1 in X_r containing L and disjoint from K. Now M_1 is the null space of a continuous real linear functional (Theorems 1.1 and 1.4). In view of the discussion above, $M_1 = \{x : \operatorname{Re} x'(x) = 0\}$ for some $x' \in X'$. The set $M = \{x : x'(x) = 0\}$ is a closed (complex) hyperplane obviously contained in M_1 and therefore disjoint from K. If $x \in L$, then $\operatorname{Re} x'(x) = 0$; also, $ix \in L \subset M_1$, and so $\operatorname{Re} x'(ix) = 0$. By (2-3), $x'(x) = 0$. Thus $L \subset M$. □

Another interesting proof of Theorem 2.5 may be found in the text by H. H. Schaefer [1]. The argument there involves some ideas we mentioned in the problem set for § II.9.

We shall now consider a number of important consequences of Theorem 2.5. The results apply to both real and complex linear spaces.

Theorem 2.6. *Let L be a closed subspace of a locally convex space X. Then L is the intersection of all the closed hyperplanes that contain it.*

Proof. If $a \in X \backslash L$, then $X \backslash L$ is an open neighborhood of a. Since X is locally convex, there exists an open convex neighborhood U of 0 such that the (open and convex) set $a + U$ is contained in $X \backslash L$. It follows from Theorem 2.5 that there is a closed hyperplane containing L but not containing a. The conclusion now follows. □

Theorem 2.7. *Let L be a closed subspace of a locally convex space X, and suppose $a \in X \backslash L$. Then there exists $x' \in X'$ such that $x'(a) = 1$ and $x'(x) = 0$ if $x \in L$.*

Proof. By Theorem 2.6, there exists a closed hyperplane M containing L and not containing a. Since $0 \in L \subset M$, we have $M = \{x : x'(a) = 0\}$ for some continuous linear functional x' (Theorems 1.1 and 1.4). Since $x'(a) \neq 0$, we may choose x' such that $x'(a) = 1$. □

Theorem 2.8. *Let X be a locally convex space. Then X is a Hausdorff space if and only if to each $x_0 \neq 0$ there corresponds an x' in X' such that $x'(x_0) = 1$.*

Proof. If X is a Hausdorff space, then the subspace $L = \{0\}$ is closed, and we may apply Theorem 2.7. On the other hand, if $x_0 \neq 0$ and x' is a continuous linear functional such that $x'(x_0) = 1$, then $U = \{x : |x'(x)| < 1\}$ is a neighborhood of 0 that does not contain x_0. If this holds for each nonzero $x_0 \in X$, then X must be a Hausdorff space, by Theorem II.9.3. □

The importance of Theorem 2.8 lies in the fact that it guarantees that there are enough continuous linear functionals to separate the points of a Hausdorff locally convex space X. For, if x and y are distinct points of X, there exists a continuous linear functional x' such that $x'(x - y) = 1 \neq 0$, and so $x'(x) \neq x'(y)$.

The characterization of closed subspaces given in Theorems 2.6 and 2.7 has a useful application to problems in approximation. For example, let S be a set of real polynomial functions defined on $[0, 1]$. A continuous function x_0 defined on $[0, 1]$ can be approximated, arbitrarily closely in the norm of $C[0, 1]$, by finite linear combinations of elements of S, if and only if x_0 is in the closed linear manifold L generated by S. It thus follows from Theorem 2.7 that x_0 can be approximated in this manner if and only if each continuous linear functional x' on $C[0, 1]$ that satisfies $x'(y) = 0$ for all $y \in S$ also satisfies $x'(x_0) = 0$. (This is because a continuous linear functional x' will be zero on S if and only if $x'(x) = 0$ for all $x \in L$.) In actual applications, this criterion involving continuous linear functionals is more useful when one knows how to represent each x' in a way that makes it possible to calculate $x'(x)$ for $x \in C[0, 1]$. Such a representation is discussed in § 5.

The following separation theorem will be needed in the proofs of several important results (e.g., Theorems 7.3 and 11.3).

Theorem 2.9. *Let K be a nonempty closed convex subset of a locally convex space X, and suppose $a \in X \backslash K$.*

(*a*) *Then there exists $x' \in X'$ such that*

$$\sup_{x \in K} \operatorname{Re} x'(x) < \operatorname{Re} x'(a). \tag{2-4}$$

(*b*) *If K is closed and absolutely convex, then x' may be chosen so that $x'(a)$ is real, and*

$$\sup_{x \in K} |x'(x)| \leq 1 < x'(a).$$

Proof. (*a*) Observe that $-a + K$ is a closed set not containing 0. Hence the complement of $-a + K$ is a neighborhood of 0. Since X is locally convex, there exists an open absolutely convex neighborhood U of 0 that is disjoint from $-a + K$. Now $-a + K + U$ is open and convex, by Lemma II.9.1 and Theorem II.10.1. It is readily verified that $0 \notin (-a + K + U)$, since U is balanced. Applying Theorem 2.4 in the real space X_r, we obtain a closed real hyperplane M_1 that contains $L = \{0\}$ and is such that $-a + K + U$ lies strictly on one side of M_1. As in the proof of Theorem 2.5, we see that there exists $x' \in X'$ such that $M_1 = \{x : \operatorname{Re} x'(x) = 0\}$. Replacing x' by $-x'$, if necessary, we may assume that $\operatorname{Re} x'(x) < 0$ for all x in $-a + K + U$.

Now U is absorbing, and so $\operatorname{Re} x'(y) \neq 0$ for some $y \in U$. (Otherwise $\operatorname{Re} x'$ would be identically zero.) Since U is balanced, there exists $y_0 \in U$ such that $\alpha = \operatorname{Re} x'(y_0) > 0$. Then, for all $x \in K$,

$$0 > \operatorname{Re} x'(-a + x + y_0) = -\operatorname{Re} x'(a) + \operatorname{Re} x'(x) + \alpha.$$

The conclusion in (2-4) follows immediately.

(*b*) Now suppose that K is also balanced. Then $x'(K)$ is a balanced set of scalars. Hence

$$\sup_{x \in K} |x'(x)| = \sup_{x \in K} \operatorname{Re} x'(x) < \operatorname{Re} x'(a) \leq |x'(a)|.$$

Multiplying x' by an appropriate scalar, we obtain the linear functional described in (*b*). □

When X is a complex linear space, we may still consider *real* half spaces of the form (2-1) and (2-2), where x' is replaced by $\operatorname{Re} x'$ for some $x' \in X^f$.

Theorem 2.10. *Let X be a locally convex space. The closed convex hull of a nonempty set S in X is the intersection of all the closed real half spaces that contain S.*

Proof. Let K be the closed convex hull of S, and let L be the intersection of all the closed real half spaces that contain S. If $a \in X \backslash K$, then by Theorem 2.9 there exists $x' \in X'$ such that $\alpha = \sup_{y \in K} \operatorname{Re} x'(y) < \operatorname{Re} x'(a)$. Thus $\{x : \operatorname{Re} x'(x) \leq \alpha\}$ is a closed real half space containing K but not a, and so $a \notin L$. Hence $L \subset K$. The reverse inclusion is obvious. □

Theorem 2.9 has an application to the problem in approximation discussed after Theorem 2.8. Let K be the closed convex hull of some set S in the real space $C[0, 1]$. A function x_0 in $C[0, 1]$ belongs to K if and only if x_0 may be approximated arbitrarily closely by convex linear combinations of elements of S (cf. Theorem II.10.2). It is easily seen that if x' is a (real) continuous linear functional on $C[0, 1]$, then $\sup_{x \in K} x'(x) = \sup_{x \in S} x'(x)$. Therefore it follows from Theorem 2.9 that x_0 may be approximated by convex linear combinations of elements of S if and only if

$$x'(x_0) \leq \sup_{x \in S} x'(x) \tag{2-5}$$

for each continuous linear functional x' on $C[0, 1]$. Note that to verify (2-5) for a particular x', we only have to find a sequence $\{x_n\}$ in S such that $x'(x_n) \to x'(x_0)$, as $n \to \infty$.

The Analytic Form of the Hahn–Banach Theorem

Theorem 2.11. *Let X be a linear space, and let p be a seminorm defined on X. Let M be a subspace of X, and let m' be a linear functional defined on M such that $|m'(x)| \leq p(x)$ if $x \in M$. Then there exists a linear functional x' defined on X such that $|x'(x)| \leq p(x)$ if $x \in X$, and $x'(x) = m'(x)$ if $x \in M$.*

Proof. We shall give two different proofs. For the case in which X is a real space, the present theorem is a direct application of Theorem I.10.4. If X is a complex space, write $m'(x) = m_1'(x) - im_1'(ix)$, with $m_1'(x)$ the real part of $m'(x)$. Then m_1' is a real linear functional on the real space M_r, and $|m_1'(x)| \leq |m'(x)| \leq p(x)$ if $x \in M$. Hence, by Theorem I.10.4, there exists a real linear functional x_1' on the real space X_r such that $x_1'(x) = m_1'(x)$ if $x \in M_r$ and $x_1'(x) \leq p(x)$ if $x \in X$. We define $x'(x) = x_1'(x) - ix_1'(ix)$. Then x' is a complex linear functional on X, and $x'(x) = m'(x)$ if $x \in M$. Given $x \in X$, choose λ such that $|\lambda| = 1$ and $\lambda x'(x) = |x'(x)|$. Then

$$|x'(x)| = x'(\lambda x) = \operatorname{Re} x'(\lambda x) = x_1'(\lambda x) \leq p(\lambda x) = p(x).$$

We now present a proof based on the geometric form of the Hahn–Banach theorem. Let X have the topology generated by the single seminorm p. Then $K = \{x \in X : p(x) < 1\}$ is an open absolutely convex neighborhood of 0. We may assume that m' is not identically zero on M, for otherwise m' would have a trivial extension to X. Thus $L = \{x \in M : m'(x) = 1\}$ is a hyperplane in M. Since $|m'(x)| \leq p(x) < 1$ for $x \in K \cap M$, we see that $K \cap L = (K \cap M) \cap L = \varnothing$. By Theorem 2.5, there exists a closed hyperplane N in X that contains the linear variety L and does not intersect K. In particular, $0 \notin N$ and there exists $x' \in X'$ such that $N = \{x : x'(x) = 1\}$. The restriction $x'|_M$ of x' to M is a linear functional on M which coincides with m' on L, since $N \supset L$. By Theorem 1.1(c), $x'|_M$ coincides with m' on M itself.

Now $x'(K)$ is an absolutely convex set of scalars containing zero. Since $K \cap N = \varnothing$, we conclude that $|x'(x)| < 1$ for $x \in K$. Given any $x \in X$, take α such that $p(x) < \alpha$. Then $\alpha^{-1}x \in K$, and so $|x'(\alpha^{-1}x)| < 1$; that is, $|x'(x)| < \alpha$. Since α was arbitrarily close to $p(x)$, we have $|x'(x)| \leq p(x)$. This holds for each $x \in X$. □

Theorem 2.12. *Let M be a subspace of a locally convex space X, and let m' be a continuous linear functional on M. Then there exists an $x' \in X'$ that is an extension of m'.*

Proof. Since m' is continuous and X is locally convex, there exists an absolutely convex neighborhood V of 0 such that $|m'(x)| < 1$ if $x \in M \cap V$. Let p be the Minkowski functional of V. Then by Theorems II.12.1 and II.12.3, p is a continuous seminorm and $\{x : p(x) < 1\} \subset V$. Thus $|m'(x)| < 1$ whenever $x \in M$ and $p(x) < 1$. The argument given at the end of the preceding proof shows that $|m'(x)| \leq p(x)$ for $x \in M$. Applying Theorem 2.11, we obtain a linear functional x' on X which is an extension of m' such that $|x'(x)| \leq p(x)$ for every x. Hence, by Theorem 1.4, x' is continuous. □

PROBLEMS

Unless otherwise noted, a space X referred to in these problems is a real topological linear space.

1. Let x' be a nonzero linear functional on X. Then the half spaces in (2-2) are closed if and only if x' is continuous.
2. Let H_1 be a real hyperplane in a complex linear space X, and suppose $H_1 = \{x : \operatorname{Re} x'(x) = 0\}$ for some $x' \in X'$. Let H be the complex hyperplane $\{x : x'(x) = 0\}$. Show that $iH_1 = \{x : -\operatorname{Re} x'(ix) = 0\}$, and deduce that $H = H_1 \cap iH_1$.
3. If K_1 and K_2 are nonempty, nonintersecting convex sets and if K_1 is open, there exists a closed hyperplane M such that K_1 is in one of the two closed half spaces determined by M and K_2 is in the other. If K_2 is also open, M can

be chosen so that K_1 and K_2 are strictly on opposite sides of M. Argument: $K = K_1 + (-1)K_2$ is convex, nonempty, open, and does not contain 0. Hence there exists a closed hyperplane through 0 and not intersecting K_2. A suitable parallel hyperplane can be chosen for M.

4. If $S \subset X$, a *support* of S is a hyperplane M such that S lies on one side of M and $S \cap M \neq \varnothing$. If $x_0 \in S \cap M$, we say that S is supported by M at x_0. A closed convex set with a nonvacuous interior is called a *convex body*. Show that a support of a convex body is closed and that the body is supported at every boundary point.

5. Let K be a convex body in X and not all of X. Consider the closed half spaces containing K and determined by the supports of K (see problem 4). The intersection of all these half spaces is K.

6. There exists a nonzero continuous linear functional x' on X if and only if in X there is at least one convex set that is open and contains 0 but is not all of X. [For the "if" part, use Theorem 2.4.] M. M. Day has shown that there exists no nonzero continuous linear functional on $L^p(a, b)$ if $0 < p < 1$ (see Köthe [1, page 158]). Use this fact to show that each continuous linear mapping of $L^p(a, b)$ into a Hausdorff locally convex space is identically zero.

7. Let X be the metric linear space S of measurable functions on $[0, 1]$ (see § II.13). Then $X' = (0)$, and hence X is not locally convex. [*Suggestion.* Suppose x' is a nonzero linear functional on X, and take $x \in X$ such that $x'(x) \neq 0$. Then find $x_1 \in X$ such that $x'(x_1) \neq 0$ and the support of x_1 is in either $[0, \frac{1}{2}]$ or $[\frac{1}{2}, 1]$. By induction, obtain $\{x_n\} \subset X$ such that $\alpha_n = x'(x_n) \neq 0$ and the measure of the support of x_n is not greater than 2^{-n}. Then use the sequence $\{z_n\}$, where $z_n = \alpha_n^{-1} x_n$, to deduce that x' is not continuous.]

8. For the next problem we require the following result. Let A and B be disjoint sets in X. If A is compact and B is closed, then there exists a neighborhood V of 0 such that $A + V$ and $B + V$ are disjoint. Outline of argument: For each $x \in A$ there exists a balanced open neighborhood V_x of 0 such that $(x + V_x + V_x + V_x) \cap B = \varnothing$. Then $(x + V_x + V_x) \cap (B + V_x) = \varnothing$. There exist $x_1, \ldots, x_n$ in A such that $A \subset \bigcup_{i=1}^{n} (x_i + V_{x_i})$. Let $V = \bigcap_{i=1}^{n} V_{x_i}$.

9. Let X be a real locally convex space. Let K_1, K_2 be nonempty, nonintersecting convex sets in X, with K_1 closed and K_2 compact. Then there exists a closed hyperplane M such that K_1 and K_2 lie strictly on opposite sides of M. State and prove a similar result for a complex locally convex space.

10. a. Let X be a (real or complex) linear space. Given $x_0 \in X$ and a seminorm p on X, there exists $x' \in X^f$ such that $x'(x_0) = p(x_0)$ and $|x'(x)| \leq p(x)$ for $x \in X$. [*Suggestion.* Define a suitable linear functional on the subspace generated by x_0.]

    b. Suppose X is a (real or complex) topological linear space, and let x_0 be a point on the boundary of an absolutely convex set V. Then there exists $x' \in X'$ such that $x'(x_0) = 1$ and $|x'(x)| \leq 1$ for $x \in V$.

11. a. Let K be a convex absorbing set in a real linear space X, let p be the

Minkowski functional of K, and let L by a hyperplane in X that does not intersect K. Show that there exists $x' \in X^f$ such that $L = \{x : x'(x) = 1\}$ and $-p(-x) \leq x'(x) \leq p(x)$ for $x \in X$.

b. Prove the geometric form of the Hahn–Banach theorem, Theorem 2.4, using Theorem I.10.4. [*Suggestion.* Assume L is a subspace of X disjoint from the open convex set K. Find a point $x_0 \in X \backslash (L \cup K)$, and let X_0 be the subspace generated by L and x_0. Then L is a hyperplane in X_0. Apply a suitable translation to X_0 and $K \cap X_0$ in order to apply part (a). Obtain a linear functional on X_0, and then use Theorem I.10.4.]

III.3 THE CONJUGATE OF A NORMED LINEAR SPACE

Throughout this section, X denotes a normed linear space. Recall from § II.4 that the conjugate space X' is a Banach space under the norm

$$\|x'\| = \sup_{\|x\|=1} |x'(x)|. \tag{3-1}$$

We shall demonstrate that there exist elements in X' having certain prescribed properties, and we shall describe the conjugate spaces of two types of normed linear spaces associated with X. Our principal tool, of course, will be the Hahn–Banach theorem. We begin by proving this theorem in a form similar to that first given by Hahn in 1927.

Theorem 3.1 (Hahn–Banach). *Suppose M is a proper subspace of the normed linear space X. If $m' \in M'$, then X' contains an element x' such that $\|x'\| = \|m'\|$ and $x'(x) = m'(x)$ if $x \in M$.*

Proof. We define $p(x) = \|m'\| \, \|x\|$, $x \in X$. Then p is a seminorm (in fact a norm if $m' \neq 0$), and $|m'(x)| \leq p(x)$ if $x \in M$. By Theorem 2.11, there exists a linear functional x' on X that is an extension of m', such that $|x'(x)| \leq \|m'\| \, \|x\|$, for $x \in X$. This implies that $x' \in X'$ and $\|x'\| \leq \|m'\|$. But $\|m'\| \leq \|x'\|$ because x' is an extension of m'. Thus $\|x'\| = \|m'\|$. □

Theorem 3.2. *If x_0 is a nonzero element of X, then there exists $x' \in X'$ such that $\|x'\| = 1$ and $x'(x_0) = \|x_0\|$. As a consequence, we have*

$$\sup_{\|x'\|=1} |x'(x)| = \|x\| \tag{3-2}$$

for each $x \in X$.

Proof. Let M be the subspace of X generated by x_0. If $x = \alpha x_0$, define $m'(x) = \alpha \|x_0\|$. It is clear that $m' \in M'$ and $\|m'\| = 1$. The existence of the required x' now follows from Theorem 3.1. Verification of (3-2) is left to the reader. □

It follows from (3-2) that if an x in X satisfies $x'(x)=0$ for every x' in X', then $x=0$.

The next theorem provides useful representations of the conjugates of subspaces and quotient spaces of X. It is convenient here to adopt the notation $M^\perp$ for the annihilator of M in X' rather than in X^f (cf. § I.13). Properties of annihilators will be discussed later in § 7.

Theorem 3.3. *Let M be a subspace of X, and let $M^\perp=\{x'\in X': x'(x)=0$ if $x\in M\}$. Then*

(*a*) *$M^\perp$ is a closed subspace of X', and the conjugate space M' is congruent to $X'/M^\perp$; we write*

$$M'\cong X'/M^\perp. \tag{3-3}$$

(*b*) *If M is a closed subspace of X, the conjugate space $(X/M)'$ is congruent to $M^\perp$; we write*

$$(X/M)'\cong M^\perp. \tag{3-4}$$

Proof. (*a*) It is evident that $M^\perp$ is a closed subspace. Given $m'\in M'$, there exists an extension x' in X', by Theorem 3.1. Let Tm' be the element $[x']$ in $X'/M^\perp$ containing x'. If y' is another extension of m', then $x'-y'\in M^\perp$, and so $[x']=[y']$. Thus T is a well-defined linear mapping from M' into $X'/M^\perp$. Also, since each y' in $[x']=Tm'$ is an extension of m' we have

$$\|m'\|\leqslant \inf\{\|y'\|: y'\in Tm'\}=\|Tm'\|.$$

However, Theorem 3.1 provides an extension y' such that $\|m'\|=\|y'\|$. It follows that $\|m'\|=\|Tm'\|$. Finally, since the restriction of any x' to M is an element of M', we conclude that T is an isometric isomorphism of M' *onto* $X'/M^\perp$.

(*b*) Let ϕ be the continuous canonical mapping of X onto X/M (cf. § II.5). Given $u'\in(X/M)'$, let $Tu'=u'\circ\phi$. Then T is a linear mapping from $(X/M)'$ into X'. Observe that if $x\in X$, then $\|x\|<1$ if and only if $\|[x]\|=\|\phi(x)\|<1$, by definition of the (quotient) norm of $\|[x]\|$. Also, note that $\|x'\|=\sup\{|x'(x)|:\|x\|<1\}$ (see problem 2). Then

$$\begin{aligned}\|Tu'\|&=\sup\{|(Tu')(x)|:\|x\|<1\}\\&=\sup\{|u'([x])|:\|x\|<1\}\\&=\sup\{|u'([x])|:\|[x]\|<1\}=\|u'\|.\end{aligned}$$

Thus T is a linear isometry. Now the range of T is obviously contained in $M^\perp$. Conversely, if $x'\in M^\perp$, define $u'([x])=x'(x)$. Then u' is a well-defined linear functional on X/M, and

$$\begin{aligned}\sup\{|u'([x])|:\|[x]\|<1\}&=\sup\{|x'(x)|:\|x\|<1\}\\&=\|x'\|<\infty,\end{aligned}$$

which proves that u' is continuous. Hence $x' = Tu' \in \mathscr{R}(T)$. We conclude that T is an isometric isomorphism of $(X/M)'$ onto $M^{\perp}$. $\square$

The next theorem is a generalization of part of Theorem 3.2. It may be proved directly from Theorem 3.1, but it is also an easy corollary of Theorems 3.2 and 3.3.

Theorem 3.4. *Let M be a proper closed subspace of X. Given an x_0 in X at a positive distance h from M, there exists x' in X' such that $\|x'\| = 1$, $x'(x_0) = h$, and $x'(x) = 0$ if $x \in M$.*

Proof. Let ϕ be the canonical mapping of X onto X/M. Then $\|\phi(x_0)\| = h$. By Theorem 3.2, there exists $u' \in (X/M)'$ such that $\|u'\| = 1$ and $u'(\phi(x_0)) = h$. If $x' = u' \circ \phi$, we see from the proof of Theorem 3.3(b) that $x' \in M^{\perp} \subset X'$, $\|x'\| = \|u'\| = 1$ and $x'(x_0) = u'(\phi(x_0)) = h$. $\square$

As an application of Theorem 3.4, we shall prove the following interesting result about the conjugate space X'.

Theorem 3.5. *If X' is separable, so is X.*

Proof. Let $\{x'_n\}$ be a countable set that is everywhere dense on the set $\{x' : \|x'\| = 1\}$ in X'. Choose $x_n \in X$ so that $\|x_n\| = 1$ and $|x'_n(x_n)| \geqslant \frac{3}{4}$. Let M be the closed linear manifold in X generated by the sequence $\{x_n\}$. If $M \neq X$, then Theorem 3.4 tells us that there exists $x' \in X'$ such that $\|x'\| = 1$ and $x'(x) = 0$ if $x \in M$. Then $x'(x_n) = 0$ if $n = 1, 2, \ldots$, and

$$\begin{aligned} \tfrac{3}{4} &\leqslant |x'_n(x_n)| = |x'_n(x_n) - x'(x_n)| \\ &\leqslant \|x'_n - x'\|\|x_n\| = \|x'_n - x'\|. \end{aligned}$$

This contradicts the fact that $\|x'_n - x'\|$ can be made as small as we please by suitable choice of n. Hence $M = X$. It then follows that linear combinations formed from $\{x_n\}$ with rational scalar coefficients constitute a countable set everywhere dense in X, so that X is separable. $\square$

The converse of Theorem 3.5 is false. For example, ℓ^1 is separable, but it turns out (see Theorem 5.2) that $(\ell^1)'$ is congruent to ℓ^{∞}, which is not separable.

The Hahn–Banach theorem can be used to prove an existence theorem for Green's function for the Dirichlet problem, under certain conditions. We present a brief sketch of the line of argument of this proof. For simplicity we consider the two-dimensional case.

Let D be a bounded, connected open set in the plane, with boundary C consisting of a finite number of smooth curves. Let B be the real Banach space of continuous functions f defined on C, with $\|f\|$ = the maximum value of $|f|$ on C. Let B_0 be the linear manifold in B consisting of those f for which the Dirichlet problem for the region D is solvable (see the discussion of Example 1, § I.5). With each point $Q \in D$ we associate a continuous linear functional ℓ defined on B_0, the value of ℓ at $f \in B_0$ being $u(Q)$, where u is the solution of the Dirichlet problem for the boundary-value function f. That is, u is harmonic in D, continuous in $\bar{D}$, and $u = f$ on C. This functional ℓ is linear. Since $u \equiv 1$ if $f \equiv 1$ and since $|u(Q)| \leq$ maximum of $|f|$ on C (by the maximum-value theorem for harmonic functions), it is evident that $\ell \in B_0'$ and $\|\ell\| = 1$. By Theorem 3.1, there exists an element L of B' that is an extension of ℓ and for which $\|L\| = 1$. We shall denote L by L_Q to exhibit its dependence on Q.

Now let P be any point not on C in the plane. If t represents a point on C, let $g_P(t) = \log \overline{tP}$. Note that $g_P \in B_0$ if P is in the complement of $\bar{D}$. The value at Q of the corresponding solution of the Dirichlet problem is $u(Q) = \ell(g_P) = \log \overline{QP}$. Now, for any fixed P not on C, g_P is an element of B, and so we can apply L_Q to g_P; we define

$$k(P, Q) = L_Q(g_P).$$

If P is on C (and Q is in D), we define

$$k(P, Q) = \log \overline{QP}.$$

We then define

$$G(P, Q) = -\log \overline{QP} + k(P, Q)$$

for $Q \in D$ and any P. We assert that G is Green's function (to be considered as a function of P with singularity, or "pole," at Q). To show this it is merely necessary to prove that, as a function of P, $k(P, Q)$ is continuous on the set $\bar{D}$ and harmonic in D. Let Δ_P denote the Laplacian operator. Then, for P in D,

$$\Delta_P k(P, Q) = L_Q(\Delta_P g_P) = L_Q(0) = 0.$$

The commuting of Δ_P and L_Q is justified by the fact that L_Q is a bounded operator, just as in differentiating under an integral sign. Hence $k(P, Q)$ is harmonic in D. All that remains is to prove continuity of $k(P, Q)$ at points P on C.

We know that $L_Q(g_P) = \log \overline{QP}$ if $Q \in D$ and P is *not* in $\bar{D}$. Hence, if $P_0 \in C$ and $R \to P_0$ from outside of $\bar{D}$, it follows that $k(R, Q) \to \log \overline{QP_0} = k(P_0, Q)$. We want to prove that $k(P, Q) \to k(P_0, Q)$ as $P \to P_0$ from inside C. It will suffice to show that, with each point P sufficiently near P_0 and in D, we can associate a point R not in $\bar{D}$ such that $R \to P_0$ and $k(P, Q) - k(R, Q) \to 0$ as $P \to P_0$. We demonstrate this as follows: With P given, draw a straight line from P to a nearest point N of C; this line is normal to C at N. Continue the line beyond N to a point R such that $\overline{PN} = \overline{NR}$. If P is sufficiently near P_0, R

will not be in $\bar{D}$. We have

$$k(P, Q) - k(R, Q) = L_Q(g_P - g_R),$$

$$g_P(t) - g_R(t) = \log(\overline{tP}/\overline{tR}).$$

It can be proved that, as $P \to P_0$, $R \to P_0$ and

$$(\overline{tP}/\overline{tR}) \to 1$$

uniformly with respect to t on C. Thus $\|g_P - g_R\| \to 0$. The desired result now follows, owing to the continuity of the linear functional L_Q.

For more details on this subject, see the papers by P. D. Lax and C. Miranda listed in the bibliography.

PROBLEMS

1. Prove Theorem 3.1 using Theorem I.10.4 instead of Theorem 2.11. [*Suggestion.* Let $p(x) = \|m'\| \|x\|$, and apply Theorem I.10.4 to Re m'. Then use the discussion preceding Theorem 2.5.]
2. Let X be a normed linear space, and let $S = \{x : \|x\| < 1\}$. Show that if $x' \in X'$, then $\|x'\| = \sup_{x \in S} |x'(x)|$. [*Suggestion.* Let $f(x) = |x'(x)|$, and observe that $f(\bar{S}) \subset \overline{f(S)}$ because f is continuous.]
3. Let Y be a subspace of a normed linear space X. For $x \in X$, show that the distance from x to Y is given by

$$\operatorname{dist}(x, Y) = \sup \{|x'(x)| : \|x'\| = 1, x'(y) = 0 \text{ for } y \in Y\}.$$

   In the special case when $Y = \{x : x'(x) = 0\}$ for some $x' \in X'$ with $\|x'\| = 1$, show that dist $(x, Y) = |x'(x)|$.
4. Let X be a normed linear space. Show that if $x' \in X'$ and $x' \neq 0$, then

$$\|x'\| = \sup_{x \in M} \frac{|x'(x)|}{\|x\|},$$

   where M is any hyperplane of the form $\{x : x'(x) = \alpha\}$, $\alpha \neq 0$.
5. Let Y be a closed subspace of a normed linear space X. For each $x \in X$, define

$$\mathcal{P}_Y(x) = \{y \in Y : \|x - y\| = \operatorname{dist}(x, Y)\}.$$

   The elements of $\mathcal{P}_Y(x)$, if such elements exist, are the best approximations to x by elements of Y.
   a. Given $x \in X \setminus Y$ and $y_0 \in Y$, show that $y_0 \in \mathcal{P}_Y(x)$ if and only if there exists $x' \in X'$ such that $\|x'\| = 1$, $x'(y) = 0$ for $y \in Y$, and $x'(x - y_0) = \|x - y_0\|$.
   b. If $\mathcal{P}_Y(x)$ is a nonempty set for each $x \in X$, we call Y an ***existence subspace*** or ***proximinal subspace***. Show that a closed hyperplane H (in X) containing 0 is an existence subspace if and only if each translate of H contains an element of minimal norm.

6. Show that the following statements about a normed linear space X are equivalent.
   a. Each $x' \in X'$ assumes its supremum on the unit ball of X, that is, given $x' \in X'$, there exists $x \in X$ such that $\|x\| = 1$ and $x'(x) = \|x'\|$.
   b. Each closed hyperplane (in X) contains an element of minimal norm.
   c. Each closed hyperplane containing 0 is an existence subspace.

III.4. THE SECOND CONJUGATE SPACE

If X is a normed linear space, we denote the conjugate of X' by X''. This space is sometimes called the *bidual* of X. We shall use x'' as a typical notation for an element of X'', and we shall often write $\langle x', x''\rangle$ in place of $x''(x')$.

Given $x \in X$, the mapping $x' \mapsto x'(x)$ is a linear functional on X' which we denote by Jx. Since $|Jx(x')| = |x'(x)| \leq \|x\| \|x'\|$, it is clear that Jx is a continuous linear functional on X'. The correspondence $x \mapsto Jx$ is called the canonical mapping of X into X''. The definition of J is expressed by the formula

$$\langle x', Jx\rangle = \langle x, x'\rangle, \qquad x \in X, x' \in X'. \tag{4-1}$$

From this it is clear that J is a linear mapping. Furthermore, from the definition of the norm in X'' and from (3-2), we have

$$\|Jx\| = \sup_{\|x'\|=1} |\langle x', Jx\rangle| = \sup_{\|x'\|=1} |\langle x, x'\rangle| = \|x\|. \tag{4-2}$$

Thus J is also an isometry and sets up a congruence between X and a subspace of X''. We shall discuss the range of J later in § 6 and § 10.

In § II.4 we mentioned the fact that an incomplete normed linear space X has an essentially unique Banach space completion $\hat{X}$ and sketched one way of obtaining $\hat{X}$. We can now present an alternate procedure. Observe that X'' is complete because it is a conjugate space (Theorem II.4.5). Hence the closure $\overline{J(X)}$ of $J(X)$ in X'' is a complete space in itself. To form $\hat{X}$, we enlarge X by adjoining to it the elements of $\overline{J(X)} \setminus J(X)$; then we carry over to $\hat{X}$ the linear space structure and norm from $\overline{J(X)}$. Because X and $J(X)$ are isometrically isomorphic, it is clear that X is dense in $\hat{X}$.

A formula like (4-1) appeared earlier in § I.7. There we defined Jx to be a functional on X^f with $Jx(x') = x'(x)$. Here we have simply restricted the domain of each functional Jx to the elements of X'. In this way J maps X into X'' rather than into X^{ff}.

Reflexive Spaces

If the range of J is all of X'', we say that X is (norm) *reflexive*. Since X'' is complete, whether or not X is, it is clear from (4-2) that X cannot be reflexive if it is not complete. On the other hand, it may even happen that X is complete and congruent to X'' and yet is not reflexive. (In this case the isometric isomorphism between X and X'' cannot be the canonical mapping J.) R. C. James [1] has constructed a Banach space X that is congruent to X'', yet the range of J is only a hyperplane in X''.

Theorem 4.1. *A Banach space X is reflexive if and only if X' is reflexive.*

Proof. Let $J_0: X \to X''$ and $J_1: X' \to X'''$ be the canonical mappings. First, suppose that X is reflexive. Take $x''' \in X'''$, and let $x' = x''' \circ J_0$. Then x' is a composition of continuous linear mappings and hence belongs to X'. Since $\mathscr{R}(J_0) = X''$, it is easy to verify that $x''' = J_1 x'$. Thus $\mathscr{R}(J_1) = X'''$, and X' is reflexive.

Conversely, suppose X' is reflexive, and suppose X is not reflexive. Then $J_0(X)$ is a proper closed subspace of X'', since X is complete and J_0 is an isometry. By Theorem 3.4, there exists $w''' \in X'''$ such that $\|w'''\| = 1$ and $w'''(x'') = 0$ for all $x'' \in J_0(X)$. Since X' is reflexive, there exists $v' \in X'$ such that $\|v'\| = \|w'''\| = 1$ and $J_1 v' = w'''$; that is,

$$\langle x'', w''' \rangle = \langle v', x'' \rangle, \ x' \in X'.$$

Now for $x \in X$,

$$0 = \langle J_0 x, w''' \rangle = \langle v', J_0 x \rangle = \langle x, v' \rangle.$$

This implies that v' is the zero functional on X, which contradicts $\|v'\| = 1$. Thus X is reflexive. □

Theorem 4.2. *If a Banach space X is reflexive, the same is true of each closed subspace of X.*

Proof. Let M be a closed subspace of X. For each $x' \in X'$, let x'_M denote the restriction of x' to M. Take m''_0 in M'', and define x''_0 by

$$x''_0(x') = m''_0(x'_M), \qquad x' \in X'. \tag{4-3}$$

Then $|x''_0(x')| \leq \|m''_0\| \|x'_M\| \leq \|m''_0\| \|x'\|$, which implies that $x''_0 \in X''$. Since X is reflexive, there exists $x_0 \in X$ such that

$$\langle x', x''_0 \rangle = \langle x_0, x' \rangle, \qquad x' \in X'. \tag{4-4}$$

In particular, whenever $x'_M = 0$ we have

$$\langle x_0, x' \rangle = \langle x', x''_0 \rangle = \langle x'_M, m''_0 \rangle = 0.$$

In view of Theorem 3.4, this implies that x_0 must be in M (since M is closed). Now for $m' \in M'$, let x' in X' be some extension of m', so that $x'_M = m'$. Then, from (4-3) and (4-4),

$$\langle m', m''_0 \rangle = \langle x'_M, m''_0 \rangle = \langle x_0, x' \rangle = \langle x_0, m' \rangle.$$

It follows that $m_0'' = J_M x_0$, where $J_M : M \to M''$ is the canonical mapping. Thus M is reflexive. □

A result similar to Theorem 4.2 holds for quotient spaces. Namely, if X is reflexive, each quotient space X/M is reflexive (when M is a closed subspace of X). A simple proof of this can be given using ideas developed in Chapter IV. See problem 5, § IV.8.

PROBLEMS

1. A finite-dimensional normed linear space is reflexive.
2. The range of J is closed in X'' if and only if X is complete.
3. Let X be a reflexive Banach space. Then each $x' \in X'$ assumes its supremum on the unit ball of X (cf. problem 6, § 3). The converse of this is true for a Banach space, but the proof is difficult (cf. R. C. James [2]).

III.5. SOME REPRESENTATIONS OF LINEAR FUNCTIONALS

In this section we shall obtain representation theorems for linear functionals on certain particular normed linear spaces. These representation theorems are of the following sort. If X is a certain given space, we find another normed linear space Y that is congruent to X' in such a way that if $x' \in X'$ and if the corresponding element of Y is y, for each $x \in X$ the value of $x'(x)$ is expressed in a well-determined way by an analytical process involving x and y. This process usually involves an infinite series or some kind of integral. The one outstanding exception to this will be the representation of continuous linear functionals on a Hilbert space X. In this case X' can be identified in a simple way with X itself.

After discussing Hilbert spaces, we shall consider the conjugates of several important normed linear spaces. Considerable attention will be given to the classical representation theorem of F. Riesz for the conjugate of $C[a, b]$. An important generalization of this, the Riesz–Kakutani theorem, will be treated briefly—its proof would require a rather long digression into measure theory. The section will conclude with a look at continuous linear functionals on two important locally convex spaces.

Continuous Linear Functionals on a Hilbert Space

Theorem 5.1 (which follows) is sometimes referred to as the Fréchet–Riesz theorem for Hilbert spaces. It was announced for the space $L^2(a, b)$ by M. Fréchet and F. Riesz simultaneously and independently in 1907; Fréchet published a detailed proof for $L^2(0, 2\pi)$ shortly thereafter. Our proof follows an argument given by Riesz in 1934.

Theorem 5.1 (Fréchet–Riesz). *Let X be a Hilbert space. Then to each $x' \in X'$ corresponds a unique $y \in X$ such that $\|y\| = \|x'\|$ and $x'(x) = (x, y)$ for every $x \in X$.*

Proof. There cannot be more than one such y for, if $(x, y_1) = (x, y_2)$ for $x \in X$, we may take $x = y_1 - y_2$ and obtain $\|y_1 - y_2\|^2 = 0$. Now, if $x' = 0$, we may take $y = 0$ in the theorem. Hence we assume $x' \neq 0$. Then $M = \{x : x'(x) = 0\}$ is a closed hyperplane in X (since x' is continuous), and $X = M \oplus M^{\perp}$, by Theorem II.7.4. Since $M \neq X$, there exists a nonzero z in $M^{\perp}$. Multiplying z by a suitable scalar, we may assume that $x'(z) = 1$. Then $x - x'(x)z \in M$ for $x \in X$. Since $z \in M^{\perp}$, this implies that $(x, z) = x'(x)(z, z)$. Setting $y = z/\|z\|^2$, we have $(x, y) = x'(x)$ for $x \in X$. In particular, for $x_0 = y/\|y\|$, we have $\|x'\| \geqslant |x'(x_0)| = |(x_0, y)| = \|y\|$. Also, by the Cauchy–Schwarz inequality,

$$\|x'\| = \sup_{\|x\|=1} |(x, y)| \leqslant \|y\|.$$

Thus $\|x'\| = \|y\|$. □

The Fréchet–Riesz theorem establishes an isometric mapping T of X' *onto* X, since each y in X determines a continuous linear functional $x \mapsto (x, y)$. It is readily verified that T is "conjugate-linear" (problem 8). Thus X' is congruent to X when X is a real Hilbert space. When the scalars are complex, X' is not technically "congruent" to X, but this situation causes no difficulties.

Theorem 5.1 will be useful in our later work on spectral theory in a Hilbert space. In a different direction, the theorem can be employed to give an elegant proof of the Radon–Nikodym theorem. See Taylor, [5, pages 360–361].

The Conjugate of ℓ^p, $1 \leqslant p < \infty$

The spaces ℓ^p were discussed in § II.2. With each p we associate p', defined as follows:

$$p' = \begin{cases} p/(p-1) & \text{if } 1 < p < \infty, \\ \infty & \text{if } p = 1, \\ 1 & \text{if } p = \infty. \end{cases}$$

Note that $(p')' = p$ and that

$$\frac{1}{p} + \frac{1}{p'} = 1 \qquad \text{if } 1 < p < \infty. \tag{5-1}$$

Note also that $p' = 2$ if and only if $p = 2$ and that $1 < p < 2$ implies $2 < p' < \infty$.

For general use, here and elsewhere, we define

$$\operatorname{sgn} c = \begin{cases} 0 & \text{if } c = 0 \\ \dfrac{c}{|c|} & \text{if } c \neq 0. \end{cases}$$

Here c may be any real or complex number. Sometimes it is convenient in printing to denote $\operatorname{sgn} \bar{c}$ by $\overline{\operatorname{sgn}}\, c$, especially if c is replaced by a lengthy or bulky expression. Observe that $c \operatorname{sgn} \bar{c} = |c|$ and that $|\operatorname{sgn} c| = 0$ or 1 according to whether c is 0 or $\neq 0$.

If $x = \{\xi_n\} \in \ell^p$, we denote by u_k that x for which $\xi_k = 1$ and $\xi_n = 0$ if $n \neq k$. If $1 \leqslant p < \infty$, we see that for any $x \in \ell^p$

$$\left\| x - \sum_{k=1}^{n} \xi_k u_k \right\| = \left(\sum_{n+1}^{\infty} |\xi_k|^p \right)^{1/p} \to 0$$

as $n \to \infty$, so that

$$x = \sum_{k=1}^{\infty} \xi_k u_k. \tag{5-2}$$

It follows that

$$x'(x) = \sum_{k=1}^{\infty} \xi_k x'(u_k) \tag{5-3}$$

if $x' \in X'$. The condition $p < \infty$ is essential for (5-3).

The problem of representing x' is now seen to be the problem of describing what sets of values are admissible for $x'(u_k)$, $k = 1, 2, \ldots$, and of giving a formula for $\|x'\|$ in terms of these values. The solution of the problem is given in the following theorem.

Theorem 5.2. *Suppose $1 \leqslant p < \infty$. Every continuous linear functional on ℓ^p is representable in one and only one way in the form*

$$x'(x) = \sum_{k=1}^{\infty} \alpha_k \xi_k, \tag{5-4}$$

where the sequence $a = \{\alpha_n\}$ is an element of $\ell^{p'}$. Every element a of $\ell^{p'}$ can be used in this way to define an element of $(\ell^p)'$, and the correspondence between x' and a is a congruence of $(\ell^p)'$ and $\ell^{p'}$. In particular,

$$\|x'\| = \begin{cases} \left(\displaystyle\sum_{k=1}^{\infty} |\alpha_k|^{p'} \right)^{1/p'} & \text{if } 1 < p < \infty \\ \sup\limits_k |\alpha_k| & \text{if } p = 1. \end{cases} \tag{5-5}$$

Proof. Suppose $x' \in (\ell^p)'$ is given. Define $\alpha_k = x'(u_k)$. Then $x'(x)$ is given by (5-3) or (5-4). First, suppose that $1 < p < \infty$. The case $p = 1$ is considered later. For a given positive integer n, choose x as follows:

$$\xi_k = \begin{cases} |\alpha_k|^{p'-1} \operatorname{sgn} \bar{\alpha}_k & \text{if } 1 \leqslant k \leqslant n \\ 0 & \text{if } n < k. \end{cases} \tag{5-6}$$

Then, if $1 \leqslant k \leqslant n$,

$$\alpha_k \xi_k = |\alpha_k|^{p'} = |\xi_k|^p,$$

and so

$$\|x\| = \left(\sum_{k=1}^{n} |\alpha_k|^{p'} \right)^{1/p}, \qquad x'(x) = \sum_{k=1}^{n} |\alpha_k|^{p'}.$$

But $|x'(x)| \leqslant \|x'\| \|x\|$, and so

$$\sum_{k=1}^{n} |\alpha_k|^{p'} \leqslant \|x'\| \left(\sum_{k=1}^{n} |\alpha_k|^{p'} \right)^{1/p}.$$

It follows that

$$\left(\sum_{k=1}^{n} |\alpha_k|^{p'} \right)^{1/p'} \leqslant \|x'\|.$$

Hence $a = \{\alpha_n\}$ is in $\ell^{p'}$, and $\|a\| \leqslant \|x'\|$.

Suppose, on the other hand, that $a = \{\alpha_n\}$ is a given element of $\ell^{p'}$. We can define $x' \in (\ell^p)'$ by (5-4). By Hölder's inequality, we have

$$|x'(x)| \leqslant \left(\sum_{k=1}^{\infty} |\alpha_k|^{p'} \right)^{1/p'} \left(\sum_{k=1}^{\infty} |\xi_k|^p \right)^{1/p} = \|a\| \|x\|,$$

so that $\|x'\| \leqslant \|a\|$. Since $x'(u_k) = \alpha_k$ in this case, it follows from the preceding work that $\|a\| \leqslant \|x'\|$. Hence $\|a\| = \|x'\|$, and all is clear when $1 < p < \infty$.

If $p = 1$, we replace (5-6) by

$$\xi_k = \begin{cases} \operatorname{sgn} \bar{\alpha}_n & \text{if } k = n \\ 0 & \text{if } k \neq n. \end{cases}$$

Then $\|x\| \leqslant 1$, and (5-4) becomes $x'(x) = |\alpha_n|$, so that $|\alpha_n| \leqslant \|x'\| \|x\| \leqslant \|x'\|$, whence $a = \{\alpha_n\} \in \ell^\infty$, with $\|a\| \leqslant \|x'\|$. The rest of the argument for $p = 1$ is left to the reader. □

Theorem 5.2 is not true for the case $p = \infty$. If it were, $(\ell^\infty)'$ would be congruent to the separable space ℓ^1. Since ℓ^∞ is not separable, this would contradict Theorem 3.5.

The Conjugate of $L^p(a, b)$

Let (a, b) be a finite or infinite interval of the real line. The space $L^p(a, b)$ was described in Example 7, § I.2, and Example 5, § II.2. Let p' be defined as above.

Theorem 5.3. *Suppose $1 \leq p < \infty$. There is an isometric isomorphism of the conjugate space of $L^p(a, b)$ and the space $L^{p'}(a, b)$ such that if $x' \in (L^p(a, b))'$ and if $y \in \mathscr{L}^{p'}(a, b)$ is a function that represents the element of $L^{p'}(a, b)$ corresponding to x', we have*

$$x'(x) = \int_a^b x(t)y(t)\,dt \tag{5-7}$$

for each $x \in \mathscr{L}^p(a, b)$. The fact that this correspondence preserves norms means that

$$\|x'\| = \left(\int_a^b |y(t)|^{p'}\,dt\right)^{1/p'}$$

if $1 < p$, and

$$\|x'\| = \sup_{t \in [a, b]}{}^0 \; |y(t)|$$

in case $p = 1$ (in which case $p' = \infty$).

Theorem 5.3 is presented in most elementary texts on measure and integration. It is becoming a common practice, however, to prove the theorem with $L^p(a, b)$ replaced by $L^p(\mu)$. In this case, one considers a measure space (T, S, μ) where S is a σ-ring of subsets of T such that $T \in \mathrm{S}$ and μ is a measure on S. Then $\mathscr{L}^p(\mu)$ is the space of all μ-measurable functions x such that $|x|^p$ is integrable on T with respect to μ. The space $L^p(\mu)$ is formed from $\mathscr{L}^p(\mu)$ by identifying functions that coincide "μ-almost everywhere" on T. If $1 < p < \infty$, then $(L^p(\mu))'$ is isometrically isomorphic to $L^{p'}(\mu)$. If $p = 1$ and if T is the union of a countable family of measurable sets of finite measure, then $(L(\mu))'$ is isometrically isomorphic to $L^\infty(\mu)$. (These ideas are discussed in Taylor [5, § 8.4.])

Theorem 5.4. *If $1 < p < \infty$, the space $L^p(a, b)$ is reflexive.*

Proof. Let $X = L^p(a, b)$, and let $T: X' \to L^{p'}(a, b)$ be the isometric isomorphism described in Theorem 5.3. Given $x'' \in X''$, let $y'' = x'' \circ T^{-1}$. Then $y'' \in (L^{p'}(a, b))'$ and $y''(Tx') = x''(x')$ for $x' \in X'$. Applying Theorem 5.3 with p' in place of p, we obtain $y \in L^p(a, b)$ corresponding to y'' such that

$$y''(z) = \int_a^b z(t)y(t)\,dt, \qquad z \in L^{p'}(a, b).$$

Here, as we shall often do, we ignore the notational distinctions between elements of $L^p(a, b)$ and $\mathscr{L}^p(a, b)$, whatever p may be. Then, if $x' \in X'$,

$$x''(x') = y''(Tx') = \int_a^b (Tx')(t)y(t)\,dt = x'(y), \tag{5-8}$$

where the last equality merely expresses the definition of T. Since (5-8) holds for all $x' \in X'$, we have proved that $x'' = Jy$, where J is the canonical mapping from X into X''. Thus X is reflexive. □

The Conjugate of $C[a, b]$

Let $[a, b]$ be a finite interval in $\boldsymbol{R}$, and let $C[a, b]$ be the space described in Example 3, § I.2 and Example 4, § II.2. The scalars may be real or complex.

Theorem 5.5. *Suppose $x' \in X'$, where $X = C[a, b]$. Then there exists a function v of bounded variation, defined on $[a, b]$ and with values in the scalar field associated with X, such that the total variation of v is $\|x'\|$ and the values of x' are given by the Riemann–Stieltjes integral*

$$x'(x) = \int_a^b x(t)\,dv(t), \qquad x \in C[a, b]. \tag{5-9}$$

Proof. Recall that $C[a, b]$ is a subspace of $B[a, b]$ (Example 3, § II.2). By Theorem 3.1, there exists a continuous linear functional f on $B[a, b]$ that is an extension of x' and such that $\|f\| = \|x'\|$. If s is any point of $(a, b]$, let x_s be the characteristic function of $[a, s]$, and let $x_a(t) \equiv 0$ (for $s = a$). Then $x_s \in B[a, b]$. We define v by

$$v(s) = f(x_s). \tag{5-10}$$

We now proceed to show that v is of bounded variation and that $V(v) \leqslant \|x'\|$, where $V(v)$ is the total variation of v. Consider any partition of $[a, b]$:

$$a = t_0 < t_1 < \cdots < t_n = b.$$

Using the $\overline{\mathrm{sgn}}$ notation explained earlier, we let

$$\varepsilon_i = \overline{\mathrm{sgn}}\,[v(t_i) - v(t_{i-1})], \qquad i = 1, \ldots, n,$$

and we construct the function

$$y(t) = \begin{cases} \varepsilon_1 & \text{if } t_0 \leqslant t \leqslant t_1 \\ \varepsilon_i & \text{if } t_{i-1} < t \leqslant t_i, \qquad i = 2, \ldots, n. \end{cases}$$

Then $y \in B[a, b]$ and $\|y\| \leqslant 1$. We can write

$$y(t) = \sum_{i=1}^{n} \varepsilon_i [y_i(t) - y_{i-1}(t)],$$

where $y_k = x_{t_k}$. Thus

$$f(y) = \sum_{i=1}^{n} \varepsilon_i[f(y_i) - f(y_{i-1})] = \sum_{i=1}^{n} \varepsilon_i[v(t_i) - v(t_{i-1})]$$

$$= \sum_{i=1}^{n} |v(t_i) - v(t_{i-1})|.$$

But $|f(y)| \leqslant \|f\| \|y\| \leqslant \|f\| = \|x'\|$. Therefore

$$\sum_{i=1}^{n} |v(t_i) - v(t_{i-1})| \leqslant \|x'\|.$$

This shows that v is of bounded variation with $V(v) \leqslant \|x'\|$.

Now we proceed to prove the validity of (5-9). With the same notation as before for an arbitrary partition and with any given x in $C[a, b]$, we define $z \in B[a, b]$ as follows:

$$z(t) = \sum_{i=1}^{n} x(t_{i-1})[y_i(t) - y_{i-1}(t)].$$

Then

$$f(z) = \sum_{i=1}^{n} x(t_{i-1})[v(t_i) - v(t_{i-1})]. \tag{5-11}$$

Also,

$$|z(t) - x(t)| = \begin{cases} |x(t_0) - x(t)| & \text{if } a \leqslant t \leqslant t_1 \\ |x(t_{i-1}) - x(t)| & \text{if } t_{i-1} < t \leqslant t_i, \qquad i = 2, \ldots, n. \end{cases}$$

Let Δ denote the partition, and let

$$|\Delta| = \max \{|t_1 - t_0|, \ldots, |t_n - t_{n-1}|\}.$$

Then we see (by the uniform continuity of x) that $\|z - x\| \to 0$ when $|\Delta| \to 0$. Since f is continuous, $f(z) \to f(x)$. But also, it is clear from (5-11) and the definition of a Stieltjes integral that

$$f(z) \to \int_a^b x(t)\, dv(t).$$

Therefore we conclude that

$$f(x) = \int_a^b x(t)\, dv(t).$$

This is the same as (5-9), since f is an extension of x'. It is a standard property of Stieltjes integrals that

$$\left|\int_a^b x(t)\,dv(t)\right| \leq \|x\| V(v). \tag{5-12}$$

Consequently, $\|x'\| \leq V(v)$. We already know that $V(v) \leq \|x'\|$, and so $V(v) = \|x'\|$. $\square$

Theorem 5.5 makes no assertion about the uniqueness of v, nor does it assert that X' is congruent to some space of which v is a member. We observe in the first place that there is no unique function of bounded variation that makes (5-9) true for all $x \in C[a, b]$. For, we can add an arbitrary constant to v without affecting (5-9). It is also easy to see that the values of v at its points of discontinuity in the interior of the interval $[a, b]$ can be altered without affecting the value of the integral in (5-9). Indeed, if w is a function of bounded variation that is equal to v at a, b and at all points t in the interior of $[a, b]$ at which v is continuous, the approximating sums that define

$$\int_a^b x(t)\,dv(t) \qquad \text{and} \qquad \int_a^b x(t)\,dw(t)$$

can be formed exclusively by use of values of v and w at points of the aforementioned kind, so that the integrals must be equal in value.

In order to obtain a congruence between the space X' and a suitable space of functions of bounded variation, we proceed in the following manner. As in Example 7, § II.2, we denote by $BV[a, b]$ the Banach space of functions of bounded variation defined on $[a, b]$, with

$$\|v\| = |v(a)| + V(v).$$

We can define an equivalence relation in $BV[a, b]$ by writing $v_1 \sim v_2$ if

$$\int_a^b x(t)\,dv_1(t) = \int_a^b x(t)\,dv_2(t)$$

for each $x \in C[a, b]$. We assert that $v \sim 0$ if and only if $v(a) = v(b)$ and $v(c-0) = v(c+0) = v(a)$ if $a < c < b$. To see this we observe first of all that $v \sim 0$ implies

$$0 = \int_a^b dv(t) = v(b) - v(a).$$

Next we observe the following. If $v \in BV[a, b]$ and $a \leq c < b$, then

$$\frac{1}{h}\int_c^{c+h} v(t)\,dt \to v(c+0) \tag{5-13}$$

as $h \to 0+$. Likewise, if $a < c \leqslant b$,

$$\frac{1}{h}\int_{c-h}^{c} v(t)\,dt \to v(c-0) \tag{5-14}$$

as $h \to 0+$. We leave the simple proofs of these facts to the reader. Now suppose that $a \leqslant c < b$, $0 < h < b-c$, and define

$$x(t) = \begin{cases} 1 & \text{if } a \leqslant t \leqslant c \\ 1-(t-c)/h & \text{if } c \leqslant t \leqslant c+h \\ 0 & \text{if } c+h \leqslant t \leqslant b. \end{cases}$$

Then $v \sim 0$ implies

$$0 = \int_a^b x(t)\,dv(t) = v(c) - v(a) + \int_c^{c+h} x(t)\,dv(t).$$

On integration by parts we obtain

$$\int_c^{c+h} x(t)\,dv(t) = -v(c) + \frac{1}{h}\int_c^{c+h} v(t)\,dt.$$

Using (5-13), we obtain the result $v(c+0) = v(a)$. A similar argument shows that $v(c-0) = v(b)$ if $a < c \leqslant b$. [We remark that $v(a+0) = v(a)$ is a consequence of $v(c+0) = v(a)$ for $a < c < b$, because there exists a sequence c_n such that $a < c_n < b$, $c_n \to a$, and $v(c_n+0) = v(c_n)$; likewise for $v(b-0) = v(b)$.]

Conversely, it is evident that if $v \in BV[a, b]$ and $v(a) = v(b) = v(c+0) = v(c-0)$ when $a < c < b$, then $v \sim 0$. For then $v(t) = v(a)$ at $t = a$, $t = b$, and all interior points of $[a, b]$ at which v is continuous, so that if $x \in C[a, b]$,

$$\int_a^b x(t)\,dv(t) = \int_a^b x(t)\,dw(t) = 0,$$

where $w(t) \equiv v(a)$.

Now let $X = C[a, b]$. We wish to associate with each $x' \in X'$ a unique $v \in BV[a, b]$ in such a way that the association makes X' congruent to a subspace of $BV[a, b]$ and such that x' and v are related by the formula (5-9). We do this by introducing the concept of a normalized function of bounded variation. We shall say that v is *normalized* if $v(a) = 0$ and if $v(t+0) = v(t)$ when $a < t < b$. Other definitions of normalization can be used; we might alternatively require $v(t-0) = v(t)$, or

$$v(t) = \tfrac{1}{2}[v(t+0) + v(t-0)].$$

Our choice is governed by consideration of some future applications.

The normalized functions form a linear manifold in $BV[a, b]$. Each equivalence class contains at most one normalized function. These facts are readily evident. From any $v \in BV[a, b]$ we can define another member v^* of $BV[a, b]$ as follows:

$$v^*(a)=0, \qquad v^*(b)=v(b)-v(a)$$
$$v^*(t)=v(t+0)-v(a), \qquad a<t<b.$$

Then one can easily verify that v^* is normalized and that $v^* \sim v$. Thus each equivalence class contains *exactly* one normalized function. Finally, we may show that $V(v^*) \leq V(v)$. For, suppose $a=t_0<t_1<\cdots<t_n=b$. If $\varepsilon>0$, we can choose points $s_1, \ldots, s_{n-1}$ at which v is continuous, with s_k so close to t_k (on the right) that

$$|v(t_k+0)-v(s_k)|<\frac{\varepsilon}{2n}.$$

It is then readily seen that if $s_0=a$ and $s_n=b$, we have

$$\sum_{k=1}^{n} |v^*(t_k)-v^*(t_{k-1})| \leq \sum_{k=1}^{n} |v(s_k)-v(s_{k-1})|+\varepsilon$$
$$\leq V(v)+\varepsilon,$$

and so $V(v^*) \leq V(v)+\varepsilon$. But then $V(v^*) \leq V(v)$.

We can now state:

Theorem 5.6 (The Riesz Representation Theorem). *Let $X=C[a, b]$. Then X' is congruent to the subspace of $BV[a, b]$ consisting of all normalized functions of bounded variation. If v is such a normalized function, the corresponding x' is given by*

$$x'(x)=\int_a^b x(t)\, dv(t). \tag{5-15}$$

Proof. Formula (5-15) defines a linear mapping $x'=Tv$, where v is normalized and $x' \in X'$. We evidently have $\|x'\| \leq V(v)$ [see (5-12)]. For a normalized v, $V(v)$ is the norm of v, because $v(a)=0$. Now consider any $x' \in X'$. Theorem 5.5 tells us that there is some $u \in BV[a, b]$ such that

$$x'(x)=\int_a^b x(t)\, du(t)$$

and $V(u)=\|x'\|$. The integral is not changed if we replace u by u^*. Then $x'=Tu^*$ and $\|x'\| \leq V(u^*)$. Also, $V(u^*) \leq V(u)=\|x'\|$. Therefore $\|x'\|=V(u^*)$. Since there is just one normalized function in each equivalence class, we see that the theorem is proved. □

The Riesz representation theorem for the conjugate of $C[0, 1]$ was proved by F. Riesz in 1909. It is the grandfather of numerous generalizations, some of which go by the same name. We shall describe below an extension given by S. Kakutani in 1941, where $[0, 1]$ is replaced by an arbitrary compact Hausdorff space T.

The Riesz–Kakutani Theorem

We begin by recalling some facts from measure theory. It is convenient to assume at first that T is a locally compact Hausdorff space which is also σ-compact (i.e., the union of a countable collection of compact subsets of T).

Let $\mathscr{B}$ be the smallest σ-ring of subsets of T that contains the compact subsets of T. The members of $\mathscr{B}$ are the *Borel* sets. Note that $T \in \mathscr{B}$ because T is σ-compact. A *Borel measure* on T is a countably additive (and non-negative) measure μ defined on $\mathscr{B}$ such that $\mu(K) < \infty$ if K is compact. We say that μ is *finite* if $\mu(T) < \infty$ and that μ is *regular* if

$$\begin{aligned} \text{(5-16)} \qquad \mu(E) &= \inf\{\mu(U) : U \text{ open Borel set, } U \supset E\} \\ &= \sup\{\mu(K) : K \text{ compact, } K \subset E\}, \end{aligned}$$

for each E in $\mathscr{B}$.

Let $C_c(T)$ denote the class of (real or complex) continuous functions on T that have compact support in T (cf. Example 3, § II.11). If μ is a regular Borel measure, then each f in $C_c(T)$ is integrable with respect to μ. (The Lebesgue theory of integration is developed, for example, in Taylor [5, Chapter 5], for real-valued functions. If f is complex-valued, the integral $\int_T f\,d\mu$ is defined in the obvious way, using the real and imaginary parts of f.)

A *finite signed Borel measure* is a difference of two finite Borel measures. Such a signed measure μ has a Jordan decomposition, $\mu = \mu^+ - \mu^-$, where μ^+ and μ^- are the upper and lower variations, respectively, of μ. For $f \in C_c(T)$, the μ-integral of f is defined by

$$\int_T f\,d\mu = \int_T f\,d\mu^+ - \int_T f\,d\mu^-.$$

A *complex Borel measure* is a set function of the form $\mu = \mu_1 + i\mu_2$, where μ_1 and μ_2 are finite signed Borel measures. For $f \in C_c(T)$ we define

$$\int_T f\,d\mu = \int_T f\,d\mu_1 + i\int_T f\,d\mu_2.$$

If μ is a signed or complex Borel measure, the *total variation* of μ is the function $|\mu|$, defined for $E \in \mathscr{B}$ by

$$|\mu|(E) = \sup \sum_{i=1}^{n} |\mu(E_i)|,$$

where the supremum is taken over all finite collections $\{E_i\}$ of disjoint Borel sets with $E_i \subset E$. It can be shown that $|\mu|$ is a Borel measure. We say that μ is *regular* if its total variation $|\mu|$ is regular.

It can be shown that the set of all regular complex Borel measures is a complex linear space under the obvious definitions of sums and scalar multiples. Further, this space is a Banach space under the "total variation norm"

$$\|\mu\| = |\mu|(T).$$

(Cf. Dunford and Schwartz [1, pages 97 and 161].)

Theorem 5.7 (The Riesz–Kakutani Theorem). *Let T be a compact Hausdorff space. Then the conjugate of the complex Banach space $C(T)$ is congruent to the Banach space of all regular complex Borel measures on T. The correspondence between a continuous linear functional L and a regular complex Borel measure μ is established by the formula*

$$L(f) = \int_T f\,d\mu, \qquad f \in C(T). \tag{5-17}$$

The conjugate of the *real* space $C(T)$ is congruent in the same way to the Banach space of all (finite) regular signed Borel measures on T.

A proof of the Riesz–Kakutani theorem is given in Dunford and Schwartz [1, pages 262–265]. We shall limit our discussion here to an outline of how each L in $C(T)'$ determines a regular complex Borel measure. We say that a linear functional L is *positive* if $L(f) \geqslant 0$ whenever $f \geqslant 0$ (i.e., whenever $f(t) \geqslant 0$ for all $t \in T$). It can be shown that each L in $C(T)'$ may be written in the form

$$L = (L_1 - L_2) + i(L_3 - L_4), \tag{5-18}$$

where the L_k are positive linear functionals. If there are regular Borel measures μ_k that correspond to the L_k as in (5-17), then it is readily verified that $\mu = (\mu_1 - \mu_2) + i(\mu_3 - \mu_4)$ will correspond to L. Thus it suffices to consider a positive linear functional L. If U is an open subset of T, let χ_U denote the characteristic function of U and define

$$\lambda(U) = \sup\{L(f) : f \in C(T), 0 \leqslant f \leqslant \chi_U\}.$$

For each subset E of T define

$$\mu^*(E) = \inf\{\lambda(U) : U \text{ open}, U \supset E\}.$$

It is easy to see that $\mu^*(U) = \lambda(U)$ if U is open. With some work it can be shown that μ^* is an "outer measure," the restriction of μ^* to the Borel sets is a regular Borel measure μ and, finally, $L(f)$ is given by (5-17). The reader may consult Rudin [1, pages 42–49] for these last details.

The Riesz–Kakutani theorem can be extended to the case where T is a locally compact Hausdorff space. See Rudin [1, pages 138–142], for example. Most measures encountered in functional analysis arise from the Riesz–Kakutani theorem for either compact or locally compact Hausdorff spaces. A variety of other generalizations of this theorem are discussed in Dunford and Schwartz [1, pages 380–381].

Radon Measures

When Ω is a locally compact Hausdorff space, a linear functional on $C_c(\Omega)$ which is continuous with respect to the supremum norm on $C_c(\Omega)$ is often called a *bounded Radon measure on* Ω. In the theory of distributions (to be discussed briefly later in this section), it is useful to consider a somewhat wider class of linear functionals on $C_c(\Omega)$. A *Radon measure on* Ω is a linear functional L on $C_c(\Omega)$ with the property that to each compact set K in Ω there corresponds a nonnegative number M such that

$$|L(f)| \leq M \cdot \sup_{x \in K} |f(x)| \tag{5-19}$$

whenever $f \in C_c(\Omega)$ and the support of f is contained in K. It can be shown that a linear functional on $C_c(\Omega)$ is a Radon measure if and only if it is continuous on the space $\mathscr{K}(\Omega)$ that $C_c(\Omega)$ becomes if given the locally convex topology defined in Example 3 of § II.11.

It can also be shown that a linear functional L on $C_c(\Omega)$ is a Radon measure if and only if it is a linear combination of positive linear functionals, as in (5-18). In this case each positive linear functional L_k in (5-18) determines a regular Borel measure μ_k. However, since Ω need not be compact, these measures need not be finite, and we have no assurance, for example, that the situation $\mu_1(E) = \mu_2(E) = \infty$ cannot arise for some Borel set E in Ω. Thus in general we cannot combine the μ_k into a regular complex Borel measure that will represent L as in (5-17).

Here, for example, is a construction of a Radon measure on Ω, where Ω is an open subset of $\boldsymbol{R}^n$. Given a continuous function g on Ω, define L_g by

$$L_g(f) = \int_{\boldsymbol{R}^n} f(x)g(x)\,dx, \qquad f \in C_c(\Omega). \tag{5-20}$$

This Lebesgue integral has a well-defined value if we agree to set $f(x)g(x) = 0$ when $x \in \boldsymbol{R}^n \backslash \Omega$. It is clear that L_g is a linear functional on $C_c(\Omega)$. Now let K be a compact set in Ω, let $|K|$ denote the Lebesgue measure of K, and let $M = |K| \cdot \sup_{x \in K} |g(x)|$. Then, if $f \in C_c(\Omega)$ and if the support of f is contained in K, we have

$$|L_g(f)| \leq \int_{\boldsymbol{R}^n} |f(x)g(x)|\,dx \leq M \cdot \sup_{x \in K} |f(x)|.$$

Thus L_g is a Radon measure. It can be shown that distinct continuous functions on Ω determine distinct Radon measures. It follows that $g \mapsto L_g$ is a one-to-one linear mapping of $C_c(\Omega)$ into $\mathscr{K}(\Omega)'$.

An example of a Radon measure on Ω that does not come from a continuous function is the *Dirac measure* δ_a, or "unit charge at a," where $a \in \Omega$. This is a positive Radon measure defined by

$$\delta_a(f) = f(a), \qquad f \in C_c(\Omega).$$

The condition (5-19) is satisfied, since $|f(a)| \leq \sup_{x \in K} |f(x)|$ whenever K is a compact set containing the support of f. It is a simple matter to describe the regular Borel measure μ that corresponds to δ_a: for each Borel set E, let $\mu(E) = 1$ if $a \in E$ and $\mu(E) = 0$ if $a \notin E$. Then

$$\delta_a(f) = \int_{R^n} f \, d\mu.$$

Distributions

Let $\mathscr{D}(\Omega)$ be the space defined in Example 6, § II.11 (Ω an open set in $\boldsymbol{R}^n$). A continuous linear functional on $\mathscr{D}(\Omega)$ is called a *distribution on* Ω. It can be shown that a linear functional L on $\mathscr{D}(\Omega)$ is a distribution if and only if for every compact subset K of Ω there exist a nonnegative constant M and a nonnegative integer m such that, for all $f \in \mathscr{D}(\Omega)$ with the support of f contained in K, we have

$$|L(f)| \leq M \cdot \max_{|p| \leq m} \sup_{x \in K} |\partial^p f(x)|. \tag{5-21}$$

Since each function in $\mathscr{D}(\Omega)$ is also in $\mathscr{K}(\Omega)$, the restriction of a Radon measure L to the functions in $\mathscr{D}(\Omega)$ is a linear functional on $\mathscr{D}(\Omega)$. In fact, if we apply (5-19) to a function f in $\mathscr{D}(\Omega)$ whose support is contained in K, we see that (5-21) holds with $m = 0$. Thus L is continuous on $\mathscr{D}(\Omega)$. In this way each Radon measure "is" a distribution. Furthermore, it can be shown that distinct Radon measures determine distinct distributions. It follows that if $g \in C(\Omega)$ and $L_g(f)$ is defined by (5-20) for $f \in \mathscr{D}(\Omega)$, then L_g is a distribution on Ω. Because of the (one-to-one) correspondence $g \mapsto L_g$, distributions on Ω are sometimes referred to as generalized functions. See Horváth [1, 2] and Treves [1].

PROBLEMS

1. Let (c) denote the space of sequences $x = \{\xi_n\}$ such that $\lim_{n \to \infty} \xi_n = \xi_0$ exists. Let $\|x\| = \sup_n |\xi_n|$. Then (c) is a subspace of ℓ^∞. Show that it is complete. Let u_0

be the x for which $\xi_n = 1$ for every n, and let $u_1, u_2, \ldots$ be as in the discussion of ℓ^p. Show that

$$x = \xi_0 u_0 + \sum_{k=1}^{\infty} (\xi_k - \xi_0) u_k.$$

If $x' \in (c)'$ and $x'(u_k) = \alpha_k$, show by special choice of x that $\Sigma_{k=1}^{\infty} |\alpha_k| < \infty$. Then, in general, we can write

$$x'(x) = \left(\alpha_0 - \sum_{k=1}^{\infty} \alpha_k\right) \xi_0 + \sum_{k=1}^{\infty} \alpha_k \xi_k.$$

Show that $(c)'$ is congruent to ℓ^1 under the correspondence $x' \leftrightarrow b = \{\beta_n\}$, where

$$x'(x) = \sum_{n=1}^{\infty} \beta_n \xi_{n-1}.$$

2. Let (c_0) be the closed subspace of (c) consisting of those $x \in (c)$ for which $\xi_n \to 0$. Show that $(c_0)'$ is congruent to ℓ^1 under the correspondence $x' \leftrightarrow a = \{\alpha_n\}$, where

$$x'(x) = \sum_{k=1}^{\infty} \alpha_k \xi_k.$$

3. If $1 < p < \infty$, then ℓ^p is reflexive
4. ℓ^1 and $L^1(a, b)$ are not reflexive.
5. (c) and (c_0) are not reflexive.
6. State and prove a representation theorem for $\ell^p(n)$, $1 \leqslant p \leqslant \infty$.
7. Let (a, b) be a finite interval of the real line.
   a. Show that each $y \in L^1(a, b)$ determines an element of $(L^\infty(a, b))'$.
   b. Show that $C[a, b]$ may be identified with a closed subspace of $L^\infty(a, b)$.
   c. The Dirac measure δ_a on $C[a, b]$ can be extended (by the Hahn–Banach theorem) to a continuous linear functional on $L^\infty(a, b)$. Show that the extension cannot be represented as in (5-7) by some y in $\mathscr{L}^1(a, b)$.
8. Let X be a Hilbert space, and let $T: X' \to X$ be the isometric mapping described in Theorem 5.1.
   a. Show that T is a "conjugate-linear" mapping; that is, show that $T(x' + y') = Tx' + Ty'$ and $T(\alpha x') = \bar{\alpha} Tx'$.
   b. For $x', y' \in X'$, define $(x', y') = (Tx', Ty')$. Show that this makes X' into a Hilbert space.
   c. Use Theorem 5.1 to show that X is reflexive.
9. Let X be the space of all continuous (real or complex) functions on $[a, b]$, and let X have the topology of pointwise convergence on $[a, b]$ (cf. Example 1, § II.11). Find a representation for the conjugate space X'.
10. a. Let X be a linear space, and let $(X_\alpha)_{\alpha \in A}$ be an indexed family of locally convex spaces. For each index $\alpha \in A$ let f_α be a linear mapping of X_α into X. Let X have the strongest locally convex topology for which every f_α is continuous, let Y be a locally convex space, and let T be a linear

mapping from X into Y. Show that T is continuous if and only if every composite mapping $T \circ f_\alpha$ is continuous (from X_α into Y).

b. Let $(K_\alpha)_{\alpha \in A}$ be an indexing of the collection of all compact subsets of an open set Ω in $\boldsymbol{R}^n$. For each $\alpha \in A$ let X_α be the space of all continuous functions on Ω whose support is contained in K_α, and let X_α have the supremum norm topology. Let $X = C_c(\Omega)$, and let $f_\alpha : X_\alpha \to X$ be the inclusion mapping. Show that the topology on X described in part (a) is the topology of $\mathscr{K}(\Omega)$ defined in Example 3, § II.11.

c. Show that a linear functional L on $C_c(\Omega)$ is a Radon measure if and only if for each $\alpha \in A$ the restriction L_α of L to X_α is continuous.

III.6 WEAK TOPOLOGIES FOR LINEAR SPACES

Suppose that X is a linear space (real or complex). We wish to define locally convex topologies on X in terms of certain seminorms. It is readily verified that if F_0 is any nonempty finite subset of X^f, then the function p defined by

$$p(x) = \max_{x' \in F_0} |x'(x)|, \qquad x \in X, \tag{6-1}$$

is a seminorm on X. Given a fixed nonempty subset F of X^f, we let $\sigma(X, F)$ denote the locally convex topology on X defined by the family $\mathscr{P}$ of all seminorms (6-1), where F_0 ranges over all nonempty finite subsets of F. A base at 0 for this topology is given by sets of the form

$$\begin{aligned} U &= \{x : |x'(x)| < r \text{ for each } x' \in F_0\} \\ &= \bigcap_{x' \in F_0} \{x : |x'(x)| < r\}, \end{aligned} \tag{6-2}$$

where $r > 0$ and F_0 is a nonempty finite subset of F.

The topology $\sigma(X, F)$ is called the *weak topology* (on X) *generated by* F. In order to justify this terminology, we recall from Theorem II.11.3 that $\sigma(X, F)$ is the weakest linear topology on X for which each seminorm in $\mathscr{P}$ is continuous. Among these seminorms are those of the form $p(x) = |x'(x)|$, where $x' \in F$. It follows from Theorem 1.4 that each $x' \in F$ is continuous with respect to $\sigma(X, F)$. On the other hand, suppose τ is any linear topology on X such that each x' in F is τ-continuous. Then the sets in (6-2) are clearly τ-neighborhoods of 0. It follows that $\sigma(X, F)$ is weaker than τ. We conclude that $\sigma(X, F)$ is the weakest linear topology on X for which all the elements of F are continuous.

The following theorem is of basic importance to an understanding of the nature of the topology $\sigma(X, F)$.

Theorem 6.1. *Let F be a nonempty subset of X^f. An element x' of X^f is continuous on X with the topology $\sigma(X, F)$ if and only if x' is in the linear manifold generated by F.*

Proof. We have observed that each $x' \in F$ is continuous with respect to $\sigma(X, F)$. It follows that every finite linear combination of elements in F is also continuous. Now suppose $y' \in X^f$ is $\sigma(X, F)$-continuous. By Theorem 1.4, $|y'(x)| \leq 1$ for x in some $\sigma(X, F)$-neighborhood V of 0. Then there exist $r > 0$ and a nonempty finite subset F_0 of F such that the set U defined by (6-2) is contained in V. Suppose $x \in X$ satisfies $x'(x) = 0$ for all $x' \in F_0$. Then for every $\alpha > 0$, $\alpha^{-1}x \in U \subset V$, which implies that $|y'(\alpha^{-1}x)| \leq 1$, or $|y'(x)| \leq \alpha$. But then $y'(x)$ must be 0. Hence, if $x'_1, \ldots, x'_n$ is a maximal linearly independent subset of F_0, it follows from Theorem 1.2 that y' is a linear combination of $x'_1, \ldots, x'_n$. □

It follows from this theorem that if M is the linear manifold generated by F, the topologies $\sigma(X, F)$ and $\sigma(X, M)$ are the same, since $\sigma(X, M)$ is the weakest topology on X for which all $x' \in M$ are continuous. Also, if M_1 and M_2 are linear manifolds in X^f, with M_1 a proper subset of M_2, then $\sigma(X, M_1)$ is strictly weaker than $\sigma(X, M_2)$.

An explanation of the conditions under which $\sigma(X, F)$ makes X a Hausdorff space depends on the concept of a *total* set of linear functionals.

Definition. A set $F \subset X^f$ is said to be *total* if to each $x \neq 0$ in X there corresponds some x' in F such that $x'(x) \neq 0$. Or, equivalently, F is total if $x'(x) = 0$ for each x' in F implies $x = 0$.

For example, if $X = C[a, b]$ and F is the set of functionals of the form $x'(x) = x(r)$, where $a \leq r \leq b$ and r is rational, then F is total.

Theorem 6.2. *The topology $\sigma(X, F)$ makes X a Hausdorff space if and only if F is total.*

Proof. Apply Theorem II.11.4. □

It is often useful to consider weak topologies on a linear space X that is already a topological linear space with a topology τ. For the subset F of X^f in $\sigma(X, F)$, let us take the conjugate X' of all continuous linear functionals on X. The resulting topology $\sigma(X, X')$ is called *the* weak topology on X. Since each $x' \in X'$ is τ-continuous, $\sigma(X, X')$ is weaker than τ (though not necessarily strictly weaker). For this reason we sometimes call τ the *initial* topology and $\sigma(X, X')$ the *weakened* topology. It is important to observe that the conjugate of X with respect to the weakened topology is the same as the conjugate of X with respect to the initial topology (cf. Theorem 6.1). Also, $\sigma(X, X')$ is the weakest linear topology on X with this property.

Whenever $\sigma(X, X')$ is strictly weaker than the initial topology on X, some sets in X will be closed in the initial topology but not closed in the weakened topology. However, if the space X (with its initial topology) is locally convex, this cannot happen to those sets in X that are convex, as the following theorem shows.

Theorem 6.3. *Let X be a locally convex space with a topology τ, and let X' be the conjugate of X. Then a convex set K in X is closed in the topology τ if and only if K is closed in the topology $\sigma(X, X')$.*

Proof. Since τ and $\sigma(X, X')$ determine the same continuous linear functionals on X, the collection of *closed* real half spaces in X is the same for both topologies. The theorem now follows easily from Theorem 2.10. ☐

Suppose X is a locally convex space. Then we can apply Theorem 6.3 to the set $\{0\}$ to conclude that $\sigma(X, X')$ is a Hausdorff topology if and only if the initial topology on X is Hausdorff (cf. Theorem II.9.3).

For our purposes, the most important examples of weak topologies will come from normed linear spaces. Except when a normed linear space X is finite dimensional, the weakened topology on X is not a norm topology (see problem 2). However, the topology will be Hausdorff.

A Topology for X'

If X is a topological linear space, let J be the canonical mapping of X into $(X')^f$ defined by

$$\langle x', Jx \rangle = \langle x, x' \rangle, \qquad x \in X, x' \in X'. \tag{6-3}$$

Then $J(X)$ is a set of linear functionals on X' that can be used to define a weak topology on X'. Each nonempty finite subset of $J(X)$ is of the form $J(A)$ where A is a finite set in X. In view of (6-3), the seminorms that define the topology $\sigma(X', J(X))$ may be written in the form

$$p(x') = \max_{x \in A} |\langle x, x' \rangle|, \qquad x' \in X'. \tag{6-4}$$

We shall habitually write $\sigma(X', X)$ in place of $\sigma(X', J(X))$.

It should be evident that the seminorms in (6-4) will be continuous with respect to some linear topology τ on X' if and only if the seminorms

$$p(x') = |\langle x, x' \rangle|, \qquad x' \in X', \tag{6-5}$$

are all continuous with respect to τ. Thus it follows that $\sigma(X', X)$ is the weakest linear topology on X' such that the seminorms (6-5) are continuous. Since the elements of X' are functions on X, we conclude that $\sigma(X', X)$ is precisely *the topology of pointwise convergence on X* (see Example 1, § II.11).

When X is a normed linear space, $\sigma(X', X)$ is often called the *weak** (weak-star) *topology*, or the w^*-topology. This terminology was introduced by writers who denoted the conjugate by X^* instead of X'. (Occasionally, "weak* topology" is used in connection with topological linear spaces.) Of course, the conjugate of a normed linear space, being a normed linear space in its own right, also has a weak topology $\sigma(X', X'')$. Since $\sigma(X', X) = \sigma(X', J(X))$ and $J(X) \subset X''$, we see that the weak* topology is weaker than $\sigma(X', X'')$. This latter topology is in turn weaker than the norm topology on X'. We shall examine this situation more closely in § 10.

For a general topological linear space X, we shall let X'_σ denote the dual (i.e., conjugate) X' with the topology $\sigma(X', X)$.

Theorem 6.4. *Let X be a topological linear space. Then the canonical mapping $J : X \to (X')^f$ maps X onto the dual $(X'_\sigma)'$ of X'_σ. Furthermore, if X' is total, then J is a one-to-one mapping.*

Proof. The range of J is a linear manifold in $(X')^f$. Since $\sigma(X', X) = \sigma(X', J(X))$, it follows from Theorem 6.1 that $J(X)$ is the dual of X'_σ. If X' is total, then for each $x \neq 0$ there exists $x' \in X'$ such that $0 \neq \langle x, x' \rangle = \langle x', Jx \rangle$. Thus Jx is not the zero functional on X'. Since J is linear, this proves that J is a one-to-one mapping. □

Suppose that X is a topological linear space such that X' is total. On the basis of Theorem 6.4 we may use J to identify X with $(X'_\sigma)'$, so that X "is" the dual of X'_σ. When this identification is made, we say that X and X'_σ are in *duality*. In this case the families of seminorms in (6-1) and (6-4) have a certain obvious symmetry. The theory of duality, which is often described in a setting more general than we have used, has applications to classical analysis as well as providing the framework for the modern theory of distributions. The reader may consult Edwards [1], Horváth [1], or Kelley and Namioka [1].

PROBLEMS

1. A topological space X is said to satisfy the "first axiom of countability" if there is a countable base of neighborhoods at each $x \in X$. Let X be a linear space, and let F be a total subset of X^f. Then the following are equivalent.
   a. The topology $\sigma(X, F)$ is metrizable.
   b. The topology $\sigma(X, F)$ satisfies the first axiom of countability.
   c. F has a countable or finite Hamel basis. [To prove that (b) implies (c), let $\{U_n\}$ be a base of neighborhoods of 0. For each n, let A_n be a finite set of elements of F such that a neighborhood of the form (6-2) is contained in U_n. Show that every element of F is a finite linear combination of elements of $\bigcup_n A_n$.]

2. Let X be an infinite-dimensional linear space, and suppose $F \subset X^f$.
   a. Every $\sigma(X, F)$-neighborhood of 0 contains a linear manifold of infinite dimension.
   b. The topology $\sigma(X, F)$ is not normable.
3. Let X be a normed linear space. The topology $\sigma(X', X)$ coincides with the norm topology on X' if and only if X is finite dimensional.
4. If X is an incomplete normed linear space and if $\hat{X}$ is its completion, then the two weak topologies $\sigma(X', X)$ and $\sigma(X', \hat{X})$ are different, even though X and $\hat{X}$ have the same normed conjugate X'.
5. Let X be a locally convex space. The topology of X is Hausdorff if and only if the canonical mapping $J: X \to (X'_\sigma)'$ is one-to-one.
6. Let $\{x_n\}$ be a sequence in a normed linear space X. Suppose that there is an $x \in X$ such that $\lim_{n\to\infty} x'(x_n) = x'(x)$, for each $x' \in X'$. Then there exists a sequence $\{y_k\}$ in X such that each y_k is a finite convex linear combination of the x_n and $\lim_{k\to\infty} \|y_k - x\| = 0$.
7. Let X be a linear space, and let τ and τ' be locally convex topologies on X. Then the two topologies coincide on an absolutely convex set A in X if and only if the τ-neighborhoods of 0 in A coincide with the τ'-neighborhoods of 0 in A. [*Suggestion.* It suffices to consider absolutely convex neighborhoods of points in A, and these have the form $\{y \in A : p(y - x) \le r\}$ and $\{y \in A : p'(y - x) \le r'\}$ for $x \in A$, $r, r' > 0$, and seminorms p, p' that are, respectively, τ-continuous and τ'-continuous. If $x, y \in A$, then $\frac{1}{2}(y - x) \in A$ since A is absolutely convex.]
8. Let X be a normed linear space, and let $S = \{x : \|x\| \le 1\}$. Let F_1 be a subset of X', and let F_2 be a (norm-) dense subset of F_1. Then the topologies $\sigma(X, F_1)$ and $\sigma(X, F_2)$ coincide on S. (Contrast this fact with the remarks following Theorem 6.1.)

III.7 POLAR SETS AND ANNIHILATORS

The study of relations between a topological linear space X and its conjugate is often simplified by using the notion of polar sets. A closely related idea, the annihilator of a set, is a valuable tool for working with linear operators on X and X'. In this section we examine some basic properties of polar sets and annihilators. These results will be needed later in this chapter and at some points in Chapters IV and V.

Definition. Let X be a topological linear space. If $A \subset X$, the polar A° of A is the set

$$A^\circ = \{x' \in X' : |\langle x, x' \rangle| \le 1, \text{ all } x \in A\}.$$

If $F \subset X'$, the polar F° of F is the set

$$F^\circ = \{x \in X : |\langle x, x' \rangle| \le 1, \text{ all } x' \in F\}.$$

To be more precise, we say that the above polars of sets in X and X' are defined *with respect to the bilinear form* $\langle \cdot , \cdot \rangle$ *on* $X \times X'$. (By "bilinear," we mean that for each $x' \in X'$ the mapping $x \mapsto \langle x, x' \rangle$ is linear on X, and for each $x \in X$ the mapping $x' \mapsto \langle x, x' \rangle$ is linear on X'.) In general, one can define polars of sets in X and X' with respect to any given bilinear form $B(\cdot , \cdot)$ on $X \times X'$. For most purposes it will suffice to consider the canonical form $B(x, x') = \langle x, x' \rangle = x'(x)$.

Given a set A in X or X', the *bipolar* $A^{\circ\circ}$ of A is the set $(A^\circ)^\circ$. As we see from part (e) of Theorem 7.1 below, $(A^{\circ\circ})^\circ = A^\circ$ and the process of forming polars terminates.

Theorem 7.1. *Let A and B be subsets of a topological linear space X.*

(a) *If $A \subset B$, then $B^\circ \subset A^\circ$.*

(b) $(A \cup B)^\circ = A^\circ \cap B^\circ$.

(c) $(\lambda A)^\circ = \lambda^{-1} A^\circ$, *for each nonzero scalar* λ.

(d) $A \subset A^{\circ\circ}$.

(e) $A^\circ = A^{\circ\circ\circ}$.

(f) *If $A \subset B \subset A^{\circ\circ}$, then $A^\circ = B^\circ$.*

The above assertions are also true when A and B are both subsets of X'.

Proof. (a) If $x' \in B^\circ$, then $|\langle x, x' \rangle| \leq 1$ for all $x \in B$ and certainly for all $x \in A$; thus $x' \in A^\circ$.

(b) The following statements are clearly equivalent: (i) $x' \in (A \cup B)^\circ$; (ii) $|\langle x, x' \rangle| \leq 1$ for all $x \in A \cup B$; (iii) $|\langle x, x' \rangle| \leq 1$ for all $x \in A$ and for all $x \in B$; and (iv) $x' \in A^\circ \cap B^\circ$.

(c) Since $\langle \lambda x, x' \rangle = \langle x, \lambda x' \rangle$, we see that $x' \in (\lambda A)^\circ$ if and only if $|\langle x, \lambda x' \rangle| \leq 1$ for all $x \in A$, and this is true if and only if $\lambda x' \in A^\circ$; that is, $x' \in \lambda^{-1} A^\circ$.

(d) If $x \in A$, then $|\langle x, x' \rangle| \leq 1$ for all $x' \in A^\circ$; that is, $x \in (A^\circ)^\circ$.

(e) Applying (a) to the relation $A \subset A^{\circ\circ}$, we have $A^{\circ\circ\circ} \subset A^\circ$. However, $A^{\circ\circ\circ}$ is the bipolar of A°, and so $A^\circ \subset A^{\circ\circ\circ}$, by (d).

(f) If $A \subset B \subset A^{\circ\circ}$, then by (a), $A^{\circ\circ\circ} \subset B^\circ \subset A^\circ$. Since $A^\circ = A^{\circ\circ\circ}$, we have $A^\circ = B^\circ$.

Interchanging x and x' in the above proofs, we obtain proofs of the corresponding statements about polars of sets in X'. □

Remarks:

1. If the scalar field is real, then the polar of a single point $x' \in X'$ is the set $\{x \in X : -1 \leq x'(x) \leq 1\}$, that is, the set of all x in X lying between the two hyperplanes $\{x : x'(x) = -1\}$ and $\{x : x'(x) = 1\}$.
2. A basis of neighborhoods of 0 for the $\sigma(X, X')$ topology is given by all sets of the form $V = \{x : |x'(x)| \leq r, x' \in F\}$, where $r > 0$ and F is a finite set in X'. Observe that $V = rF^\circ = (r^{-1}F)^\circ$. Since $r^{-1}F$ is again a finite

set, it follows that a basis of $\sigma(X, X')$-neighborhoods of 0 is given by all the polars of finite subsets of X'.

3. If X is a normed linear space and $S=\{x:\|x\|\leq 1\}$, then $S^\circ=\{x'\in X':\|x'\|\leq 1\}$. This follows from Theorem 3.2.

Theorem 7.2. *Let X be a topological linear space. If $A\subset X$, then A° is absolutely convex and closed in the topology $\sigma(X', X)$. If $F\subset X'$, then F° is absolutely convex and closed in the topology $\sigma(X, X')$.*

Proof. For each $x\in A$, let p_x be the seminorm defined by $p_x(x')=|\langle x, x'\rangle|$, $x'\in X'$. Then p_x is $\sigma(X', X)$-continuous, and hence the set $\{x'\in X' : p_x(x')\leq 1\}$ is $\sigma(X', X)$-closed and absolutely convex (cf. Lemma II.11.1). Since

$$A^\circ=\bigcap_{x\in A}\{x'\in X' : p_x(x')\leq 1\},$$

it follows that A° is also $\sigma(X', X)$-closed and absolutely convex. The statement about a set F in X' is proved in a similar manner. □

If A is a (nonempty) set in X, then $(A^\circ)^\circ$ is absolutely convex and $\sigma(X, X')$-closed, by Theorem 7.2. Thus, among the sets B with the property that $A\subset B\subset A^{\circ\circ}$ are the balanced hull of A, the closure of A, the $\sigma(X, X')$-closure of A, and the closed convex hull of A. By Theorem 7.1 (f), the polars of these sets all coincide with A°. Similar remarks apply, of course, to a nonempty set A in X'.

The next theorem is the main result of this section.

Theorem 7.3 (The Bipolar Theorem). *Let X be a topological linear space. If A is a nonempty subset of X, then $A^{\circ\circ}$ is the $\sigma(X, X')$-closed absolutely convex hull of A. If F is a nonempty subset of X', then $F^{\circ\circ}$ is the $\sigma(X', X)$-closed absolutely convex hull of F.*

Proof. First, we consider F in X'. Let G be the $\sigma(X', X)$-closed absolutely convex hull of F. Then $G\subset F^{\circ\circ}$ because $F^{\circ\circ}$ is $\sigma(X', X)$-closed, absolutely convex and contains F. To prove the reverse inclusion, we take $a'\in X'\backslash G$. Applying Theorem 2.9(b) to the locally convex space X'_σ (as defined just prior to Theorem 6.4), we obtain an $x''\in(X'_\sigma)'$ such that $|\langle x', x''\rangle|\leq 1$ for $x'\in G$ and $\langle a', x''\rangle>1$. By Theorem 6.4, there exists $x\in X$ such that $x''=Jx$. Then $|\langle x, x'\rangle|=|\langle x', x''\rangle|\leq 1$ for $x'\in G$, from which it follows that $x\in G^\circ\subset F^\circ$. Also, $\langle x, a'\rangle=\langle a', x''\rangle>1$, and so $a'\notin F^{\circ\circ}$. Thus $F^{\circ\circ}\subset G$. This proves the second part of the theorem. The proof of the first part is similar, except that the step involving Theorem 6.4 is not needed. □

Annihilators

Annihilators were introduced in § I.13 in order to display certain properties of linear operators on a linear space X and on its algebraic conjugate X^f. Now we consider a topological linear space X and its (topological) conjugate X'. The discussion below parallels that of § I.13, with X^f replaced by X'. Henceforth the term "annihilator" will refer to the definition given here.

Definition. If $A \subset X$, the annihilator $A^\perp$ of A is the set

$$A^\perp = \{x' \in X' : \langle x, x' \rangle = 0, \text{ all } x \in A\}.$$

If $F \subset X'$, the annihilator $F^\perp$ of F is the set

$$F^\perp = \{x \in X : \langle x, x' \rangle = 0, \text{ all } x' \in F\}.$$

As in the case of polars, the annihilators of sets in X and X' are defined, strictly speaking, with respect to the canonical bilinear form $\langle \cdot, \cdot \rangle$ on $X \times X'$.

The following properties of annihilators are easily verified. If A and B are both subsets of X or X', then

1. $A^\perp = A^\circ$ whenever A is a linear manifold,
2. $A^\perp$ is a linear manifold whether or not A is,
3. $A \subset B$ implies $B^\perp \subset A^\perp$,
4. $A \subset A^{\perp\perp} = (A^\perp)^\perp$,
5. $A^\perp = A^{\perp\perp\perp}$,
6. $A \subset B \subset A^{\perp\perp}$ implies $A^\perp = B^\perp$.

One difference between annihilators of sets in X and annihilators of sets in X' should be noted. While it is obvious that $X^\perp = (0)$, it is not always true that the annihilator of X' is the zero vector in X. In fact, $(X')^\perp = (0)$ if and only if X' is a total set of functionals. This will be the case, for instance, when X is a locally convex Hausdorff space. (See Theorem 6.2 and the remark following Theorem 6.3.)

Now suppose that A is a subset of X. Then $A^{\perp\perp}$ is a linear manifold, and so $(A^{\perp\perp})^\perp = (A^{\perp\perp})^\circ$. Since $A^\perp = A^{\perp\perp\perp}$, we conclude from Theorem 7.2 that $A^\perp$ is $\sigma(X', X)$-closed. Similarly, if $F \subset X'$, then $F^\perp$ is a $\sigma(X, X')$-closed linear manifold. From this it follows that if $A \subset X$ and if B is the $\sigma(X, X')$-closed linear manifold generated by A, then $A \subset B \subset A^{\perp\perp}$, and so $A^\perp = B^\perp$. Similarly, if $F \subset X'$ and if G is the $\sigma(X', X)$-closed linear manifold generated by F, then $F^\perp = G^\perp$.

Theorem 7.4. *Let X be a topological linear space. If A is a nonempty subset of X, then $A^{\perp\perp}$ is the $\sigma(X, X')$-closed linear manifold generated by A. If*

F is a nonempty subset of X', then $F^{\perp\perp}$ is the $\sigma(X', X)$-closed linear manifold generated by F.

Proof. Let B be the $\sigma(X, X')$-closed linear manifold generated by A. Then, by properties (1), (2) and the remarks above, $A^{\perp\perp} = (A^{\perp})^{\circ} = (B^{\perp})^{\circ} = (B^{\circ})^{\circ}$. But $B^{\circ\circ} = B$ by the bipolar theorem, since B is absolutely convex and $\sigma(X, X')$-closed. The proof for $F^{\perp\perp}$ when $F \subset X'$ is similar. □

If X is a locally convex space and if A is a closed linear manifold in X, then A is $\sigma(X, X')$-closed (Theorem 6.3), and hence $A = A^{\perp\perp}$.

Now suppose that X is a normed linear space, and let F be a norm-closed linear manifold in X'. Then it may happen that F is a proper subspace of $F^{\perp\perp}$. For an example we take $X = (c)$. By problem 1 of § 5, every $x' \in X'$ is representable in the form

$$x'(x) = \sum_{k=1}^{\infty} \beta_k \xi_{k-1}, \qquad \text{with} \qquad \sum_{k=1}^{\infty} |\beta_k| < \infty, \tag{7-1}$$

where $x = \{\xi_k : k = 1, 2, \ldots\}$ and $\xi_0 = \lim_{k\to\infty} \xi_k$. Now let F be the set of those x' for which $\beta_1 = 0$. This is a proper closed subspace of X'. But since F contains all of the "coordinate" functionals, $x \mapsto \xi_n$, it follows that $F^{\perp} = (0)$. Hence $F^{\perp\perp} = X' \neq F$. It can be shown that if X is a normed linear space, then every norm-closed linear manifold in X' is also $\sigma(X', X)$-closed if and only if X is reflexive (problem 1).

PROBLEMS

1. a. Let X be a Hausdorff locally convex space. A linear manifold F in X' is total if and only if F is $\sigma(X', X)$-dense.
   b. If X is a reflexive Banach space, a linear manifold F in X' is total if and only if F is norm-dense.
   c. A normed linear space X is a reflexive Banach space if and only if every norm-closed linear manifold in X' is $\sigma(X', X)$-closed. [*Hint.* Given $x'' \in X''$, consider the hyperplane $\{x' : x''(x') = 0\}$ in X'.]
2. Suppose M is a linear manifold in a normed linear space X, and suppose $x'_0 \in X' \backslash M^{\perp}$.
   a. The distance from x'_0 to $M^{\perp}$ is given by

   $$\operatorname{dist}(x'_0, M^{\perp}) = \sup\{|x'(x)| : x \in M, \|x\| \leq 1\}.$$

   b. There exists $x' \in M^{\perp}$ such that $\|x'_0 - x'\| = \operatorname{dist}(x'_0, M^{\perp})$. Hence $M^{\perp}$ is a proximinal subspace of X' in the sense of problem 5, § 3.
3. If X is a reflexive Banach space, then every closed subspace of X is proximinal and every closed subspace of X' is proximinal. [*Note.* It is also true that if every closed subspace of a Banach space X is proximinal, then X is reflexive. The proof is difficult.]

4. Suppose X is a Hilbert space, and in this problem denote the orthogonal complement of a set A in X by $A^{\perp}$. Let $T: X' \to X$ be the isometric mapping described by Theorem 5.1. If $A \subset X$, then

$$A^{\perp} = T(A^{\perp}), \qquad A^{\perp} = T^{-1}(A^{\perp})$$

and $A^{\perp\perp} = A^{\perp\perp}$.

III.8 EQUICONTINUITY AND $\mathfrak{S}$-TOPOLOGIES

The concept of equicontinuity is useful in many areas of analysis. It was first discussed by Giulio Ascoli in 1883 for families of continuous real-valued functions defined on a compact set of real numbers. Several important generalizations of Ascoli's definition and of the well-known Arzelà–Ascoli theorem are discussed in Taylor [5, pages 166–169]. The appropriate definition for our purposes concerns families of linear functionals.

Definition. Let X be a topological linear space. A family $\mathscr{F}$ of linear functionals on X is said to be *equicontinuous* if to each $\varepsilon > 0$ there corresponds a neighborhood U of 0 such that $|\langle x, x' \rangle| < \varepsilon$ for all $x \in U$ and all $x' \in \mathscr{F}$.

Each x' in an equicontinuous family is necessarily continuous, because it is bounded on a neighborhood of 0 (cf. Theorem 1.4).

Theorem 8.1. *Let X be a topological linear space. A subset F of the dual X' is equicontinuous if and only if F is contained in the polar U° of a neighborhood of 0 in X. In particular, F is equicontinuous if and only if F° is a neighborhood of 0 in X.*

Proof. If F is equicontinuous, then there exists a neighborhood U of 0 in X such that $|\langle x, x' \rangle| \leq 1$ for all $x \in U$ and $x' \in F$; that is, $F \subset U^{\circ}$. We note also that, in this case, $F^{\circ} \supset U^{\circ\circ} \supset U$, so that F° is a neighborhood of 0. Now suppose that $F \subset U^{\circ}$, where U is a neighborhood of 0 in X. Given $\varepsilon > 0$, take $y \in \varepsilon U$. Then $\varepsilon^{-1} y \in U$, which implies that $|\langle \varepsilon^{-1} y, x' \rangle| \leq 1$ for $x' \in F$; that is, $|\langle y, x' \rangle| \leq \varepsilon$ for $x' \in F$. Thus F is equicontinuous. Finally, if F° is a neighborhood of 0, then F must be equicontinuous, since $F \subset (F^{\circ})^{\circ}$. $\square$

Example. When X is a normed linear space, a subset F of X' is equicontinuous if and only if it is bounded in the sense of the norm. To prove this we observe that a base of neighborhoods of 0 for the norm topology of X is given by sets of the form λS, where $\lambda > 0$ and $S = \{x : \|x\| \leq 1\}$. Furthermore, polars of these sets are just balls in X', since $(\lambda S)^{\circ} = \lambda^{-1} S^{\circ} = \lambda^{-1}\{x' : \|x'\| \leq 1\} = \{x' : \|x'\| \leq \lambda^{-1}\}$. It now follows from Theorem 8.1 that F is equicontinuous if and only if $F \subset (\lambda S)^\circ$ for some $\lambda > 0$, that is, if and only if F is norm-bounded. A similar result holds in topological linear spaces. See problem 1.

The next result will be used in the proof of Theorem 8.3.

Theorem 8.2. *Let X be a topological linear space, and let F be an equicontinuous subset of X'. Then the Minkowski functional p of F° is a continuous seminorm, and*

$$F^{\circ\circ} = \{x' \in X^f : |\langle x, x'\rangle| \leqslant p(x), x \in X\}.$$

Proof. Since F° is an absolutely convex neighborhood of 0 (Theorems 7.2 and 8.1), the Minkowski functional p of F° is a continuous seminorm (Theorems II.12.1 and II.12.3). Suppose $x' \in F^{\circ\circ}$. Given $x \in X$, let α be such that $p(x) < \alpha$. Then $\alpha^{-1}x \in F^\circ$, by definition of p, and so $|\langle \alpha^{-1}x, x'\rangle| \leqslant 1$, and $|\langle x, x'\rangle| \leqslant \alpha$. Since α was arbitrarily close to $p(x)$, we have $|\langle x, x'\rangle| \leqslant p(x)$. Conversely, suppose $x' \in X^f$ and $|\langle x, x'\rangle| \leqslant p(x)$ for all $x \in X$. Then x' is continuous, by Theorem 1.4. Furthermore, $|\langle x, x'\rangle| \leqslant p(x) \leqslant 1$ for $x \in F^\circ$ (cf. Theorem II.12.2). Hence $x' \in F^{\circ\circ}$. □

Before coming to the next theorem we need to recall that the set of all scalar-valued functions on X can be written as the Cartesian product $\Pi_{x\in X} \boldsymbol{K}_x$, where $\boldsymbol{K}_x$ is the field $\boldsymbol{K}$ of scalars ($\boldsymbol{R}$ or $\boldsymbol{C}$) for each x. For each $y \in X$, let $\pi_y : \Pi \boldsymbol{K}_x \to \boldsymbol{K}$ be the evaluation map defined by $\pi_y(f) = f(y)$. The product topology on $\Pi \boldsymbol{K}_x$ may be characterized as the weakest topology for which each π_y is continuous (cf. Taylor [5, pages 136–138]). Thus the product topology on $\Pi \boldsymbol{K}_x$ is the topology of pointwise convergence on X (Example 1, § II.11). Since X' is obviously a linear manifold in $\Pi \boldsymbol{K}_x$, it is clear that the $\sigma(X', X)$ topology is the relative topology that X' inherits as a subspace of $\Pi \boldsymbol{K}_x$.

The space X^f of all linear functionals on X is also a linear manifold in $\Pi \boldsymbol{K}_x$. Furthermore, X^f is the intersection of all sets of the form

$$\{f \in \Pi \boldsymbol{K}_x : (\pi_{\alpha x+\beta y} - \alpha\pi_x - \beta\pi_y)(f) = 0\},$$

where $x, y \in X$ and $\alpha, \beta \in \boldsymbol{K}$. These sets are all closed because the maps π_Y used to define the sets are continuous. Thus X^f is a closed set in $\Pi \boldsymbol{K}_x$.

The fundamental result about equicontinuous sets is given in the next theorem.

Theorem 8.3 (Alaoglu–Bourbaki). *Let X be a topological linear space. Then every equicontinuous subset F of X' is relatively compact in the topology $\sigma(X', X)$; that is, $\bar{F}$ is $\sigma(X', X)$-compact.*

Proof. Let p be the Minkowski functional of F° and for each $x \in X$, let $A_x = \{\alpha \in \boldsymbol{K} : |\alpha| \leqslant p(x)\}$. Since each A_x is compact in $\boldsymbol{K}_x = \boldsymbol{K}$, the set $A =$

$\Pi_{x\in X} A_x$ must be compact in the product topology, by the Tychonoff theorem (cf. Taylor [5, page 137]). Now A is the set of all functions g defined on X such that $g(x) \in A_x$ for each $x \in X$, while $F^{\circ\circ}$ is a set of linear functions with the same property. In fact, Theorem 8.2 says that

$$F^{\circ\circ} = X^f \cap A.$$

Since X^f is closed and A is compact, $F^{\circ\circ}$ is compact in $\Pi\, K_x$. However, $F^{\circ\circ}$ is a subset of X' and, as we remarked above, the product topology on subsets of X' is just the $\sigma(X', X)$-topology. Hence $F^{\circ\circ}$ is $\sigma(X', X)$-compact. Since $F \subset F^{\circ\circ}$, we conclude that F must be relatively $\sigma(X', X)$-compact. □

$\mathfrak{S}$-topologies

Theorem 8.1 describes a connection between the topology of X and the equicontinuous sets in X'. This connection will become even clearer after we discuss a general method for defining locally convex topologies on X and X'. We present this method mainly as an illustration of the use of polars of sets. However, it is of fundamental importance in the general theory of topological linear spaces. A more detailed exposition may be found, for example, in Horváth [1, pages 195–202].

Let us recall from Chapter II that a balanced set A in a topological linear space X is absorbing if to each $x \in X$ there corresponds an $\alpha > 0$ such that $\alpha x \in A$. A set B is bounded if to each balanced neighborhood U of 0 there corresponds an $r > 0$ such that $B \subset rU$. Useful characterizations of $\sigma(X, X')$-bounded sets in X and $\sigma(X', X)$-bounded sets in X' are given in problem 3.

Theorem 8.4. *Let X be a topological linear space. A set A in X is bounded in the topology $\sigma(X, X')$ if and only if A° is absorbing. A set F in X' is bounded in the topology $\sigma(X', X)$ if and only if F° is absorbing.*

Proof. We shall prove only the statement about a set A in X. Given $x' \in X'$, note that $|\alpha x'(x)| \leq 1$ if and only if $|x'(x)| \leq \alpha^{-1}$, for $x \in A$ and $\alpha > 0$. It follows that $\alpha x' \in A^\circ$ for some $\alpha > 0$ if and only if the set of scalars $\{x'(x) : x \in A\}$ is bounded. Since A° is balanced, we see that A° is absorbing in X' if and only if $\{x'(x) : x \in A\}$ is bounded for each $x' \in X'$. The desired conclusion now follows from problem 3. □

Now, given a collection $\mathfrak{S}$ of $\sigma(X, X')$-bounded subsets of X, we can define a topology on X' as follows: Note that if $A \in \mathfrak{S}$, then A° is absolutely convex and absorbing, by Theorems 7.2 and 8.4. Hence the collection of all finite intersections of sets of the form λA°, $\lambda > 0$ and $A \in \mathfrak{S}$, is a base at 0 for a locally convex topology on X', called the $\mathfrak{S}$-topology (cf. Theorem II.9.2). For $A \in \mathfrak{S}$, the Minkowski functional of A° is a seminorm, which we shall denote by p_A. The $\mathfrak{S}$-topology is sometimes called *the topology of uniform convergence on sets belonging to* $\mathfrak{S}$, because the topology is generated by the seminorms $\{p_A : A \in \mathfrak{S}\}$ and because for $x' \in X'$,

$$\begin{aligned} p_A(x') &= \inf\{\rho : x' \in \rho A^\circ\} \\ &= \inf\{\rho : |\langle x, x'\rangle| \leq \rho, \text{ all } x \in A\} \\ &= \sup_{x \in A} |\langle x, x'\rangle|. \end{aligned}$$

As one might expect, a given $\mathfrak{S}$-topology on X' may arise from several different collections of $\sigma(X, X')$-bounded subsets of X. For example, since $A^\circ = (A^{\circ\circ})^\circ$, we may replace the sets in $\mathfrak{S}$ by their $\sigma(X, X')$-closed absolutely convex hulls and leave the $\mathfrak{S}$-topology on X' unchanged.

The topology $\sigma(X', X)$ is an example of an $\mathfrak{S}$-topology. We may take $\mathfrak{S}$ to be the collection of all subsets of X containing exactly one element or the collection of all subsets of X containing only a finite number of elements. We leave to the reader the verification that the $\sigma(X', X)$ topology is also an $\mathfrak{S}$-topology when $\mathfrak{S}$ is the collection of (1) all subsets of the form

$$\left\{\sum_{i=1}^{n} \lambda_i x_i : \sum |\lambda_i| \leq \alpha, \quad \text{some} \quad \alpha > 0, \quad x_i \in X\right\};$$

or (2) all absolutely convex, $\sigma(X, X')$-closed, $\sigma(X, X')$-bounded, finite-dimensional subsets of X; or even (3) all $\sigma(X, X')$-bounded finite-dimensional subsets of X.

If $\mathfrak{S}$ is the collection of *all* $\sigma(X, X')$-bounded subsets of X, the $\mathfrak{S}$-topology on X' is called the *strong topology* and is sometimes denoted by $\beta(X', X)$. If X is a normed linear space, it can be shown that $\beta(X', X)$ coincides with the norm topology on X', for in this case the class of $\sigma(X, X')$-bounded sets coincides with the class of norm-bounded sets (cf. Theorem 9.3).

$\mathfrak{S}$-topologies on X itself may be constructed by taking $\mathfrak{S}$ to be a collection of $\sigma(X', X)$-bounded subsets of X'. A base at 0 in X is then constructed from finite intersections of sets λF°, $\lambda > 0$ and $F \in \mathfrak{S}$. Using the bipolar theorem, we can prove that *every* locally convex topology on a linear space X is an $\mathfrak{S}$-topology for a suitable collection $\mathfrak{S}$ of $\sigma(X', X)$-bounded sets.

Theorem 8.5. *Let X be a linear space endowed with a locally convex topology τ, and let $\mathfrak{S}$ be the collection of all equicontinuous sets in X'. Then the elements of $\mathfrak{S}$ are $\sigma(X', X)$-bounded and the $\mathfrak{S}$-topology on X coincides with τ.*

Proof. If $F \in \mathfrak{S}$, then F° is absorbing because it is a neighborhood of 0 (Theorem 8.1). It follows from Theorem 8.4 that F is $\sigma(X'; X)$-bounded. If U

is any τ-neighborhood of 0, then there exists a closed absolutely convex τ-neighborhood V of 0 such that $V \subset U$ (problem 3 of § II.11). Since V is convex, it is $\sigma(X, X')$-closed, by Theorem 6.3. Hence $V = V^{\circ\circ}$ by the bipolar theorem. Since $V^\circ \in \mathfrak{S}$ (by Theorem 8.1), and $(V^\circ)^\circ$ is a τ-neighborhood of 0 contained in U, we see that the polars of the sets in $\mathfrak{S}$ form a base at 0 for the topology τ. Thus τ is the $\mathfrak{S}$-topology. □

PROBLEMS

1. Let X be a topological linear space. A subset F of X' is equicontinuous if and only if there exists a continuous seminorm p such that $|\langle x, x'\rangle| \leq p(x)$ for all $x \in X$ and $x' \in F$.
2. Construct a family of continuous real-valued functions defined on $[0, 1]$ that is not equicontinuous.
3. Let X be a topological linear space.
   a. A subset A of X is $\sigma(X, X')$-bounded if and only if for each $x' \in X'$ the family of scalars $\{x'(x) : x \in A\}$ is bounded.
   b. A subset F of X' is $\sigma(X', X)$-bounded if and only if for each $x \in X$ the family of scalars $\{x'(x) : x' \in F\}$ is bounded.
4. Let X be a locally convex space. A subset F of X' is $\sigma(X', X)$-bounded if and only if F is contained in the polar B° of a barrel B in X. (A barrel is an absorbing, absolutely convex and closed set.)
5. Why must the sets in $\mathfrak{S}$ be $\sigma(X, X')$-bounded in order to define an $\mathfrak{S}$-topology on X'?
6. Let X be a Hausdorff locally convex space, and let τ be a locally convex topology on the dual X' such that X "is" the dual of X' with respect to τ; that is, the set of τ-continuous linear functionals on X' is exactly $J(X)$, where J is the canonical map of X into $(X')^f$. Show that τ is an $\mathfrak{S}$-topology, where $\mathfrak{S}$ is some collection of absolutely convex $\sigma(X, X')$-compact subsets of X.

III.9 THE PRINCIPLE OF UNIFORM BOUNDEDNESS

One of the three or four most important results about linear mappings is the principle of uniform boundedness or, as it is often called, the Banach–Steinhaus theorem. The general formulation of this principle for families of linear operators is given later in Theorem IV.1.2. Here we are concerned with the special case of families of linear functionals. The main result is Theorem 9.1. The other results follow easily from this. Illustrations of the use of the principle of uniform boundedness are given in the problems and at several points later in the text.

Theorem 9.1. *Let X be a Banach space, and let F be a subset of X'. Suppose for each $x \in X$ the family of scalars $\{x'(x) : x' \in F\}$ is bounded. Then there is a uniform bound $M > 0$ such that $\|x'\| \leq M$ for all $x' \in F$.*

Proof. For each positive integer n let $S_n = \{x : |x'(x)| \leqslant n \text{ when } x' \in F\}$. The continuity of each x' ensures that S_n is closed. By hypothesis, each x belongs to some S_n, so $X = \bigcup_{n=1}^{\infty} S_n$. Thus, by Baire's category theorem, at least one of the S_n, say S_N, contains a nonempty open set and hence contains a closed ball $B_\rho(z)$ of positive radius ρ centered at some z in S_N. That is, $|x'(x)| \leqslant N$ for all $x \in B_\rho(z)$ and all $x' \in F$. Now if y is an arbitrary element of X with $\|y\| \leqslant 1$, then we have $\rho y + z \in B_\rho(z)$, and thus $|x'(\rho y + z)| \leqslant N$ for every $x' \in F$. It follows that

$$|x'(\rho y)| \leqslant |x'(\rho y + z)| + |x'(z)| \leqslant 2N.$$

Letting $M = 2N/\rho$, we have

$$\|x'\| = \sup_{\|y\| \leqslant 1} |x'(y)| \leqslant \frac{2N}{\rho} = M,$$

for all $x' \in F$. □

Theorem 9.2. *Let X be a normed linear space, and let A be a subset of X. Suppose that for each $x' \in X'$ the family of scalars $\{x'(x) : x \in A\}$ is bounded. Then the set A is bounded in X; that is, there is a constant $M > 0$ such that $\|x\| \leqslant M$ for all $x \in A$.*

Proof. Let J be the canonical mapping of X into X'' (see § 4). Applying the previous theorem to the Banach space X' and the set $\{Jx : x \in A\}$ of continuous linear functionals on X', we obtain a uniform bound $M > 0$ such that $\|Jx\| \leqslant M$ for all $x \in A$. Since $\|Jx\| = \|x\|$, the result follows. □

When X is a real normed linear space, Theorem 9.2 can be given a geometric interpretation. If for a fixed $x' \in X'$ the set $\{x'(x) : x \in A\}$ is bounded, then there exists $\alpha > 0$ (depending on x') such that $A \subset \{x : -\alpha < x'(x) < \alpha\}$. That is, the set A lies between the parallel closed hyperplanes $\{x : x'(x) = -\alpha\}$ and $\{x : x'(x) = \alpha\}$. Theorem 9.2 states that if for every closed hyperplane L through 0 the set A lies in some "slab" between two closed hyperplanes parallel to L, then A is norm-bounded.

The condition in Theorem 9.2 that the set of scalars $\{x'(x) : x \in A\}$ is bounded for each $x' \in X'$ is equivalent to the condition that A is bounded in the topology $\sigma(X, X')$. Similarly, a set F in X' is bounded in the topology $\sigma(X', X)$ if and only if $\{x'(x) : x' \in F\}$ is bounded for each $x \in X$. These facts follow easily from the definitions of the topologies $\sigma(X, X')$ and $\sigma(X', X)$ (problem 3 in § 8). Since the converses of Theorems 9.1 and 9.2 are obviously true, we can reformulate them as follows:

Theorem 9.3. *Let X be a normed linear space. Then a subset A of X is $\sigma(X, X')$-bounded if and only if it is norm-bounded. If X is complete, then a* *subset F of X' is $\sigma(X', X)$-bounded if and only if F is bounded in the norm of X'.*

As a direct application of the principle of uniform boundedness, we have:

Theorem 9.4. (*a*) *Let $\{x_n\}$ be a sequence in a normed linear space X such that $\lim_{n\to\infty} x'(x_n)$ exists for each $x' \in X'$. Then the sequence of norms $\|x_n\|$ is bounded.*

(*b*) *Let X be a Banach space, and let $\{x_n'\}$ be a sequence in X' such that $\lim_{n\to\infty} x_n'(x)$ exists for each $x \in X$. Then the sequence of norms $\|x_n'\|$ is bounded. Furthermore, if the function x' is defined on X by*

$$x'(x) = \lim_{n\to\infty} x_n'(x),$$

then $x' \in X'$.

Proof. (*a*) This follows from Theorem 9.2. (*b*) By Theorem 9.1, there exists an M such that $\|x_n'\| \leq M$ for $n = 1, 2, \ldots$. It is readily verified that the pointwise limit x' of the sequence $\{x_n'\}$ is a linear functional. Now for each $x \in X$ and all n, $|x_n'(x)| \leq M\|x\|$. Letting $n \to \infty$, we obtain $|x'(x)| \leq M\|x\|$, which shows that x' is continuous. □

References to other formulations of the principle of uniform boundedness are given in Dunford and Schwartz [1, pages 80–82]. Generalizations of Theorem 9.1 to certain topological linear spaces are discussed in Edwards [1, Chapter 7] and Horváth [1, pages 211–214].

PROBLEMS

1. It is essential that X be complete in Theorem 9.1. To show this, let X be the subset of those $x \in \ell^2$ for which $x = \{\xi_k\}$ and $\xi_k = 0$ if k exceeds some integer depending on x. Let $x_n'(x) = n\xi_n$, and let F be the countable set of the elements $x_1', x_2', \ldots$.
2. Construct a sequence $\{f_n\}$ in $C[0, 1]$ such that the family of real numbers $\{f_n(t)\}$ is bounded for each $t \in [0, 1]$ and yet the sequence of norms $\|f_n\|$ is unbounded. Discuss the relation of this example to Theorem 9.2.
3. Suppose $1 < p < \infty$. Let $\{\alpha_k\}$ be a sequence of scalars with the property that $\Sigma_1^\infty \alpha_k\xi_k$ is convergent whenever $\Sigma_1^\infty |\xi_k|^p < \infty$. Then $\Sigma_1^\infty |\alpha_k|^{p'} < \infty$, where $p' = p/(p-1)$. [*Hint.* Consider the family of sequences of the form $y_n = (\alpha_1, \ldots, \alpha_n, 0, 0, \ldots)$.]
4. If $\Sigma_1^\infty \alpha_k\xi_k$ is convergent whenever $\Sigma_1^\infty |\xi_k| < \infty$, then $\sup_k |\alpha_k| < \infty$.
5. If $\Sigma_1^\infty \alpha_k\xi_k$ is convergent whenever $\lim_{k\to\infty} \xi_k = 0$, then $\Sigma_1^\infty |\alpha_k| < \infty$.

6. Suppose $1 \leq p < \infty$. Let g be a measurable function on $\boldsymbol{R}$. Suppose that gf is integrable on $\boldsymbol{R}$ for every $f \in \mathcal{L}^p(-\infty, \infty)$. Then $g \in \mathcal{L}^{p'}(-\infty, \infty)$. (If $p = 1$, let $p' = \infty$.)

7. a. A Banach space X is said to have a *countable basis* $\{u_n\}$, $n = 1, 2, \ldots$, if each $x \in X$ can be represented in one and only one way as a series $x = \Sigma_1^\infty \omega_k u_k$. Show that no u_n is 0 and that $\{u_n/\|u_n\|\}$ is also a countable basis.

   b. Let W be the class of all sequences $w = \{\omega_n\}$ such that $\Sigma_1^\infty \omega_k u_k$ is convergent. Define

$$\|w\| = \sup_n \left\|\sum_1^n \omega_k u_k\right\|.$$

   Show that with this norm, W is a Banach space. [First show that $|\omega_n| \leq 2(\|w\|/\|u_n\|)$.] Now define $Tw = \Sigma_1^\infty \omega_k u_k$. Show that T is a linear homeomorphism of W onto all of X.

   c. Define $u'_k \in X'$ by $u'_k(x) = \omega_k$. Show that $\|u'_k\| \leq 2\|T^{-1}\|/\|u_k\|$.

   d. Let Y be the space of all scalar sequences $y = \{\eta_k\}$ such that $\Sigma_1^\infty \omega_k \eta_k$ is convergent whenever $w = \{\omega_k\} \in W$. Define

$$\|y\| = \sup_{\|x\| \leq 1} \left|\sum_1^\infty \omega_k \eta_k\right|, \qquad \text{where } x = Tw.$$

   Show that Y is congruent to X', with $y \leftrightarrow x'$, where $x'(x) = \Sigma_1^\infty \omega_k \eta_k$.

8. Show that a Banach space with a countable basis is separable.

9. Let X be a normed linear space. A linear manifold $M \subset X'$ is *norm-determining* for X if there is an $\varepsilon > 0$ (depending only on M) such that

$$\varepsilon\|x\| \leq \sup\{|x'(x)| : x' \in M, \|x'\| = 1\}.$$

   a. If M is norm-determining, then $|||x||| = \sup\{|x'(x)| : x' \in M, \|x'\| = 1\}$ defines a new norm on X, equivalent to the original norm.

   b. Use Theorem 9.1 to obtain the following generalization of Theorem 9.2: Let A be a subset of a normed linear space X, and let M be a closed subspace of X' that is norm-determining for X. Suppose that $\sup_{x \in A} |x'(x)| < \infty$ for each $x' \in M$. Then A is a bounded set in X.

10. A locally convex space X is said to be barreled if every barrel in X is a neighborhood of 0 (cf. problem 4, § II.11).

   a. A locally convex space X is barreled if and only if every $\sigma(X', X)$-bounded subset of X' is equicontinuous.

   b. A Fréchet space is barreled.

III.10 WEAK TOPOLOGIES FOR NORMED LINEAR SPACES

Our interest in weak topologies lies primarily in their usefulness in normed linear spaces. In this section we employ the general theorems of § 6 to § 9 to obtain results in a form most readily adapted to our study of linear operators in Chapter IV.

Let X be a normed linear space, with X' its normed conjugate and X'' its second normed conjugate. Most of the important topologies for X, X', and X'' have already been examined in theorems and examples. The two most useful topologies on X are the norm topology and the *weak topology*, $\sigma(X, X')$. On X' there are three important topologies: (1) the norm topology, defined in terms of the norm on X; (2) the weak topology, $\sigma(X', X'')$, defined by considering the space X' and its dual X''; and (3) the weak* topology $\sigma(X', X)$. Since the weak* topology is the topology $\sigma(X', J(X))$, it is weaker than the weak topology $\sigma(X', X'')$, by the remarks following Theorem 6.1. The two topologies coincide if and only if $J(X) = X''$, that is, if and only if X is reflexive. Of course, the weak topology $\sigma(X', X'')$ is weaker than the norm topology on X'. Both $\sigma(X', X)$ and $\sigma(X', X'')$ are Hausdorff topologies because $J(X)$ and X'' are obviously total sets of functionals on X'. Three topologies may be defined for X'' just as for X', but of these three the norm topology is by far the most useful. Occasionally the topology $\sigma(X'', X')$ is of interest.

Closed Convex Sets

When X is an infinite-dimensional normed linear space, the norm topology is strictly stronger than the weak topology (problem 2, § 6), and yet the collection of all closed convex sets is the same for both topologies (Theorem 6.3). Similarly, a convex set in X' is norm-closed if and only if it is $\sigma(X', X'')$-closed. Such a set, however, is not necessarily weak*-closed unless the space X is reflexive (cf. problem 1, § 7). (*Any* weak*-closed set in X' is, of course, $\sigma(X', X'')$-closed.) It follows from the bipolar theorem (Theorem 7.3) that an absolutely convex set F in X' is weak*-closed if and only if $F = F^{\circ\circ}$. In particular, if F is a linear manifold, then F is weak*-closed if and only if to each $x' \in X' \backslash F$ there corresponds an $x \in F^{\perp}$ for which $x'(x) \neq 0$.

A characterization of weak*-closed convex sets in X' is given in the following important theorem.

Theorem 10.1. (Krein–Šmulian) *Let X be a Banach space. A convex set F in X' is weak*-closed if and only if $F \cap \{x' : \|x'\| \leqslant r\}$ is weak*-closed for each $r > 0$.*

We omit the rather lengthy proof of this theorem. The reader may consult Dunford and Schwartz [1, pages 427–429]. As a special case of this theorem, we note that a linear manifold M in X' is weak*-closed if and only if $M \cap \{x' : \|x'\| \leqslant 1\}$ is weak*-closed.

Bounded Sets

When X is a normed linear space, the notions of topological boundedness (§ II.12) and metric boundedness coincide for the norm topology because the positive multiples of the unit ball form a base of neighborhoods of 0. Furthermore, by the principle of uniform boundedness (Theorem 9.3), the weakly bounded sets coincide with the norm-bounded sets.

For the weak* topology on X', we have the following remarkable and powerful theorem.

Theorem 10.2 (Alaoglu). *The closed unit ball $S'=\{x':\|x'\|\leq 1\}$ in the conjugate of a normed linear space X is $\sigma(X',X)$-compact.*

Proof. We observed in the remarks following Theorem 7.1 that S' is the polar of the unit ball in X. Also, S' is $\sigma(X',X)$-closed (Theorem 7.2). It follows from Theorems 8.1 and 8.3 that S is $\sigma(X',X)$-compact.

For the convenience of the reader who may have skipped over the discussion of equicontinuity, we shall give a separate proof here for the case of a normed linear space. For each $x\in X$ let $A_x=\{\alpha:|\alpha|\leq\|x\|\}$, and let $A=\Pi_{x\in X}A_x$. The set A consists of all functions g on X such that $g(x)\in A_x$, that is, $|g(x)|\leq\|x\|$ for $x\in X$. Since S' is a set of linear functions with the same property, we have $S'\subset X^f\cap A$. In fact, each x' in $X^f\cap A$ has the property that $|x'(x)|\leq\|x\|$ for all x, so that $\|x'\|\leq 1$. Thus

$$S'=X^f\cap A.$$

By the Tychonoff theorem, A is compact in the product topology on $\Pi_{x\in X}\boldsymbol{K}_x$, where each $\boldsymbol{K}_x$ is the field of scalars. Also, X^f is a closed subset of $\Pi_{x\in X}\boldsymbol{K}_x$ (see § 8 for a proof). Thus S' is compact. Now $S'\subset X'\subset\Pi_{x\in X}\boldsymbol{K}_x$ and the restriction of the product topology to X' is just the topology $\sigma(X',X)$. Hence S' is $\sigma(X',X)$-compact. □

When X is a Banach space, the bounded sets in X' have several useful characterizations.

Theorem 10.3. *Let X be a Banach space, and let F be a subset of X'. The following statements are equivalent:*

(*a*) *F is $\sigma(X',X)$-bounded,*
(*b*) *F is $\sigma(X',X'')$-bounded,*
(*c*) *F is norm-bounded,*
(*d*) *F is equicontinuous,*
(*e*) *F is relatively compact in the topology $\sigma(X',X)$.*

Proof. The fact that (*c*) implies (*b*) and (*b*) implies (*a*) is evident from the relative strengths of the topologies. Of course, (*a*) implies (*c*) by the principle of uniform boundedness (Theorem 9.3). The equivalence of (*c*) and (*d*) was demonstrated in the example following Theorem 8.1. The fact that (*c*) implies (*e*) follows easily from Alaoglu's theorem. Finally, suppose (*e*) is true. For each x in X, the map $x' \mapsto x'(x)$ is $\sigma(X', X)$-continuous, and so the image set $\{x'(x) : x' \in F\}$ is relatively compact and therefore bounded. It follows that (*a*) is true (cf. problem 3 of § 8). □

Weak Convergence

In the older literature dealing with Banach spaces, before the development of the weak topologies, a good deal of attention was given to the concept of weak convergence of sequences, even though the weak topologies themselves were not investigated. The concept of weak convergence remains a valuable tool today in current research in functional analysis and in its applications to the calculus of variations and the general theory of differential equations.

If $\{x_n\}$ is a sequence in X, we shall say that $\{x_n\}$ converges weakly to $x \in X$ if $\{x_n\}$ converges to x in the sense of the weak topology $\sigma(X, X')$. This happens if and only if $x'(x_n) \to x'(x)$ for each $x' \in X'$. If $\{x'_n\}$ is a sequence in X', we shall say that $\{x'_n\}$ is weak* convergent to $x' \in X'$ if $\{x'_n\}$ converges to x' in the topology $\sigma(X', X)$, that is, if $x'_n(x) \to x'(x)$ for each $x \in X$.

In the special case when X is an inner-product space, a sequence $\{x_n\}$ in X converges weakly to x if and only if $(x_n, y) \to (x, y)$ for each $y \in X$, by the Fréchet–Riesz theorem (Theorem 5.1).

It can be shown that the weak topology on an infinite-dimensional normed linear space never has a countable base of neighborhoods of 0 (cf. problem 1). It follows that not all topological questions can be discussed using sequences alone. For example, a point can be in the weak closure of a set and yet not be the weak limit of a sequence of elements of the set (cf. problem 2). On the other hand, we shall see in Theorem 10.10 that sequences are sufficient to characterize weak compactness.

The following theorem is very useful for checking the weak convergence of sequences. Applications are given in the problems at the end of the section.

Theorem 10.4. *Let X be a normed linear space. A sequence $\{x_n\}$ in X converges weakly to $x \in X$ if and only if $\sup_n \|x_n\| < \infty$ and $x'(x_n) \to x'(x)$ for each $x' \in F$, where F generates a linear manifold that is (norm-) dense in X'.*

Proof. The necessity of the conditions follows from Theorem 9.4. Therefore suppose that $\{x_n\}$ satisfies the conditions in the theorem, and let $M = \sup_n \|x_n\|$. Given $x' \in X'$ and $\varepsilon > 0$, there exists $y' \in X'$ such that y' is in the linear manifold generated by F and $\|x'-y'\|<\varepsilon/4M$. It follows for each n that

$$\begin{aligned}|x'(x_n)-x'(x)| &\leqslant |x'(x_n)-y'(x_n)|+|y'(x_n)-y'(x)|+|y'(x)-x'(x)|\\ &\leqslant \|x'-y'\|\|x_n\|+|y'(x_n)-y'(x)|+\|y'-x'\|\|x\|\\ &<\varepsilon/2+|y'(x_n)-y'(x)|.\end{aligned}$$

Since y' is a finite linear combination of elements of F, a routine calculation shows that $y'(x_n)\to y'(x)$. Hence there exists N such that $|y'(x_n)-y'(x)|<\varepsilon/2$ for $n\geqslant N$. This proves that $|x'(x_n)-x'(x)|<\varepsilon$ for $n\geqslant N$. Since x' was an arbitrary element of X', we have proved that $\{x_n\}$ converges weakly to x. □

Theorem 10.5. *If X is a separable normed linear space, every bounded sequence in X' contains a weak* convergent subsequence.*

Proof. Let $\{x'_n\}$ be a bounded sequence in X', and let $\{x_k\}$ be a sequence that is dense in X. Since $\{x'_n(x_1)\}$ is a bounded sequence of scalars, it contains a convergent subsequence, which we denote by $\{x'_{n1}(x_1)\}$. Likewise, $\{x'_{n1}(x_2)\}$ contains a convergent subsequence, which we denote by $\{x'_{n2}(x_2)\}$. Continuing by induction, we obtain a "diagonal sequence" $\{x'_{nn}\}$ such that $\lim_{n\to\infty} x'_{nn}(x_k)$ exists for each k. It is then easy to see that $\{x'_{nn}(x)\}$ is a Cauchy sequence for each x, thus defining $x'\in X'$ such that $\{x'_{nn}\}$ is weak* convergent to x'. □

The similarity between Theorem 10.5 and Alaoglu's theorem (Theorem 10.2) becomes more apparent when we use the following terminology.

Definition. If a set A in a topological space X has the property that every sequence in A has a subsequence that converges to a point in X, then A is said to be *conditionally sequentially compact.* If every sequence in A has a subsequence that converges to a point in A, then A is called *sequentially compact.*

Theorem 10.5 thus says that when X is a separable normed linear space, the bounded sets in X' are weak* conditionally sequentially compact. This fact does not follow immediately from Alaoglu's theorem because, in the most general case, it is possible for a set to be weak*-compact and yet not be weak* sequentially compact. This cannot happen when X is separable or reflexive [see problems 3(c) and 3(d)].

Reflexivity and Weak Compactness

When X is a reflexive Banach space, we obtain in Theorem 10.6 the "dual" of Theorem 10.5. The proof illustrates a technique that sometimes makes it possible to eliminate from theorems the requirement that X be separable.

We observe that if X is given the topology $\sigma(X, X')$ and the canonical image $J(X)$ of X in X'' is given the topology induced on it by $\sigma(X'', X')$, then J is a linear homeomorphism of X onto $J(X)$. This is clear directly from the way in which the bases at 0 are defined for the topologies in question.

Theorem 10.6. *If X is a reflexive Banach space, each bounded sequence in X contains a weakly convergent subsequence. In particular, if $\{x_n\}$ is a sequence for which $\|x_n\| \leq 1$, it contains a subsequence converging weakly to a limit x for which $\|x\| \leq 1$.*

Proof. Let $\{x_n\}$ be a bounded sequence in X, and let X_0 be the closed linear manifold generated by $x_1, x_2, \ldots$. It is easy to see that X_0 is separable. It is also reflexive (Theorem 4.2). Hence X_0'' is separable, and therefore X_0' is separable (Theorem 3.5). The canonical mapping J of X_0 onto X_0'' carries $\{x_n\}$ into a bounded sequence in X_0''. By Theorem 10.5, this latter sequence contains a subsequence $\{Jx_{n_k}\}$ which is $\sigma(X_0'', X_0')$-convergent to a limit. We may write this limit as Jx, for some $x \in X_0$, since X_0 is reflexive. It follows that $\{x_{n_k}\}$ is $\sigma(X_0, X_0')$-convergent to x because J is a linear homeomorphism. Now each element of X', when restricted to X_0, determines an element of X_0', so it is clear that $\{x_{n_k}\}$ is $\sigma(X, X')$-convergent. In particular, if $\|x_n\| \leq 1$ for all n, then for any $x' \in X'$ we have $|x'(x_{n_k})| \leq \|x'\| \|x_{n_k}\| \leq \|x'\|$. Thus $|x'(x)| \leq \|x'\|$ and $\|x\| = \sup_{\|x'\|=1} |x'(x)| \leq 1$, by (3·2). $\square$

An important characterization of reflexivity is given in Theorem 10.8. This theorem should be compared and contrasted with Alaoglu's theorem (Theorem 10.2) and also with Theorem II.3.6. The proof of Theorem 10.8 requires the following result.

Theorem 10.7. *For a normed linear space X, consider $S \subset X$ and $S'' \subset X''$ defined as follows:*

$$S = \{x : \|x\| \leq 1\}, \qquad S'' = \{x'' : \|x''\| \leq 1\}.$$

If $J(S)$ denotes the canonical image of S in X'', then S'' is the $\sigma(X'', X')$-closure of $J(S)$.

Proof. Since X' is a topological linear space and X'' is its topological dual, we may consider polars of sets in X' and X'' with respect to the canonical bilinear form B on $X' \times X''$ given by $B(x', x'') = \langle x', x'' \rangle$. Then

$$\begin{aligned} J(S)^\circ &= \{x' \in X' : |\langle x', y'' \rangle| \leq 1, \quad \text{all} \quad y'' \in J(S)\} \\ &= \{x' \in X' : |\langle x, x' \rangle| \leq 1, \quad \text{all} \quad x \in S\}. \end{aligned}$$

That is, $J(S)^\circ$ is the closed unit ball S' in X'. It follows that

$$\begin{aligned} J(S)^{\circ\circ} &= (S')^\circ = \{x'' \in X'' : |\langle x', x''\rangle| \leq 1, \quad \text{all} \quad x' \in S'\} \\ &= S'', \end{aligned}$$

by the definition of $\|x''\|$. Applying the bipolar theorem (Theorem 7.3) in the spaces X' and X'', we have that $J(S)^{\circ\circ}$ is the closure of $J(S)$ in the $\sigma(X'', X')$ topology. □

Theorem 10.8. *The normed linear space X is reflexive if and only if the set $X = \{x : \|x\| \leq 1\}$ is compact with respect to the weak topology $\sigma(X, X')$.*

Proof. If S is compact, then $J(S)$ is $\sigma(X'', X')$-compact since J is a linear homeomorphism, as noted above. Since $\sigma(X'', X')$ is a Hausdorff topology (cf. Theorem 6.2), $J(S)$ is $\sigma(X'', X')$-closed. Thus, by Theorem 10.7, $J(S) = S''$. It now follows easily that $J(X) = X''$, so that X is reflexive.

On the other hand, if X is reflexive, then $J^{-1}(S'') = S$. Since S'' is $\sigma(X'', X')$-compact and J^{-1} is continuous, S must be $\sigma(X, X')$-compact. □

It is remarkable that the compactness condition in Theorem 10.8 can be replaced by sequential compactness. Half of this characterization of reflexivity is already contained in Theorem 10.6: if X is reflexive, then bounded sets are weakly conditionally sequentially compact. We state the converse as follows:

Theorem 10.9. *If X is a Banach space such that every bounded sequence contains a weakly convergent subsequence, then X is reflexive.*

We shall not give a complete proof of this theorem. It was first proved under the additional assumption that X is separable (Banach [1, pages 189–191]). After the subject of weak topologies had developed considerably, W. F. Eberlein proved Theorem 10.10 below. Theorem 10.9 is a consequence of this result and Theorem 10.8, since $\{x : \|x\| \leq 1\}$ is weakly closed (Theorem 6.3).

Theorem 10.10. (Eberlein) *Let X be a Banach space. A set in X is weakly compact if and only if it is weakly closed and weakly conditionally sequentially compact.*

A partial outline of a proof of Eberlein's result runs as follows: If A is weakly conditionally sequentially compact, then A is weakly bounded, and hence A is norm-bounded (Theorem 9.3). Then the canonical image $J(A)$ is norm-bounded in X''. If B is the $\sigma(X'', X')$ closure of $J(A)$, then it follows that B is $\sigma(X'', X')$-compact, by Theorem 10.3. It is difficult to prove that $B \subset J(X)$.

Once this is done, it follows that $J^{-1}(B)$ is a $\sigma(X, X')$-compact subset of X, since J is a linear homeomorphism of X onto $J(X)$. Finally, A itself is weakly compact if it is weakly closed because $A \subset J^{-1}(B)$. A proof that $B \subset J(X)$ can be given using among other things the Krein–Šmulian theorem (Theorem 10.1). A more elementary proof that $B \subset J(X)$ has been provided by R. Whitley [1]. A fairly short argument can be given to prove that a weakly compact set is weakly sequentially compact (cf. Whitley [1]). This proof uses the fact that if X is a separable Banach space, then the weak topology on a weakly compact subset is a metric topology (see problem 10). The hypothesis of separability is sidestepped as in the proof of Theorem 10.6.

Theorem 10.10 is commonly linked with a similar theorem by V. Šmulian which states that a subset A of a Banach space X is weakly conditionally sequentially compact if to every sequence in A there corresponds a point in X such that every weak neighborhood of the point contains an element of the sequence. The combination of these two results is called the Eberlein–Šmulian theorem (cf. Dunford and Schwartz [1, Chapter V, § 6] and Whitley [1]).

PROBLEMS

1. a. Let X be a Banach space, and let H be a Hamel basis for X (see § I.11). If H is an infinite set, then H is uncountable.
   b. If X is an infinite-dimensional normed linear space, then the topology $\sigma(X, X')$ on X is not metrizable.
2. In $X = \ell^2$ let $A = \{x_{mn} : 1 \leq m < n < \infty\}$, where x_{mn} is the element of ℓ^2 whose mth term is 1, nth term is m, and other terms are zero. Then no sequence of elements of A converges weakly to the origin, yet the origin is in the weak closure of A.
3. Let X be a normed linear space.
   a. If $F \subset X'$ is weak* conditionally sequentially compact, then the weak* closure of F is weak* compact.
   b. If X is a separable normed linear space, then the weak* topology on $\{x' : \|x'\| \leq 1\}$ is metrizable. [Suppose $\{y_n : n = 1, 2, \ldots\}$ is dense in X, and define seminorms p_n by $p_n(x') = |x'(y_n)|$. The family $\{p_n\}$ generates a locally convex T_1-topology on X' which has a countable base of neighborhoods of 0. This topology is metrizable, by Theorem II.13.1, and coincides with the weak* topology on $\{x' : \|x'\| \leq 1\}$.]
   c. Use (b) to show that if $F \subset X'$ is weak*-compact and if X is separable, then F is weak* sequentially compact.
   d. If $F \subset X'$ is weak*-compact and if X is reflexive, then F is weak* sequentially compact. (Prove this without using Theorem 10.10.)
4. Let X be a Banach space, and let A be a subset of X that generates a linear manifold dense in X. Then $\{x'_n\}$ is weak* convergent to some $x' \in X'$ if and only if $\sup_n \|x'_n\| < \infty$ and $\lim_n x'_n(x)$ exists for each $x \in A$.

5. If $x_n = \{\xi_k^{(n)}\} \in \ell^p$, where $1 < p < \infty$, $\{x_n\}$ is weakly convergent to $x = \{\xi_k\}$ if and only if $\sup_n \|x_n\| < \infty$ and $\lim_{n\to\infty} \xi_k^{(n)} = \xi_k$ for each k.
6. In $X = \ell^1$, $\{x_n\}$ converges weakly to x if and only if $\|x_n - x\| \to 0$ (see Banach [1, page 137]). In spite of this, the topology $\sigma(X, X')$ is not the same as the topology generated by the norm.
7. If $X = C[a, b]$, a sequence $\{x_n\}$ is weakly convergent to x if and only if $\sup_n \|x_n\| < \infty$ and $x_n(t) \to x(t)$ for each t.
8. For $X = L^p(a, b)$, where $1 < p < \infty$, $\{x_n\}$ converges weakly to x if and only if $\sup_n \|x_n\| < \infty$ and $\int_E x_n(t)\, dt \to \int_E x(t)\, dt$ for each measurable set E of finite measure in $[a, b]$. [Use Theorem 10.4 and the known representation of elements of X' in this case.] For $p = 1$ the result breaks down, the conditions being necessary but not sufficient for weak convergence. This is because the finite linear combinations of characteristic functions of sets of finite measure are not dense in $L^\infty(a, b)$. For a further discussion see Banach [1, page 136].
9. A space X is called weakly sequentially complete if the existence of $\lim_{n\to\infty} x'(x_n)$ for each $x' \in X'$ implies the existence of $x \in X$ such that $\{x_n\}$ converges weakly to x. The space $C[a, b]$ is not weakly sequentially complete. However, any reflexive space is weakly sequentially complete. The nonreflexive spaces $L^1(a, b)$ and ℓ^1 are weakly sequentially complete. See Banach [1, page 141].
10. Let X be a separable Banach space.
    a. X' contains a countable total subset.
    b. Let A be a weakly compact subset of X. Then the weak topology on A is a metric topology. [*Hint.* A continuous mapping of a compact topological space onto a Hausdorff topological space is a homeomorphism.]
11. Let X be a normed linear space. Find a compact Hausdorff topological space T and a mapping $\psi: X \to C(T)$ such that ψ is an isometric isomorphism from X onto a linear manifold in $C(T)$.
12. If X is a Banach space, then the weak*-closed convex hull of a weak*-compact subset of X' is again weak*-compact.
13. Let X be a normed linear space, and let M be a linear manifold in X'. If M, considered as a normed linear space, is reflexive, then $M^{\perp\perp} = M$. [*Hint.* Show that the unit ball of M is $\sigma(X', X'')$-compact. Then use Theorem 10.1.]
14. A nonempty closed convex set K in a reflexive Banach space X contains a point of minimum norm. That is, if $d = \inf\{\|x\| : x \in K\}$, then there exists $x \in K$ such that $\|x\| = d$. [*Hint.* Consider a sequence $\{x_n\}$ in K such that $\|x_n\| \to d$.]
15. A normed linear space X is called *uniformly convex* if to each $\varepsilon > 0$ corresponds $\delta(\varepsilon) > 0$ such that $\|(x+y)/2\| \leq 1 - \delta(\varepsilon)$ when $\|x\| \leq 1$, $\|y\| \leq 1$, and $\|x - y\| \geq \varepsilon$. Every uniformly convex Banach space is reflexive. [*Hint.* Take $x'' \in X''$ with $\|x''\| = 1$, and obtain a net $\{x_\alpha\}$ in X such that $\|x_\alpha\| \leq 1$ and $\{J(x_\alpha)\}$ converges to x'' in the $\sigma(X'', X')$ topology. Nets were defined in

problem 9, § II.9. Define a new net $\{x_\alpha + x_{\alpha'}\}$, with the indexing over all α, α'. Show that $\|x_\alpha + x_{\alpha'}\| \to 2$, and use the uniform convexity of X to deduce that $\{x_\alpha\}$ is a Cauchy net. Then $x_\alpha \to x$ for some $x \in X$ and $Jx = x''$.]

III.11 THE KREIN–MILMAN THEOREM

We conclude this chapter with an important result concerning compact convex sets due to M. Krein and D. Milman [1]. Their theorem has many interesting uses. Among the applications we shall give here will be a proof of the Stone–Weierstrass theorem.

Suppose that K is a subset of a (real or complex) linear space X. An *extremal subset* of K is a nonempty subset E of K such that if $a, b \in K$ and if a proper convex combination of a and b [say, $ta + (1-t)b$, where $0 < t < 1$] is in E, then $a, b \in E$. An *extremal point* (or *extreme point*) of K is an extremal subset that consists of just one point. It is easily seen that if E_1 is an extremal subset of K and if E_2 is an extremal subset of E_1, then E_2 is an extremal subset of K.

In the real plane $\boldsymbol{R}^2$, the vertices of a closed square are its only extremal points, but the entire boundary is an extremal subset. Every point on the boundary of a closed disc in $\boldsymbol{R}^2$ is an extremal point, and any nonempty subset of the boundary is an extremal subset. Clearly an open convex set in a locally convex space cannot have extremal points. But even some bounded closed convex sets have no extremal points (see problems 1 and 2).

Lemma 11.1. *Let K be a nonempty closed compact set in a topological linear space X. Given $x' \in X'$, let $\beta = \sup \{\operatorname{Re} x'(x) : x \in K\}$. Then $E = \{x \in K : \operatorname{Re} x'(x) = \beta\}$ is a nonempty compact extremal subset of K.*

Proof. Since $\operatorname{Re} x'$ is continuous and K is closed and compact, the set E is closed (hence compact) and nonempty. Suppose $a, b \in K$ and $x_0 = ta + (1-t)b \in E$, where $0 < t < 1$. Then

$$\beta = \operatorname{Re} x'(x_0) = t \operatorname{Re} x'(a) + (1-t) \operatorname{Re} x'(b).$$

Since $\operatorname{Re} x'(a) \leq \beta$ and $\operatorname{Re} x'(b) \leq \beta$, it follows that $\operatorname{Re} x'(a) = \operatorname{Re} x'(b) = \beta$, and so $a, b \in E$. □

Note in Lemma 11.1 that E is a proper subset of K unless $\operatorname{Re} x'$ is constant on K.

Theorem 11.2. *Let K be a nonempty compact set in a Hausdorff locally convex space X. Then K has at least one extremal point.*

Proof. Let $\mathscr{S}$ be the collection of all nonempty compact extremal subsets of K. Then $\mathscr{S}$ is nonempty, because $K \in \mathscr{S}$. Partially order $\mathscr{S}$ by reversed set inclusion; that is, define $E_1 < E_2$ if $E_1 \supset E_2$. If $\mathscr{L}$ is any completely ordered subset of $\mathscr{S}$, then the intersection E of all the members of $\mathscr{L}$ is nonempty since K is compact. It is readily checked that E is an upper bound of $\mathscr{L}$ in $\mathscr{S}$. By Zorn's lemma, $\mathscr{S}$ contains a maximal element E_0. Then E_0 is nonempty and compact and cannot properly contain a nonempty compact extremal subset of K. It follows from Lemma 11.1 that $\operatorname{Re} x'$ must be constant on E_0 for each $x' \in X'$. By the Hahn–Banach theorem, the real parts of the functionals in X' separate the points of X (see the remark following Theorem 2.8). Hence E_0 contains only one point. □

Theorem 11.3 (Krein–Milman). *Let X be a Hausdorff locally convex space, and let K be a nonempty compact subset of X. Then K is contained in the closed convex hull of the set of extremal points of K. In particular, if K is a compact convex set in X, then it is the closed convex hull of the set of its extremal points.*

Proof. Let K_1 be the closed convex hull of the set of extremal points of K, and suppose there exists $a \in K \backslash K_1$. Then by Theorem 2.9 there exists $x' \in X'$ such that $\sup \{\operatorname{Re} x'(x) : x \in K_1\} < x'(a)$. Hence if $\beta = \sup \{\operatorname{Re} x'(x) : x \in K\}$ and $E = \{x \in K : \operatorname{Re} x'(x) = \beta\}$, then $K_1 \cap E = \varnothing$. Now E is nonempty and compact and so contains an extremal point x_0, by Theorem 11.2. But E is an extremal subset of K by Lemma 11.1, so x_0 is an extremal point of K. This is impossible since $E \cap K_1 = \varnothing$. Hence $K \subset K_1$; if K is also convex, then $K_1 \subset K$, and so $K = K_1$. □

The original version of the Krein–Milman theorem was about weak*-compact convex sets in the conjugate of a normed linear space. This is still the setting for many applications of the theorem. Frequently, Alaoglu's theorem (or Theorem 10.3) is used to prepare the way for the Krein–Milman theorem. The next result follows immediately from these two theorems.

Theorem 11.4. *The closed unit ball S' in the conjugate of a normed linear space is the weak*-closed convex hull of the set of its extremal points.*

Some of the usefulness of Theorem 11.4 stems from the fact that for some important Banach spaces there exist simple characterizations of the extremal points of the closed unit ball in the conjugate space (see Theorem 11.10 and problem 10).

We note for later use that any extremal point in the closed unit ball of a nonzero normed linear space must have norm 1 for, if $0 < \|x\| < 1$, then $x = \|x\| \cdot (x/\|x\|) + (1 - \|x\|) \cdot 0$. Obviously, the zero vector is not an extremal point unless it is the only point in the unit ball.

The Krein–Milman theorem has proved to be a useful tool in the study of best approximation in normed linear spaces. The next two theorems were given by I. Singer for a special class of spaces and then were extended to arbitrary normed linear spaces by G. Choquet (cf. Singer [1, pages 58–68]).

Theorem 11.5. *Let Y be a proper closed subspace of a normed linear space X. Given $a \in X$, suppose there exists a best approximation to a by elements of Y, that is, suppose there exists $b \in Y$ such that $\|a-b\| = \operatorname{dist}(a, Y)$. Then there exists an extremal point x' of the closed unit ball S' in X such that*

$$\operatorname{Re} x'(b) \geqslant 0,$$
$$x'(a-b) = \|a-b\| = \operatorname{dist}(a, Y).$$

Proof. If $a \in Y$, take $b = a$, and let z' be any extremal point of S'. Choose $|\alpha| = 1$ such that $\operatorname{Re} \alpha z'(b) \geqslant 0$. It is readily verified that $x' = \alpha z'$ is an extremal point of S' with the desired properties. Now suppose $a \in X \backslash Y$, and let

$$K = \{x' \in S' : x'(a-b) = \|a-b\|\}.$$

Define $x'' \in X''$ by $x''(x') = x'(b)$ for $x' \in X'$, and let

$$\beta = \sup\{\operatorname{Re} x''(x') : x' \in K\} = \sup\{\operatorname{Re} x'(b) : x' \in K\}.$$

Then K is nonempty and $\beta \geqslant 0$, because Theorem 3.4 implies there exists $z' \in S'$ such that $z'(b) = 0$ and $z'(a-b) = z'(a) = \operatorname{dist}(a, Y) = \|a-b\|$. Now K is obviously $\sigma(X', X)$-closed and hence is $\sigma(X', X)$-compact, by Alaoglu's theorem. Since x'' is $\sigma(X', X)$-continuous, we may apply Lemma 11.1 to conclude that the set

$$E = \{x' \in K : \operatorname{Re} x''(x') = \beta\}$$

is a nonempty $\sigma(X', X)$-compact extremal subset of K. By Theorem 11.2, E contains an extremal point x', which must, consequently, be an extremal point of K. Now it is easily verified that K is an extremal subset of S'. (Use the fact that for $x' \in S'$, $x'(a-b)$ is a point in the disc $|z| \leqslant \|a-b\|$, and $\|a-b\|$ is itself an extremal point of this disc.) Hence x' is an extremal point of S'. Since $x' \in E$ and $\beta \geqslant 0$, we are finished. □

Theorem 11.6. *Let Y be a proper closed subspace of a normed linear space X. Given $a \in X$ and $b \in Y$, then b is a best approximation to a by elements of Y if and only if to each y in Y there corresponds an extremal point x' of the closed unit ball S' in X' such that*

$$\operatorname{Re} x'(b-y) \geqslant 0,$$
$$x'(a-b) = \|a-b\|.$$

Proof. Suppose that b has the stated property. Then for any $y \in Y$, the corresponding x' has the property that $\operatorname{Re} x'(b) \geq \operatorname{Re} x'(y)$. Hence

$$\begin{aligned}\|a-b\| &= \operatorname{Re} x'(a-b) \leq \operatorname{Re} x'(a-y)\\ &\leq |x'(a-y)| \leq \|a-y\|.\end{aligned}$$

(Note that we have not used the fact that x' is an extremal point of S'.) It follows that $\|a-b\| = \operatorname{dist}(a, Y)$. Conversely, suppose that $\|a-b\| = \operatorname{dist}(a, Y)$, and take $y \in Y$. Then $\|(a-y)-(b-y)\| = \operatorname{dist}(a, Y) = \operatorname{dist}(a-y, Y)$. Replacing a and b in Theorem 11.5 by $a-y$ and $b-y$, respectively, we obtain the appropriate extremal point x' in S'. □

The following result is a special case of Theorem 11.5, where Y is the zero subspace of X.

Theorem 11.7. *Let E be the set of extremal points in the closed unit ball of the conjugate of a normed linear space X. To each $x \in X$ corresponds an $x' \in E$ such that $x'(x) = \|x\|$. Hence*

$$\|x\| = \sup_{x' \in E} |x'(x)|.$$

As another application of the Krein–Milman theorem, we shall prove the Stone–Weierstrass theorem for the complex Banach space $C(T)$, where T is a compact Hausdorff topological space. Recall from the Riesz–Kakutani theorem (Theorem 5.7) that the conjugate of $C(T)$ may be identified with the Banach space $\mathcal{M}(T)$ of all regular complex Borel measures μ on T, with the total variation norm $\|\mu\| = |\mu|(T)$. With this identification, the annihilator of a set A in $C(T)$ becomes the set

$$A^{\perp} = \left\{\mu \in \mathcal{M}(T) : \int_T g\, d\mu = 0, g \in A\right\}.$$

The *support* of a measure μ in $\mathcal{M}(T)$ is the set of all $t \in T$ such that $|\mu|(U) > 0$ for all open neighborhoods U of t in T. It follows easily that the support of μ is closed. If the support is empty, then for each $t \in T$ there exists an open set U containing t such that $|\mu|(U) = 0$. Since T is compact, it must be the union of finitely many such open sets. It follows that $|\mu|(T) = 0$. Thus the support of any nonzero measure in $\mathcal{M}(T)$ is a nonempty compact subset of T.

Our proof of the Stone–Weierstrass theorem is essentially due to L. de Branges [1]. We begin with a preliminary theorem of de Branges.

Theorem 11.8. *Let A be a subalgebra of $C(T)$, and let $K = \{\mu \in A^{\perp} : \|\mu\| \leq 1\}$. Let μ be an extremal point of K, and let f be a real-valued function in A such that $0 < f < 1$. Then f is constant on the support of μ.*

Proof. If $\mu = 0$, then the support of μ is empty, and the theorem is trivially true. If $\mu \neq 0$, then $\|\mu\| = 1$. Define regular complex Borel measures ν and λ by

$$\nu(E) = \int_E f\,d\mu, \qquad \lambda(E) = \int_E (1-f)\,d\mu,$$

for Borel sets E. Since $f \in A$ and since A is an algebra, it follows that $\nu, \lambda \in A^\perp$. Also, ν and λ are nonzero, because $0 < f < 1$. We have

$$\mu = \|\nu\| \cdot \frac{\nu}{\|\nu\|} + \|\lambda\| \cdot \frac{\lambda}{\|\lambda\|},$$

and this is a convex combination of elements of K, because

$$\begin{aligned} \|\nu\| + \|\lambda\| &= \int_T f\,d|\mu| + \int_T (1-f)\,d|\mu| \\ &= |\mu|(T) = \|\mu\| = 1. \end{aligned}$$

Since μ is an extremal point, we conclude that $\mu = \nu/\|\nu\|$, and so $\nu = \|\nu\| \cdot \mu$. Hence

$$\int_E f\,d\mu = \int_E \|\nu\|\,d\mu$$

for all Borel sets E. This implies that $f(t) = \|\nu\|$ for all t except on a set of $|\mu|$-measure zero. Since f is continuous, $f(t) = \|\nu\|$ on the support of μ. □

Theorem 11.9 (Stone–Weierstrass). *Let T be a compact Hausdorff topological space, and let A be a closed subalgebra of $C(T)$ with the property that (a) A separates the points of T, (b) if f is in A, then the complex conjugate $\bar{f}$ is in A, and (c) the constant functions belong to A. Then $A = C(T)$.*

Proof. Let $K = \{\mu \in A^\perp : \|\mu\| \leq 1\}$. Then K is obviously nonempty, convex, and weak*-closed (cf. § 7). Also, K is weak*-compact, by Alaoglu's theorem. Hence K contains an extremal point μ, by the Krein–Milman theorem (actually Theorem 11.2 suffices here). Suppose the support of μ contains at least two distinct points s and t. Then it follows easily from the properties of A that A contains a real-valued function f such that $0 < f < 1$ and $f(s) \neq f(t)$. This is impossible, by Theorem 11.8. So the support of μ consists of only one point, say, t. Hence μ must be a scalar multiple of a Dirac measure, $\mu = \alpha \cdot \delta_t$, and

$$\int_T f\,d\mu = \alpha \cdot f(t), \qquad f \in A.$$

However, μ belongs to $A^\perp$, and so $\alpha \cdot f(t) = 0$ for $f \in A$. Since A contains the constant functions, this implies $\alpha=0$. Hence $\mu=0$, which shows that the only extremal point of K is the zero measure. It follows that $K=(0)$ and $A^{\perp}=(0)$. Since A is closed, we have $A=C(T)$, as a consequence of the Hahn–Banach theorem (Theorem 3.4). □

The proofs of Theorems 11.8 and 11.9 lead naturally to the following characterization of the extremal points of the closed unit ball of the conjugate of $C(T)$.

Theorem 11.10. *Let T be a compact Hausdorff topological space. A measure μ in $\mathcal{M}(T)$ is an extremal point of the closed unit ball B in $\mathcal{M}(T)$ if and only if $\mu=\alpha\cdot\delta_t$ for some $t\in T$ and some scalar α with $|\alpha|=1$.*

Proof. Suppose μ is an extremal point of B. Then $\|\mu\|=1$. An argument like that in the proof of Theorem 11.8 shows that each real-valued f in $C(T)$ such that $0<f<1$ must be constant on the support of μ. Using Urysohn's lemma (cf. Dunford and Schwartz [1, page 15]), we see immediately that this is impossible unless the support of μ is a single point. Thus $\mu=\alpha\cdot\delta_t$ for some α and t, and $|\alpha|=1$ because $\|\mu\|=1$.

Now, if $|\alpha|=1$ and $t\in T$, then clearly $\alpha\cdot\delta_t\in B$. Suppose that

$$\alpha\cdot\delta_t=\gamma\cdot\mu_1+(1-\gamma)\cdot\mu_2, \tag{11-1}$$

where $\mu_1,\mu_2\in B$ and $0<\gamma<1$. Applying these measures to the set $\{t\}$, we have

$$\alpha=\gamma\cdot\mu_1\{t\}+(1-\gamma)\cdot\mu_2\{t\}. \tag{11-2}$$

Now, for $i=1,2$, we have

$$|\mu_i\{t\}|\leqslant|\mu_i|\{t\}\leqslant|\mu_i|(T)=\|\mu_i\|\leqslant 1. \tag{11-3}$$

But since α is an extremal point of the unit circle in the complex plane, (11-2) can hold only if

$$\mu_1\{t\}=\mu_2\{t\}=\alpha. \tag{11-4}$$

Then $|\mu_i\{t\}|=|\alpha|=1$, and so (11-3) implies that $|\mu_i|\{t\}=|\mu_i|(T)=1$, $i=1,2$. Since $|\mu_i|$ is a positive measure, we conclude that the support of μ_i is the set $\{t\}$. Hence $\mu_i=\lambda_i\cdot\delta_t$ for some scalars λ_i, $i=1,2$. In view of (11-4), we have $\lambda_i=\alpha$ and $\mu_i=\alpha\cdot\delta_t$. Thus each measure $\alpha\cdot\delta_t$ with $|\alpha|=1$ is an extremal point of B. □

PROBLEMS

1. Let X be the space of real sequences $x=\{\xi_n\}$ such that $\lim_{n\to\infty}\xi_n=0$, and let $\|x\|=\sup_n|\xi_n|$. Then the closed unit ball in X has no extremal points. [If

$x = \{\xi_n\}$ and $\|x\| = 1$, construct $y = \{\alpha_n\}$ and $z = \{\beta_n\}$ with $\|y\| \leq 1$, $\|z\| \leq 1$ such that $x = (y + z)/2$. You can let $\alpha_n = \beta_n = \xi_n$ for all but one n.]

2. The closed unit ball in the complex space $L^1(0, 1)$ has no extremal points.
3. Describe the extremal points of the closed unit ball in $C[0, 1]$. [Consider separately the cases of real and complex scalars.]
4. The set E of extremal points of a compact set K is not necessarily compact. As an example, let K be the convex hull in $\boldsymbol{R}^3$ of the points $(1, 0, 1)$, $(1, 0, -1)$ and $(\cos\theta, \sin\theta, 0)$, $0 \leq \theta \leq 2\pi$.
5. A normed linear space X is said to be *strictly convex* if $\|x + y\| = \|x\| + \|y\|$ only if x and y are linearly dependent.
   a. Show that X is strictly convex if and only if every boundary point of the closed unit ball S in X is an extremal point.
   b. Show that X is strictly convex if and only if for each $x' \in X'$ there is at most one $x \in S$ such that $\operatorname{Re} x'(x) = \sup\{\operatorname{Re} x'(w) : w \in S\}$.
6. If $1 < p < \infty$, then ℓ^p is strictly convex.
7. Let X and Y be Hausdorff locally convex spaces (over the same scalar field), and let T be a continuous linear mapping from X into Y.
   a. If K is a nonempty compact convex subset of X, then every extremal point of $T(K)$ is the image under T of some extremal point of K.
   b. If K is a convex set in X and E is an extremal subset of $T(K)$, then $\{x \in X : Tx \in E\}$ is an extremal subset of K.
   c. Show how part (b) is related to Lemma 11.1.
8. Prove the following generalization of the Krein–Milman theorem: Let X be a Hausdorff locally convex space with a topology τ, and let K be a $\sigma(X, X')$-compact convex subset of X. Then K is the τ-closed convex hull of the set of its extremal points.
9. Let K be a compact convex set in a Hausdorff locally convex space X. Then every continuous real linear functional assumes its maximum on K at an extremal point of K.
10. A point in the conjugate of $L^1(0, 1)$ is an extremal point of the closed unit ball if and only if it is represented by an f in $L^\infty(0, 1)$ such that $|f(t)| = 1$ for almost all $t \in [0, 1]$.
11. Suppose the closed unit ball of a reflexive Banach space X contains only a finite number of extremal points. Then X is finite dimensional.
12. Let M be a proper subspace of a normed linear space X.
   a. Given $x \in X$, there exists an extremal point x' of the set $\{x' \in M^\perp : \|x'\| \leq 1\}$ such that $x'(x) = \operatorname{dist}(x, M)$.
   b. Show that $x \in \bar{M}$ if and only if $x'(x) = 0$ for all extremal points of the unit ball in $M^\perp$.
13. Explain why $L^1(0, 1)$ cannot be isometrically isomorphic to the conjugate of a normed linear space.
14. Use Theorem 11.10 to restate Theorem 11.6 for the specific case of the real space $X = C[0, 1]$.

IV GENERAL THEOREMS ON LINEAR OPERATORS

The basic facts about closed linear operators in normed linear spaces are presented in this chapter. Completeness of the spaces is assumed only where it seems to be essential for decisive results. Throughout the chapter, theoretical results are accompanied by concrete illustrations and applications.

The first four sections concern continuous linear operators. The main results here are the principle of uniform boundedness (Theorem 1.2) and Theorems 1.4 and 1.5 on the existence of continuous inverses of linear operators. Sections 2, 3, and 4 present applications of the theory in § 1.

The highly important open mapping theorem and closed graph theorem are proved in § 5 (Theorems 5.5 and 5.7). These results were first given by S. Banach in 1932 for operators in complete metric linear spaces. Since then, more general settings for these theorems have been found. For our purposes, however, it is enough to consider only Banach spaces.

A large block of material in this chapter deals with a closed linear operator and its conjugate. The theorems in § 9 are more profound than those in § 8. The pooling of all the results about an operator and its conjugate, using theorems from §§ 5, 8, and 9, leads to the "state diagram" in § 10, which is a complete tabulation of all the relations of a certain kind that can exist between an operator and its conjugate. Additional information for operators with closed range is given in Theorem 10.1, the closed range theorem. Adjoint operators are discussed in § 11. Their theory closely resembles the theory of conjugate operators but has certain special features because the conjugate of a Hilbert space X may be identified with X itself via the Fréchet–Riesz representation theorem.

A special class of operators, projections, are dealt with in § 12. Such operators will arise naturally in the spectral theory to be discussed in Chapters V and VI.

The chapter concludes with an introduction to the theory of Fredholm operators. Such operators are found in many applications, some to be discussed later in Chapter V. The main "perturbation" theorems, Theorems

13.5, 13.6, and 13.8, are important generalizations of Theorems 1.4 and 1.5, mentioned above.

IV.1 SPACES OF LINEAR OPERATORS

Linear operators were defined in § I.3. If X and Y are linear spaces (with the same scalar field), the set of all operators on X into Y is a linear space if we define addition of operators and multiplication of operators by scalars in the natural way, namely,

$$(A+B)x = Ax + Bx, \qquad (\alpha A)x = \alpha(Ax).$$

The operator that maps every $x \in X$ into the zero element of Y is the zero of the linear space of operators; without serious danger of confusion with other zeros, we denote this zero operator by 0.

If X and Y are topological linear spaces, those linear operators on X into Y that are continuous on X form a subspace of the space of *all* linear operators on X into Y. We shall denote the linear space of all *continuous* linear operators on X into Y by $L(X, Y)$.

If X and Y are normed linear spaces, a linear operator A on X into Y is continuous on X (i.e., $A \in L(X, Y)$) if $\sup_{\|x\| \le 1} \|Ax\| < \infty$. For such an A we define

$$\|A\| = \sup_{\|x\| \le 1} \|Ax\|. \tag{1-1}$$

These matters were discussed in § II.1, particularly in Theorem II.1.1; formula (1-4) of that section gives useful alternative formulas for the norm of an operator. When X and Y are normed linear spaces, it is a common practice to call the elements of $L(X, Y)$ *bounded linear operators*. The use of the adjective "bounded" as an equivalent for "continuous" is explained by the remarks preceding Theorem II.1.2.

Theorem 1.1. *If X and Y are normed linear spaces, $L(X, Y)$ is a normed linear space with the norm defined by* (1-1). *If Y is complete, so is $L(X, Y)$.*

Proof. We leave to the reader the simple verification that $\|A\|$ is a norm on $L(X, Y)$. When Y is complete, the proof of the completeness of $L(X, Y)$ is practically identical to the proof that the normed conjugate space X' is complete (Theorem II.4.5): instead of considering a Cauchy sequence $\{x'_n\}$ in X', one considers a Cauchy sequence $\{A_n\}$ in $L(X, Y)$; the absolute values such as $|x'_n(x)|$, which appeared in the proof of Theorem II.4.5, are now replaced everywhere by norms, $\|A_n x\|$. No other changes are required. □

Suppose that A_n and A are in $L(X, Y)$ and that $\|A_n - A\| \to 0$. This is equivalent to the statement that $\|A_n x - Ax\| \to 0$ uniformly for all $x \in X$ such that $\|x\| \leqslant 1$. For this reason the topology for $L(X, Y)$ defined by the norm (1-1) is often called the *uniform topology* for $L(X, Y)$.

Another topology on $L(X, Y)$ of importance is the *strong operator topology*—the locally convex topology defined by the family of all seminorms of the form

$$p_x(A) = \|Ax\|,$$

where $x \in X$. The third most commonly used topology is the *weak operator topology*. It is defined by the family of all seminorms of the form

$$p_{x,y'}(A) = |y'(Ax)|,$$

where $x \in X$ and $y' \in Y'$. Since $\|Ax\| \leqslant \|A\| \, \|x\|$ and $|y'(Ax)| \leqslant \|y'\| \, \|Ax\|$, it is evident that the uniform operator topology is stronger than the strong operator topology and that the strong operator topology is stronger than the weak operator topology. We shall always use the uniform operator topology on $L(X, Y)$, unless specific reference is made to another topology. (We note in passing that $L(X, Y)$, as a normed linear space, has a conjugate space that can be used to define a "weak" topology on $L(X, Y)$. This topology should not be confused with the weak operator topology.)

The Principle of Uniform Boundedness

Although we shall not investigate the relations among the various topologies on $L(X, Y)$, the uniform boundedness theorem of Chapter III has a very important analogue in the present setting. It says essentially that the class of bounded sets in $L(X, Y)$ is the same for all three operator topologies described above.

Theorem 1.2 (The Principle of Uniform Boundedness). *Suppose that X and Y are normed linear spaces and X is complete. Let G be a subset of $L(X, Y)$.*

(*a*) *If* $\sup_{A \in G} \|Ax\| < \infty$ *for each* $x \in X$, *then* $\sup_{A \in G} \|A\| < \infty$.

(*b*) *If* $\sup_{A \in G} |y'(Ax)| < \infty$ *for each* $x \in X$, $y' \in Y'$, *then* $\sup_{A \in G} \|A\| < \infty$.

Proof. The proof of Theorem III.9.1 applies to part (*a*), when linear functionals and absolute values are replaced by linear operators and norms. An alternate proof of (*a*) will be given later in § 5. To prove (*b*) we note that, for $x \in X$, $\{Ax : A \in G\}$ is a bounded set in Y, by a form of the old uniform boundedness principle (Theorem III.9.2). It then follows from (*a*) that $\{A : A \in G\}$ is a bounded set in $L(X, Y)$. □

As a particular case of Theorem 1.2, we note the following (with assumption on X and Y as in the theorem). *If $A_n \in L(X, Y)$ and if $Ax = \lim_{n \to \infty} A_n x$ exists* *for each* $x \in X$, *then* $A \in L(X, Y)$; *that is,* A *is not only linear, but also continuous.*

The Algebra $L(X)$

When $Y = X$, we shall write $L(X)$ in place of $L(X, X)$. For $A, B \in L(X)$ we define the product AB as the linear operator such that $(AB)x = A(Bx)$ for $x \in X$. Then $AB \in L(X)$, because the composition of continuous mappings is continuous. With this definition of multiplication, $L(X)$ becomes an *algebra.* That is, $L(X)$ is a linear space together with an operation of multiplication having the following properties.

1. $(AB)C = A(BC)$.
2. $A(B + C) = AB + AC$.
3. $(A + B)C = AC + BC$.
4. $(\alpha A)(\beta B) = (\alpha\beta)(AB)$, $\quad \alpha, \beta$ scalars.

The algebra is called real or complex according as the scalar field is real or complex.

In the next theorem and throughout the rest of the text, we let I denote the identity operator on X; that is, $Ix = x$ for all $x \in X$.

Theorem 1.3. *Suppose* X *is a normed linear space. Then*

$$\|AB\| \leq \|A\| \, \|B\| \tag{1-2}$$

for each pair A, B *in* $L(X)$. *Also,*

$$\|I\| = 1. \tag{1-3}$$

Proof. For each x, $\|(AB)x\| = \|A(Bx)\| \leq \|A\| \, \|Bx\| \leq \|A\| \, \|B\| \, \|x\|$, by (1-3) in § II.1. Formula (1-2) follows immediately. The truth of (1-3) is immediate from the definition of the norm. $\square$

From (1-2) it follows that the product AB is a continuous function of the pair A, B.

When an operator A in $L(X)$ is injective, the domain of the inverse A^{-1} is understood to be the range $\mathscr{R}(A)$ of A (which may not be all of X). If A is not only injective, but $\mathscr{R}(A) = X$ and A^{-1} is continuous, then A^{-1} belongs to $L(X)$ and we say that A is *invertible in the algebra* $L(X)$. The next two theorems furnish important information about the existence and nature of inverse operators under certain conditions. Formula (1-4) is closely related to an old result in the theory of integral equations, generally known as the Neumann expansion (named after C. Neumann, cf. Hellinger and Toeplitz [1, page 1347]), though the method goes back as far as Liouville.

Theorem 1.4. *Let X be a Banach space. Suppose that an operator A in $L(X)$ has the property that the series $\Sigma_0^\infty A^n$ converges in the uniform operator topology (where $A^0 = I$). Then $I - A$ is invertible in the algebra $L(X)$ and*

$$(I-A)^{-1} = \sum_0^\infty A^n. \tag{1-4}$$

A sufficient condition for (1-4) *is that* $\|A\| < 1$. *In this case*

$$\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}. \tag{1-5}$$

Proof. If we let $B = \Sigma_0^\infty A^n$, we may use the continuity of multiplication to obtain

$$AB = BA = \sum_0^\infty A^{n+1},$$

and hence

$$(I-A)B = B(I-A) = I.$$

This implies that $(I-A)^{-1}$ exists and is equal to B. When $\|A\| < 1$, the series $\Sigma_0^\infty \|A\|^n$ converges to $(1-\|A\|)^{-1}$. But $\|A^n\| \leq \|A\|^n$, by (1-2). Since $L(X)$ is complete, it follows that the series $\Sigma_0^\infty A^n$ converges in $L(X)$ (cf. Theorem II.4.3). The inequality (1-5) is now evident. $\square$

A useful geometric interpretation of this theorem may be given by setting $T = I - A$. Then T is invertible in $L(X)$ whenever $\|I - T\| = \|A\| < 1$. Thus the open unit ball centered at the identity I, $\{T \in L(X) : \|I - T\| < 1\}$, consists entirely of invertible elements of $L(X)$.

Theorem 1.4 has great theoretical value—it will be used many times in this chapter and the next—but it also provides a practical *iterative method* for computing approximate solutions of equations of the form $x - Ax = y$. We let

$$\begin{aligned} x_0 &= y \\ x_1 &= y + Ax_0 \\ x_2 &= y + Ax_1 = y + Ay + A^2 y \\ &\;\vdots \\ x_n &= y + Ax_{n-1} = S_n y, \qquad \text{where} \qquad S_n = \sum_{k=0}^{n} A^k. \end{aligned}$$

If A satisfies the condition of Theorem 1.4 (or problem 2), then $x_n \to x = (I-A)^{-1}y$. When $\|A\| < 1$, the error of the nth approximation can be estimated by

$$\|x - x_n\| \leq \|(I-A)^{-1} - S_n\| \, \|y\|$$
$$\leq \Big(\sum_{n+1}^{\infty} \|A\|^k \Big) \|y\| = \frac{\|A\|^{n+1}}{1 - \|A\|} \|y\|.$$

Theorem 1.5. *Let X be a Banach space, and let G be the set of all invertible elements in the algebra $L(X)$. If $A \in G$, $B \in L(X)$, and $\|A - B\| < 1/\|A^{-1}\|$, then $B \in G$, and*

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \, \|A - B\|}, \tag{1-6}$$

$$\|B^{-1} - A^{-1}\| \leq \frac{\|A^{-1}\|^2 \|A - B\|}{1 - \|A^{-1}\| \, \|A - B\|}. \tag{1-7}$$

Thus, in particular, G is an open set in $L(X)$ and the mapping $A \mapsto A^{-1}$ is a continuous function on G.

Proof. Theorem 1.4 shows that $A^{-1}B$ is invertible in $L(X)$, because $\|I - A^{-1}B\| = \|A^{-1}(A-B)\| \leq \|A^{-1}\| \, \|A - B\| < 1$. Also, from (1-5),

$$\|(A^{-1}B)^{-1}\| \leq \frac{1}{1 - \|I - A^{-1}B\|} \leq \frac{1}{1 - \|A^{-1}\| \, \|A - B\|}. \tag{1-8}$$

Since $B = A(A^{-1}B)$, we see that B has an inverse in $L(X)$, namely, $(A^{-1}B)^{-1}A^{-1}$. Also, $\|B^{-1}\| \leq \|(A^{-1}B)^{-1}\| \, \|A^{-1}\|$. The estimate (1-6) follows from this and (1-8). To verify (1-7), we note that

$$B^{-1} - A^{-1} = (I - A^{-1}B)B^{-1} = A^{-1}(A - B)B^{-1}.$$

Thus

$$\|B^{-1} - A^{-1}\| \leq \|A^{-1}\| \, \|A - B\| \, \|B^{-1}\|,$$

which combines with (1-6) to yield (1-7). □

It is worth mentioning that the proof of this theorem applies verbatim to elements of $L(X, Y)$, where Y is also a Banach space and where the (open) set G is replaced by the (open) set $\{A \in L(X, Y) : A^{-1}$ exists and is in $L(Y, X)\}$. Also, (1-6) and (1-7) are valid for $A, B \in L(X, Y)$ when $\|A - B\| < 1/\|A^{-1}\|$.

Approximate Solutions of Equations

Functional analysis has an important application in the investigation of approximate solutions of linear equations. Frequently, a problem involves a linear equation that cannot be solved explicitly, even though it is known that a unique solution exists. We have already mentioned an iteration method that is often useful in such situations. Another method of attack is to simplify the equation in some fashion and solve the "approximate" equation. This solution can then be taken as an approximation to the solution of the original equation.

To be more precise, let us consider the equation

$$Ax = y, \tag{1-9}$$

where $A \in L(X, Y)$, for some Banach spaces X and Y. It is reasonable to feel that some alternate equation

$$B\tilde{x} = \tilde{y} \tag{1-10}$$

will be a useful approximation to (1-9) if B in $L(X, Y)$ and $\tilde{y}$ can be chosen so that $\|A - B\|$ and $\|y - \tilde{y}\|$ are sufficiently small. Of course, we would hope to choose B and $\tilde{y}$ in a way that would make it easier to solve (1-10) for $\tilde{x}$ than to solve (1-9) for x.

Suppose we know that $\mathscr{R}(A) = Y$ and that A^{-1} exists and is continuous; then (1-9) has a unique solution for each $y \in Y$. Theorem 1.5 tells us that when $\|A - B\|$ is small enough, the alternate equation (1-10) also has a unique solution $\tilde{x}$ for each $\tilde{y} \in Y$. The number $\|x - \tilde{x}\|$ is a measure of the error that arises when $\tilde{x}$ is used as an approximation for x. We have

$$\begin{aligned} \|x - \tilde{x}\| &\leq \|A^{-1}y - B^{-1}y\| + \|B^{-1}y - B^{-1}\tilde{y}\| \\ &\leq \|A^{-1} - B^{-1}\| \, \|y\| + \|B^{-1}\| \, \|y - \tilde{y}\|. \end{aligned} \tag{1-11}$$

From (1-11) in combination with (1-6) and (1-7) we see that, for a given $\varepsilon > 0$, it is possible to choose $\delta_1 > 0$ and $\delta_2 > 0$ in such a way that $\|x - \tilde{x}\| < \varepsilon$ whenever B and $\tilde{y}$ satisfy $\|A - B\| < \delta_1$ and $\|y - \tilde{y}\| < \delta_2$. Thus the error can be made as small as is desired.

Once B and $\tilde{y}$ have been chosen so as to guarantee that $\|x - \tilde{x}\|$ will be suitably small, it is possible to obtain a sharper estimate of the error.

Theorem 1.6. *Let X and Y be Banach spaces, and let A and B be operators in $L(X, Y)$ such that both A^{-1} and B^{-1} exist and belong to $L(Y, X)$. For any fixed y, $\tilde{y} \in Y$, let x and $\tilde{x}$ be the (unique) solutions of the equations*

$$Ax = y \qquad \text{and} \qquad B\tilde{x} = \tilde{y}.$$

Then

$$\|x - \tilde{x}\| \leq \|A^{-1}\| \, \|A - B\| \, \|\tilde{x}\| + \|A^{-1}\| \, \|y - \tilde{y}\|. \tag{1-12}$$

Proof. We have

$$\begin{aligned} \tilde{x} - x &= (\tilde{x} - A^{-1}\tilde{y}) + (A^{-1}\tilde{y} - A^{-1}y) \\ &= A^{-1}(A - B)\tilde{x} + A^{-1}(\tilde{y} - y) \end{aligned}$$

The desired inequality follows immediately. $\square$

Example. Let (a_{ij}) be an $n \times n$ matrix. Then (a_{ij}) determines a linear operator A on $\ell^1(n)$, and it is not difficult to show that

$$\|A\| = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|. \tag{1-13}$$

Suppose that the values of the a_{ij} are not known exactly due to errors in measurement or perhaps due to computer round-off error. Denote the approximate value of a_{ij} by b_{ij}. If we assume that $|a_{ij} - b_{ij}| < \varepsilon$ for $1 \leq i, j \leq n$, then the operator B determined by (b_{ij}) satisfies

$$\|A - B\| < n\varepsilon,$$

for the norm (1-13). Suppose that the matrix (b_{ij}) is found to be nonsingular, and suppose that ε satisfies $n\varepsilon < 1/\|B^{-1}\|$. Then A must have an inverse in $L(\ell^1(n))$ and

$$\|A^{-1}\| \leq \frac{\|B^{-1}\|}{1 - n\varepsilon\|B^{-1}\|}, \tag{1-14}$$

by (1-6). If we solve $B\tilde{x} = y$ for $\tilde{x}$ and use this as an approximation for the true solution of $Ax = y$, then the error is estimated from (1-12):

$$\|x - \tilde{x}\| < n\varepsilon \|A^{-1}\| \, \|\tilde{x}\|.$$

Unfortunately, in this situation we have no way of estimating $\|A^{-1}\|$ directly. So we use (1-14) to obtain

$$\|x - \tilde{x}\| < \frac{n\varepsilon\|B^{-1}\|}{1 - n\varepsilon\|B^{-1}\|} \|\tilde{x}\|.$$

Another application of Theorem 1.6 will be given in § 2.

PROBLEMS

1. Suppose that X, Y, and Z are normed linear spaces such that X is a subspace of Z dense in Z and Y is complete. Show that $L(X, Y)$ is congruent to $L(Z, Y)$ by mapping $T \in L(Z, Y)$ into $A \in L(X, Y)$, where A is the restriction of T to X.
2. Let X be a Banach space. Suppose that $A \in L(X)$ has the property that $\Sigma_0^\infty A^n x$ converges for each $x \in X$. Show that $I - A$ is injective and $\mathscr{R}(I - A) = X$. [By a result in § 5, this will guarantee that $I - A$ is invertible in the algebra $L(X)$.]
3. Let X, Y be Banach spaces, and let

$$G = \{A \in L(X, Y) : A^{-1} \text{ exists and is in } L(Y, X)\}.$$

Using the uniform operator topology on G, show that the mapping $(A, y) \mapsto A^{-1}y$ is continuous from the Cartesian product space $G \times Y$ into X. [See (1-11).]

4. Let X, Y be normed linear spaces, and let A, B be operators in $L(X, Y)$. Suppose that A has a continuous inverse A^{-1} (whose domain is $\mathcal{R}(A)$), and let $m = 1/\|A^{-1}\|$. If $\|A - B\| < m$, then B has a continuous inverse and B^{-1} satisfies the inequalities (1-6) and (1-7). [*Hint.* See Theorem II.1.2.]

IV.2 INTEGRAL EQUATIONS OF THE SECOND KIND

We start out by referring back to Example 2 in § I.5 where integral equations of Fredholm and Volterra type were defined. Such equations can be considered in various function spaces; first, we shall deal with the space of continuous functions. Let X be the Banach space $C[a, b]$, and let K be the element of $L(X)$ defined by $Kx = y$, where

$$y(s) = \int_a^b k(s, t)x(t)\, dt, \tag{2-1}$$

and the function $k(s, t)$ (called the kernel) is subjected to certain restrictions. There are various sets of restrictions that are adequate for the purposes of the developments that are to follow. The simplest restrictions are as follows:

1. $k(s, t)$ is continuous in both variables when $a \leqslant s \leqslant b$ and $a \leqslant t \leqslant b$.
2. If M is the maximum of $|k(s, t)|$ on the square $[a, b] \times [a, b]$, then $M(b - a) < 1$.

These restrictions may be made less severe, but at the cost of some complication in statements and proofs. Our main purpose is to give an exposition of the formal aspects of part of the theory of integral equations of the second kind, and the two foregoing conditions on the kernel will serve conveniently for the rigorous justification of the formal procedures.

Fredholm Equations of the Second Kind

We are concerned with the equation

$$y(s) = x(s) - \int_a^b k(s, t)x(t)\, dt;$$

as an equation in the space X it can be written

$$y = x - Kx. \tag{2-2}$$

According to Theorem 1.4, this equation has a unique solution x for each given y if $\|K\| < 1$. (We do not assert that $\|K\| < 1$ is a *necessary* condition; only that it is sufficient.) The solution is

$$x = y + Ky + K^2y + \cdots + K^ny + \cdots. \tag{2-3}$$

Now, from (2-1) and the definition of M in condition (2), we see that

$$\|K\| \leq M(b-a). \tag{2-4}$$

For a precise value of $\|K\|$, which we do not need here, the reader may see Liusternik and Sobolev [1, page 96]. Condition (2) then assures us that $\|K\| < 1$.

We are going to show that each of the operators $K^2, K^3, \ldots$ is of the same sort as K, with a suitable kernel. Let us define

$$\begin{cases} k_1(s,t) = k(s,t) \\ k_n(s,t) = \displaystyle\int_a^b k_1(s,u) k_{n-1}(u,t)\, du, \qquad n > 1. \end{cases} \tag{2-5}$$

By definition, $z = K^2 y$ means $z = K(Ky)$, or

$$z(s) = \int_a^b k(s,t)\left[\int_a^b k(t,u) y(u)\, du\right] dt.$$

We invert the order of integration, obtaining

$$z(s) = \int_a^b y(u)\left[\int_a^b k(s,t)k(t,u)\, dt\right] du$$

$$= \int_a^b k_2(s,u) y(u)\, du.$$

Thus K^2 is defined in terms of the kernel k_2 just as K is defined in terms of k_1. It is easily proved by induction that $z = K^n y$ is expressed by

$$z(s) = \int_a^b k_n(s,t) y(t)\, dt, \tag{2-6}$$

so that k_n is the kernel used to determine K^n. For $n > 1$, k_n is called the nth iterated kernel.

The series (2-3) now takes the form

$$x(s) = y(s) + \sum_{n=1}^{\infty} \int_a^b k_n(s,t) y(t)\, dt. \tag{2-7}$$

The convergence of (2-3) as a series of elements in the Banach space $C[a,b]$ means that (2-7) is uniformly convergent on $[a,b]$. Now consider the function h defined by

$$h(s,t) = \sum_{n=1}^{\infty} k_n(s,t). \tag{2-8}$$

It is easily proved by induction from (2-5) that

$$|k_n(s,t)| \leq M^n (b-a)^{n-1}, \qquad n > 1.$$

Since we have assumed $M(b-a)<1$, it follows that the series (2-8) converges uniformly; thus h is a continuous function of s and t. Moreover, we can exchange the order of summation and integration in (2-7) and obtain

$$x(s)=y(s)+\int_a^b h(s,t)y(t)\,dt. \tag{2-9}$$

Let us define an operator H with h as kernel, so that $z=Hy$ means

$$z(s)=\int_a^b h(s,t)y(t)\,dt.$$

Then (2-9) becomes

$$x=y+Hy.$$

We thus arrive at the conclusion that when conditions (1) and (2) are satisfied, there is an operator H determined by a kernel h in such a way that

$$(I-K)^{-1}=I+H. \tag{2-10}$$

The kernel h is sometimes called the *reciprocal* kernel corresponding to k. It has also been called the *resolvent* kernel, but in modern operator theory the word resolvent is used in a slightly different way.

An operator K defined by a kernel k as in (2-1) has been called a Fredholm-type operator, after I. Fredholm, who developed a comprehensive theory for integral equations of the second kind at the beginning of the twentieth century.

An Approximate Solution of a Fredholm Equation

Let us consider the equation

$$y(s)=x(s)-\int_0^{1/2}\sin st\, x(t)\,dt, \tag{2-11}$$

where $x, y \in C[0,\frac{1}{2}]$. This equation has a unique solution x for each $y \in C[0,\frac{1}{2}]$. For, if we let K be the integral operator whose kernel is $\sin st$, then $0 \leq \sin st \leq \sin\frac{1}{4} \leq \frac{1}{4}$ when $0 \leq s, t \leq \frac{1}{2}$, and so $\|K\| \leq \frac{1}{8}$, by (2-4). Thus $I-K$ has a bounded inverse, and the solution of (2-11) is given by (2-9), where $h(s,t)$ must be calculated. Using (2-5), we have

$$k_2(s,t)=\int_0^{1/2}\sin su\,\sin ut\,du=\frac{\sin\frac{1}{2}(s-t)}{2(s-t)}-\frac{\sin\frac{1}{2}(s+t)}{2(s+t)},$$

when $s \neq t$, $k_2(s,s)=(s-\sin s)/4s$ if $s>0$, and $k_2(0,0)=0$. The expression for $k_3(s,t)$, however, involves an integral that cannot be expressed in closed form using elementary functions.

Instead of trying to find the exact solution of (2-11) for a given y, let us consider an approximate solution. Our objective is to replace (2-11) by an equation that is much easier to solve and whose solution is a good approximation to the solution of (2-11). An approximation to the operator $I-K$ of (2-11) can be written in the form $I-K_1$ (and we want $\|K-K_1\|$ to be small). To construct K_1 we expand the kernel of K in a Taylor series:

$$\sin st = st - \frac{1}{6}s^3t^3 + \frac{1}{120}s^5t^5 - \cdots.$$

If we define $K_1 \in L(X)$ by $K_1x = y$, where

$$y(s) = \int_0^{1/2} st\, x(t)\, dt,$$

then

$$\begin{aligned}\|Kx - K_1x\| &\leq \sup_s \int_0^{1/2} |\sin st - st|\, |x(t)|\, dt \\ &\leq \sup_s \int_0^{1/2} \frac{1}{6}s^3t^3\|x\|\, dt \\ &= \sup_s \frac{s^3\|x\|}{384} = \frac{1}{3072}\|x\|.\end{aligned}$$

Thus $\|K-K_1\| \leq \frac{1}{3072} < 3.26 \times 10^{-4}$. Therefore we shall study the following equation as an approximation to (2-11).

(2-12) $$y(s) = \tilde{x}(s) - \int_0^{1/2} st\, \tilde{x}(t)\, dt.$$

For a fixed y, the solution $\tilde{x}$ of (2-12) will be a very good approximation to the solution x of (2-11). In fact, if $A = I-K$ and $B = I - K_1$, then $\|A-B\| < 3.26 \times 10^{-4}$ and $\|A^{-1}\| < 1/(1-\frac{1}{8}) = \frac{8}{7}$, by (1-5), so that

(2-13) $$\|x - \tilde{x}\| < (\tfrac{8}{7})(3.26 \times 10^{-4})\|\tilde{x}\| < (3.73 \times 10^{-4})\|\tilde{x}\|,$$

by Theorem 1.6.

The "approximate" equation (2-12) is easily solved. We first rewrite it in the form

(2-14) $$\tilde{x}(s) = y(s) + Cs,$$

where

(2-15) $$C = \int_0^{1/2} t\tilde{x}(t)\, dt.$$

Substituting (2-14) into (2-15), we have

$$C = \int_0^{1/2} [ty(t) + Ct^2]\, dt.$$

Simplifying, we obtain

$$C = \frac{24}{23} \int_0^{1/2} ty(t)\, dt. \tag{2-16}$$

The solution of (2-12) is given by (2-14) and (2-16). Whenever a particular y makes (2-16) difficult to evaluate, we can replace y by an approximation $\tilde{y}$ and still use Theorem 1.6 to estimate $\|x - \tilde{x}\|$.

If necessary, we can improve our approximation to $x(s)$ by replacing K_1 with an operator K_n whose kernel is the first n terms of the Taylor expansion of sin st. (Since the Taylor series converges uniformly, it is easy to see that $\|K - K_n\| \to 0$ as $n \to \infty$.) The equation $y = (I - K_n)x$ is readily solved because $(I - K_n)^{-1} \in L(X)$ and K_n has what is known as a "degenerate" kernel.

In general, a kernel is said to be *degenerate* if it has the form

$$k(s, t) = \sum_{i=1}^{n} \phi_i(s)\psi_i(t),$$

where $\{\phi_1, \ldots, \phi_n\}$ and $\{\psi_1, \ldots, \psi_n\}$ are both linearly independent sets of functions. An integral equation with such a kernel is easily converted into a system of n linear equations in n unknowns. A discussion of integral equations with degenerate kernels may be found in Riesz and Sz.-Nagy [1, pages 161–165].

Volterra Equations of the Second Kind

The theory of Volterra integral equations may be developed independently, or it may be regarded as a special case of the theory of Fredholm integral equations, with important special features. A Volterra-type operator K is defined by $y = Kx$, where

$$y(s) = \int_a^s k(s, t)x(t)\, dt.$$

We shall assume that the kernel $k(s, t)$ is continuous in both variables in the triangular region defined by the inequalities $a \leqslant t \leqslant s \leqslant b$. If we define k to have the value 0 in the part of the square $[a, b] \times [a, b]$ where $t > s$, we can regard K as a Fredholm-type operator with this special feature of the kernel k. Since $k(s, s)$ need not be 0, there may be discontinuities of the kernel at points along the diagonal $t = s$ of the square. These discontinuities have no adverse effects on the earlier developments in this section, however.

It may be verified by induction from (2-5) that when k is a Volterra kernel, the iterated kernels k_n are also Volterra kernels; that is, $k_n(s, t) = 0$ if $t > s$. The iterative formula for k_n becomes

$$k_n(s, t) = \int_t^s k_1(s, u) k_{n-1}(u, t)\, du, \qquad n > 1, t \leq s. \tag{2-17}$$

From this it can readily be proved by induction that

$$|k_n(s, t)| \leq M^n \frac{(s-t)^{n-1}}{(n-1)!}, \tag{2-18}$$

where M is the maximum of $|k(s, t)|$. Hence certainly

$$|k_n(s, t)| \leq M^n \frac{(b-a)^{n-1}}{(n-1)!}.$$

This inequality shows that the series (2-8) converges uniformly. The function $h(s, t)$ defined by (2-8) is also a Volterra kernel; it is continuous in the triangle $a \leq t \leq s \leq b$, and $h(s, t) = 0$ if $t > s$.

From (2-18) and the formula for K^n in terms of k_n it follows that

$$\|K^n\| \leq M^n \frac{(b-a)^n}{n!}. \tag{2-19}$$

Consequently, the series $\sum_{n=0}^{\infty} K^n$ converges in the uniform topology of the space $L(X)$. By Theorem 1.4, the equation $y = x - Kx$ has a unique solution for each y, the solution being

$$x = \sum_{n=0}^{\infty} K^n y = y + \sum_{n=1}^{\infty} K^n y.$$

Furthermore, this solution may be expressed in the form

$$x(s) = y(s) + \int_a^s h(s, t) y(t)\, dt.$$

The very important fact about Volterra integral equations of the second kind is, then, that such an equation is always uniquely solvable. The special nature of the Volterra-type operator K ensures that the Neumann expansion is always convergent, regardless of the magnitude of the kernel k. This is not true, in general, in the case of a Fredholm-type operator.

IV.3 $\mathscr{L}^2$ KERNELS

In this section we consider the space $L^2(a, b)$ (hereafter in this section written simply as L^2). Suppose $k(s, t)$ belongs to the class $\mathscr{L}^2(E)$, where E is the square $[a, b]\times[a, b]$. Then we call k an $\mathscr{L}^2$ kernel. Such a kernel defines a bounded operator K on L^2 in the following way. If $x \in L^2$, let $y = Kx$, where this means

$$y(s) =^0 \int_a^b k(s, t)x(t)\, dt.$$

Since

$$|y(s)|^2 \leqslant^0 \int_a^b |k(s, t)|^2\, dt \int_a^b |x(t)|^2\, dt,$$

by the Schwarz inequality, we see that

$$\int_a^b |y(s)|^2\, ds \leqslant \int_a^b \int_a^b |k(s, t)|^2\, dt\, ds \int_a^b |x(t)|^2\, dt.$$

This shows that

$$\|y\| = \|Kx\| \leqslant \left(\int_a^b \int_a^b |k(s, t)|^2\, ds\, dt\right)^{1/2} \|x\|,$$

and hence that

$$\|K\| \leqslant \left(\int_a^b \int_a^b |k(s, t)|^2\, ds\, dt\right)^{1/2}. \tag{3-1}$$

As a member of $\mathscr{L}^2(E)$, $k(s, t)$ determines an element of the Hilbert space $L^2(E)$. We denote this element by k and its norm in $L^2(E)$ by $\|k\|$:

$$\|k\| = \left(\int_a^b \int_a^b |k(s, t)|^2\, ds\, dt\right)^{1/2}. \tag{3-2}$$

From (3-1) we see that $\|K\| \leqslant \|k\|$.

The equation $y = x - Kx$ in L^2 corresponds to a Fredholm integral equation

$$y(s) =^0 x(s) - \int_a^b k(s, t)x(t)\, dt$$

in $\mathscr{L}^2$. By Theorem 1.4, we know that $y = x - Kx$ has a unique solution x in L^2 for each choice of y in L^2, provided that $\|K\| < 1$. This will certainly be the case if $\|k\| < 1$.

In the case of $\mathscr{L}^2$ kernels we can define the iterated kernels k_n and the reciprocal kernel h in a manner formally the same as we did in § 2 with continuous kernels. The details are somewhat different, however. The iterated kernels are defined essentially as before: $k_1(s, t) = k(s, t)$ and

$$k_n(s, t) =^0 \int_a^b k_1(s, u)k_{n-1}(u, t)\, du,$$

where $=^0$ refers to the exception of a set of two-dimensional measure zero. It is not difficult to show by induction that we also have

$$k_n(s,t) =^0 \int_a^b k_{n-1}(s,u)k(u,t)\,du. \tag{3-3}$$

Also, by induction and the Schwarz inequality, we find that k_n is an $\mathscr{L}^2$ kernel, with

$$\int_a^b \int_a^b |k_n(s,t)|^2\,ds\,dt \le \left(\int_a^b \int_a^b |k(s,t)|^2\,ds\,dt\right)^n,$$

or

$$\|k_n\| \le \|k\|^n.$$

It is a more delicate matter to estimate the numerical magnitude of $k_n(s,t)$. Let us write

$$\begin{aligned} u(s) &=^0 \left(\int_a^b |k(s,t)|^2\,dt\right)^{1/2} \\ v(t) &=^0 \left(\int_a^b |k(s,t)|^2\,ds\right)^{1/2}. \end{aligned} \tag{3-4}$$

Now, for each t except on a set of measure 0, $k(s,t)$ as a function of s determines an element k_t of L^2. Likewise, $k_n(s,t)$ determines an element $k_t^{(n)}$. Formula (3-3) shows that

$$k_t^{(n)} = K^{n-1}k_t = Kz,$$

where $z = K^{n-2}k_t$. Thus

$$k_n(s,t) =^0 \int_a^b k(s,u)z(u)\,du,$$

$$|k_n(s,t)| \le^0 \left(\int_a^b |k(s,u)|^2\,du\right)^{1/2}\left(\int_a^b |z(u)|^2\,du\right)^{1/2},$$

$$|k_n(s,t)| \le^0 u(s)\|z\|.$$

But

$$\|z\| = \|K^{n-2}k_t\| \le \|K\|^{n-2}\|k_t\| = \|K\|^{n-2}v(t),$$

and so

$$|k_n(s,t)| \le^0 \|K\|^{n-2}u(s)v(t), \qquad n \ge 2. \tag{3-5}$$

This inequality shows that the series (2-8) converges and defines a function $h(s,t)$ almost everywhere on $E = [a,b]\times[a,b]$, provided that $\|K\| < 1$. Furthermore, the partial sums of (2-8) form a monotone sequence of positive

$\mathscr{L}^2$ functions on E, which are dominated by $u(s)v(t)$. Using Lebesgue's monotone convergence theorem, we see that h is an $\mathscr{L}^2$ kernel, that the series in (2-7) converges almost everywhere, and that we may reverse the order of summation and integration to obtain

$$x(s) = {}^0y(s) + \int_a^b h(s, t)y(t)\,dt$$

as the solution of the integral equation in $\mathscr{L}^2$ (all this when $\|K\| < 1$).

Observe that in this case the Banach space theory tells us merely that

$$x = y + Ky + K^2y + \cdots + K^ny + \cdots,$$

the series being convergent in L^2. The additional information that

$$x(s) = y(s) + \int_a^b k(s, t)y(t)\,dt + \cdots + \int_a^b k_n(s, t)y(t)\,dt + \cdots$$

is convergent almost everywhere, with its terms dominated as in (3-5) is more than was obtained directly from Theorem 1.4.

Approximation by Degenerate Kernels

In §2, in connection with (2-11), we presented a useful method of approximating certain integral operators by operators with degenerate kernels. Operators on L^2 with $\mathscr{L}^2$ kernels may also be approximated by operators with degenerate kernels. The proof of the following theorem describes how to construct these degenerate operators.

Theorem 3.1. *Let K be an integral operator on L^2 with an $\mathscr{L}^2$ kernel $k(s, t)$. Given $\varepsilon > 0$, there exists an integral operator M on L^2, with a degenerate kernel $m(s, t)$, such that $\|K - M\| < \varepsilon$.*

Proof. Let $\{u_n\}$ be a sequence of functions that determine a complete orthonormal set in L^2. For each t except on a set of measure 0, $k(s, t)$ determines an element k_t of L^2, and we define

$$a_n(t) = (k_t, u_n) = \int_a^b k_t(s)\overline{u_n(s)}\,ds = \int_a^b k(s, t)\overline{u_n(s)}\,ds. \tag{3-6}$$

The Fourier series of k_t with respect to $\{u_n\}$ converges in L^2 to k_t (cf. Theorem II.7.3), so

$$\int_a^b \Big|k(s, t) - \sum_1^n a_i(t)u_i(s)\Big|^2 ds = \Big\|k_t - \sum_1^n a_i(t)u_i\Big\|^2 = \Big\|\sum_{n+1}^{\infty} a_i(t)u_i\Big\|^2 = \sum_{n+1}^{\infty} |a_i(t)|^2 < \infty, \tag{3-7}$$

by Parseval's formula, for $n = 1, 2, \ldots$. Also,

$$\sum_1^\infty |a_n(t)|^2 = \|k_t\|^2 = \int_a^b |k(s, t)|^2 \, ds =^0 [v(t)]^2 \tag{3-8}$$

(see (3-4)). For a fixed n, $a_n(t)$ is defined almost everywhere on $[a, b]$. If we set $a_n(t) = 0$ on the rest of $[a, b]$, it is clear from the last integral in (3-6) that a_n is a measurable function of t. The partial sums of the series in (3-8) converge almost everywhere to the $\mathscr{L}^1$ function $[v(t)]^2$. By Lebesgue's monotone convergence theorem, each $|a_n|^2$ is in $\mathscr{L}^1(a, b)$ and we may integrate the series term by term. Thus (3-7) becomes

$$\int_a^b \int_a^b \left| k(s, t) - \sum_1^n a_i(t) u_i(s) \right|^2 ds \, dt = \sum_{n+1}^\infty \int_a^b |a_i(t)|^2 \, dt < \infty. \tag{3-9}$$

We now choose n so that the series on the right of (3-9) is less than ε^2. If we define the kernel of M by

$$m(s, t) = \sum_{i=1}^n u_i(s) a_i(t),$$

then (3-1) and (3-9) show that $\|K - M\| < \varepsilon$. Since each $a_i \in \mathscr{L}^2(a, b)$, m is an $\mathscr{L}^2$ kernel. $\square$

IV.4 DIFFERENTIAL EQUATIONS AND INTEGRAL EQUATIONS

In order to indicate some of the reasons that integral equations are of significant interest, we shall show how certain types of questions concerning differential equations can be recast as questions about integral equations. For simplicity we confine attention to second-order linear ordinary differential equations. The discussion could be extended to cover ordinary equations of higher order.

Consider the equation

$$y''(s) + a_1(s) y'(s) + a_2(s) y(s) = x(s), \tag{4-1}$$

where a_1 and a_2 are fixed continuous functions of s, defined when $a \leq s \leq b$. Let $X = C[a, b]$. The *initial-value problem* for (4-1) is the problem of finding $y \in X$, with y' and y'' also in X such that (4-1) is satisfied and such that $y(a)$ and $y'(a)$ have preassigned values, say, $y(a) = \alpha_0$, $y'(a) = \alpha_1$. The function x is to be an arbitrarily assigned member of X. The fact that this initial-value problem always has a solution (and a unique one), no matter how α_0, α_1, and x are chosen, is a consequence of standard existence theorems in the theory of differential equations. But we can also demonstrate the fact by using what we have learned in § 2 about Volterra integral equations.

We write $y''(s) = z(s)$, assuming for the moment that y is a solution of the initial-value problem. Then

$$\int_a^u z(t)\,dt = y'(u) - \alpha_1,$$

$$\int_a^s du \int_a^u z(t)\,dt = y(s) - \alpha_0 - \alpha_1(s-a). \tag{4-2}$$

The iterated integral in this last formula can be expressed in the form

$$\int_a^s z(t)\,dt \int_t^s du = \int_a^s (s-t)z(t)\,dt.$$

On replacing y, y', and y'' in (4-1) by their expression in terms of z, we obtain

$$\begin{aligned} z(s) + a_1(s)\left\{\alpha_1 + \int_a^s z(t)\,dt\right\} \\ + a_2(s)\left\{\alpha_0 + \alpha_1(s-a) + \int_a^s (s-t)z(t)\,dt\right\} = x(s). \end{aligned} \tag{4-3}$$

This equation is satisfied by z if y is a solution of the initial-value problem. Conversely, if z satisfies (4-3) and y is defined by (4-2), y is a solution of the initial-value problem.

If we define

$$\begin{aligned} k(s,t) &= -a_1(s) - a_2(s)(s-t), \\ w(s) &= x(s) - \alpha_0 a_2(s) - \alpha_1\{a_1(s) + a_2(s)(s-a)\}, \end{aligned}$$

the equation (4-3) can be written in the form

$$z(s) - \int_a^s k(s,t)z(t)\,dt = w(s).$$

This is a Volterra integral equation of the second kind. As we know from § 2, it has a unique solution $z \in X$ for each $w \in X$. *Therefore the initial value problem for* (4-1) *always has a unique solution.*

There is another important class of problems related to the differential equation (4-1), namely, the two-point problems. For a *two-point problem* we ask for a solution of the differential equation that satisfies a certain specified condition at $s = a$ and another condition at $s = b$. A common form of condition is one in which we specify the value at an end point of some predetermined linear combination of y and y'. We shall illustrate by discussing the problem with the conditions $y(a) = \alpha$, $y(b) = \beta$. We refer back to Example 3 of § I.3 for certain facts that we need. There it was shown (we change the notation slightly) that if $x \in C[a, b]$ and if

$$c(s,t)=\begin{cases}\dfrac{(b-s)(t-a)}{b-a} & \text{when } a \leqslant t \leqslant s\\[2ex] \dfrac{(s-a)(b-t)}{b-a} & \text{when } s \leqslant t \leqslant b,\end{cases}$$

then the function

$$z(s) = -\int_a^b c(s,t)x(t)\,dt$$

satisfies the conditions

$$z''(s) = x(s), \qquad z(a) = z(b) = 0.$$

Consequently, the function

$$y(s) = \alpha\frac{b-s}{b-a} + \beta\frac{s-a}{b-a} - \int_a^b c(s,t)x(t)\,dt \tag{4-4}$$

satisfies the conditions

$$y''(s) = x(s), \qquad y(a) = \alpha, \qquad y(b) = \beta.$$

In (4-4) let us replace $x(t)$ by

$$x(t) - a_1(t)y'(t) - a_2(t)y(t).$$

In this way we see that y will be a solution of our two-point problem if and only if it has a continuous derivative and satisfies the equation

$$y(s) = \alpha\frac{b-s}{b-a} + \beta\frac{s-a}{b-a} - \int_a^b c(s,t)\{x(t) - a_1(t)y'(t) - a_2(t)y(t)\}\,dt. \tag{4-5}$$

The discussion of this equation is much simplified if we assume that $a_1(t) \equiv 0$. For this special case let us write

$$k(s,t) = c(s,t)a_2(t),$$

$$w(s) = \alpha\frac{b-s}{b-a} + \beta\frac{s-a}{b-a} - \int_a^b c(s,t)x(t)\,dt.$$

Then (4-5) becomes

$$y(s) - \int_a^b k(s,t)y(t)\,dt = w(s).$$

This is a Fredholm integral equation of the second kind. It will certainly have a unique solution $y \in C[a, b]$ if $\|K\| < 1$, where K is the Fredholm-type operator with kernel k. (The condition $\|K\| < 1$ is sufficient; it may not be necessary.) We know from § 2 that $\|K\| \leq M(b-a)$, where M is the maximum value of $|k(s, t)|$. Now the maximum value of $(b-a)|c(s, t)|$ is readily found to be $(b-a)^2/4$. Let A be the maximum value of $|a_2(t)|$. Then certainly $\|K\| < 1$ if $(b-a)^2 A < 4$. We summarize the conclusions.

The two-point problem for the equation

$$y''(s) + a_2(s)y(s) = x(s)$$

with the end conditions $y(a) = \alpha$, $y(b) = \beta$, is always uniquely solvable, with arbitrarily assigned α, β, x, provided that

$$(b-a)^2 \max |a_2(s)| < 4.$$

(For what may happen if this inequality is not satisfied, see problem 20 in § V.7.)

We return to the case when $a_1(t) \not\equiv 0$. One way to handle this is to make a change of variable. If $a_1'(s)$ is continuous, we can define

$$q(s) = \exp\left(-\tfrac{1}{2}\int_a^s a_1(t)\,dt\right)$$

and set $y = uq$. Then, with u as the new unknown, we have the equation

$$u''(s) + \{a_2(s) - \tfrac{1}{2}a_1'(s) - \tfrac{1}{4}[a_1(s)]^2\}u(s) = \frac{x(s)}{q(s)}.$$

The new end conditions are

$$u(a) = \frac{\alpha}{q(a)} = \alpha, \qquad u(b) = \frac{\beta}{q(b)}.$$

Since the $u'(s)$ term does not appear in the differential equation, the previous method may be used to recast the problem as an integral equation problem.

If a_1 cannot be eliminated, so that the y' term is actually present in (4-5), we have what is called an *integro-differential* equation. It is a more complicated matter to discuss the possibility of solution of such an equation, and we shall leave the matter at this point.

In some later sections (§ VI.2 and § VI.5) we shall discuss another phase of the relation between differential and integral equations, with applications to eigenvalue problems of differential equations of Sturm–Liouville type.

IV.5 CLOSED LINEAR OPERATORS

For the purposes of applications to analysis it is essential to consider some linear operators that are not continuous. However, many of the most important discontinuous linear operators have a property that in some respects compensates for the absence of the property of continuity. This property is most naturally described in terms of the graph of a function.

Definition. Let X and Y be topological spaces. A function f with domain $\mathscr{D}$ in X and range in Y is said to be *closed* if its graph, $\{(x, f(x)) : x \in \mathscr{D}\}$, is a closed set in the product topology of $X \times Y$.

If f is injective, the graph of f^{-1} is the set $\{(f(x), x) : x \in \mathscr{D}\}$ in $Y \times X$. It follows readily that f^{-1} is closed if and only if f is closed.

When X and Y are metric spaces, the topology of $X \times Y$ can also be defined by a metric. (In particular, a sequence $\{(x_n, y_n)\}$ in $X \times Y$ converges to a point (x, y) if and only if $x_n \to x$ and $y_n \to y$.) In this case, the function f is closed if and only if the situation

$$x_n \in \mathscr{D}, \qquad x_n \to x, \qquad f(x_n) \to y$$

implies that

$$x \in \mathscr{D} \qquad \text{and} \qquad f(x) = y.$$

We shall frequently use this criterion for a closed function when studying linear operators on normed linear spaces.

Example 1. *A discontinuous but closed operator:* Let $X = C[0, 1]$, and let $X = Y$. Let $\mathscr{D}$ be the set of $x \in X$ such that the derivative $x'(s)$ is defined and continuous on $[0, 1]$; let T be the operator with domain $\mathscr{D}$ defined by $Tx = x'$. It is evident that T is linear. If $x_n(s) = s^n$, then $\|x_n\| = 1$, $x'_n(s) = ns^{n-1}$, and so $\|Tx_n\| = \|x'_n\| = n$, $n = 1, 2, \ldots$. Thus T is not continuous, because $\sup_{\|x\|=1} \|Tx\| = \infty$. But T is closed. Indeed, suppose $x_n \in \mathscr{D}$, $x_n \to x$, $Tx_n \to y$. Then $x'_n(s)$ converges uniformly to $y(s)$, and y is continuous. It follows by a standard convergence theorem that x is differentiable, with derivative equal to y. Therefore $x \in \mathscr{D}$, $Tx = y$, and T is closed.

Operators defined by means of ordinary or partial differentiation are frequently discontinuous, but in the applications it is often possible to arrange matters so that the differential operators to be dealt with are closed.

Example 2. Let $X = L^2(0, 1)$. Let T be an operator with domain and range in X, defined as follows: $\mathscr{D}(T)$ is the class of x in X such that x is absolutely continuous on $[0, 1]$ and its derivative $x' = dx/ds$ (which exists almost everywhere on $[0, 1]$) is again in X. This means that x may be represented in the form

$$x(s) = \alpha + \int_0^s u(t)\, dt,$$

where $u \in X$ and α is a scalar. We then define $Tx = x'$. [Here, as often elsewhere, we ignore the distinction between $L^2(0, 1)$ and $\mathscr{L}^2(0, 1)$. To say that x in $L^2(0, 1)$ is absolutely continuous means that x is represented by an absolutely continuous function in $\mathscr{L}^2(0, 1)$.]

We shall demonstrate that T is closed. Assume that $x_n \in \mathscr{D}$, $x, y \in X$, $x_n \to x$, $Tx_n \to y$. We can write

$$x_n(s) = x_n(0) + \int_0^s x'_n(t)\, dt. \tag{5-1}$$

Now

$$\left|\int_0^s x'_n(t)\, dt - \int_0^s y(t)\, dt\right| \leq \int_0^s |x'_n(t) - y(t)|\, dt \leq \int_0^1 |x'_n(t) - y(t)|\, dt$$
$$\leq \left(\int_0^1 |x'_n(t) - y(t)|^2\, dt\right)^{1/2} \left(\int_0^1 1^2\, dt\right)^{1/2}.$$

Therefore $\int_0^s x'_n(t)\, dt$ converges to $\int_0^s y(t)\, dt$; the convergence on $[0, 1]$ is uniform with respect to s. Next,

$$x_n(0) - x_m(0) = x_n(s) - x_m(s) - \int_0^s [x'_n(t) - x'_m(t)]\, dt,$$

and

$$|x_n(0) - x_m(0)| = \left(\int_0^1 |x_n(0) - x_m(0)|^2\, dt\right)^{1/2}$$
$$\leq \left(\int_0^1 |x_n(s) - x_m(s)|^2\, ds\right)^{1/2} + \left\{\int_0^1 \left|\int_0^s [x'_n(t) - x'_m(t)]\, dt\right|^2 ds\right\}^{1/2}.$$

It then follows that $\{x_n(0)\}$ is a Cauchy sequence, with some limit α. Returning to (5-1), we see that $x_n(s)$ converges uniformly to the limit $z(s)$, where

$$z(s) = \alpha + \int_0^s y(t)\, dt.$$

Also, $z'(s) \overset{0}{=} y(s)$ and $y \in L^2(0, 1)$. Hence $z \in \mathscr{D}(T)$ and $Tz = y$. The uniform convergence of $x_n(s)$ to $z(s)$ implies that

$$\int_0^1 |x_n(s) - z(s)|^2\, ds \to 0.$$

That is, $x_n \to z$. Since $x_n \to x$, we must have $x = z$. But then $x \in \mathscr{D}(T)$ and $Tx = y$. Hence T is closed.

We now present two elementary theorems to have on record for later use.

Theorem 5.1. *Let X be a topological space, and let Y be a Hausdorff space. Suppose f is a continuous function, with range in Y, whose domain $\mathscr{D}$ is a closed set in X. Then f is closed.*

Proof. Let W be the complement in $X \times Y$ of the graph of f, and take $(x_0, y_0) \in W$. If $x_0 \notin \mathscr{D}$, then $X \setminus \mathscr{D}$ is an open neighborhood of x_0 and $(X \setminus \mathscr{D}) \times Y$ is an open neighborhood of (x_0, y_0) that lies in W. If $x_0 \in \mathscr{D}$, then $f(x_0) \neq y_0$, and so there exist disjoint open sets V_1 and V_2 in Y such that $f(x_0) \in V_1$ and $y_0 \in V_2$. Since f is continuous, there is an open neighborhood U of x_0 in X such that $f(x) \in V_1$ if $x \in \mathscr{D} \cap U$. It follows that $U \times V_2$ is an open neighborhood of (x_0, y_0) that lies in W. We conclude that W is open and hence the graph of f is closed. □

Theorem 5.2. *Let X and Y be normed linear spaces, with Y complete. Let T be a linear operator with domain $\mathscr{D}$ in X and range in Y. Suppose that T is both closed and continuous. Then $\mathscr{D}$ is closed.*

Proof. Suppose $x \in \bar{\mathscr{D}}$. Then there exists a sequence $\{x_n\}$ from $\mathscr{D}$ with $x_n \to x$. The sequence $\{Tx_n\}$ is a Cauchy sequence, for $\|Tx_n - Tx_m\| \leq C\|x_n - x_m\|$, where C is the norm of T as an operator on the space $\mathscr{D}$. Hence $\{Tx_n\}$ has some limit $y \in Y$. But then $x \in \mathscr{D}$ and $Tx = y$, since T is closed. □

There are certain conditions under which it may be concluded that a closed linear operator is continuous. One of the most important theorems of this type is that if X and Y are Banach spaces, a linear operator whose domain is all of X and whose range is in Y is continuous if it is closed. This theorem is a consequence of the series of theorems that we shall now consider.

To simplify the notation used in the proofs of theorems about an operator T whose domain $\mathscr{D}$ is a subspace of X, we shall usually write $T(W)$ for the set $\{Tx : x \in \mathscr{D} \cap W\}$.

Lemma 5.3. *Let X and Y be topological linear spaces. Let T be a linear operator whose domain is a subspace of X and whose range is a set of the second category in Y. Then $\overline{T(U)}$ is a neighborhood of 0 in Y whenever U is a neighborhood of 0 in X.*

Proof. First, suppose that X is a normed linear space. If U is a neighborhood of 0 in X, there is an $\varepsilon > 0$ such that the ball $\|x\| \leq 2\varepsilon$ lies in U. Let W be the ball $\|x\| \leq \varepsilon$. Then nW is the ball $\|x\| \leq n\varepsilon$, and the union of the sets $T(nW)$ for $n = 1, 2, \ldots$ is the range of T. If $T(W)$ were nowhere dense, then each (homeomorphic) set $T(nW) = nT(W)$ would be nowhere dense, which would make the range of T a set of the first category, contrary to the hypothesis. Thus $\overline{T(W)}$ has a nonempty interior; that is, there exists some point $y_0 \in T(W)$ that is an interior point of $\overline{T(W)}$. If we let $V = \overline{T(W)} - y_0$, it follows that V is a (closed) neighborhood of 0. Now

$$T(W)-y_0 \subset T(W)-T(W)$$
$$\subset T(W-W) \subset T(U),$$

because of the way in which W and U are related. Since V is the closure of $T(W)-y_0$, we have $V \subset \overline{T(U)}$.

For the case in which X is an arbitrary topological linear space the only difference in the proof concerns the way in which W is chosen. We choose for W a balanced neighborhood of 0 such that $W+W \subset U$ (see Theorem II.9.2). The fact that W is balanced and absorbing assures that X is the union of the sets nW, $n=1,2,\ldots$, and also that $W-W=W+W \subset U$. The rest of the argument is unchanged. □

Lemma 5.4. *Let X and Y be normed linear spaces, and let X be complete. Let T be a closed linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T) \subset Y$. Let B_1 be the closed ball $\|x\| \leqslant 1$ in X. If $\overline{T(B_1)}$ is a neighborhood of 0 in Y, then so is $T(B_1)$.*

Proof. For $\alpha>0$ let B_α be the closed ball $\|x\| \leqslant \alpha$ in X, and let C_α be the closed ball $\|y\| \leqslant \alpha$ in Y. Choose $\alpha>0$ such that $C_\alpha \subset \overline{T(B_1)}$. It is easy to see that

$$C_{\alpha 2^{-n}} \subset \overline{T(B_{2^{-n}})}, \qquad n=0,1,\ldots \tag{5-2}$$

(since scalar multiplication is continuous). We shall prove that $C_{\alpha/2} \subset T(B_1)$. Suppose $y \in C_{\alpha/2}$. Then y is a limit of points of the form Tz_n where $\|z_n\| \leqslant 1$. Our objective is to choose the z_n in such a way that they converge to some $z \in B_1$. Then, since T is closed, it will follow that $Tz=y$; that is, $y \in T(B_1)$.

From (5-2) we have $y \in C_{\alpha/2} \subset \overline{T(B_{1/2})}$. Given any $\varepsilon>0$, there exists an $x_1 \in \mathscr{D}(T) \cap B_{1/2}$ such that $\|y-Tx_1\|<\varepsilon$. It suffices to use $\varepsilon=\alpha/4$, for this will make $y-Tx_1 \in C_{\alpha/4} \subset \overline{T(B_{1/4})}$, by (5-2). Now, if we take $\varepsilon=\alpha/8$ there exists $x_2 \in \mathscr{D}(T) \cap B_{1/4}$ such that $\|(y-Tx_1)-Tx_2\|<\frac{1}{8}\alpha$, and hence $y-Tx_1-Tx_2 \in C_{\alpha/8} \subset \overline{T(B_{1/8})}$. Continuing this procedure, we obtain a sequence $\{x_n\}$ in $\mathscr{D}(T)$ such that

$$\|x_n\| \leqslant 2^{-n}, \qquad \text{and} \qquad \left\|y-\sum_1^n Tx_k\right\| \leqslant \alpha 2^{-n-1}. \tag{5-3}$$

Let $z_n=\sum_{k=1}^n x_k$. Then (5-3) shows that $Tz_n \to y$. Furthermore, $\sum_{k=1}^\infty \|x_k\| \leqslant \sum_{k=1}^\infty 2^{-k}=1$, which implies that $\{z_n\}$ converges to some $z \in X$ with $\|z\| \leqslant 1$, since X is a Banach space. (Cf. Theorem II.4.3). Finally, T is a closed operator, so $z \in \mathscr{D}(T)$ and $Tz=y$. This proves that $C_{\alpha/2} \subset T(B_1)$. □

Theorem 5.5 (The Open Mapping Theorem). *Let X and Y be normed linear spaces, and let X be complete. Let T be a closed linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T)$ a set of the second category in Y. Then T is an open* *mapping; that is, if W is an open set in X, then $T(\mathscr{D}(T)\cap W)$ is open in Y. Also, $\mathscr{R}(T)=Y$.*

Proof. For each $\alpha>0$ let B_α be the closed ball $\|x\|\leq\alpha$ in X. By Lemma 5.3, $\overline{T(B_1)}$ is a neighborhood of 0 in Y, and hence $T(B_1)$ is also a neighborhood of 0, by Lemma 5.4. It follows that $T(B_\alpha)$ is a neighborhood of 0 for each $\alpha>0$, since $T(B_\alpha)=T(\alpha B_1)=\alpha T(B_1)$. Now suppose that W is open in X and $y\in T(W)$. Then $y=Tx$, for some $x\in\mathscr{D}(T)\cap W$. Let $\alpha>0$ be such that $x+B_\alpha\subset W$. Then $T(W)\supset T(x+B_\alpha)=y+T(B_\alpha)$. This shows that $T(W)$ contains a neighborhood of y (since $T(B_\alpha)$ is a neighborhood of 0). Thus $T(W)$ is open.

To prove that $\mathscr{R}(T)=Y$, let $\beta>0$ be such that $T(B_1)$ contains the closed ball $\|y\|\leq\beta$. If $y\in Y$ and $y\neq 0$, then $(\beta\|y\|^{-1})y\in T(B_1)$. Thus $(\beta\|y\|^{-1})y=Tw$ for some $w\in\mathscr{D}(T)\cap B_1$ and

$$y=T((\beta^{-1}\|y\|)w), \tag{5-4}$$

whence $y\in\mathscr{R}(T)$. □

Corollary 5.6. *Let X, Y, and T satisfy the hypotheses of Theorem 5.5. Then there exists some $m>0$ such that each $y\in Y$ is of the form $y=Tx$, for some $x\in\mathscr{D}(T)$ with $\|x\|\leq m\|y\|$. If T^{-1} exists, it is continuous.*

Proof. It follows from (5-4) that $m=\beta^{-1}$ works. If T^{-1} exists and if $y=Tx$, then $\|T^{-1}y\|=\|x\|\leq m\|y\|$. Hence T^{-1} is continuous and $\|T^{-1}\|\leq m$. □

The open mapping theorem is usually applied to a situation where $\mathscr{R}(T)=Y$ and Y is a *complete* normed linear space. In this case $\mathscr{R}(T)$ must be of the second category by Baire's category theorem. We shall use this fact in the proof of the closed graph theorem.

Theorem 5.7 (The Closed Graph Theorem). *Let X and Y be Banach spaces. Let T be a closed linear operator whose domain is all of X and whose range is in Y. Then T is continuous.*

Proof. Since X and Y are complete metric spaces, the product space $X\times Y$ is a Banach space when the norm is given by

$$\|(x,y)\|=\|x\|+\|y\|,\qquad (x,y)\in X\times Y.$$

The graph of T, $G(T)$, is a closed linear manifold in $X\times Y$ and can, therefore, be regarded as a Banach space by itself. We define a linear operator A, with $G(T)$ as its domain and with the Banach space X as its range, as follows:

$$A(x,Tx)=x.$$

Since $\|A(x,Tx)\|=\|x\|\leq\|(x,Tx)\|$, A is continuous. By Theorem 5.1, A is closed. Evidently A^{-1} exists and is defined by $A^{-1}x = (x, Tx)$. Then A^{-1} is continuous, by Corollary 5.6. The linear operator B defined by $B(x, Tx) = Tx$ is clearly continuous from $G(T)$ into Y. Consequently, $T = BA^{-1}$ is continuous from X into Y. □

The closed graph theorem may be used to prove the principle of uniform boundedness (Theorem 1.2). Suppose that X is a Banach space, Y is a normed linear space, and G is a subset of $L(X, Y)$ with the property that

$$\sup_{A \in G} \|Ax\| < \infty, \qquad \text{for each } x \in X.$$

Let us prove that $\sup\{\|A\| : A \in G\} < \infty$. First of all, we observe that it suffices to prove this with the added assumption that Y is complete. For, if Y is not complete and $\hat{Y}$ is its completion, we can just as well look upon G as a subset of $L(X, \hat{Y})$. This does not affect $\|Ax\|$ or $\|A\|$. Now, assuming that Y is complete, we consider the class $B(G, Y)$ of all bounded functions g with domain G and range in Y, with $\|g\| = \sup\{\|g(A)\| : A \in G\}$. This space is complete. For each $x \in X$, let Tx be the function on G defined by $Tx(A) = Ax$. Then T is a linear operator from X into $B(G, Y)$. Since $x_n \to x$ implies that $Ax_n \to Ax$ for each $A \in G$, it is a simple matter to verify that T is closed. By the closed graph theorem, T must be continuous; that is, $\|Tx\| \leq \|T\|\|x\|$ for all $x \in X$. Now

$$\|Tx\| = \sup_{A \in G} \|Tx(A)\| = \sup_{A \in G} \|Ax\|.$$

Thus, for each $A \in G$ and all $x \in X$, $\|Ax\| \leq \|T\|\|x\|$. This implies that $\|A\| \leq \|T\|$ for each $A \in G$, which concludes our proof.

It is perhaps surprising that a special case of the closed graph theorem can be deduced from the principle of uniform boundedness, with the help of the Hahn–Banach theorem. See problem 6 of § 8.

Occasionally, it is possible to prove a theorem in two different ways, one proof using the closed graph theorem and the other using the principle of uniform boundedness. Problems 3 to 6 of § III.9 have this property.

We now consider two concrete applications of Corollary 5.6 and the closed graph theorem.

Example 3. Let Y be the Banach space $C[a, b]$, and let X be a linear space consisting of those functions in $C[a, b]$ that have continuous first and second derivatives on $[a, b]$ and that satisfy some definitely specified linear end conditions (e.g., $x(a) = 0$ and $x'(a) = 0$, or $x(a) = x(b) = 0$). Let T be the linear differential operator with domain X and range in Y defined by $Tx = y$, where

$$a_0(s)x''(s) + a_1(s)x'(s) + a_2(s)x(s) = y(s). \tag{5-5}$$

The fixed coefficients a_0, a_1, a_2 are assumed to be in Y. We can make X into a Banach space by defining $\|x\|$ to be the greatest of the maximum values on $[a, b]$ of $|x(s)|$, $|x'(s)|$, $|x''(s)|$, respectively. The operator T is continuous, for if

$$M = \|a_0\| + \|a_1\| + \|a_2\|,$$

where $\|a_k\|$ is the norm of a_k as an element of Y, we see that $\|y\| = \|Tx\| \leq M\|x\|$. Also, T is closed by Theorem 5.1.

Now let us suppose that the differential equation (5-5) has a unique solution $x \in X$ (i.e., satisfying the specified end conditions) for each choice of $y \in Y$. That is, we assume $\mathscr{R}(T) = Y$ and that T^{-1} exists. It then follows by Corollary 5.6 that T^{-1} is continuous. The continuous dependence of x on y offers a certain assurance that "perturbation-theory" methods for approximating the solution of the differential equation are satisfactory. That is, a small perturbation of the function y will result in a small perturbation of the solution x and its first and second derivatives. The striking thing about this result is that it is obtained without any detailed knowledge of the behavior of the coefficients a_0, a_1, a_2 or of the nature of the end conditions. There is, of course, the strong assumption that the problem always has a unique solution.

The principle involved in the foregoing example can be applied to partial differential equations also. See Friedman [1].

Example 4. Suppose $1 \leq p \leq \infty$ and $1 \leq q \leq \infty$. Let $\{a_{ij}\}$ be an infinite matrix $(i, j = 1, 2, \ldots)$ of scalars with the property that for each $x = \{\xi_j\}$ in ℓ^p the series

$$\eta_i = \sum_{j=1}^{\infty} \alpha_{ij}\xi_j,$$

is convergent for each i, and the sequence $y = \{\eta_i\}$ belongs to ℓ^q. Then the operator A defined by $y = Ax$ is a continuous linear mapping of ℓ^p into ℓ^q.

To prove this we observe first of all that

$$x_i'(x) = \sum_{j=1}^{\infty} \alpha_{ij}\xi_j$$

defines a continuous linear functional x_i' on ℓ^p. If $1 < p < \infty$, this follows from problem 3 of § III.9, for

$$|x_i'(x)| \leq \left(\sum_{j=1}^{\infty} |\alpha_{ij}|^{p'}\right)^{1/p'} \left(\sum_{j=1}^{\infty} |\xi_j|^p\right)^{1/p},$$

so that x_i' is continuous, with

$$\|x_i'\| \leq \left(\sum_{j=1}^{\infty} |\alpha_{ij}|^{p'}\right)^{1/p'} < \infty.$$

(Actually, the last $\leqslant$ can be replaced by $=$, but we do not need this fact here.) If $p = 1$, we have

$$\|x_i'\| \leqslant \sup_j |\alpha_{ij}| < \infty;$$

if $p = \infty$, we have

$$\|x_i'\| \leqslant \sum_{j=1}^{\infty} |\alpha_{ij}| < \infty.$$

These latter facts result from problems 4 and 5 of § III.9. Now, to prove that A is continuous, it suffices to prove that it is closed, by Theorem 5.7. Suppose that $x_n \to x$ and $Ax_n \to y$, where $x_n, x \in \ell^p$ and $y \in \ell^q$. From $x_n \to x$ we conclude $x_i'(x_n) \to x_i'(x)$ as $n \to \infty$, since x_i' is continuous. Now, if $z = \{\zeta_i\}$ is any member of ℓ^q, ζ_i is a continuous function of z, as we see from the fact that $|\zeta_i| \leqslant \|z\|$. Since $Ax_n = \{x_i'(x_n)\}$ and $y = \{\eta_i\}$, we conclude from $Ax_n \to y$ that $x_i'(x_n) \to \eta_i$. But then $\eta_i = x_i'(x)$. This means that $y = Ax$. Thus A is closed, and the argument is complete.

Closed Operators with Closed Range

In many applications it is useful to know if an operator T has closed range. The case when T is injective is particularly important.

Theorem 5.8. *Let X and Y be Banach spaces, and let T be a closed linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T) \subset Y$. Suppose that T^{-1} exists. Then T^{-1} is continuous if and only if $\mathscr{R}(T)$ is closed in Y.*

Proof. The inverse T^{-1} is closed. If T^{-1} is also continuous, then its domain $\mathscr{R}(T)$ is closed, by Theorem 5.2. Conversely, if $\mathscr{R}(T)$ is closed, it is complete and hence is of the second category (by Baire's category theorem). Thus T^{-1} is continuous by the open mapping theorem (Corollary 5.6). □

When T is not injective, we may consider the operator $\hat{T}$ from $X/\mathscr{N}(T)$ into Y defined by

$$\hat{T}([x]) = Tx,$$

for all $[x] \in \mathscr{D}(\hat{T}) \equiv \{[x] : x \in \mathscr{D}(T)\}$. Clearly $\hat{T}$ is an injective operator with the same range as T. (Notice that $\mathscr{D}(\hat{T})$ is well defined because $\mathscr{D}(T) \supset \mathscr{N}(T)$.) When T is closed, $\mathscr{N}(T)$ is closed (problem 3), and hence $X/\mathscr{N}(T)$ is a normed linear space. Furthermore, $\hat{T}$ is closed (problem 4). We call $\hat{T}$ the *one-to-one operator induced by T.*

Now suppose that X and Y are Banach spaces and T is closed. Then $X/\mathscr{N}(T)$ is a Banach space. Applying Theorems 5.8 and II.1.2 to $\hat{T}$, we conclude that the subspace $\mathscr{R}(T)=\mathscr{R}(\hat{T})$ is closed if and only if there is a constant $m>0$ such that

$$m\|[x]\|\leqslant\|Tx\|, \qquad x\in\mathscr{D}(T);$$

that is,

$$\text{(5-6)} \qquad m\cdot\operatorname{dist}(x,\mathscr{N}(T))\leqslant\|Tx\|, \qquad x\in\mathscr{D}(T).$$

Definition. If T is closed, the *lower bound* (or *minimum modulus*) of T is the supremum $\gamma(T)$ of all $m\geqslant 0$ such that (5-6) holds. If $\gamma(T)$ is finite, it satisfies the inequality

$$\gamma(T)\cdot\operatorname{dist}(x,\mathscr{N}(T))\leqslant\|Tx\|, \qquad x\in\mathscr{D}(T).$$

The discussion above is summarized in the following theorem.

Theorem 5.9. *Let X and Y be Banach spaces, and let T be a closed linear operator with $\mathscr{D}(T)\subset X$ and $\mathscr{R}(T)\subset Y$. Then $\mathscr{R}(T)$ is closed if and only if $\gamma(T)>0$.*

The next theorem will be needed in §§ 12, 13, and V.10.

Theorem 5.10. *Let X, Y and T be as in Theorem 5.9. Suppose there exists a closed subspace M of Y such that $\mathscr{R}(T)\cap M=(0)$ and such that the subspace $\mathscr{R}(T)\oplus M$ is closed in Y. Then $\mathscr{R}(T)$ is closed.*

Proof. Let $\mathscr{D}(T)\times M$ be the domain of a linear operator T_1 from $X\times Y$ into Y defined by

$$T_1(x,m)=Tx+m, \qquad (x,m)\in\mathscr{D}(T)\times M.$$

Since T is closed, it is readily verified that T_1 is closed. Furthermore, $\mathscr{R}(T_1)$ is the closed subspace $\mathscr{R}(T)\oplus M$ in Y. Hence $\gamma(T_1)>0$. Now $\mathscr{N}(T_1)=\mathscr{N}(T)\times\{0\}$, because $\mathscr{R}(T)\cap M=(0)$. Thus, for $x\in\mathscr{D}(T)$,

$$\begin{aligned}\|Tx\|=\|T_1(x,0)\|&\geqslant\gamma(T_1)\cdot\operatorname{dist}((x,0),\mathscr{N}(T_1))\\&=\gamma(T_1)\cdot\operatorname{dist}(x,\mathscr{N}(T)).\end{aligned}$$

Hence $\gamma(T)\geqslant\gamma(T_1)>0$. By Theorem 5.9, $\mathscr{R}(T)$ is closed. □

PROBLEMS

1. Let T be the operator on $L^2(0,1)$ described in Example 2. Let T_1 be the restriction of T to $\mathscr{D}(T_1)\equiv\{x\in\mathscr{D}(T): x(0)=x(1)\}$, and let T_2 be the restriction of T to $\mathscr{D}(T_2)\equiv\{x\in\mathscr{D}(T): x(0)=x(1)=0\}$. Show that T_1 and T_2 are closed.

Prove that the operator A of Example 4, § I.5, is closed, with domain and range in $L^1(0, \infty)$. It suffices to deal with the case when $\lambda = 0$, since the addition of a scalar multiple of I to a closed operator leaves a closed operator.

Let X and Y be normed linear spaces, and let T be a closed linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T) \subset Y$. Then $\mathscr{N}(T)$ is closed. However, $\mathscr{R}(T)$ need not be closed, even when $T \in L(X, Y)$. [For an example, take $X = Y = \ell^1$, and let $T(\xi_1, \xi_2, \xi_3, \ldots) = (\xi_1, \frac{1}{2}\xi_2, \frac{1}{3}\xi_3, \ldots)$. Then $\overline{\mathscr{R}(T)} = X$, but the vector $(1, 2^{-2}, 3^{-2}, 4^{-2}, \ldots)$ is in $X \backslash \mathscr{R}(T)$.]

Let T be a closed linear operator. Then the one-to-one operator $\hat{T}$ induced by T is also closed. Also, $\hat{T}$ is continuous if and only if T is continuous, in which case, $\|\hat{T}\| = \|T\|$.

Use the closed graph theorem to prove the following form of the open mapping theorem: "Let X and Y be Banach spaces, and let T be a closed linear operator with domain in X and range equal to Y. Then T is an open mapping." [*Hint.* Consider the one-to-one operator induced by T.]

Prove the closed graph theorem from Theorem 5.8.

Let X be a linear space with two norms $\|x\|_1$, $\|x\|_2$, and suppose X is complete with respect to each norm. If there exists M such that $\|x\|_2 \leq M\|x\|_1$ for every x, then there is also an m such that $\|x\|_1 \leq m\|x\|_2$.

Prove the closed graph theorem by defining another norm on X, with $\|x\|_1 = \|x\| + \|Tx\|$. Then use problem 7.

Let X and Y be normed linear spaces, and let T be a linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T) \subset Y$. If there exists a closed linear extension of T, that is, a closed linear operator T_1 such that $\mathscr{D}(T_1) \supset \mathscr{D}(T)$ and $T_1 x = Tx$ for $x \in \mathscr{D}(T)$, then T is said to be *closable*. Show that T is closable if and only if for any $y \neq 0$ in Y, $(0, y)$ is not in the closure $\overline{G(T)}$ of the graph of T. In this case, $\overline{G(T)}$ is the graph of a closed linear extension of T.

Let X and Y be Banach spaces, and let T be a linear operator with $\mathscr{D}(T) = X$ and $\mathscr{R}(T) \subset Y$. Let F be a total set of linear functionals in Y', and suppose that $y' \circ T$ is continuous for each $y' \in F$. Then T is continuous.

Prove the following generalization of the closed graph theorem: Let X and Y be Banach spaces, and let T be a closed linear operator with $\mathscr{D}(T) \subset X$ and $\mathscr{R}(T) \subset Y$. Suppose there exists a closed subspace M of X such that $\mathscr{D}(T) \cap M = (0)$ and $\mathscr{D}(T) \oplus M$ is closed. Then T is continuous.

Suppose that S and T are closed linear operators from Banach spaces X and Y, respectively, into a Banach space Z, such that $Z = \mathscr{R}(S) \oplus \mathscr{R}(T)$. Then $\mathscr{R}(S)$ and $\mathscr{R}(T)$ are both closed.

Let X and Y be Banach spaces, and let T be a closed linear operator from X into Y such that $\mathscr{R}(T)$ is closed. If M is a linear manifold in X such that $M + \mathscr{N}(T)$ is closed, then $T(M)$ is closed. In particular, if M is closed and $\mathscr{N}(T)$ is finite dimensional, then $T(M)$ is closed. [*Suggestion.* Consider the operator $\hat{T}$ defined after the proof of Theorem 5.8.]

Prove that $L^q(0, 1)$ is of the first category in $L^p(0, 1)$ when $1 \leq p < q$.

Suggestion. Show that the inclusion mapping of $L^q(0, 1)$ into $L^p(0, 1)$ is continuous. See Taylor [5, page 278].

15. For each $x \in L^1(0, 2\pi)$ let

$$\xi_n = \frac{1}{2\pi}\int_0^{2\pi} x(t)\, e^{-int}\, dt, \qquad n = 0, \pm 1, \pm 2, \ldots.$$

Suppose X is some closed subspace of $L^1(0, 2\pi)$ such that $\Sigma_{-\infty}^{\infty} |\xi_n| < \infty$ for each $x \in X$. Show that there is some constant M such that

$$\sum_{-\infty}^{\infty} |\xi_n| \leqslant M \int_0^{2\pi} |x(t)|\, dt$$

for each $x \in X$.

16. Suppose $x \in L^1(0, 2\pi)$ and

$$y(t) = \frac{1}{2\pi}\int_0^{2\pi} \frac{x(\theta)}{1 - te^{-i\theta}}\, d\theta,$$

where t is complex and $|t| < 1$. The two following propositions are true (A. E. Taylor [4, pages 41 and 45]).

a. If $1 < p < \infty$ and $x \in L^p(0, 2\pi)$, then $y \in H^p$ (see Example 8 of § II.2 for the definition of H^p).

b. There exists a constant M_p, depending only on p, such that

$$\left(\frac{1}{2\pi}\int_0^{2\pi} |y(re^{i\theta})|^p\, d\theta\right)^{1/p} \leqslant M_p\left(\int_0^{2\pi} |x(\theta)|^p\, d\theta\right)^{1/p}$$

for all $x \in L^p(0, 2\pi)$ and for all r, $0 \leqslant r \leqslant 1$.

Show that (b) is a consequence of (a) by means of the closed graph theorem.

IV.6 SOME REPRESENTATIONS OF BOUNDED LINEAR OPERATORS

In this section we discuss the representation of bounded linear operators on X into Y, especially for the case in which X and Y are chosen in various ways from the spaces ℓ^p, (c) and (c_0). The discussion is based on a theorem of very general scope about the representation of linear operators.

For the purpose of our first general theorem we make the following assumptions: X is an arbitrary Banach space; Z is an arbitrary normed linear space; S is an arbitrary nonempty set of elements of unspecified nature; Y is a Banach space that is a subspace of the linear space of all functions defined on S, with values in Z. If $s \in S$ and $y \in Y$, we call $y(s)$ the component of y at s, and we assume that $y(s)$ is a continuous function of y. A possible choice for Y is ℓ^p $(1 \leqslant p \leqslant \infty)$, with S the set of positive integers and Z the space of scalars. Again, the space $B(S)$ of bounded scalar-valued functions on S, for an arbitrary S, is a possible choice for Y.

In what follows we refer to the space $L(X, Z)$. This is X' if Z happens to be the space of scalars; it is this choice of Z that we shall use mainly in the applications of these general considerations.

Theorem 6.1. *Assume X, Y, Z, S as in the foregoing remarks. Then, to each $A \in L(X, Y)$ there corresponds a function on S to $L(X, Z)$, whose value at s we denote by $a(s)$ such that if $y = Ax$, then $y(s) = a(s)x$ for each s. Conversely, if $a(\cdot)$ is any function with domain S and range in $L(X, Z)$ such that for each $x \in X$ the formula $y(s) = a(s)x$ defines an element of Y, this dependence of y on x defines an element A of $L(X, Y)$.*

Proof. If $A \in L(X, Y)$ is given and $y = Ax$, we define $a(s)$ by $a(s)x = y(s)$. Then $a(s) \in L(X, Z)$ for each s. The linearity of $a(s)$ is evident, and the continuity of $a(s)x$ with respect to x follows from the fact that $y(s)$ is a continuous function of y. For the converse part of the proof, the linearity of A is evident, and its continuity will follow from Theorem 5.7 if we prove that A is closed. Suppose that $x_n \to x$. Then $a(s)x_n \to a(s)x$, because $a(s) \in L(X, Z)$. Suppose also that $Ax_n \to y$. Then $(Ax_n)(s) \to y(s)$ for each s, by the continuous dependence of components on vectors. But $(Ax_n)(s) = a(s)x_n$. Hence $a(s)x = y(s)$, so that $y = Ax$; that is, A is closed. □

A situation of some special interest is that in which the space Y is of such a nature that the norm of an element in Y is given by

$$\|y\| = \sup_{s \in S} \|y(s)\|_Z. \tag{6-1}$$

Here $\|y(s)\|_Z$ denotes the norm of $y(s)$ as an element of Z, s being fixed. Among the examples of spaces Y of this type, all with Z the space of scalars, we cite ℓ^∞, (c), (c_0), $B(S)$ and, if S is a topological space, the subspace $C(S)$ of $B(S)$.

Theorem 6.2. *With the hypotheses as in Theorem* 6.1, *suppose the norm in Y is given by* (6-1). *Then $a(\cdot)$ is a bounded function on S to $L(X, Z)$, and*

$$\|A\| = \sup_{s \in S} \|a(s)\|. \tag{6-2}$$

If Z is complete and if Y consists of all bounded functions on S to Z, any bounded function $a(\cdot)$ on S to $L(X, Z)$ determines an operator belonging to $L(X, Y)$.

The proof is left to the reader. The assumption in the last part, that Z is complete, is made to ensure that Y is complete, this being needed in Theorem 6.1.

Let us denote by $B(S, Z)$ the Banach space of all bounded functions on S to the Banach space Z, with norm as in (6-1). Using this notation for arbitrary Z, we see that Theorem 6.2 implies that $L(X, B(S, Z))$ is congruent to $B(S, L(X, Z))$ in a natural way. If Y is a closed subspace of $B(S, Z)$, the theorem states that $L(X, Y)$ is congruent to a subspace of $B(S, L(X, Z))$, but the theorem does not describe this subspace. A certain description of the subspace may be inferred from Theorem 6.1.

We can now give representation theorems for elements of $L(X, Y)$, where X is one of the spaces (c_0), (c), or ℓ^p $(1 \leq p < \infty)$ and Y is one of the spaces ℓ^∞, (c), or (c_0). For this purpose we need Theorem III.5.2 and the results stated in problems 1 and 2 of § III.5. We take Z as the space of scalars and S as the set of positive integers.

Theorem 6.3. *Each bounded linear operator A on (c_0) into ℓ^∞, (c), or (c_0) determines and is determined by an infinite matrix of scalars $\alpha_{ij}(i, j = 1, 2, \ldots)$, $y = Ax$ being expressed by the equations*

$$\eta_i = \sum_{j=1}^{\infty} \alpha_{ij}\xi_j, \qquad i = 1, 2, \ldots. \tag{6-3}$$

The norm of A is given by

$$\|A\| = \sup_i \sum_{j=1}^{\infty} |\alpha_{ij}|. \tag{6-4}$$

For the case of mapping into ℓ^∞ the only restriction on the matrix (α_{ij}) is that the expression defining $\|A\|$ be finite. For the case of mapping into (c) the only additional requirement is that the limits

$$\alpha_j = \lim_{i\to\infty} \alpha_{ij}, \qquad j = 1, 2, \ldots \tag{6-5}$$

must exist. Finally, the mapping is into (c_0) if and only if, in addition, $\alpha_1 = \alpha_2 = \cdots = 0$.

Proof. For the case of mapping into ℓ^∞, the theorem results immediately from problem 2, § III.5 and Theorems 6.1, 6.2. For the case of mapping into (c) the further requirement is that $\lim_{i\to\infty} \eta_i$ exist for each choice of x in (6-3). An obvious special choice of x shows the necessity for the existence of the limits in (6-5). On the other hand, if these limits exist, it is easy to prove that

$$\lim_{i\to\infty} \eta_i = \sum_{j=1}^{\infty} \alpha_j\xi_j. \tag{6-6}$$

For, we know that $\Sigma_{j=1}^{n} |\alpha_{ij}| \leq \|A\|$ for all i and n. Letting first i and then n become infinite, we see that $\Sigma_{j=1}^{\infty} |\alpha_j| \leq \|A\|$. Then, for each n,

$$\left|\eta_i - \sum_{j=1}^{\infty} \alpha_j \xi_j\right| = \left|\sum_{j=1}^{\infty} (\alpha_{ij} - \alpha_j)\xi_j\right| \leq \sum_{j=1}^{n} |\alpha_{ij} - \alpha_j|\,|\xi_j| + 2\|A\| \sup_{j>n} |\xi_j|.$$

From this it follows that (6-6) is true, so that $y \in (c)$ if $x \in (c_0)$. We leave to the reader the proof of the last assertion in the theorem. □

The next theorem concerns operators $A \in L((c), Y)$, where Y is either ℓ^∞, (c), or (c_0). It is slightly different from the foregoing theorem, because of the intervention of the limit $\xi_0 = \lim_{j\to\infty} \xi_j$ when $x = \{\xi_j\} \in (c)$.

Theorem 6.4. *Each bounded linear operator A on (c) into ℓ^∞, (c), or (c_0) determines and is determined by a matrix of scalars α_{ij}, $i = 1, 2, \ldots$, $j = 0, 1, 2, \ldots$, $y = Ax$ being expressed by the equations*

$$\eta_i = \sum_{j=0}^{\infty} \alpha_{ij}\xi_j, \qquad i = 1, 2, \ldots. \tag{6-7}$$

The norm of A is given by

$$\|A\| = \sup_i \sum_{j=0}^{\infty} |\alpha_{ij}|. \tag{6-8}$$

For $A \in L((c), \ell^\infty)$ the sole condition on the matrix (α_{ij}) is that the expression defining $\|A\|$ be finite. For $A \in L((c), (c))$ there is the additional requirement that the limit

$$\alpha = \lim_{i\to\infty} \sum_{j=0}^{\infty} \alpha_{ij} \tag{6-9}$$

exist and that the limits

$$\alpha_j = \lim_{i\to\infty} \alpha_{ij} \tag{6-10}$$

exist if $j = 1, 2, \ldots$ (no requirement when $j = 0$). Finally, the range of A is in (c_0) if and only if, in addition, $\alpha = 0$ and $\alpha_1 = \alpha_2 = \cdots = 0$.

Proof is left to the reader, who should refer to problem 1, § III.5. We note, incidentally, that, if $A \in L((c), (c))$ and $\lim_{i\to\infty} \eta_i = \eta_0$, then

$$\eta_0 = \xi_0\alpha + \sum_{j=1}^{\infty} (\xi_j - \xi_0)\alpha_j. \tag{6-11}$$

For, from problem 1, § III.5 we have

$$Ax = \xi_0 A u_0 + \sum_{j=1}^{\infty} (\xi_j - \xi_0) A u_j.$$

If u_0' is the continuous linear functional defined by $u_0'(x) = \xi_0$, we have

$$u_0'(Ax) = \xi_0 u_0'(Au_0) + \sum_{j=1}^{\infty} (\xi_j - \xi_0) u_0'(Au_j),$$

and a little checking shows that this result is equivalent to (6-11). We see from (6-11) that $\eta_0 = \xi_0$ for every choice of x if and only if $\alpha = 1$ and $\alpha_1 = \alpha_2 = \cdots = 0$.

PROBLEMS

1. Each bounded linear operator A on ℓ^p $(1 \le p < \infty)$ into ℓ^∞, (c), or (c_0) is representable by an infinite matrix (α_{ij}) of scalars, where $y = Ax$ is expressed by equations (6-3). The norm of A is

$$\|A\| = \begin{cases} \sup\limits_i \left(\sum\limits_{j=1}^{\infty} |\alpha_{ij}|^{p'} \right)^{1/p'} & \text{if } 1 < p < \infty \\ \sup\limits_i \sup\limits_j |\alpha_{ij}| & \text{if } p = 1. \end{cases}$$

   Except for the difference in the expression for $\|A\|$, the rést of the assertion here is just like that in Theorem 6.3.

2. A matrix (α_{ij}), $i, j = 1, 2, \ldots$, defines a bounded linear operator A on ℓ^∞ into ℓ^∞, by means of equations (6-3) if the following expression, which then defines the norm of A, is finite:

$$\|A\| = \sup_i \sum_{j=1}^{\infty} |\alpha_{ij}|.$$

   This is not, however, the most general form for an element of $L(\ell^\infty, \ell^\infty)$, for elements of $(\ell^\infty)'$ are not all expressible by an infinite series.

3. If X is any Banach space, the general form of $A \in L(X, \ell^q)$, where $1 \le q < \infty$, is $y = Ax$, where $y = \{\eta_i\}$ and $\eta_i = x_i'(x)$, $\{x_i'\}$ being a sequence of elements of X' such that $\{x_i'(x)\} \in \ell^q$ for each x.

4. Each bounded linear operator A on ℓ^1 into ℓ^q $(1 \le q < \infty)$ is representable as in (6-3) by a matrix (α_{ij}). The only condition on the matrix is that the following expression, defining $\|A\|$, be finite.

$$\|A\| = \sup_j \left(\sum_{i=1}^{\infty} |\alpha_{ij}|^q \right)^{1/q}.$$

   Conversely, any matrix satisfying this condition determines an A in $L(\ell^1, \ell^q)$.

IV.7 THE M. RIESZ CONVEXITY THEOREM

Consider a matrix (α_{ij}) of complex numbers, not all 0, with m rows and n columns. Let $\xi_1, \ldots, \xi_n$ and $\zeta_1, \ldots, \zeta_m$ be complex numbers, and let $x = (\xi_1, \ldots, \xi_n)$, $z = (\zeta_1, \ldots, \zeta_m)$. If $\mu \geqslant 0$, define

$$M_\mu(x) = \begin{cases} \left(\sum_{j=1}^{n} |\xi_j|^{1/\mu} \right)^{\mu} & \text{if } \mu > 0 \\ \sup_j |\xi_j| & \text{if } \mu = 0. \end{cases}$$

If $\nu \geqslant 0$, define $N_\nu(z)$ in a similar manner. Let

$$M(\mu, \nu) = \sup \left| \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ij} \zeta_i \xi_j \right|, \tag{7-1}$$

the supremum being taken for all x and z such that $M_\mu(x) \leqslant 1$ and $N_\nu(z) \leqslant 1$. Then $\log M(\mu, \nu)$ *is a convex function of the point* (μ, ν) *in the part of the* (μ, ν)*-plane defined by the inequalities* $\mu \geqslant 0$, $\nu \geqslant 0$. This means that if (μ_1, ν_1) and (μ_2, ν_2) are admissible points,

$$\log M[(1-t)\mu_1 + t\mu_2, (1-t)\nu_1 + t\nu_2]$$

is a convex function of t when $0 \leqslant t \leqslant 1$.

We shall refer to this result as the M. Riesz convexity theorem. M. Riesz's proof of this theorem was given in his paper [1], subject to the more restrictive conditions $0 \leqslant \mu \leqslant 1$, $0 \leqslant \nu \leqslant 1$, $\mu + \nu \geqslant 1$. See also Hardy, Littlewood and Polya [1, pages 214–219]. In the more general form the result is due to Thorin [1]. If the matrix elements α_{ij} are real and if the vectors x, z are assumed to have real components, the convexity property of $\log M(\mu, \nu)$ obtains when $0 \leqslant \mu$, $0 \leqslant \nu$, and $\mu + \nu \geqslant 1$. This was shown partly by Riesz and partly by Thorin.

We shall not prove the Riesz convexity theorem. Our concern is to discuss its application to continuous linear mappings of ℓ^p into ℓ^q. For a more far-reaching discussion of the development of such ideas, with various important applications to integrals, Fourier coefficients, and other topics, see Dunford and Schwartz [1, §§ 10, 11 in Chapter VI] and Zygmund [1, Chapter XI, especially pages 93–105].

Suppose $1 \leqslant p \leqslant \infty$, $1 \leqslant q \leqslant \infty$, and consider the m by n matrix (α_{ij}) as representing a linear operator A mapping $\ell^p(n)$ into $\ell^q(m)$, with $y = Ax$ meaning $x = (\xi_1, \ldots, \xi_n)$, $y = (\eta_1, \ldots, \eta_m)$, and

$$\eta_i = \sum_{j=1}^{n} \alpha_{ij} \xi_j.$$

With the aid of Theorem III.3.2 we see that $\|A\| = \sup |y'(Ax)|$, the supremum being taken for all $x \in \ell^p(n)$ with $\|x\| \leqslant 1$ and all linear functionals y' on $\ell^q(m)$ with $\|y'\| \leqslant 1$. Since $y'(y)$ has a representation

$$y'(y) = \sum_{i=1}^{m} \zeta_i \eta_i,$$

we see that

$$\|A\| = \sup \left| \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_{ij} \zeta_i \xi_j \right|, \tag{7-2}$$

the supremum being taken for all x such that $\|x\| \leqslant 1$ and all $(\zeta_1, \ldots, \zeta_m)$ determining a functional y' with $\|y'\| \leqslant 1$. Now the normed conjugate of $\ell^q(m)$ is congruent to $\ell^{q'}(m)$. Hence, comparing (7-1) and (7-2), we see that

$$\|A\| = M(1/p, 1/q'),$$

where $1/p$ is understood to be 0 if $p = \infty$, and likewise for $1/q'$. To emphasize the dependence of $\|A\|$ on p and q we shall write it as $\|A\|_{p,q}$.

It is clear that when $\nu \leqslant 1$, convexity as a function of (μ, ν) is the same as convexity as a function of $(\mu, 1-\nu)$. Hence, by Riesz's theorem, $\log \|A\|_{p,q}$ is convex as a function of $(1/p, 1/q)$. In the complex case this is true for $p \geqslant 1$, $q \geqslant 1$. In the real case there is the additional restriction $(1/p)+(1/q') \geqslant 1$, which is the same as $q \geqslant p$. We summarize.

Theorem 7.1. *If* $A \in L(\ell^p(n), \ell^q(m))$, *then* $\log \|A\|_{p,q}$ *is a convex function of* $(1/p, 1/q)$ *for all* $p \geqslant 1$, $q \geqslant 1$ *in the complex case and for* $q \geqslant p \geqslant 1$ *in the real case.*

In order to display more directly the implication of this theorem, suppose $\psi(t)$ is a positive function such that $\log \psi(t)$ is convex when $0 \leqslant t \leqslant 1$. In particular, when $0 < t < 1$, the point $(t, \log \psi(t))$ on the graph of $\log \psi(t)$ is on or below the straight line joining the points $(0, \log \psi(0))$ and $(1, \log \psi(1))$. When the condition for this is worked out, it is found to be

$$\psi(t) \leqslant [\psi(0)]^{1-t}[\psi(1)]^t. \tag{7-3}$$

Suppose now that we are given pairs (p_1, q_1), (p_2, q_2) that are admissible under the conditions stated in Theorem 7.1. Let p and q be defined in terms of t by

$$\frac{1}{p} = \frac{1-t}{p_1} + \frac{t}{p_2}, \qquad \frac{1}{q} = \frac{1-t}{q_1} + \frac{t}{q_2}, \qquad 0 < t < 1. \tag{7-4}$$

Let $\psi(t) = \|A\|_{p,q}$. Then (7-3) becomes

$$\|A\|_{p,q} \leqslant (\|A\|_{p_1,q_1})^{1-t}(\|A\|_{p_2,q_2})^t. \tag{7-5}$$

As t varies from 0 to 1, p varies from p_1 to p_2; likewise for q. Formula (7-5) expresses the content to Theorem 7.1. For example, if $p_1=1$, $q_1=2$, $p_2=2$, $q_2=\infty$, we find that

$$p=\frac{2}{2-t}, \qquad q=\frac{2}{1-t}=\frac{2p}{2-p}.$$

In this case (7-5) becomes

$$\|A\|_{p,2p/(2-p)} \leqslant (\|A\|_{1,2})^{(2-p)/p}(\|A\|_{2,\infty})^{2(p-1)/p}.$$

Here p may vary from 1 to 2.

Now let us consider an infinite matrix (α_{ij}), $i, j=1, 2, \ldots$. Let $A_{n,m}$ be the mapping of $\ell^p(n)$ into $\ell^q(m)$ defined by considering α_{ij} for $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$. It is easily proved that the infinite matrix defines a continuous linear mapping A of ℓ^p into ℓ^q if and only if $\|A_{n,m}\|_{p,q}$ is bounded as m and n vary and that, when such is the case,

$$\|A\|_{p,q}=\sup_{m,n}\|A_{n,m}\|_{p,q}. \tag{7-6}$$

We write $\|A\|_{p,q}$ rather than $\|A\|$, because we now wish to consider the possibility that A may be regarded as belonging to $L(\ell^p, \ell^q)$ for various values of p and q.

Theorem 7.2. *Suppose that (α_{ij}) is an infinite matrix of scalars. Let (p_1, q_1) and (p_2, q_2) be admissible pairs according to the conditions of Theorem 7.1. Suppose the matrix defines an A belonging both to $L(\ell^{p_1}, \ell^{q_1})$ and to $L(\ell^{p_2}, \ell^{q_2})$. Then A belongs to $L(\ell^p, \ell^q)$ for each (p, q) given by (7-4), and the norms in these various cases satisfy the inequality (7-5).*

The proof is immediate, by applying (7-5) to $A_{n,m}$ and taking note of the assertion made in connection with (7-6).

PROBLEMS

1. If the infinite matrix (α_{ij}) defines an operator $A \in L(\ell^p)$, denote its norm by $\|A\|_p$. Show that if the operator defined by the matrix belongs to $L(\ell^1)$ and $L(\ell^\infty)$, it belongs to $L(\ell^p)$ for $1<p<\infty$, and $\|A\|_p \leqslant (\|A\|_1)^{1/p}(\|A\|_\infty)^{1-(1/p)}$.
2. Carry out the proof of (7-6).

IV.8 CONJUGATES OF LINEAR OPERATORS

The notion of the conjugate of a linear operator is closely related to the notion of the transpose of a linear operator, as defined in § I.12. We first consider the case of a bounded operator $A \in L(X, Y)$, where X and Y are normed linear spaces.

If $y' \in Y'$, the linear functional $x \mapsto y'(Ax)$ is the composition $y' \circ A$ of two continuous linear mappings and hence is an element of X'. We denote this functional by $A'y'$ to indicate its dependence on y' and A. Now the correspondence $y' \mapsto A'y'$ is itself linear and so defines a linear operator A' that maps Y' into X'. The operator A' is called the *conjugate* of A; it is defined by the formula

$$\langle x, A'y' \rangle = \langle Ax, y' \rangle, \qquad x \in X, \; y' \in Y'. \tag{8-1}$$

We shall see in Theorem 8.2 that A' is continuous, so that $A' \in L(Y', X')$.

If $Y' = Y^f$ (which is the case if Y is finite dimensional), A' is the same as the transpose A^T. However, Y' is a proper subspace of Y^f in general, so A^T is an extension of A'; that is, A' is the restriction of A^T to the space Y'.

Other notations and terminologies have been used for A'. It has been denoted by $\bar{A}$ or A^*; it has been called the adjoint of A and also the dual of A. In this book the notation A^* and the name "adjoint" are reserved for use in connection with linear operators in Hilbert space. In that situation A^* is closely related to A', but is not the same as A'. We shall consider adjoints in § 11.

We sometimes need some formal algebraic rules for handling conjugate operators. If $A, B \in L(X, Y)$, then

$$(A + B)' = A' + B' \qquad \text{and} \qquad (\alpha A)' = \alpha A'. \tag{8-2}$$

If $B \in L(X, Y)$ and $A \in L(Y, Z)$, then

$$(AB)' = B'A'. \tag{8-3}$$

If $A \in L(X, Y)$ and if A^{-1} exists and belongs to $L(Y, X)$, then $(A')^{-1}$ exists, belongs to $L(X', Y')$, and is the same as $(A^{-1})'$. Verification of these assertions is left to the reader.

If we are studying an operator A, a knowledge of certain facts about A' can be helpful to us. In general, A' is more amenable to investigation than A^T. In studying A' we usually need representation theorems for elements of Y' and X' (as in § III.5 for example). We also need representation theorems for bounded linear operators. For instance, suppose $X = Y = \ell^p$, where $1 \leq p < \infty$. We know from Theorem III.5.2 that X' is congruent in a natural way to $\ell^{p'}$. It turns out that, if an operator A in $L(\ell^p)$ is represented by an infinite matrix, then A', which we identify with an element of $L(\ell^{p'})$, is represented by the transposed matrix. For a somewhat more general result of this type, see problem 7.

The Conjugate of an Integral Operator

Let $X = L^p(a, b)$, $Y = L^q(a, b)$, where $1 \leq p, q < \infty$ and $[a, b]$ is a finite interval. Suppose that k is a bounded measurable function on $[a, b] \times [a, b]$. This "kernel" defines a linear operator K from X into Y: if $x \in X$, let $y = Kx$, where

$$y(s) = \int_a^b k(s, t)x(t)\, dt.$$

By Hölder's inequality,

$$\tag{8-4} \begin{aligned} |y(s)| &\leq \int_a^b |k(s, t)|\, |x(t)|\, dt \\ &\leq \left(\int_a^b |k(s, t)|^{p'}\, dt\right)^{1/p'} \left(\int_a^b |x(t)|^p\, dt\right)^{1/p} \\ &\leq M(b-a)^{1/p'}\|x\|_p, \end{aligned}$$

where M is a bound for $|k(s, t)|$ on $[a, b] \times [a, b]$ and where $\|x\|_p$ is the norm of x in $L^p(a, b)$. It follows from (8-4) that $\|y\|_q \leq M(b-a)^{1/p'+1/q}\|x\|_p$, which shows that $K \in L(X, Y)$.

We shall now show that K' may be represented as an integral operator when X' and Y' are identified with $L^{p'}(a, b)$ and $L^{q'}(a, b)$, respectively. To be precise, consider the composite mapping

$$L^{q'}(a, b) \to Y' \xrightarrow{K'} X' \to L^{p'}(a, b),$$

where the first and third maps are the identifications given by Theorem III.5.3. This composite mapping is again denoted by K', and we shall show that it may be represented as an integral operator.

Given $f \in L^{q'}(a, b)$, let y' be the element of Y' corresponding to f. Then, for $x \in X$,

$$\tag{8-5} \langle x, K'y' \rangle = \langle Kx, y' \rangle = \int_a^b \left[\int_a^b k(s, t)x(t)\, dt\right] f(s)\, ds.$$

By (8-4), the (measurable) function

$$g(s) = \int_a^b |k(s, t)|\, |x(t)|\, dt$$

is bounded. When we use Hölder's inequality, a computation similar to that in (8-4) shows that

$$\int_a^b \left[\int_a^b |k(s, t)|\, |x(t)|\, |f(s)|\, dt\right] ds = \int_a^b g(s)|f(s)|\, ds < \infty.$$

Since this iterated integral is finite and since $k(s, t)x(t)f(s)$ is measurable on

$[a, b]\times[a, b]$, we may interchange the order of integration in (8-5). This fact, known as Tonelli's theorem, is deducible from Fubini's theorem. See Dunford and Schwartz [1, page 194]. Thus

$$\langle x, K'y'\rangle = \int_a^b x(t)\left[\int_a^b k(s, t)f(s)\,ds\right]dt \tag{8-6}$$
$$= \int_a^b x(s)\left[\int_a^b k(t, s)f(t)\,dt\right]ds.$$

Now $\int_a^b k(t,\cdot)f(t)\,dt$ is an element of $L^{p'}(a, b)$, by Hölder's inequality. Since (8-6) is true for all $x\in X$, it follows from Theorem III.5.3 that $\int_a^b k(t,\cdot)f(t)\,dt$ corresponds to the linear functional $K'y'\in X'$. Identifying f with y', we may write $K'f$ instead of $K'y'$ and represent K' as an element of $L(L^{q'}, L^{p'})$, defined by

$$(K'f)(s) = \int_a^b k(t, s)f(t)\,dt.$$

Conjugates of Densely Defined Linear Operators

Throughout the rest of this section we shall assume that T is a linear operator whose domain is a dense linear manifold in a normed linear space X and whose range is in a normed linear space Y. For brevity, we say that T is densely defined. If $y'\in Y'$, then $y'\circ T$ is a linear functional on $\mathscr{D}(T)$. If T is not continuous, then $y'\circ T$ may not be continuous. However, if it *is* continuous, it has a unique extension to a continuous linear functional on $\overline{\mathscr{D}(T)} = X$. (The proof of this resembles the proof of Theorem II.4.1.) We denote this element of X' by $T'y'$. Let

$$\mathscr{D}(T') = \{y'\in Y' : y'\circ T \text{ is continuous on } \mathscr{D}(T)\}.$$

The conjugate of T is the linear operator T' defined by

$$\langle x, T'y'\rangle = \langle Tx, y'\rangle, \qquad x\in\mathscr{D}(T),\ y'\in\mathscr{D}(T'). \tag{8-7}$$

Note that (8-7) completely determines T' since it specifies, for each $y'\in\mathscr{D}(T')$, the values of the linear functional $T'y'$ on the dense subspace $\mathscr{D}(T)$ of X.

The definition of T' says nothing about the "size" of $\mathscr{D}(T')$ when T is not continuous. For instance, there exists a linear operator T with $\mathscr{D}(T) = X = \ell^2$ and $\mathscr{R}(T)\subset\ell^2$ such that $\mathscr{D}(T')$ consists only of the zero functional! (See Goldberg [2, page 53].) Fortunately, this situation cannot occur when T is closed or when T is closable, as in problem 9, § 5.

Theorem 8.1. *If T is closed, then $\mathscr{D}(T')$ is total.*

Proof. To show that $\mathscr{D}(T')$ is total we must show that given $y_0 \neq 0$ there exists $y' \in \mathscr{D}(T')$ such that $y'(y_0) \neq 0$. Let $G(T)$ be the graph of T. Clearly $(0, y_0) \notin G(T)$. Since $G(T)$ is closed in $X \times Y$, there exists (by Theorem III.3.4) $z' \in (X \times Y)'$ such that $z'(0, y_0) \neq 0$ and $z'(x, Tx) = 0$ for all $(x, Tx) \in G(T)$. Define $y' \in Y'$ by $y'(y) = z'(0, y)$, so that $y'(y_0) \neq 0$. If we define $x' \in X'$ by $x'(x) = z'(x, 0)$, then $0 = z'(x, Tx) = x'(x) + y'(Tx)$, for $x \in \mathscr{D}(T)$. This shows that $y' \in \mathscr{D}(T')$. $\square$

An equivalent formulation of Theorem 8.1 is that $\mathscr{D}(T')$ is weak*-dense in Y' when T is closed. (See problem 1 of § III.7.) $\mathscr{D}(T')$ may fail to be norm-dense, however, even when X and Y are complete and T is closed. This is not so when Y is reflexive, for in this case $\mathscr{D}(T')$ is total if and only if $\overline{\mathscr{D}(T')} = Y'$ (problem 1 of § III.7).

Theorem 8.2. *$\mathscr{D}(T') = Y'$ if and only if T is continuous. In this case T' is also continuous and $\|T'\| = \|T\|$.*

Proof. Let $S = \{x \in \mathscr{D}(T) : \|x\| \leq 1\}$. If T is continuous, then $x \mapsto y'(Tx)$ is continuous on $\mathscr{D}(T)$ for each $y' \in Y'$. So $\mathscr{D}(T') = Y'$. Also, $\|T'y'\| = \sup_{\|x\| \leq 1} |\langle x, T'y'\rangle| = \sup_{x \in S} |\langle Tx, y'\rangle| \leq \sup_{x \in S} \|y'\| \, \|Tx\| = \|y'\| \, \|T\|$. This implies that T' is continuous and $\|T'\| \leq \|T\|$.

Suppose that $\mathscr{D}(T') = Y'$. Then $\sup_{x \in S} |\langle Tx, y'\rangle| \leq \|T'y'\|$ for each $y' \in Y'$. By the principle of uniform boundedness (Theorem III.9.2), $\sup_{x \in S} \|Tx\| < \infty$; that is, T is continuous. Finally, for $x \in S$ and $\|y'\| = 1$, $|\langle Tx, y'\rangle| \leq \|T'y'\| \leq \|T'\|$. Using Theorem III.3.2, we have $\|Tx\| \leq \|T'\|$, so that $\|T\| \leq \|T'\|$. $\square$

Some results that can be stated for arbitrary (densely defined) linear operators depend on the following theorem.

Theorem 8.3. *T' is a closed linear operator.*

Proof. Let $\{y'_n\}$ be a sequence in $\mathscr{D}(T')$ such that $y'_n \to y'$ and $T'y'_n \to x'$, for some $x' \in X'$, $y' \in Y'$. Then for $x \in \mathscr{D}(T)$, $y'_n(Tx) \to y'(Tx)$ and $y'_n(Tx) = T'y'_n(x) \to x'(x)$. Hence $y'(Tx) = x'(x)$, $x \in \mathscr{D}(T)$. Since x' is continuous, we see that $y' \in \mathscr{D}(T')$ and $T'y' = x'$. Thus T' is closed. $\square$

The Conjugate of a Differential Operator

Let $X = L^p(a, b)$, where $1 \leq p < \infty$ and $[a, b]$ is a finite interval. Let $D(p)$ be the linear manifold of x in X such that x is absolutely continuous on $[a, b]$

and its derivative $x' = dx/ds$ is again in X. The set $D(p)$ is dense in X because it contains the polynomial functions. Let T be a linear operator on X, with domain $D(p)$, defined by $Tx = x'$. We shall represent the conjugate of T as a differential operator on $X' = L^{p'}(a, b)$ whose domain is the set $D_0(p')$ of all $y \in D(p')$ such that $y(a) = y(b) = 0$. [Note that $y(a)$ and $y(b)$ are well defined because y can be represented by an absolutely continuous function in $\mathscr{L}^{p'}(a, b)$.]

Suppose that $x \in \mathscr{D}(T)$ and $y \in D_0(p')$. Then an integration by parts shows that

$$\langle Tx, y\rangle = \int_a^b x'(t)y(t)\,dt = -\int_a^b x(t)y'(t)\,dt = \langle x, -y'\rangle.$$

It follows that $y \in \mathscr{D}(T')$ and $T'y = -y' = -dy/ds$. Thus $D_0(p') \subset \mathscr{D}(T')$. Now suppose that $y \in \mathscr{D}(T')$. Let $w = T'y$ and define z by

$$z(s) = \int_a^s w(t)\,dt.$$

Then z is absolutely continuous and $z' = w$. We shall show that $z \in D_0(p')$ and $y = -z$. From this it will follow that $\mathscr{D}(T') = D_0(p')$. If $x \in \mathscr{D}(T)$, then

$$\int_a^b x(t)w(t)\,dt = \langle x, T'y\rangle = \langle Tx, y\rangle = \int_a^b x'(t)y(t)\,dt. \tag{8-8}$$

By choosing $x(t) \equiv 1$, we see that $z(b) = 0$. From the definition of z it is then apparent that $z \in D_0(p')$. Because $z' = w$, integration by parts leads to

$$\int_a^b x(t)w(t)\,dt = \int_a^b x(t)z'(t)\,dt = -\int_a^b x'(t)z(t)\,dt. \tag{8-9}$$

Combining (8-8) and (8-9), we see that

$$\int_a^b x'(t)[y(t) + z(t)]\,dt = 0$$

for each $x \in \mathscr{D}(T)$. The arbitrariness of x here means that x' can be an arbitrary element of $L^p(a, b)$, and therefore we can conclude that $y + z = 0$ in $L^{p'}(a, b)$, which is what we wished to show.

When $1 < p < \infty$, it can be shown that the operator T defined above is the conjugate of T'. Hence T is closed, by Theorem 8.3.

Conjugates of rather general types of differential operators are discussed in Goldberg [2, pages 126–157].

Annihilators, Ranges, and Null Spaces

We now turn to an examination of theorems analogous to those of § I.13, with X', Y' replacing X^f, Y^f and T' replacing the transpose of T. The differences are due mainly to the intervention of topological matters: an annihilator (as defined in § III.7) is always closed, and a linear manifold and its closure possess the same annihilator. Also, unlike § I.13, we assume only that $\mathscr{D}(T)$ is dense in X.

Theorem 8.4.
(*a*) $\overline{\mathscr{R}(T)}^{\perp} = \mathscr{R}(T)^{\perp} = \mathscr{N}(T')$
(*b*) $\overline{\mathscr{R}(T)} = \mathscr{N}(T')^{\perp}$
(*c*) $\overline{\mathscr{R}(T)} = Y$ *if and only if* $(T')^{-1}$ *exists.*

The proof is left to the reader. For part (*b*) see the remark following Theorem III.7.4.

When we exchange the roles of T and T' we do not get quite as much information as in Theorem 8.4.

Theorem 8.5. *Suppose that T is closed. Then*
(*a*) $\mathscr{R}(T')^{\perp} \cap \mathscr{D}(T) = \mathscr{N}(T)$
(*b*) $\overline{\mathscr{R}(T')} \subset \mathscr{N}(T)^{\perp}$
(*c*) $\mathscr{R}(T')$ *is total in X' if and only if T^{-1} exists.*

Proof. It is easy to see that

$$\mathscr{R}(T')^{\perp} \cap \mathscr{D}(T) \supset \mathscr{N}(T). \tag{8-10}$$

Take $x \in \mathscr{R}(T')^{\perp} \cap \mathscr{D}(T)$. Then, for $y' \in \mathscr{D}(T')$, we have $\langle Tx, y'\rangle = \langle x, T'y'\rangle = 0$. Since $\mathscr{D}(T')$ is total by Theorem 8.1, $Tx = 0$. Hence $\mathscr{R}(T')^{\perp} \cap \mathscr{D}(T) \subset \mathscr{N}(T)$. This proves (*a*). Part (*b*) follows easily from (*a*). Now suppose $\mathscr{R}(T')$ is total. Then $\mathscr{R}(T')^{\perp} = (0)$, and (8-10) implies that T^{-1} exists. Conversely, suppose T^{-1} exists, and take $x_0 \neq 0$ in X. Then $(x_0, 0)$ does not belong to the graph of T. Since the graph of T is closed, there exists $z' \in (X \times Y)'$ such that $z'(x_0, 0) \neq 0$ and $z'(x, Tx) = 0$ for $x \in \mathscr{D}(T)$. Define x' and y' as in the proof of Theorem 8.1, and observe that $x'(x_0) = z'(x_0, 0) \neq 0$, while $0 = z'(x, Tx) = x'(x) + y'(Tx)$ for $x \in \mathscr{D}(T)$. It follows that $y' \in \mathscr{D}(T')$ and $x' = -T'y' \in \mathscr{R}(T')$. Thus $\mathscr{R}(T')$ is total in X'. This concludes the proof of (*c*). ☐

In connection with Theorem 8.5(*b*), we observe that if $\mathscr{D}(T) = X$ and $\overline{\mathscr{R}(T')}$ is weak*-closed in X', then $\overline{\mathscr{R}(T')} = \mathscr{N}(T)^{\perp}$. This happens, for example, if X is reflexive and $T \in L(X, Y)$. (See problem 1 in § III.7.) Later, in Theorem 10.1, we shall prove that if X and Y are complete, T is closed, and $\mathscr{R}(T')$ is closed, then $\mathscr{R}(T') = \mathscr{N}(T)^{\perp}$.

PROBLEMS

1. Let X and Y be normed linear spaces. If $T \in L(X, Y)$ and if T maps X congruently onto all of Y, then T' maps Y' congruently onto all of X'.
2. Suppose $T \in L(X, Y)$. Then $T'' \equiv (T')' \in L(X'', Y'')$. Let J_1 and J_2 be the operators defining the canonical mappings of X and Y into X'' and Y'', respectively. Then the diagram is commutative; that is, $T''J_1 = J_2T$. If we identify X with $J_1(X)$ and Y with $J_2(Y)$, then T'' is an extension of T. In particular, T'' coincides with T if X is reflexive.

$$\begin{tikzcd} X'' \arrow[r, "T''"] & Y'' \\ X \arrow[u, "J_1"] \arrow[r, "T"'] & Y \arrow[u, "J_2"'] \end{tikzcd}$$

3. Let X and Y be Banach spaces. If X is reflexive and if there exists $T \in L(X, Y)$ such that $\mathscr{R}(T) = Y$, then Y is reflexive. [*Hint.* First consider the case when $\mathscr{N}(T) = (0)$.]
4. Let M be a closed subspace of a normed linear space X, and let Φ be the canonical quotient mapping of X onto X/M. Determine the null space and range of Φ', and show that $(X/M)'$ is congruent to $M^\perp$ (cf. Theorem III.3.3). Give a similar proof that M' is congruent to $X'/M^\perp$ by using the operator $\Psi: M \to X$ defined by $\Psi x = x$, $x \in M$.
5. Suppose that M is a closed subspace of a reflexive space X. Then X/M is reflexive.
6. The principle of uniform boundedness may be used to prove the following special case of the closed graph theorem. Let T be a closed linear operator with $\mathscr{D}(T) = X$, $\mathscr{R}(T) \subset Y$, where X is a Banach space and Y is a reflexive Banach space. Then T is continuous. [*Hint.* Show that $\mathscr{D}(T') = Y'$.]
7. Let X and Y be Banach spaces with countable bases $\{u_n\}$ and $\{v_n\}$, respectively, $n = 1, 2, \ldots$. See problem 7, § III.9. Let $\{u'_n\}$ and $\{v'_n\}$ be the corresponding sequences of coefficient functionals, so that

$$x = \sum_1^\infty \xi_j u_j \qquad u'_j(x) = \xi_j, \qquad x \in X$$

$$y = \sum_1^\infty \eta_i v_i \qquad v'_i(y) = \eta_i, \qquad y \in Y.$$

By the problem referred to, elements of X' and Y' are representable in the respective forms

$$x'(x) = \sum_1^\infty \lambda_j \xi_j, \qquad y'(y) = \sum_1^\infty \mu_i \eta_i.$$

Suppose $A \in L(X, Y)$, $\alpha_{ij} = v'_i(Au_j)$, so that $y = Ax$ is expressed by

$$\eta_i = \sum_{j=1}^\infty \alpha_{ij}\xi_j, \qquad i = 1, 2, \ldots.$$

Show that $x' = A'y'$ is expressed by

$$\lambda_j = \sum_{i=1}^{\infty} \alpha_{ij}\mu_i \qquad j = 1, 2, \ldots .$$

Note that this result means that if A is represented by a certain infinite matrix, A' is represented by the transposed matrix.

8. Suppose $A \in L((c), (c))$ is defined by (6-7). Let

$$\alpha_{00} = \alpha - \sum_{j=1}^{\infty} \alpha_j$$

$$\alpha_{0j} = \alpha_j, \qquad j = 1, 2, \ldots,$$

where α and α_j are defined by (6-9) and (6-10). If x' and y' are elements of $(c)'$ defined by

$$x'(x) = \sum_{j=0}^{\infty} \lambda_j \xi_j, \qquad y'(y) = \sum_{i=0}^{\infty} \mu_i \eta_i,$$

show that $A'y' = x'$ means

$$\lambda_j = \sum_{i=0}^{\infty} \alpha_{ij}\mu_i, \qquad j = 0, 1, 2, \ldots .$$

9. Let X, Y be normed linear spaces, and let T be a linear operator from X into Y, with $\mathscr{D}(T) = X$. Then T is continuous if and only if T is continuous with respect to the weak topologies, $\sigma(X, X')$ and $\sigma(Y, Y')$, on X and Y. [*Hint.* See Theorem II.11.5.]

10. Let X and Y be normed linear spaces, and let S be a linear operator from Y' into X', with $\mathscr{D}(S) = Y'$. Let J_1 and J_2 be the canonical mappings of X and Y into X'' and Y'', respectively. Then the following statements are equivalent:
    a. S is continuous with respect to the weak* topologies $\sigma(Y', Y)$ and $\sigma(X', X)$;
    b. $S \in L(Y', X')$ and the conjugate S' maps $J_1(X)$ into $J_2(Y)$;
    c. There is an operator $T \in L(X, Y)$ such that $S = T'$.

11. Let X and Y be normed linear spaces. Every operator in $L(Y', X')$ is the conjugate of some operator in $L(X, Y)$ if and only if Y is reflexive.

IV.9 THEOREMS ABOUT CONTINUOUS INVERSES

Throughout this section we assume that X and Y are normed linear spaces and that T is a linear operator whose domain is dense in X and whose range is in Y. In some cases we assume that a space is complete or an operator is closed; these assumptions are made explicit in each case. The conjugate spaces X', Y' are complete, in any event.

Lemma 9.1. *If T does not have a continuous inverse, then there exists a sequence $\{x_n\}$ in $\mathscr{D}(T)$ such that $\|x_n\| \to \infty$ and $Tx_n \to 0$.*

Proof. It follows from Theorem II.1.2 that there is a sequence $\{u_n\}$ in $\mathscr{D}(T)$ such that $u_n \neq 0$ and $\|Tu_n\|/\|u_n\| \to 0$. Consequently, $T(u_n/\|u_n\|) \to 0$, and so we may assume that $\|u_n\| = 1$ for each n. A sequence with the desired properties is given by

$$x_n = \begin{cases} \dfrac{u_n}{\|Tu_n\|^{1/2}} & \text{if } Tu_n \neq 0, \\ nu_n & \text{if } Tu_n = 0. \quad \square \end{cases}$$

Theorem 9.2. *$\mathscr{R}(T') = X'$ if and only if T^{-1} exists and is continuous.*

Proof. Suppose that $\mathscr{R}(T') = X'$ but that T does not have a continuous inverse. Let $\{x_n\}$ be the sequence given in Lemma 9.1. Then $\langle x_n, T'y' \rangle = \langle Tx_n, y' \rangle \to 0$ for each $y' \in Y'$. Since $\mathscr{R}(T') = X'$, this implies that $x'(x_n) \to 0$ for each $x' \in X'$. But then the uniform boundedness principle (Theorem III.9.4) implies that $\{\|x_n\|\}$ is bounded, which is a contradiction.

Now suppose, conversely, that T has a continuous inverse. If x' is fixed in X', then $y \mapsto x'(T^{-1}y)$ is a continuous linear functional on $\mathscr{R}(T)$. By the Hahn–Banach theorem, this functional can be extended to a continuous linear functional on all of Y. That is, there exists $y' \in Y'$ such that $y'(y) = x'(T^{-1}y)$ when $y \in \mathscr{R}(T)$, which implies $y'(Tx) = x'(x)$ for each $x \in \mathscr{D}(T)$. Thus $y' \in \mathscr{D}(T')$ and $T'y' = x'$. Since x' was arbitrary, $\mathscr{R}(T') = X'$. $\square$

Theorem 9.3. *Suppose that Y is complete and that $\mathscr{R}(T) = Y$. Then T' has a continuous inverse.*

A proof of this theorem can be given using an argument very much like that in the first part of the proof of Theorem 9.2. (We leave this to the reader.) It is possible in each case to avoid the use of Lemma 9.1 by applying the principle of uniform boundedness directly instead of arguing by contradiction. See problem 1.

The next theorem is in the nature of a converse to Theorem 9.3. But the proof is more complicated, and we need to assume that X is complete and T is closed.

Theorem 9.4. *Suppose that X is complete, T is closed, and T' has a continuous inverse. Then $\mathscr{R}(T) = Y$. Also, T is an open mapping, and hence T^{-1} is continuous if it exists.*

Proof. Let $S = \{x \in \mathscr{D}(T) : \|x\| \leq 1\}$. Since S is absolutely convex, the same is true for $T(S)$ and $\overline{T(S)}$. Suppose $y_0 \in Y \setminus \overline{T(S)}$. By Theorem

III.2.9, there exists $y' \in Y'$ such that

$$y'(y_0) > \sup\{|y'(y)| : y \in \overline{T(S)}\}$$
$$\geq \sup\{|y'(Tx)| : x \in S\}.$$

Clearly $y' \circ T$ is bounded on S and hence is continuous on $\mathscr{D}(T)$ (cf. Theorem III.1.4). Thus $y' \in \mathscr{D}(T')$, and we have

$$\|T'y'\| < y'(y_0) = |y'(y_0)| \leq \|y'\|\,\|y_0\|. \tag{9-1}$$

Since T' has a continuous inverse, there exists $m > 0$ such that $m\|y'\| \leq \|T'y'\|$ for every $y' \in \mathscr{D}(T')$ (Theorem II.1.2). It follows from (9-1) that $\|y_0\| > m$. This shows that $\overline{T(S)}$ contains the ball $\{y : \|y\| \leq m\}$.

We now use the fact that X is complete and T is closed to conclude, from Lemma 5.4, that $T(S)$ itself contains a ball $\{y : \|y\| \leq \beta\}$ for some $\beta > 0$. Since the range of T is a subspace containing $T(S)$, it follows by homogeneity that $\mathscr{R}(T) = Y$. The rest of the theorem follows from the arguments given in the last part of the proof of Theorem 5.5 and in the proof of Corollary 5.6. □

PROBLEMS

1. a. Fill in the details of the following proof of Theorem 9.3. Let $S = \{x' \in \mathscr{R}(T') : \|x'\| \leq 1\}$, and let $F = \{y' \in \mathscr{D}(T') : T'y' = x', \text{ some } x' \in S\}$. Since $\mathscr{R}(T) = Y$, T' has an inverse. To show that $(T')^{-1}$ is continuous it suffices to show that F is bounded in the norm of Y'.
   b. Prove the first half of Theorem 9.2 using an argument analogous to that given in part (a).
2. Let X and Y be normed linear spaces. Let T be a linear operator with domain dense in X and range in Y. Then $\overline{\mathscr{R}(T)} = Y$ and T has a continuous inverse if and only if $\mathscr{R}(T') = X'$ and T' has a continuous inverse.
3. a. Fill in the details of the following proof of Theorem 9.4. First of all, $\overline{T(S)} = T(S)^{\infty}$. Since T' has a continuous inverse, there exists $m > 0$ such that $(1/m)\|T'y'\| \geq \|y'\|$ for all $y' \in \mathscr{D}(T')$. Then $\overline{T(S)} = T(S)^{\circ\circ} \supset \{y : \|y\| \leq m\}$. The rest of the proof is the same as that given in the text.
   b. If T, X, and Y are the same as in Theorem 9.4 except that $\mathscr{D}(T) = X$, prove that Y must be complete.
4. Let (a_{ij}) be a matrix that determines an operator in $L(\ell^2)$. Then the system of equations

$$\sum_{j=1}^{\infty} a_{ij}\xi_j = \eta_i, \qquad i = 1, 2, \ldots$$

has a solution $x = (\xi_1, \xi_2, \ldots)$ in ℓ^2 for each $y = (\eta_1, \eta_2, \ldots)$ in ℓ^2 if and only if there exists $M > 0$ such that

$$\sum_{j=1}^{\infty} |a_{ji}\zeta_j|^2 \geq M \sum_{j=1}^{\infty} |\zeta_j|^2, \qquad i = 1, 2, \ldots$$

for each $z = (\zeta_1, \zeta_2, \ldots)$ in ℓ^2.

IV.10 THE STATES OF AN OPERATOR AND ITS CONJUGATE

In this section we shall consider a linear operator whose domain is a dense subspace of a normed linear space X and whose range is in a normed linear space Y. The theorems in § 9, together with some of those in § 5 and § 8 make it possible for us to tabulate a large number of interesting and useful implications involving T and T'. In order to record this information concisely, we make a ninefold classification of what we shall call the *state* of an operator. We list three possibilities for $\mathcal{R}(T)$.

I. $\mathcal{R}(T) = Y$.
II. $\overline{\mathcal{R}(T)} = Y$, but $\mathcal{R}(T) \neq Y$.
III. $\overline{\mathcal{R}(T)} \neq Y$.

As regards T^{-1}, we also list three possibilities.

1. T^{-1} exists and is continuous.
2. T^{-1} exists but is not continuous.
3. T^{-1} does not exist.

By combining these possibilities we obtain nine different situations. For instance, it may be that $\overline{\mathcal{R}(T)} = Y$, that $\mathcal{R}(T) \neq Y$, and that T^{-1} exists but is discontinuous. We shall describe this as state II_2 for T; alternatively, we shall say that T is in state II_2.

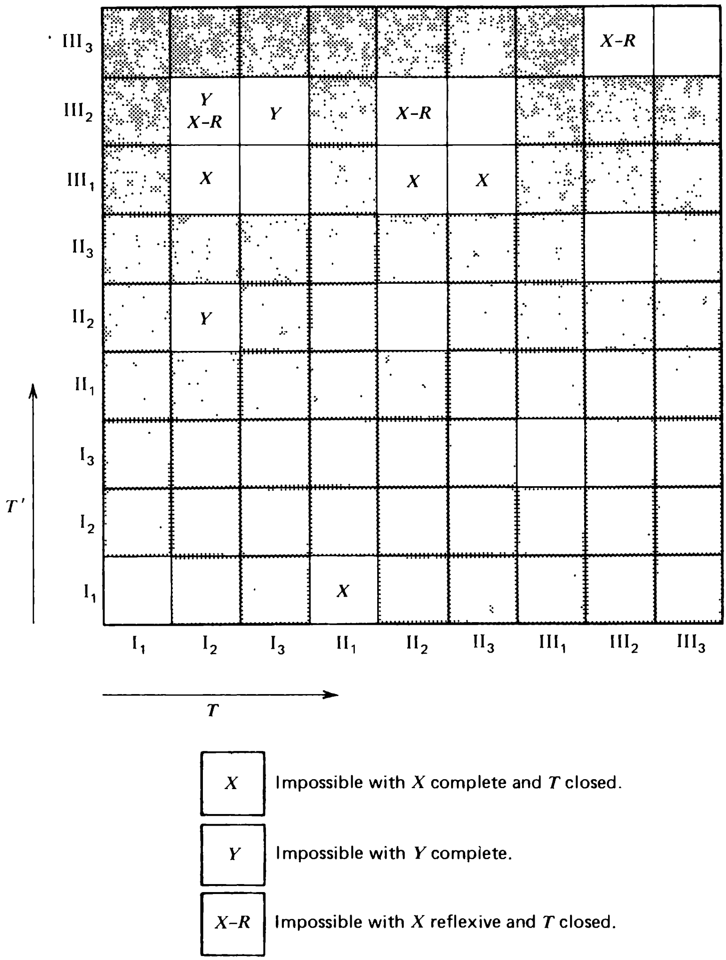
The notion of the state of an operator can be applied, in particular, to T'. Here T' replaces T and X' replaces Y in the description of the states; to say that T' is in state I_3 means that $\mathcal{R}(T') = X'$ and T' does not have an inverse.

It is possible for T to be in any one of the nine states if no restrictions are placed on X and Y. However, since X' and Y' are both complete and T' is closed, it is impossible for T' to be in state I_2 or II_1, according to Theorem 5.8.

Next, we define the *state of the pair* (T, T'); this is the ordered pair of the states of T and T', respectively, with the state of T listed first. We consider all eighty-one possibilities in the state diagram on page 238, each small square corresponding to a state of the pair (T, T').

Various theorems in § 5, § 8, and § 9 show that many states of the pair (T, T') can never occur. The shaded squares in the diagram correspond to states that are impossible by virtue of our theorems, *without requiring X or Y to be complete or T to be closed.* The reader should now examine the state diagram, as we review the facts related to the shaded squares. All squares in the I_2 and II_1 rows are shaded, since T' can never be in these states; this was explained above. Theorem 8.4 says that if T is in a I or II state, T' must be in a 1 or 2 state, and vice versa. Thus any state of the pair with I or II for T and 3 for T' is impossible; also III for T and 1 or 2 for T' is impossible. It is easily verified that $\mathcal{R}(T')^{\perp} \supset \mathcal{N}(T)$, even when T is not closed (cf. formula (8-10)).

The state diagram for linear operators.



This implies that a state of the pair (T, T') with 3 for T and I or II for T' is impossible. Theorem 9.2 says that we have 1 for T if and only if we have I for T'. This eliminates states with 1 for T and II or III for T' and also those with 2 or 3 for T and I for T'. All but sixteen of the eighty-one squares are shaded as a result of the foregoing considerations. These sixteen remaining states for the pair (T, T') are all possible, if no restrictions are placed on X or Y; this has been shown by the construction of examples, all of which have T in $L(X, Y)$ (see Taylor and Halberg [1]). Some of these examples are considered in the problems. Many implications can now be read from the state diagram.

Samples: The state of T is III_1 if and only if the state of T' is I_3; if the state of T is III_2, the state of T' is either II_3 or III_3.

Next, we consider the effect of assuming that Y is complete. Theorem 9.3 tells us, in this case, that I for T makes 2 and 3 impossible for T'. Of the sixteen states not previously eliminated, this rules out $(\mathrm{I}_2, \mathrm{II}_2)$, $(\mathrm{I}_2, \mathrm{III}_2)$, and $(\mathrm{I}_3, \mathrm{III}_2)$. These exclusions are indicated on the state diagram by putting the letter Y in the corresponding squares.

Now consider the effect of assuming that T is closed and X is complete but making no completeness assumption about Y. Theorem 9.4 says, in this case, that any state except I_1 and I_3 for T is impossible if we have 1 for T'. This rules out $(\mathrm{I}_2, \mathrm{III}_1)$, $(\mathrm{II}_1, \mathrm{I}_1)$, $(\mathrm{II}_2, \mathrm{III}_1)$, and $(\mathrm{II}_3, \mathrm{III}_1)$, in addition to some of the shaded squares. These exclusions are indicated by putting an X in the appropriate squares on the state diagram.

With X and Y both complete and T closed, there are nine states still left as apparently possible for the pair (T, T'). Examples exist that show that these states are actually possible.

If we assume that T is closed and X is reflexive (but make no restriction on Y), then a linear manifold in X' is total if and only if it is norm-dense (problem 1 of § III.7), and it follows from Theorem 8.5 that we must have I or II for T' if we have 1 or 2 for T. Hence, in addition to the states ruled out because X is complete and T is closed, $(\mathrm{I}_2, \mathrm{III}_2)$, $(\mathrm{II}_2, \mathrm{III}_2)$, and $(\mathrm{III}_2, \mathrm{III}_3)$ are eliminated from the sixteen states corresponding to the unshaded squares. These three exclusions are indicated by $X - R$ in the corresponding square on the state diagram.

It is now interesting to observe on the state diagram that with T closed, X reflexive, and Y complete, there are seven squares left blank. Also, there is at most one blank square in any one row or column. Hence, with these conditions on $T, X,$ and Y, the state of either T or T' determines uniquely the state of the other. These seven states for the pair (T, T') do actually occur; examples can be constructed with $X = Y = \ell^2$ and with T a bounded linear operator.

The state diagram was constructed first for continuous linear operators defined on all of X (Taylor and Halberg [1]). Later S. Goldberg demonstrated that the same diagram holds for closed linear operators (Goldberg [1]), and he obtained the state diagram on page 238. Tabulations similar to the state diagram have also been constructed for general linear operators on normed spaces (Webb [1]) and for linear operators on various topological linear spaces (Krishnamurthy [1]). Also see Halberg and Samuelsson [1].

As an illustration of how the state diagram can help organize our thinking, let us examine $\mathscr{R}(T)$ and $\mathscr{R}(T')$, with Theorem 5.8 and the state diagram as our principal tools. If we assume that X and Y are complete and T is closed, then states I_1, I_3, and III_1 are possible only for an operator with closed range, while states II_2, II_3, and III_2 are possible only for an operator whose range is *not* closed (see Theorem 5.8). Examining the open squares in the state diagram (but ignoring "$X-R$"), we see that $(\mathrm{I}_1, \mathrm{I}_1)$, $(\mathrm{I}_3, \mathrm{III}_1)$, and $(\mathrm{III}_1, \mathrm{I}_3)$ all correspond to $\mathscr{R}(T)$ and $\mathscr{R}(T')$ *both closed*; the states $(\mathrm{II}_2, \mathrm{II}_2)$, $(\mathrm{II}_2, \mathrm{III}_2)$, $(\mathrm{II}_3, \mathrm{III}_2)$, and $(\mathrm{III}_2, \mathrm{II}_3)$ all correspond to $\mathscr{R}(T)$ and $\mathscr{R}(T')$ *both nonclosed*; only states $(\mathrm{III}_2, \mathrm{III}_3)$ and $(\mathrm{III}_3, \mathrm{III}_3)$ seem to need more examination. At least at this point we can say that if $\overline{\mathscr{R}(T)} = Y$, then $\mathscr{R}(T)$ is closed if and only if $\mathscr{R}(T')$ is closed. This observation leads us to a somewhat stronger result, essentially due to Banach.

Theorem 10.1 (The closed range theorem). *Let T be a closed densely defined linear operator from a Banach space X into a Banach space Y. Then the following assertions are equivalent*:

(*a*) $\mathscr{R}(T)$ *is closed*

(*b*) $\mathscr{R}(T')$ *is closed*

(*c*) $\mathscr{R}(T')$ *is weak*-closed*

(*d*) $\mathscr{R}(T) = \mathscr{N}(T')^{\perp}$

(*e*) $\mathscr{R}(T') = \mathscr{N}(T)^{\perp}$.

Proof. $(a)\Leftrightarrow(d)$, by Theorem 8.4(b). Since the implications $(e)\Rightarrow(c)$ and $(c)\Rightarrow(b)$ are elementary (cf. § III.7), we shall prove only that $(a)\Rightarrow(e)$ and $(b)\Rightarrow(a)$.

Suppose that $\mathscr{R}(T)$ is closed, and take $x' \in \mathscr{N}(T)^{\perp}$. Then x' induces a continuous linear functional u' on $X/\mathscr{N}(T)$, where

$$u'([x]) = x'(x).$$

(See the proof of Theorem III.3.3(b).) Let $\hat{T}$ be the one-to-one operator induced by T. Since $\mathscr{R}(\hat{T}) = \mathscr{R}(T)$ is closed, $\hat{T}$ has a bounded inverse that maps $\mathscr{R}(T)$ into $X/\mathscr{N}(T)$ (Theorem 5.8). Therefore the composition $u' \circ \hat{T}^{-1}$ is a continuous linear functional on $\mathscr{R}(T)$, given by

$$(u' \circ \hat{T}^{-1})(Tx) = x'(x), \qquad x \in \mathscr{D}(T). \tag{10-1}$$

Extend $u' \circ \hat{T}^{-1}$ to an element y' in Y', by the Hahn–Banach theorem. Then (10-1) implies that $y' \in \mathscr{D}(T')$ and $x' = T'y' \in \mathscr{R}(T')$. This shows that $\mathscr{N}(T)^{\perp} \subset \mathscr{R}(T')$. The reverse inclusion was given in Theorem 8.5. Thus (a) implies (e).

Now suppose that $\mathscr{R}(T')$ is closed, let $Y_0 = \overline{\mathscr{R}(T)}$, and let T_0 be the operator T considered as a mapping into the Banach space Y_0. Clearly T_0 is closed. It is easy to see that $\mathscr{R}(T_0') = \mathscr{R}(T')$ since every element of Y_0' has an extension to an element of Y', by the Hahn–Banach theorem. Thus $\mathscr{R}(T_0')$ is closed. From the discussion preceding the theorem it follows that $\mathscr{R}(T_0)$ is closed. But then $\mathscr{R}(T) = Y_0 = \overline{\mathscr{R}(T)}$. Hence ($b$) implies ($a$). □

PROBLEMS

Examples of operators T are given in the first thirteen problems that provide illustrations of various states for the pair (T, T'). In these examples X and Y are chosen from the spaces $\ell^p (1 \le p < \infty)$ and (c_0). In each of these spaces we denote the vector $(0, \ldots 0, 1, 0, \ldots)$ (with 1 in the kth place) by u_k. In ℓ^p, with $1 \le p < \infty$, and (c_0) the set $\{u_k\}$ is a countable basis, and when we define T on these spaces, it suffices to define Tu_k for each k. If X and Y are both spaces in which $\{u_k\}$ is a countable basis, T' is determined by the transpose of the infinite matrix that represents T. In each problem it is left for the reader to show that the state of the pair (T, T') is as indicated.

1. $X = Y = \ell^2$. $Tu_k = 2^{1-k} u_k$. The state is $(\mathrm{II}_2, \mathrm{II}_2)$.
2. $X = \ell^1$, $Y = \ell^2$, T as in problem 1. The state is $(\mathrm{II}_2, \mathrm{III}_2)$.
3. $X = Y = \ell^2$. $Tu_1 = 0$, $Tu_k = u_k$ if $k \ge 2$. The state is $(\mathrm{III}_3, \mathrm{III}_3)$.
4. $X = Y = \ell^2$. $Tu_1 = 0$, $Tu_k = u_{k-1}$, $k \ge 2$. The state is $(\mathrm{I}_3, \mathrm{III}_1)$.
5. $X = Y = \ell^2$. $Tu_k = u_{k+1}$. The state is $(\mathrm{III}_1, \mathrm{I}_3)$.
6. $X = Y = \ell^2$. $Tu_1 = 0$, $Tu_k = \Sigma_{i=k-1}^{\infty} 2^{1-i} u_i$. The state is $(\mathrm{II}_3, \mathrm{III}_2)$. Note that $T(u_k - u_{k+1}) = 2^{2-k} u_{k-1}$ if $k \ge 2$. Also note that $\Sigma_1^{\infty} 2^{1-k} u_{2k}$ is not in $\mathscr{R}(T)$.
7. $X = Y = \ell^2$. $Tu_k = 2^{1-k}(u_2 + \cdots + u_{k+1})$. The state is $(\mathrm{III}_2, \mathrm{II}_3)$. We can also take $X = Y = (c_0)$ in this case.
8. $X = Y = \ell^2$. $Tu_1 = u_2$, $Tu_k = u_{k-1} + u_{k+1}$ if $k \ge 2$. The state is $(\mathrm{II}_2, \mathrm{II}_2)$. Note that u_{2k} is in $\mathscr{R}(T)$, but u_{2k-1} is not. But it can be shown that $u_{2k-1} \in \overline{\mathscr{R}(T)}$.
9. Suppose $A \in L(X, Y)$ and $\mathscr{R}(A) = Z \ne Y$. Define $B \in L(X, Z)$ by setting $Bx = Ax$ when $x \in X$. Suppose also that Y is complete and $\overline{\mathscr{R}(A)} = Y$. Then we can identify Z' with Y' and B' with A'. The states of A' and B' are the same but, if the state of A is II_i, $i = 1, 2, 3$, that of B is I_i. In this way we can produce examples of states $(\mathrm{I}_2, \mathrm{II}_2)$, $(\mathrm{I}_2, \mathrm{III}_2)$, and $(\mathrm{I}_3, \mathrm{III}_2)$ from some of the earlier examples.
10. X the subspace of ℓ^2 generated by $\{u_k\}$, $Y = \ell^2$. T as in problem 4, but with this change in X. The state is $(\mathrm{II}_3, \mathrm{III}_1)$.
11. If Y is a complete space and X is a proper but dense subspace of Y, let $Tx = x$ for $x \in X$. Then $T \in L(X, Y)$ and the state of the pair is $(\mathrm{II}_1, \mathrm{I}_1)$.
12. X the subspace of ℓ^1 generated by $\{u_k\}$, Y either X or ℓ^1 itself. $Tu_1 = \Sigma_1^{\infty} 2^{-k} u_k$, $Tu_k = u_{k-1}$ if $k \ge 2$. The state is $(\mathrm{II}_2, \mathrm{III}_1)$ or $(\mathrm{I}_2, \mathrm{III}_1)$ according to whether Y is ℓ^1 or X.
13. Let Y be any Banach space of infinite dimension, and let H be a Hamel basis for Y with all elements h of H such that $\|h\| \le 1$. If $y \in Y$ and $y = \alpha_1 h_1 + \cdots + \alpha_n h_n$, let $N(y) = \Sigma_{i=1}^{n} |\alpha_i|$. Let X be the space with the same elements as Y, but with N as a norm. Let $Ay = y$, and consider A as an element of $L(X, Y)$. The state is $(\mathrm{I}_2, \mathrm{III}_1)$.
14. Suppose $A \in L(X, X')$. Assume that $\overline{\mathscr{R}(A)} = \overline{\mathscr{R}(A')}$. If X is reflexive, show that the only possible states are $(\mathrm{I}_1, \mathrm{I}_1)$, $(\mathrm{II}_2, \mathrm{II}_2)$, $(\mathrm{III}_3, \mathrm{III}_3)$.
15. Suppose $A \in L(X)$. Let B be a topological isomorphism of X onto X', and suppose $A = B^{-1} A' B$. Then the only possible states for A are I_1, II_2, III_3.
16. Let X, Y be Banach spaces, and let A, B be elements of $L(X, Y)$. Suppose that A has a continuous inverse, and let $m = 1/\|A^{-1}\|$.
    a. If $\|A - B\| < m/2$, then $\mathscr{R}(A)$ is not a proper subset of $\mathscr{R}(B)$. [*Hint.* Assume the contrary, and use Riesz's lemma, where θ is chosen so that $\|A - B\|/(m - \|A - B\|) < \theta < 1$. See problem 4 of § 1.]

b. The set $\{A \in L(X, Y): A \text{ is in state } \mathrm{III}_1\}$ is open in $L(X, Y)$.
c. The set $\{A \in L(X, Y): A \text{ is in state } \mathrm{I}_3\}$ is also open in $L(X, Y)$.
d. If X and Y are only normed linear spaces and $\|A - B\| < m/2$, then it can be shown that $\overline{\mathscr{R}(A)}$ is not a proper subset of $\overline{\mathscr{R}(B)}$.

IV.11 ADJOINT OPERATORS

Throughout this section X and Y will denote arbitrary Hilbert spaces. We use the same symbol (parentheses) for the inner product in both spaces. If $T \in L(X, Y)$, we have defined T' so that $T' \in L(Y', X')$. However, because of the Fréchet–Riesz representation theorem (Theorem III.5.1), it is possible to identify X' with X and Y' with Y. In so doing, we are led to an operator belonging to $L(Y, X)$, which we may consider in place of T'. A similar procedure applies to unbounded operators whose domains are dense in X. To be precise, let E_X be the operator that associates to each $y \in X$ the linear functional $x \mapsto (x, y)$. Then E_X is a "conjugate-linear" isometry of X onto X' (problem 8 of § III.5). Let E_Y be the corresponding isometry of Y onto Y'.

Definition. Let T be a densely defined linear operator from X into Y. The *adjoint* of T is the operator T^* defined by

$$T^* = E_X^{-1} T' E_Y, \tag{11-1}$$

where the domain of T^* is the set of all y for which $(E_X^{-1} T' E_Y)(y)$ is defined.

Clearly $\mathscr{D}(T^*) = \{y : E_Y(y) \in \mathscr{D}(T')\} = \{y : x \mapsto (Tx, y) \text{ is continuous on } \mathscr{D}(T)\}$. By the Fréchet–Riesz representation theorem, we see that

$y \in \mathscr{D}(T^*)$ if and only if there
exists some $w \in X$ such that
$(Tx, y) = (x, w)$ for all $x \in \mathscr{D}(T)$.

Such a w, if it exists, must be unique (since $\overline{\mathscr{D}(T)} = X$) and, then, $w = T^*y$ and

$$(Tx, y) = (x, T^*y), \qquad x \in \mathscr{D}(T),\ y \in \mathscr{D}(T^*). \tag{11-2}$$

It should be clear from (11-1) that T^* is linear (since its definition involves *two* conjugate-linear mappings) and that if $T \in L(X, Y)$, then $T^* \in L(Y, X)$ and (see Theorem 8.2)

$$\|T^*\| = \|T\|. \tag{11-3}$$

In general, if $\overline{\mathscr{D}(T)} = X$, T^* is a closed linear operator. While this fact is a consequence of (11-1) and Theorem 8.3, it may also be established directly from (11-2). Using the continuity of the inner product (Theorem II.6.3), we have

$$(Tx, y) = \lim_{n\to\infty} (Tx, y_n) = \lim_{n\to\infty} (x, T^*y_n) = (x, z)$$

whenever $\{y_n\}$ is a sequence in $\mathscr{D}(T^*)$ such that $y_n \to y$ and $T^*y_n \to z$; from this it follows that T^* is closed.

Properties of Adjoints

The following formal properties of adjoints are easily verified, either by using (11-1) and properties of conjugates of operators or by arguing directly from (11-2):

$$(S+T)^* = S^* + T^*, \qquad (\alpha S)^* = \bar{\alpha} S^*, \qquad 0^* = 0,$$

if $S, T \in L(X, Y)$. If I is the identity operator on X, then $I^* = I$. If $T \in L(X, Y)$ and $S \in L(Y, Z)$, where Z is also a Hilbert space, then $ST \in L(X, Z)$ and

$$(ST)^* = T^*S^*. \tag{11-4}$$

We write T^{**} for $(T^*)^*$. When $T \in L(X, Y)$, it follows from (11-2) that

$$T^{**} = T. \tag{11-5}$$

This fact is needed when proving the following useful theorem.

Theorem 11.1. *If $T \in L(X, Y)$, then*

$$\|T^*T\| = \|TT^*\| = \|T^*\|\,\|T\| = \|T\|^2.$$

Proof. For $x \in X$, we use (11-2) and the Cauchy–Schwarz inequality to obtain $\|Tx\|^2 = (Tx, Tx) = (x, T^*Tx) \leqslant \|T^*T\|\,\|x\|^2$, whence $\|T\|^2 \leqslant \|T^*T\| \leqslant \|T^*\|\,\|T\| = \|T\|^2$, by (11-3). Thus $\|T^*T\| = \|T^*\|\,\|T\| = \|T\|^2$. Replacing T by T^*, we have $\|TT^*\| = \|T^{**}T^*\| = \|T^*\|^2 = \|T\|^2$. □

Algebraic rules for adjoints of unbounded operators are more complicated because they involve the domains of the operators in question. This situation is examined in problem 1.

Example 1. Let X be the complex space $L^2(a, b)$, and let $K \in L(X)$ be an integral operator with an $\mathscr{L}^2$ kernel $k(s, t)$ (see § 3). We shall use (11-2) to compute K^*. Given $x, y \in X$,

$$\begin{aligned}(x, K^*y) = (Kx, y) &= \int_a^b \left[\int_a^b k(s, t)x(t)\,dt\right]\overline{y(s)}\,ds \\ &= \int_a^b x(t)\left[\int_a^b k(s, t)\overline{y(s)}\,ds\right]dt \\ &= \int_a^b x(t)\overline{\left[\int_a^b \overline{k(s, t)}y(s)\,ds\right]}\,dt \\ &= \int_a^b x(s)\overline{\left[\int_a^b \overline{k(t, s)}y(t)\,dt\right]}\,ds.\end{aligned} \tag{11-6}$$

The justification for interchanging the order of integration comes from the Schwarz inequality and Tonelli's theorem. (The details of this argument are similar to those given in § 8.) Since (11-6) holds for all $x \in X$, we conclude that

$$(K^*y)(s) = \int_a^b \overline{k(t, s)}y(t)\,dt.$$

It is interesting to note that if $k(s, t) = \overline{k(t, s)}$ for all $(s, t) \in [a, b] \times [a, b]$; then K^* is exactly the same operator as K. Examples of kernels with this property are

$$\sin st, \qquad e^{i(s-t)}, \qquad \text{and} \qquad (s-t)^2.$$

A linear operator K such that $K^* = K$ is said to be self-adjoint. We shall spend the major portion of Chapter VI studying self-adjoint operators. Examples of adjoints of unbounded operators will be given there.

Orthogonal Complements, Ranges, and Null Spaces

The relations between annihilators, ranges, and null spaces, as set forth in § 8, have counterparts in relations among orthogonal complements, ranges, and null spaces. The situation is simpler than that of § 8, owing to the identification of X' and X. A Hilbert space is reflexive, and every closed subspace of X' is weak*-closed. We list the following results, leaving verification to the reader (cf. problem 4 of § III.7 and problem 2). We assume here that T is a densely defined closed linear operator from X into Y.

$$\mathscr{R}(T)^\perp = \mathscr{N}(T^*), \qquad \overline{\mathscr{R}(T)} = \mathscr{N}(T^*)^\perp \tag{11-7}$$

$$\mathscr{R}(T^*)^1 = \mathscr{N}(T), \qquad \overline{\mathscr{R}(T^*)} = \mathscr{N}(T)^\perp \tag{11-8}$$

Theorem 11.2. *If $T \in L(X, Y)$, then*

(*a*) $\mathscr{N}(T) = \mathscr{N}(T^*T), \qquad \mathscr{N}(T^*) = \mathscr{N}(TT^*)$

(*b*) $\overline{\mathscr{R}(T)} = \overline{\mathscr{R}(TT^*)}, \qquad \overline{\mathscr{R}(T^*)} = \overline{\mathscr{R}(T^*T)}$

(*c*) *If any one of the four subspaces $\mathscr{R}(T)$, $\mathscr{R}(T^*)$, $\mathscr{R}(TT^*)$, $\mathscr{R}(T^*T)$ is closed, then so are the others.*

Proof. (*a*) Obviously $\mathscr{N}(T) \subset \mathscr{N}(T^*T)$. If $T^*Tx = 0$, then $(Tx, Tx) = (T^*Tx, x) = 0$, and so $Tx = 0$. Hence $\mathscr{N}(T^*T) = \mathscr{N}(T)$. The second part of (*a*) follows by substituting T^* for T.

(*b*) Observe that $(TT^*)^* = T^{**}T^* = TT^*$. Hence, by (11-7) and part (*a*), $\overline{\mathscr{R}(T)} = \mathscr{N}(T^*)^\perp = \mathscr{N}(TT^*)^\perp = \overline{\mathscr{R}(TT^*)}$. The second part follows by substituting T^* for T.

(*c*) The closed range theorem (Theorem 10.1) implies that $\mathscr{R}(T)$ is closed if and only if $\mathscr{R}(T^*)$ is closed. It thus suffices to show that $\mathscr{R}(T)$ is closed if and only if $\mathscr{R}(TT^*)$ is closed, for the other implication will follow by substituting T^* for T. First, suppose that $\mathcal{R}(T)$ is closed. Then $\mathcal{R}(T^*)$ is closed, and hence $X = \mathcal{N}(T) \oplus \mathcal{R}(T^*)$, by (11-8) and Theorem II.7.4. It follows that $\mathcal{R}(T) = T(X) = T(\mathcal{R}(T^*)) = \mathcal{R}(TT^*)$, and so $\mathcal{R}(TT^*)$ is closed. On the other hand, if $\mathcal{R}(TT^*)$ is closed, then by (*b*), $\overline{\mathcal{R}(T)} = \mathcal{R}(TT^*) \subset \mathcal{R}(T)$, and so $\mathcal{R}(T)$ is closed. $\square$

States of T and T^*

Let us now recall from § 10 the terminology about the nine possible states of a densely defined closed linear operator T from X into Y. It is easy to see from (11-1) that the state of T^* is always the same as the state of T'. As a consequence, T^* may replace T' in the state diagram of § 10. Since X and Y are reflexive, there are just seven possible states for the pair (T, T^*); the state of either T or T^* completely determines the state of the other.

An operator T in $L(X)$ is said to be *normal* if $TT^* = T^*T$. Clearly a self-adjoint operator in $L(X)$ is normal, but the converse is not true. For a normal operator the number of possible states is severely limited.

Theorem 11.3. *For a normal operator T in $L(X)$ the only possible states are* I_1, II_2, *and* III_3. *The state of T^* is the same as that of T.*

Proof. Since T is normal, we have $\overline{\mathcal{R}(T)} = \overline{\mathcal{R}(TT^*)} = \overline{\mathcal{R}(T^*T)} = \overline{\mathcal{R}(T^*)}$, by Theorem 11.2. It follows that neither of the operators T, T^* can be in a III state unless the other is also. The conclusions of the theorem follow from an examination of the state diagram. $\square$

The results of this section will be used in Chapter VI.

PROBLEMS

In the problems below, X and Y are Hilbert spaces and S, T are densely defined closed linear operators from X into Y.

1. a. If $\mathcal{D}(S+T) \equiv \mathcal{D}(S) \cap \mathcal{D}(T)$ is dense in X, then $(S+T)^*$ is an extension of $S^* + T^*$.
   b. If $X = Y$ and I is the identity on X, then $(\alpha I + T)^* = \bar{\alpha} I + T^*$, for any scalar α.
   c. For any scalar α, $(\alpha T)^* = \bar{\alpha} T^*$.
   d. If V is a densely defined closed linear operator from Y into a Hilbert space Z and if the domain of VT, $\{x \in \mathcal{D}(T) : Tx \in \mathcal{D}(V)\}$, is dense in X, then $(VT)^*$ is an extension of T^*V^*.
2. a. Show that $\mathcal{D}(T^*)$ is dense in Y, so that $(T^*)^*$ is well-defined.

b. Show that $T^{**} = T$. [*Hint*. The product space $X \times Y$ is a Hilbert space when an inner product is defined by

$$([x_1, y_1], [x_2, y_2]) = (x_1, x_2) + (y_1, y_2),$$

for $[x_i, y_i] \in X \times Y$. It is easy to see that the graph of T^{**}, $G(T^{**})$, contains the graph of $T, G(T)$. If $[x_0, T^{**}x_o] \in G(T^{**})\backslash G(T)$, then there exists $[u, v] \in G(T^{**})$ such that $([x_0, T^{**}x_0], [u, v]) \neq 0$, and yet $([x, Tx], [u, v]) = 0$ for all $x \in \mathscr{D}(T)$. (Why?)]

3. Verify (11-7), and use problem 2 to verify (11-8).
4. $\mathscr{R}(T) = Y$ if and only if T^* has a continuous inverse. $\mathscr{R}(T^*) = X$ if and only if T has a continuous inverse.
5. Given $T \in L(X)$, there exists $S \in L(X)$ such that $ST = I$ if and only if T has a continuous inverse.
6. Suppose that A is a linear operator with $\mathscr{D}(A) = X$, $\mathscr{R}(A) \subset X$, but do not suppose that A is closed. If $(Ax, y) = (x, Ay)$ for all $x, y \in X$, then A is continuous.

IV.12 PROJECTIONS

Operators of a special type, called projections, play an important role in the systematic study of linear operators. The main topic of this section is the connection between projections and direct sums of linear manifolds.

Let X be a linear space. A linear operator P with domain X and range in X is called a *projection* (of X) if $P^2 = P$. Each projection P determines a direct sum decomposition of X (cf. § I.6), namely,

$$X = \mathscr{R}(P) \oplus \mathscr{N}(P). \tag{12-1}$$

It is clear that $X = \mathscr{R}(P) + \mathscr{N}(P)$, since each x in X may be written in the form $x = Px + (x - Px)$. Furthermore, elements x of $\mathscr{R}(P)$ are characterized by the fact that $Px = x$. So, if $x \in \mathscr{R}(P) \cap \mathscr{N}(P)$, then $x = Px = 0$; that is, $\mathscr{R}(P) \cap \mathscr{N}(P) = (0)$. This proves (12-1)

Conversely, every direct sum decomposition of X determines a projection. Indeed, if M_1 and M_2 are complementary linear manifolds in X such that

$$X = M_1 \oplus M_2,$$

then each element x in X may be written uniquely in the form $x = x_1 + x_2$, with $x_1 \in M_1$ and $x_2 \in M_2$ (cf. § I.6). If we define P by $Px = x_1$, then it is clear that P is a linear operator such that $\mathscr{R}(P) = M_1$, $\mathscr{N}(P) = M_2$, and $P^2 = P$. We call P the *projection of* X *onto* M_1 *along* M_2. The operator $I - P$ is also a projection—the projection of X onto M_2 along M_1.

Thus far nothing has been said about topology in connection with projections.

Theorem 12.1. *Let X be a Hausdorff topological linear space, and let P be a continuous projection of X. Then $\mathcal{R}(P)$ and $\mathcal{N}(P)$ are closed.*

Proof. Since (0) is closed, so is the inverse image $\{x : Px = 0\} = \mathcal{N}(P)$. Similarly, $\mathcal{R}(P)$ is the null space of the continuous operator $I - P$. □

Theorem 12.2. *Let X be a Banach space, and let M_1 and M_2 be closed subspaces such that $X = M_1 \oplus M_2$. Then the projection P of X onto M_1 along M_2 is continuous.*

Proof. Because of the closed graph theorem it suffices to prove that P is a closed operator. Suppose $x_n \to x$ and $Px_n \to y$. Then $x_n - Px_n \to x - y$. Since $Px_n \in M_1$ and $x_n - Px_n \in M_2$, it follows that $y \in M_1$ and $x - y \in M_2 = \mathcal{N}(P)$. Then $Px - Py = 0$, and $Px = Py = y$. Thus P is closed. □

The decomposition $X = M_1 \oplus M_2$ in Theorem 12.2 is said to be a *topological* direct sum because M_1 and M_2 are both closed subspaces.

It follows from Theorems 12.1 and 12.2 that if M_1 is a closed subspace of a Banach space X, then there is a one-to-one correspondence between continuous projections onto M_1 and closed subspaces complementary to M_1. However, there may be no continuous projection onto M_1. In fact, J. Lindenstrauss and I. Tzafriri [1] have proved that in every Banach space that is not topologically isomorphic to a Hilbert space, there exists a closed subspace that is not the range of a continuous projection. (Cf. Kadets and Mityagin [1] and Pelczynski [1].) For a concrete example of this, take $X = \ell^\infty$ and $M_1 = (c_0)$. (See Whitley [2].)

Theorem 12.3. *Let M be a finite-dimensional subspace of a Hausdorff locally convex space X. Then there exists a continuous projection of X onto M.*

Proof. Let $x_1, \ldots, x_n$ be a basis for M, and let M_i be the $(n-1)$-dimensional subspace generated by $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$. Then each M_i is closed. (For normed linear spaces this is Theorem II.3.3. For the general case we must appeal to problem 9 of § II.9.) By the Hahn–Banach theorem (Theorem III.2.7), there exist $x_i' \in X'$ such that $x_i'(x_i) = 1$ and $x_i'(x) = 0$ for all $x \in M_i$. It is readily checked that the operator P defined by

$$Px = \sum_{i=1}^{n} x_i'(x) x_i,$$

is a continuous projection of X onto M. □

Theorem 12.4. *Let X be a Banach space, and let P and Q be projections in $L(X)$ such that $\|P - Q\| < 1$. Then $\mathcal{R}(P)$ is isomorphic to $\mathcal{R}(Q)$. In fact, the* *operator T defined by*

$$T = QP + (I - Q)(I - P)$$

is bijective and maps $\mathcal{R}(P)$ onto $\mathcal{R}(Q)$ and $\mathcal{N}(P)$ onto $\mathcal{N}(Q)$.

Proof. Let $S = PQ + (I - P)(I - Q)$. A simple computation shows that

$$ST = TS = I - P - Q + PQ + QP = I - (P - Q)^2.$$

Since $\|(P-Q)^2\| \leq \|P-Q\|^2 < 1$, it follows from Theorem 1.4 that ST and TS are both bijective. Hence T must be bijective. Now it is clear that T maps $\mathcal{R}(P)$ into $\mathcal{R}(Q)$ and $\mathcal{R}(I-P)$ into $\mathcal{R}(I-Q)$. Since $X = \mathcal{R}(P) \oplus \mathcal{R}(I-P)$ and $X = \mathcal{R}(Q) \oplus \mathcal{R}(I-Q)$ and since T is bijective, it follows easily that T maps $\mathcal{R}(P)$ *onto* $\mathcal{R}(Q)$ and $\mathcal{R}(I-P)$ *onto* $\mathcal{R}(I-Q)$. □

The connection between direct sums and projections extends easily to finite families of linear manifolds. Let $M_1, \ldots, M_n$ $(n \geq 2)$ be linear manifolds in a linear space X. We say that this family is linearly independent if no M_i contains a nonzero vector that is in the subspace determined by the remaining $n-1$ linear manifolds. An equivalent condition is that, if $x_i \in M_i$ and $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$, then $x_i = 0$ if $\alpha_i \neq 0$. The subspace generated by the elements of $M_1 \cup \cdots \cup M_n$ is denoted by $M_1 \oplus \cdots \oplus M_n$ and called the *direct sum* of $M_1, \ldots, M_n$. Elements x of the direct sum are representable uniquely in the form

$$x = x_1 + \cdots + x_n, \qquad x_i \in M_i. \tag{12-2}$$

Whenever we use the $\oplus$ notation, it is to be understood implicitly that the manifolds in question are linearly independent.

If $X = M_1 \oplus \cdots \oplus M_n$, the representation (12-2) determines projections $P_1, \ldots, P_n$, defined on X by $P_i x = x_i$. These projections satisfy the relations

$$\begin{aligned} &P_i^2 = P_i, \qquad P_i P_j = 0 \qquad \text{for } i \neq j, \\ &I = P_1 + \cdots + P_n. \end{aligned} \tag{12-3}$$

Conversely, if the linear operators $P_1, \ldots, P_n$ are given with domain X and ranges $M_1, \ldots, M_n$ in X and if they satisfy the conditions (12-3), then it is readily checked that $X = M_1 \oplus \cdots \oplus M_n$.

Theorem 12.5. *Let X be a Banach space, let $M_1, \ldots, M_n$ be linear manifolds in X such that $X = M_1 \oplus \cdots \oplus M_n$, and let $P_1, \ldots, P_n$ be the associated projections. Then $M_1, \ldots, M_n$ are closed if and only if $P_1, \ldots, P_n$ are continuous.*

Proof. If the P_i are continuous, then the M_i are closed because $M_i = \mathcal{N}(I - P_i)$. Now suppose that the M_i are closed. Define a new norm on X by

$$|||x||| = \|P_1 x\| + \cdots + \|P_n x\|.$$

(This is just a product norm on the Cartesian product $M_1 \times \cdots \times M_n$.) Since each M_i is closed, it is easy to see that X is complete under the new norm. Furthermore,

$$\|x\| = \|P_1 x + \cdots + P_n x\| \leq \|P_1 x\| + \cdots + \|P_n x\| = |||x|||$$

for each x in X. Thus the identity mapping from $(X, ||| \cdot |||)$ onto $(X, \| \cdot \|)$ is continuous. Then the inverse mapping is closed, and hence is continuous, by the closed graph theorem. Consequently, there exists $m > 0$ such that

$$|||x||| \leq m\|x\|, \qquad x \in X.$$

But then

$$\|P_i x\| \leq \sum_{j=1}^{n} \|P_j x\| = |||x||| \leq m\|x\|,$$

and so each P_i is continuous. □

Projections in Inner-Product Spaces

Let X be an inner-product space. We say that two sets M and N in X are *orthogonal* if $(x, y) = 0$ whenever $x \in M$ and $y \in N$. If the range and null space of a projection P are orthogonal, we call P an *orthogonal projection.*

Theorem 12.6. *An orthogonal projection P of an inner-product space X is continuous. If $P \neq 0$, then $\|P\| = 1$.*

Proof. Given $x \in X$, the orthogonality of P implies $x = Px + u$, where $(u, Px) = 0$. Then $\|x\|^2 = \|Px\|^2 + \|u\|^2$, so that $\|Px\|^2 \leq \|x\|^2$; therefore P is continuous and $\|P\| \leq 1$. From $P = P^2$ we see that $\|P\| \leq \|P\|^2$, whence $1 \leq \|P\|$ if $P \neq 0$. □

Theorems 12.6 and 12.1 imply that if P is an orthogonal projection, the subspaces in the decomposition $X = \mathscr{R}(P) \oplus \mathscr{N}(P)$ are closed. In fact, we claim that $\mathscr{N}(P) = \mathscr{R}(P)^\perp$ and $\mathscr{R}(P) = \mathscr{N}(P)^\perp$. Clearly $\mathscr{N}(P) \subset \mathscr{R}(P)^\perp$ because P is orthogonal. If $x \in \mathscr{R}(P)^\perp$ and if we write $x = u + v$ where $u \in \mathscr{R}(P)$ and $v \in \mathscr{N}(P)$, then $v \in \mathscr{R}(P)^\perp$, and hence $u = x - v \in \mathscr{R}(P)^\perp$. This implies that $u = 0$ and $x = v$. Consequently, $\mathscr{N}(P) = \mathscr{R}(P)^\perp$. Applying this fact to $I - P$, we have $\mathscr{R}(P) = \mathscr{N}(I - P) = \mathscr{R}(I - P)^\perp = \mathscr{N}(P)^\perp$.

The relation $\mathscr{N}(P) = \mathscr{R}(P)^\perp$ shows that there can be at most one orthogonal projection P onto a given closed subspace M, for in this case $\mathscr{N}(P)$ must be $M^\perp$. When X is a Hilbert space such a projection must always exist. This fact is the essential content of Theorems II.7.2 and II.7.4. (Observe that the correspondence $x \mapsto x_M$ of Theorem II.7.2 defines the desired orthogonal projection.) Thus there is a one-to-one correspondence between the closed subspaces of a Hilbert space and the orthogonal projections onto them.

The following characterization of orthogonal projections will be needed in Chapter VI.

Theorem 12.7. *A projection P in $L(X)$ is orthogonal if and only if*

$$(Px, y) = (x, Py) \qquad \text{for } x, y \in X. \tag{12-4}$$

Proof. Write $x = Px + u$, $y = Py + v$, with $u, v \in \mathcal{N}(P)$. Then

$$(Px, y) = (Px, Py) + (Px, v),$$
$$(x, Py) = (Px, Py) + (u, Py).$$

If $\mathcal{R}(P)$ and $\mathcal{N}(P)$ are orthogonal, we see that $(Px, y) = (Px, Py) = (x, Py)$. On the other hand, if (12-4) holds and $x \in \mathcal{R}(P)$, $y \in \mathcal{N}(P)$, then $Px = x$, $Py = 0$, so that $(x, y) = (Px, y) = (x, Py) = (x, 0) = 0$, and hence P is orthogonal. □

An operator P that satisfies (12-4) is said to be *symmetric*. When X is complete and $P \in L(X)$, this condition is equivalent to P being self-adjoint, as we see from (11-2). (Cf. problem 6, § 11.)

A family of linear manifolds is called an orthogonal family if each pair of distinct manifolds from the family are orthogonal. If $M_1, \ldots, M_n$ is such an orthogonal family, then the M_i are linearly independent. Indeed, if $\alpha_1 x_1 + \cdots + \alpha_n x_n = 0$, with $x_i \in M_i$, the orthogonality shows that $0 = (\alpha_1 x_1 + \cdots + \alpha_n x_n, x_i) = \alpha_i \|x_i\|^2$, and hence $x_i = 0$ if $\alpha_i \neq 0$. Furthermore, if X is complete and the M_i are closed, then the direct sum $M_1 \oplus \cdots \oplus M_n$ is necessarily closed. This follows from the next theorem.

Theorem 12.8. *Let M and N be closed subspaces of a Hilbert space X, and let P_M and P_N be the orthogonal projections of X onto M and N, respectively. If M and N are orthogonal, then $M \oplus N$ is closed, and $P_M + P_N$ is the orthogonal projection of X onto $M \oplus N$ along $M^\perp \cap N^\perp$.*

Proof. The completeness of X is needed only to guarantee the existence of the orthogonal complements $M^\perp, N^\perp$ and the projections P_M and P_N. Since M and N are orthogonal, we have $N \subset M^\perp$ and $M \subset N^\perp$; that is, $\mathcal{R}(P_N) \subset \mathcal{N}(P_M)$ and $\mathcal{R}(P_M) \subset \mathcal{N}(P_N)$. Thus

$$P_M P_N = P_N P_M = 0.$$

From this it is easily verified that $P_M + P_N$ is a projection of X onto $M \oplus N$ along $M^\perp \cap N^\perp$ (cf. problem 1). Thus $M \oplus N$ is closed. Finally, $P_M + P_N$ is an orthogonal projection, because $M^\perp \cap N^\perp$ is obviously orthogonal to $M \oplus N$. □

Pseudoinverses of Linear Operators

As an application of Theorems 12.1 and 12.2, we shall characterize the class of all T in $L(X, Y)$ such that there exist continuous projections of X and Y onto $\mathcal{N}(T)$ and $\mathcal{R}(T)$, respectively. This result will be used in § 13.

Let X and Y be Banach spaces. Given T in $L(X, Y)$, we say that an operator S in $L(Y, X)$ is a *pseudoinverse* of T if

$$TST = T. \tag{12-5}$$

Obviously, if T has an inverse T^{-1} in $L(Y, X)$, then T^{-1} is a pseudoinverse of T. In this case T^{-1} is the only pseudoinverse for, if we multiply both sides of (12-5) on the left and right by T^{-1}, we obtain $S = T^{-1}$. In general, a pseudoinverse provides a substitute for a true inverse when solving the equation

$$Tx = y.$$

Indeed, if at least one solution x exists, then Sy is also a solution, since

$$T(Sy) = TS(Tx) = Tx = y.$$

Theorem 12.9. *Let X and Y be Banach spaces, and suppose $T \in L(X, Y)$. Then the following conditions are equivalent:*

(a) *There exist projections $P \in L(X)$ and $Q \in L(Y)$ such that*

$$\mathcal{R}(P) = \mathcal{N}(T), \qquad \mathcal{R}(Q) = \mathcal{R}(T); \tag{12-6}$$

(b) *There exist closed subspaces W and Z such that*

$$X = \mathcal{N}(T) \oplus W, \qquad Y = Z \oplus \mathcal{R}(T); \tag{12-7}$$

(c) *T has a pseudoinverse.*

Proof. It is clear that (a) implies (b). Suppose (b) holds, and let T_0 be the restriction of T to W. Then T_0 is a bijective map of W onto $\mathcal{R}(T)$ (Theorem I.6.3). Furthermore, $\mathcal{R}(T)$ is a Banach space, by Theorem 5.10. Thus T_0^{-1} must be continuous, by Corollary 5.6. Let Q be the (continuous) projection of Y onto $\mathcal{R}(T)$ along Z. It is easy to see that $T_0^{-1}Q$ is a pseudoinverse of T. Thus (b) implies (c). Finally, suppose S is a pseudoinverse of T. Then ST is obviously a projection and $\mathcal{N}(ST) \supset \mathcal{N}(T)$. Also, if $STx = 0$, then $Tx = T(STx) = 0$, which proves that $\mathcal{N}(ST) = \mathcal{N}(T)$. Likewise, (12-5) implies that TS is a projection and $\mathcal{R}(TS) = \mathcal{R}(T)$. So ($c$) implies ($a$), with $P = I - ST$ and $Q = TS$. □

When X and Y are Hilbert spaces and T is an operator in $L(X, Y)$ with closed range, then we may take $W = \mathcal{N}(T)^{\perp}$ and $Z = \mathcal{R}(T)^{\perp}$ in (12-7). In this case the pseudoinverse of T constructed in the proof of Theorem 12.9 is commonly known as the *Moore–Penrose inverse* (cf. Ben-Israel and Greville

[1]). Moore–Penrose inverses occur in numerous applications, particularly in matrix theory and statistics. See Nashed [1].

PROBLEMS

1. Let P and Q be projections of a linear space X.
   a. If $PQ = QP$, then PQ is a projection and
   $$\mathcal{R}(PQ) = \mathcal{R}(P) \cap \mathcal{R}(Q)$$
   $$\mathcal{N}(PQ) = \mathcal{N}(P) + \mathcal{N}(Q).$$
   b. $P+Q$ is a projection if and only if $PQ = QP = 0$. In this case
   $$\mathcal{R}(P+Q) = \mathcal{R}(P) \oplus \mathcal{R}(Q)$$
   $$\mathcal{N}(P+Q) = \mathcal{N}(P) \cap \mathcal{N}(Q).$$
2. Let M be a closed subspace of a Banach space X such that the dimension of the quotient space X/M is finite. Then every projection of X onto M is continuous.
3. Suppose X is a Banach space, $M_1, \ldots, M_n$ are closed and linearly independent subspaces of X, and $M = M_1 \oplus \cdots \oplus M_n$. Let $P_1, \ldots, P_n$ be the associated projections of M onto $M_1, \ldots, M_n$, respectively. Then M is closed if and only if $P_1, \ldots, P_n$ are continuous.
4. If X is a Banach space and M_1, M_2 are closed and linearly independent subspaces of X, then $M_1 \oplus M_2$ is closed if and only if there exists a $d > 0$ such that $\|x_1 - x_2\| \geq d$ whenever $x_1 \in M_1$, $x_2 \in M_2$ and $\|x_1\| = \|x_2\| = 1$.
5. Let P and Q be continuous projections of a Banach space X such that $\|P - Q\| < 1$, and let $T = I - P + QP$. Then T is bijective, maps $\mathcal{R}(P)$ onto $\mathcal{R}(Q)$ and $Tx = x$ for $x \in \mathcal{N}(P)$. Hence $X = \mathcal{R}(Q) \oplus \mathcal{N}(P)$.
6. Let P be a continuous projection of a normed linear space X. Let $M = \mathcal{R}(P)$ and $N = \mathcal{N}(P)$. Then P' is a continuous projection of X' with $\mathcal{R}(P') = N^\perp$ and $\mathcal{N}(P') = M^\perp$. Hence $X' = N^\perp \oplus M^\perp$.
7. Let X be a normed linear space. If P is a continuous projection of X having a finite-dimensional range, then P' also has a finite-dimensional range.
8. If P is an orthogonal projection of a Hilbert space X onto a subspace M, then
   $$\|(I-P)x\| = \operatorname{dist}(x, M), \qquad x \in X.$$
9. Let P be a continuous projection on a Hilbert space X, with $\|P\| \leq 1$. Then P is an orthogonal projection. [*Hint.* If $x \in \mathcal{N}(P)^\perp$, then $\|Px\|^2 = \|x\|^2 + \|Px - x\|^2$. Deduce that $\mathcal{N}(P)^\perp \subset \mathcal{R}(P)$. Given $x \in \mathcal{R}(P)$, write $x = u + v$ with $u \in \mathcal{N}(P)^\perp$, $v \in \mathcal{N}(P)$. Then $x = Px = Pu = u$, whence $\mathcal{R}(P) \subset \mathcal{N}(P)^\perp$.]
10. Let P be a continuous projection on a Hilbert space X such that $P^*P = PP^*$. Then P is an orthogonal projection.
11. Let M and N be closed subspaces of a Hilbert space X, and let P_M and P_N be the orthogonal projections of X onto M and N, respectively. Show that $M \subset N$ if and only if $P_N P_M = P_M$. In this case, show also that $P_M P_N = P_M$.

12. Let X and Y be Hilbert spaces, and let T be an operator in $L(X, Y)$ with closed range. Let S be the Moore–Penrose inverse of T (see page 251).
   a. For any $y \in Y$, the element Sy is the best approximate solution of the equation $Tx = y$ in the sense that $\|T(Sy) - y\| = \inf\{\|Tx - y\| : x \in X\}$.
   b. For any $y \in \mathscr{R}(T)$, the element Sy is the best solution of the equation $Tx = y$ in the sense that $\|Sy\| = \inf\{\|x\| : Tx = y\}$.

IV.13 FREDHOLM OPERATORS

When X and Y are Banach spaces, we say that an operator T in $L(X, Y)$ is a *Fredholm operator* if $\dim \mathscr{N}(T) < \infty$ and $\dim Y/\mathscr{R}(T) < \infty$, and we denote the class of such operators by $\Phi(X, Y)$. If $X = Y$, we write simply $\Phi(X)$. Every bijective operator in $L(X, Y)$ is clearly a Fredholm operator. In § V.7 we shall show that if $I - K$ is the operator associated with a Fredholm integral equation of the second kind (cf. § 2), then $I - K \in \Phi(X)$. This fact accounts for our use of the term "Fredholm operator."

Given T in $L(X, Y)$, we let $n(T) = \dim \mathscr{N}(T)$ and $d(T) = \dim Y/\mathscr{R}(T)$. Note that $d(T)$ equals the dimension of any subspace of Y complementary to $\mathscr{R}(T)$ (cf. Theorem I.6.5). We call $n(T)$ the *nullity* of T and $d(T)$ the *defect* of T. If at least one of these two quantities is finite, we define the *index* of T to be the extended integer

$$\kappa(T) = n(T) - d(T).$$

Theorem 13.1. *Let X, Y, Z be Banach spaces.*
(*a*) *If $T \in \Phi(X, Y)$ and $S \in \Phi(Y, Z)$, then $ST \in \Phi(X, Z)$ and*

$$\kappa(ST) = \kappa(S) + \kappa(T). \tag{13-1}$$

(*b*) *If $T \in L(X, Y)$, $S \in L(Y, Z)$ and $ST \in \Phi(X, Z)$, then $T \in \Phi(X, Y)$ if and only if $S \in \Phi(Y, Z)$.*

Proof. If $T \in L(X, Y)$ and $S \in L(Y, Z)$, then of course $ST \in L(X, Z)$, From here on the proof is entirely algebraic. We rely heavily on the results of § I.6. Let Y_1 be a linear manifold in $\mathscr{R}(T)$ complementary to $\mathscr{R}(T) \cap \mathscr{N}(S)$ (in the space $\mathscr{R}(T)$). Similarly, let Y_2 and Y_3 be linear manifolds in $\mathscr{N}(S)$ and Y, respectively, such that

$$Y = \overbrace{Y_1 \oplus \underbrace{[\mathscr{R}(T) \cap \mathscr{N}(S)] \oplus Y_2}_{\mathscr{N}(S)}}^{\mathscr{R}(T)} \oplus Y_3. \tag{13-2}$$

The existence of Y_1, Y_2, Y_3 follows from suitable applications of Theorem I.11.2. Now $\mathscr{N}(ST) = \{x : Tx \in \mathscr{N}(S)\} = \{x : Tx \in \mathscr{R}(T) \cap \mathscr{N}(S)\}$. Thus the restriction of T to $\mathscr{N}(ST)$ maps $\mathscr{N}(ST)$ onto $\mathscr{R}(T) \cap \mathscr{N}(S)$. From this it follows that $\dim \mathcal{N}(ST) = \dim \mathcal{N}(T) + \dim [\mathcal{R}(T) \cap \mathcal{N}(S)]$. (Apply Theorem I.6.4 with X replaced by $\mathcal{N}(ST)$.) That is,

$$n(ST) = n(T) + \dim [\mathcal{R}(T) \cap \mathcal{N}(S)]. \tag{13-3}$$

Since $Y_1 \oplus Y_3$ is a complement of $\mathcal{N}(S)$ in Y, it follows that SY_1 and SY_3 are linearly independent and

$$\mathcal{R}(S) = SY_1 \oplus SY_3 = \mathcal{R}(ST) \oplus SY_3.$$

(Cf. Theorem I.6.3.) Let Z_1 be a complement of $\mathcal{R}(S)$ in Z. Then

$$Z = Z_1 \oplus \mathcal{R}(S) = Z_1 \oplus \mathcal{R}(ST) \oplus SY_3.$$

Hence $d(ST) = \dim Z_1 + \dim SY_3 = d(S) + \dim SY_3$. However, $\dim SY_3 = \dim Y_3$, because S is a one-to-one mapping of Y_3 onto SY_3. Thus we have

$$d(ST) = d(S) + \dim Y_3. \tag{13-4}$$

Using (13-3) and (13-2), we see that

$$\begin{aligned} n(ST) + d(T) &= n(T) + \dim [\mathcal{R}(T) \cap \mathcal{N}(S)] + \dim Y_2 + \dim Y_3 \\ &= n(T) + n(S) + \dim Y_3. \end{aligned}$$

(Addition is associative in the set of nonnegative integers and $+\infty$.) By adding $d(S)$ to both sides and using (13-4), we obtain

$$n(ST) + d(T) + d(S) = n(T) + n(S) + d(ST). \tag{13-5}$$

Now if T and S are Fredholm operators, then (13-3) implies that $n(ST) \leq n(T) + n(S) < \infty$. It follows from (13-5) that $d(ST) < \infty$ and $\kappa(ST) = \kappa(T) + \kappa(S)$. This proves part (*a*). To prove (*b*) we suppose that $n(ST) < \infty$ and $d(ST) < \infty$. Then (13-3) and (13-4) imply that $n(T) < \infty$ and $d(S) < \infty$. From (13–5) it is clear that, in this case, $d(T) < \infty$ if and only if $n(S) < \infty$. Hence $T \in \Phi(X, Y)$ if and only if $S \in \Phi(Y, Z)$. □

Recall from § 12 that a pseudoinverse of T is an operator S in $L(Y, X)$ such that $TST = T$. Recall also that if such an S exists, then ST and TS are continuous projections, with $\mathcal{R}(I - ST) = \mathcal{N}(ST) = \mathcal{N}(T)$ and $\mathcal{R}(TS) = \mathcal{R}(T)$. (See the proof of Theorem 12.9.)

Theorem 13.2. *If $T \in \Phi(X, Y)$, then $\mathcal{R}(T)$ is closed and T has a pseudo-inverse in $L(Y, X)$.*

Proof. Theorem 12.3 implies that $X = \mathcal{N}(T) \oplus W$ for some closed subspace W, since $\dim \mathcal{N}(T) < \infty$. Let Z be a complement of $\mathcal{R}(T)$ in Y. Then $\dim Z = \dim Y/\mathcal{R}(T) < \infty$, and so Z is a closed subspace. By Theorem 12.9, T has a pseudoinverse. □

Theorem 13.3 *(a) If $T \in \Phi(X, Y)$ and if S is any pseudoinverse of T, then $S \in \Phi(Y, X)$ and $\kappa(S) = -\kappa(T)$. (b) If $T \in \Phi(X, Y)$, then $\kappa(T) = 0$ if and only if T has a bijective pseudoinverse.*

Proof. The identity $T(ST) = T$ implies that $ST \in \Phi(X)$, by Theorem 13.1(b), and this in turn implies that $S \in \Phi(Y, X)$. The rest of (a) follows by applying Theorem 13.1(a) to the identity $TST = T$. To prove (b) observe that if S is a bijective pseudoinverse of T, then (a) implies that $\kappa(T) = -\kappa(S) = 0$. For the converse, let W be a closed complement of $\mathcal{N}(T)$ in X, and let Z be a closed complement of $\mathcal{R}(T)$ in Y. If $\kappa(T) = 0$, then $\dim Z = \dim Y/\mathcal{R}(T) = \dim \mathcal{N}(T) < \infty$, and so there exists a continuous one-to-one linear mapping F_0 of Z onto $\mathcal{N}(T)$. (Consider a basis for Z and a basis for $\mathcal{N}(T)$.) Let T_0 be the restriction of T to W, and let S be the "direct sum" of F_0 and T_0^{-1}.

$$\begin{array}{ccc} X & & Y \\ \| & & \| \\ \mathcal{N}(T) & \xleftarrow{F_0} & Z \\ \oplus & & \oplus \\ W & \xleftarrow{T^{-1}} & \mathcal{R}(T) \end{array} \tag{13-8}$$

To be precise, let $S = F_0(I - Q) + T_0^{-1}Q$, where Q is the projection of Y onto $\mathcal{R}(T)$ along Z. It is readily checked that S is bijective and is a pseudoinverse of T. □

In the next theorem we shall use the fact that if T is a linear operator from a finite-dimensional space X into itself, then $\kappa(T) = 0$. Indeed, if $X = \mathcal{R}(T) \oplus Z$, then $d(T) = \dim Z = \dim X - \dim \mathcal{R}(T) = \dim \mathcal{N}(T)$, by Theorems I.6.2 and I.6.4.

Theorem 13.4. *Let T be an operator in $\Phi(X, Y)$ of index zero. If F is in $L(X, Y)$ and* $\dim \mathcal{R}(F) < \infty$, *then $T + F \in \Phi(X, Y)$ and $\kappa(T + F) = 0$.*

Proof. First, suppose that $X = Y$ and T is the identity operator I on X. Let X_1, X_2 and X_3 be linear manifolds such that

$$X = \overbrace{X_1 \oplus [\mathcal{N}(F) \cap \mathcal{R}(F)]}^{\mathcal{N}(F)} \oplus X_2 \oplus X_3, \qquad X = X_1 \oplus \underbrace{[\mathcal{N}(F) \cap \mathcal{R}(F)] \oplus X_2}_{\mathcal{R}(F)} \oplus X_3.$$

Note that $\dim X_2 \oplus X_3 = \dim X/\mathcal{N}(F)$. Since $X/\mathcal{N}(F)$ is isomorphic to $\mathcal{R}(F)$ (Theorem I.6.6), we conclude that X_2 and X_3 are finite dimensional. Hence the space $X_4 = \mathcal{R}(F) \oplus X_3$ is finite dimensional. It is easy to see that $I + F$

maps X_4 back into itself. Let A be the restriction of $I+F$ to X_4. Then A acts in a finite-dimensional space, and so

$$n(A)-d(A)=0. \tag{13-9}$$

It is easy to see that $\mathcal{N}(I+F)\subset\mathcal{R}(F)\subset X_4$. Hence

$$n(I+F)=n(A). \tag{13-10}$$

Now $X=X_1\oplus X_4$ and $I+F$ is the identity on X_1. Also, $(I+F)X_4=\mathcal{R}(A)\subset X_4$. Thus

$$\mathcal{R}(I+F)=(I+F)X_1\oplus(I+F)X_4=X_1\oplus\mathcal{R}(A).$$

If we let M be a complement of $\mathcal{R}(A)$ in X_4, then

$$X=X_1\oplus X_4=X_1\oplus\mathcal{R}(A)\oplus M=\mathcal{R}(I+F)\oplus M.$$

Thus M is also a complement of $\mathcal{R}(I+F)$ in X; that is, $d(I+F)=d(A)$. This fact, together with (13-9) and (13-10), proves the theorem when $T=I$. For the general case we use the fact that T has a bijective pseudoinverse S. Let $F_1=I-ST$. Then F_1 is a projection of X onto $\mathcal{N}(T)$, and so $\dim\mathcal{R}(F_1)<\infty$. Then

$$T+F=S^{-1}(ST+SF)=S^{-1}(I+SF-F_1).$$

Since $\dim\mathcal{R}(SF-F_1)<\infty$, the argument above shows that $I+SF-F_1$ is a Fredholm operator of index zero. Of course, S^{-1} has index zero because S is bijective. Thus $\kappa(T+F)=0$, by Theorem 13.1. □

Theorem 13.5. *Let T be in $\Phi(X, Y)$, and let S be any pseudoinverse of T. If B is any operator in $L(X, Y)$ such that $I+SB$ is a Fredholm operator of index zero, then $T+B\in\Phi(X, Y)$ and*

$$\kappa(T+B)=\kappa(T). \tag{13-11}$$

If $I+SB$ is bijective, then

$$n(T+B)\leqslant n(T), \qquad d(T+B)\leqslant d(T). \tag{13-12}$$

Proof. Let $F=I-ST$. Then $\dim\mathcal{R}(F)=\dim\mathcal{N}(T)<\infty$ and

$$S(T+B)=I+SB-F.$$

If $I+SB$ is a Fredholm operator of index zero, the same is true for $(I+SB)-F$, by Theorem 13.4. Hence $S(T+B)\in\Phi(X)$. Since $S\in\Phi(Y,X)$, by Theorem 13.3, it follows from Theorem 13.1 that $T+B\in\Phi(X, Y)$ and $0=\kappa(S(T+B))=\kappa(S)+\kappa(T+B)=-\kappa(T)+\kappa(T+B)$. This proves (13-11).

Now, if $I+SB$ is bijective, then $\mathcal{N}(T+B)\cap\mathcal{N}(F)=(0)$. For, if $(T+B)x=0=Fx$, then $0=S(T+B)x=(I+SB-F)x=(I+SB)x$ and, consequently, $x=$

0. Thus $\mathcal{N}(T+B)$ is contained in some subspace complementary to $\mathcal{N}(F)$. It follows that

$$\dim \mathcal{N}(T+B) \leq \dim X/\mathcal{N}(F) = \dim \mathcal{R}(F) = n(T).$$

This proves the first part of (13-12); the second part now follows immediately from (13-11). □

If T is a Fredholm operator and B is in $L(X, Y)$ with $\dim \mathcal{R}(B) < \infty$, then $\dim \mathcal{R}(SB) < \infty$, and so $I+SB$ is a Fredholm operator of index zero, by Theorem 13.4. Thus Theorem 13.5 shows that $T+B$ is a Fredholm operator and $\kappa(T+B) = \kappa(T)$. Generalizations of this result are discussed in the problems and in § V.7.

Theorem 13.6. *Let T be in $\Phi(X, Y)$, and let S be any pseudoinverse of T. Then* (13-11) *and* (13-12) *hold for all B in $L(X, Y)$ such that* $\|B\| < 1/\|S\|$.

Proof. If $\|B\| < 1/\|S\|$, then $\|SB\| < 1$, and hence $I+SB$ is bijective, by Theorem 1.4. Thus Theorem 13.5 applies here. □

Theorems 13.5 and 13.6 are often referred to as perturbation theorems. The operator $T+B$ is said to be a *perturbation* of T by the operator B. We have just proved above that the property of being a Fredholm operator of a certain fixed index is stable under perturbations by operators with finite-dimensional ranges and by operators of sufficiently small norm. In particular, $\Phi(X, Y)$ is an open subset of $L(X, Y)$.

Example 1. Let $X = Y = \ell^1$, and define T in $L(X)$ by $Tx = y$, where $x = (\xi_1, \xi_2, \ldots)$ and $y = (\xi_1 - \xi_2, \xi_3, \xi_4, \ldots)$. If $x \in \mathcal{N}(T)$, then $\xi_1 = \xi_2$ and $0 = \xi_3 = \xi_4 = \cdots$, and so $n(T) = 1$. Clearly $d(T) = 0$. Let us define S in $L(X)$ by $Sx = y$, where $y = (0, -\xi_1, \xi_2, \xi_3, \ldots)$. It is readily verified that S is a pseudoinverse of T and $\|S\| = 1$. From Theorem 13.6 it follows that if B in $L(X)$ satisfies $\|B\| < 1 = 1/\|S\|$, then $d(T+B) = d(T) = 0$ and $n(T+B) = \kappa(T+B) = \kappa(T) = n(T) = 1$. For this particular operator T, the conclusions of Theorem 13.6 do not hold for any perturbation constant larger than $1 = 1/\|S\|$. See problem 9.

Perturbations of Unbounded Fredholm Operators

It is important to have a result like Theorem 13.6 for unbounded operators. Let T be a linear operator with domain in a Banach space X and range in a Banach space Y. We shall call T a Fredholm operator if T is a closed operator with $n(T) < \infty$ and $d(T) < \infty$. In the rest of the section, B will denote a linear operator with domain in X and range in Y such that $\mathcal{D}(B) \supset \mathcal{D}(T)$; this will make $\mathcal{D}(T+B)$ the same as $\mathcal{D}(T)$.

Example 2. As an example of the situation we have in mind, let $X = Y = L^2(a, b)$ and let n be a positive integer. Let $\mathcal{D}(T)$ be the set of x in X such that the derivative $x^{(n-1)}$ exists and is absolutely continuous on $[a, b]$ and $x^{(n)}$ is in $L^2(a, b)$, and for such x let $Tx = x^{(n)}$. Then $\mathcal{R}(T) = Y$ and $\mathcal{N}(T)$ is the n-dimensional space of polynomials of degree at most $n-1$. It can be shown that T is a closed operator; thus T is a Fredholm operator. A typical perturbing operator B is given by a formal differential expression involving lower order derivatives, such as

$$\nu = \sum_{k=0}^{n-1} b_k D^k,$$

where $D = d/ds$ and $b_0, \ldots, b_{n-1} \in L^1(a, b)$. Let $\mathcal{D}(B) = \{x : x^{(n-2)}$ is absolutely continuous and $\nu x \in L^2(a, b)\}$ and, for $x \in \mathcal{D}(B)$, let $Bx = \nu x$. Then clearly $\mathcal{D}(B) \supset \mathcal{D}(T)$. Furthermore, it can be shown that B acts like a continuous operator when compared with T (cf. Goldberg [2, p. 169]); that is, B is "T-bounded" in the sense of the following definition.

Definition. If there exist nonnegative constants α, β such that

$$\|Bx\| \leqslant \alpha\|x\| + \beta\|Tx\|, \qquad x \in \mathcal{D}(T), \tag{13-13}$$

then we say that B is *relatively bounded with respect to T* or, simply, *T-bounded.*

If B is continuous, it is obviously T-bounded. A sufficient condition for B to be T-bounded is that there exists a closed linear operator B_0 that is an extension of B (problem 10).

Suppose now that B is T-bounded. Without loss of generality we may assume that α and β are both positive. Let D_1 be the linear space $\mathcal{D}(T)$ endowed with a new norm $\|\cdot\|_1$, given by

$$\|x\|_1 = \alpha\|x\| + \beta\|Tx\|, \qquad x \in \mathcal{D}(T). \tag{13-14}$$

The fact that T is closed implies that D_1 is a Banach space (problem 10). If T_1 denotes the operator T considered as a mapping of D_1 into Y, then $\|T_1x\| = \beta^{-1}(\beta\|Tx\|) \leqslant \beta^{-1}\|x\|_1$ for $x \in D_1$, so that T_1 is continuous and $\|T_1\| \leqslant \beta^{-1}$. Moreover, it is apparent from (13-13) and (13-14) that the restriction B_1 of B to D_1 is also continuous, with $\|B_1\| \leqslant 1$. This justifies the use of the term "T-bounded."

By now it should be evident how we are going to generalize Theorem 13.6. However, there still is one small problem: $T + B$ may not be closed even when T and B are both closed. The following theorem, due to B. Sz.-Nagy, provides a solution to this difficulty.

Theorem 13.7. *If T is closed and if there exist positive numbers α and β such that*

$$\beta<1 \qquad \text{and} \qquad \|Bx\|\leq \alpha\|x\|+\beta\|Tx\|$$

for all $x \in \mathscr{D}(T)$, then $T+B$ is closed.

Proof. For $x \in \mathscr{D}(T)$ we have

$$\|(T+B)x\|\leq \|Tx\|+\|Bx\|\leq \alpha\|x\|+(1+\beta)\|Tx\| \tag{13-15}$$

and

$$\|(T+B)x\|\geq \|Tx\|-\|Bx\|\geq (1-\beta)\|Tx\|-\alpha\|x\|. \tag{13-16}$$

Suppose $x_n \in \mathscr{D}(T+B)=\mathscr{D}(T)$, $x_n \to x$ and $(T+B)x_n \to y$. Then (13-16) implies that $\{Tx_n\}$ is a Cauchy sequence, since $\beta<1$. Thus there is some $z \in Y$ such that $Tx_n \to z$. Since T is closed, $x \in \mathscr{D}(T)=\mathscr{D}(T+B)$ and $Tx=z$; that is, $Tx_n \to Tx$. But we also have $x_n \to x$, and so replacing x by $x_n - x$ in (13-15), we conclude that $(T+B)x_n \to (T+B)x$. Thus $y=(T+B)x$, which proves that $T+B$ is closed. □

Now let T be a Fredholm operator. Then there exist closed subspaces W in X and Z in Y such that $X=\mathscr{N}(T)\oplus W$ and $Y=\mathscr{R}(T)\oplus Z$. Also, $\mathscr{R}(T)$ is closed (Theorem 5.10). Since $\mathscr{N}(T)\subset \mathscr{D}(T)$, it is easy to see that

$$\mathscr{D}(T)=\mathscr{N}(T)\oplus[W\cap \mathscr{D}(T)].$$

The restriction T_0 of T to the Banach space W is a closed operator with a bounded inverse T_0^{-1} defined on $\mathscr{R}(T)$, by Theorem 5.8. Let Q be the (continuous) projection of Y onto $\mathscr{R}(T)$ along Z. It is easy to see that if $S=T_0^{-1}Q$, then $S \in L(Y,X)$, $\mathscr{R}(S)\subset W\cap \mathscr{D}(T)$ and

$$TSTx=Tx, \qquad x \in \mathscr{D}(T). \tag{13-17}$$

If S is any operator in $L(Y,X)$ such that $\mathscr{R}(S)\subset \mathscr{D}(T)$ and (13-17) holds for all $x \in \mathscr{D}(T)$, then we shall call S a pseudoinverse of T. In this case, TS is defined on all of Y. It is easy to verify that TS is a closed operator. Hence TS is continuous, by the closed graph theorem. Furthermore, (13-17) implies that TS is a projection of Y onto $\mathscr{R}(T)$.

Theorem 13.8. *Let T be a Fredholm operator, and let S be any pseudoinverse of T. Suppose that B satisfies*

$$\|Bx\|\leq \alpha\|x\|+\beta\|Tx\|, \qquad x \in \mathscr{D}(T),$$

for some α, β satisfying $\beta<1$ and $\alpha\|S\|+\beta\|TS\|<1$. Then $T+B$ is a Fredholm operator, with $n(T+B)\leq n(T)$, $d(T+B)\leq d(T)$, and $\kappa(T+B)=\kappa(T)$.

Proof. Without loss of generality, we may assume that $\alpha, \beta > 0$. Let T_1 and B_1 be the operators on D_1 described earlier, where the norm on D_1 is given by (13-14). Since T_1 is continuous, we have $T_1 \in \Phi(D_1, Y)$. For $y \in Y$,

$$\|Sy\|_1 = \alpha\|Sy\| + \beta\|TSy\| \leq (\alpha\|S\| + \beta\|TS\|)\|y\|.$$

This shows that S is still continuous when regarded as a mapping S_1 of Y into D_1; furthermore,

$$\|S_1\| \leq \alpha\|S\| + \beta\|TS\| < 1. \tag{13-18}$$

It is clear from (13-17) that S_1 is a pseudoinverse of T_1. From the definition of the norm in D_1 and from (13-18) we see that

$$\|B_1\| \leq 1 < 1/\|S_1\|.$$

Thus B_1 is a sufficiently small perturbation of T_1, and the conclusions of Theorem 13.6 apply to T_1 and $T_1 + B_1$. But $T_1 + B_1$ and $T + B$ have the same null space and range, because $\mathscr{D}(T + B) = \mathscr{D}(T)$. Hence the conclusions of Theorem 13.6 apply to T and $T + B$. Finally, $T + B$ is a Fredholm operator because the condition $\beta < 1$ ensures that $T + B$ is a closed operator. □

Example 3. Let us apply Theorem 13.8 to the nth-order differential operator T considered in Example 2. For $x \in \mathscr{D}(T)$ and $a \leq s \leq b$ we have

$$x(s) = x(a) + x'(a)(s-a) + \cdots + \frac{x^{(n-1)}(a)}{(n-1)!}(s-a)^{n-1}$$
$$+ \int_a^s \frac{(s-t)^{n-1}}{(n-1)!} x^{(n)}(t)\, dt$$

(by Taylor's theorem). Therefore we define S and F by

$$Sx(s) = \int_a^s \frac{(s-t)^{n-1}}{(n-1)!} x(t)\, dt, \qquad x \in L^2(a, b),$$

$$Fx(s) = x(a) + \cdots + \frac{x^{(n-1)}(a)}{(n-1)!}(s-a)^{n-1}, \qquad x \in \mathscr{D}(T),$$

so that

$$x = Fx + STx,$$

for $x \in \mathscr{D}(T)$. Since Fx is a polynomial of degree not exceeding $n-1$, we have $Tx = TFx + TSTx = TSTx$. Thus S is a pseudoinverse of T. Furthermore, since $\mathscr{R}(T) = L^2(a, b)$, it follows that TS is the identity operator and $\|TS\| = 1$.

We now estimate $\|S\|$. For $x \in L^2(a, b)$, the Schwarz inequality implies that

$$|Sx(s)| \leq \frac{1}{(n-1)!}\left(\int_a^s |x(t)|^2\, dt\right)^{1/2}\left(\int_a^s (s-t)^{(n-1)2}\, dt\right)^{1/2}$$

$$\leq \frac{\|x\|}{(n-1)!}\left(\frac{(s-a)^{2n-1}}{2n-1}\right)^{1/2}.$$

From this we obtain

$$\|S\| \leq \frac{1}{(n-1)!}\left(\int_a^b \frac{(s-a)^{2n-1}}{2n-1}\, ds\right)^{1/2}$$

$$\leq \frac{(b-a)^n}{(n-1)!\,(2n-1)^{1/2}(2n)^{1/2}}$$

This estimate of $\|S\|$ may be used in Theorem 13.8 to provide information about perturbations of T. Of course, given some particular operator B, it may require hard analysis to decide if there exist constants α, β satisfying the hypotheses of the theorem.

Improvements can be made on Theorem 13.8, but they require arguments considerably different from those given here. An excellent account of this work, originally due to Kato [1] and Gohberg and Krein [1] appears in Goldberg [2, Chapter V]. In some cases the relative size of the perturbing operator B can be larger than that allowed by our theorem. However, the constants $\|S\|$ and $\|TS\|$ are often easier to estimate than the constant appearing in more general theorems.

Theorem 13.8 is only one of several useful generalizations of Theorem 13.6. Some others are considered in the problems below. In these problems we let $\Phi^l(X, Y)$ be the set of all T in $L(X, Y)$ with the property that $n(T) < \infty$ and there exists a continuous projection of Y onto $\mathcal{R}(T)$, and we let $\Phi^r(X, Y)$ be the set of all T in $L(X, Y)$ with the property that $d(T) < \infty$ and there exists a continuous projection of X onto $\mathcal{N}(T)$. From the proof of Theorem 13.2 it follows that each operator in $\Phi^l(X, Y)$ and $\Phi^r(X, Y)$ has a pseudoinverse. The following simple result is sometimes useful in dealing with these operators.

Theorem 13.9. *If $T \in L(X, Y)$, $U \in L(Y, X)$ and if $TUT - T$ has a pseudoinverse, then so does T.*

Proof. Suppose that R satisfies $(TUT - T)R(TUT - T) = TUT - T$. Rearranging this, we have

$$T = TUT - T(UT - I)R(TU - I)T$$

$$= T(U - UTRTU + RTU + UTR - R)T.$$

Thus T has a pseudoinverse. $\square$

PROBLEMS

In the problems below, X, Y, and Z are Banach spaces.

1. An operator T in $L(X, Y)$ is a Fredholm operator of index zero if and only if there exists $F \in L(X, Y)$ such that $\dim \mathscr{R}(F) < \infty$ and $T + F$ is bijective.
2. Show that an operator T in $L(X, Y)$ is in $\Phi^l(X, Y)$ if and only if there exists $L \in L(Y, X)$ such that $LT = I_X - F$ where $\dim \mathscr{R}(F) < \infty$. Also, show that T is in $\Phi^r(X, Y)$ if and only if there exists $L \in L(Y, X)$ such that $TL = I_Y - F$ where $\dim \mathscr{R}(F) < \infty$.
3. a. If $S \in \Phi^l(Y, Z)$ and $T \in \Phi^l(X, Y)$, show that $ST \in \Phi^l(X, Z)$ and $\kappa(ST) = \kappa(S) + \kappa(T)$. [*Suggestion.* Use the fact that if U and V are pseudoinverses of S and T, respectively, then $VUST = I_X + (VT - I_X) + V(US - I_Y)T$.]
   b. If $S \in \Phi^r(Y, Z)$ and $T \in \Phi^r(X, Y)$, then $ST \in \Phi^r(X, Z)$ and $\kappa(ST) = \kappa(S) + \kappa(T)$.
   c. If $S \in L(Y, Z)$, $T \in L(X, Y)$, and $ST \in \Phi^l(X, Z)$, then $T \in \Phi^l(X, Y)$.
4. If $T \in \Phi^l(X, Y)$, $F \in L(X, Y)$, and $\dim \mathscr{R}(F) < \infty$, show that $T + F \in \Phi^l(X, Y)$ and $\kappa(T + F) = \kappa(T)$.
5. Suppose T is in $\Phi^l(X, Y)$ and S is a pseudoinverse of T. If B is in $L(X, Y)$ and $\|B\| < 1/\|S\|$, then $T + B \in \Phi^l(X, Y)$ and $\kappa(T + B) = \kappa(T)$. In particular, $\Phi^l(X, Y)$ and $\Phi^l(X, Y) \backslash \Phi(X, Y)$ are open subsets of $L(X, Y)$.
6. a. If $T \in L(X, Y)$, then $T \in \Phi(X, Y)$ if and only if $T' \in \Phi(Y', X')$.
   b. If $T \in L(X, Y)$ and $T \in \Phi^l(X, Y)$, then $T' \in \Phi^r(Y', X')$. The converse is not true in general. For example, the range of the inclusion mapping T of (c) into ℓ^∞ is not complemented in ℓ^∞, and yet the null space $(c)^\perp$ of T' is complemented in the conjugate space of ℓ^∞. See Pietsch [1, page 366].
7. Suppose T in $L(X, Y)$ has closed range.
   a. Suppose $n(T) < \infty$, and let W be a closed subspace of X such that $X = \mathscr{N}(T) \oplus W$. The restriction T_0 of T to W has a bounded inverse. If $B \in L(X, Y)$ satisfies $\|B\| < 1/\|T_0^{-1}\|$, then $T + B$ has closed range and $n(T + B) \leqslant n(T)$.
   b. Suppose $d(T) < \infty$ (but do not assume that $\mathscr{N}(T)$ has a closed complement in X). Then there exists $\varepsilon > 0$ such that if $B \in L(X, Y)$ satisfies $\|B\| < \varepsilon$, then $\mathscr{R}(T + B)$ is closed and $d(T + B) \leqslant d(T)$.
8. The condition in the first part of Theorem 13.5 is necessary as well as sufficient. Suppose that $T \in \Phi(X, Y)$, $B \in L(X, Y)$, $T + B \in \Phi(X, Y)$, and $\kappa(T + B) = \kappa(T)$. Show that $I + SB$ is a Fredholm operator of index zero, where S is any pseudoinverse of T.
9. Let T be the operator on ℓ^1 discussed in Example 1. Define B in $L(\ell^1)$ by $Bx = y$, where $x = (\xi_1, \xi_2, \ldots)$, $y = (\eta_1, \eta_2, \ldots)$, $\eta_1 = \xi_2 - \xi_1$, and $\eta_k = \xi_{k+2} - \xi_{k+1}$ for $k \geqslant 2$.
   a. Show that $\|B\| = 2$.
   b. If $\lambda \neq 0, 1$ show that $x \in \mathscr{N}(T + \lambda B)$ if and only if $\xi_1 = \xi_2$ and $\xi_{k+1} = \lambda^{-1}(\lambda - 1)\xi_k$ for $k \geqslant 3$. Show that $n(T + \lambda B) = 2$ if λ is a real number such

that $\frac{1}{2} < \lambda < 1$. Since $\|\lambda B\|$ can be made arbitrarily close to 1 for such λ, this shows that the perturbation constant $\|S\|^{-1} = 1$ in Example 1 cannot be enlarged.

c. Show that $d(T + \frac{1}{2}B) = \infty$.

10. Let T be a linear operator with domain in a Banach space X and range in a Banach space Y.

a. T is closed if and only if D_1 is a Banach space, where D_1 is the set $\mathscr{D}(T)$ with the norm (13-14) for any $\alpha, \beta > 0$.

b. If T is closed, all norms (13-14) on $\mathscr{D}(T)$, with $\alpha, \beta > 0$, are equivalent.

c. Suppose T is closed, and let B be a linear operator from X into Y with $\mathscr{D}(B) \supset \mathscr{D}(T)$. If there exists a closed linear operator B_0 that is an extension of B, then B is T-bounded.

11. Let Δ be a bounded open set in C, and let X be the space of all complex-valued functions that are continuous on the closure $\bar{\Delta}$ and locally analytic on Δ (i.e., analytic in a neighborhood of each point of Δ). For $f \in X$, define $T_f \in L(X)$ by $(T_f h)(z) = f(z)h(z)$. Note that if $f, g \in X$, $T_f T_g = T_g T_f = T_{fg}$ and $T_f + T_g = T_{f+g}$.

a. If $|f(x)| > 0$ for $z \in \bar{\Delta}$, then T_f has an inverse in $L(X)$.

b. If f has only isolated zeros on Δ, then T_f is one-to-one.

c. If $f(z) \neq 0$ on $\bar{\Delta} \backslash \Delta$ and if f has only a finite number of zeros on Δ, then $\kappa(T_f) = d(T_f)$ is the number of zeros of f, counted according to multiplicities. [First prove that if $f(z) = (z - a)^p$, $a \in \Delta$, then $X = \mathscr{R}(T_f) \oplus N$, where N is the linear space generated by the set $\{1, (z-a), \ldots, (z-a)^{p-1}\}$.]

V SPECTRAL ANALYSIS OF LINEAR OPERATORS

As an introduction, we shall indicate the main trend of ideas to be developed in this chapter. Throughout the chapter X will denote a normed linear space that contains some nonzero elements. The scalar field may be either real or complex, except as explicitly stated otherwise. Given a linear operator T with domain and range in X, we shall consider operators of the form $\lambda I - T$, where λ is a scalar and I is the identity operator on X. For convenience we usually suppress the I and write $\lambda - T$ in place of $\lambda I - T$.

Definition. The *resolvent set* of T is the set $\rho(T)$ of all λ such that the range of $\lambda - T$ is dense in X and $\lambda - T$ has a continuous inverse. For $\lambda \in \rho(T)$, the operator $(\lambda - T)^{-1}$ is called the *resolvent operator* and is often denoted by R_λ. The *spectrum* of T is the set $\sigma(T)$ of all scalar values not in $\rho(T)$.

In a broad sense, spectral theory, or spectral analysis, of linear operators is the systematic study of the relations among the operators T, R_λ, the sets $\rho(T)$, $\sigma(T)$, and various other operators and subspaces of X that enter the picture naturally.

For nearly all of the extensive developments of spectral theory we assume that X is complete and that T is closed. In this case, if $\lambda \in \rho(T)$, the fact that $\lambda - T$ has a continuous inverse implies (see Theorem IV.5.2) that $\mathscr{R}(\lambda - T)$ is closed; since it is dense in X, it is all of X. Hence in this case, when $\lambda \in \rho(T)$ the resolvent operator R_λ belongs to $L(X)$. It turns out that $\rho(T)$ is an open set in the space of scalars and that R_λ is developable as a power series in $\lambda - \lambda_0$ about each point $\lambda_0 \in \rho(T)$. This power series has coefficients in $L(X)$ and it converges in the norm of $L(X)$. Thus we can regard R_λ as an analytic function defined on $\rho(T)$, with values in $L(X)$. When the scalars are complex, we can utilize much of the machinery of the classical theory of functions of a complex variable. The interplay between complex analysis and linear algebra underlies many developments in this chapter, particularly in the later sections. For this reason, we begin the chapter with a discussion of analytic functions whose values lie in a complex Banach space.

If X is of finite dimension n and if $\mathscr{D}(T)=X$, we can represent T by an $n\times n$ matrix; then $\lambda - T$ is also represented by a matrix, and the determinant of this matrix, $\det(\lambda - T)$, is a polynomial of degree n in λ. The spectrum of T is composed of those scalars λ that are roots of the equation

$$\det(\lambda - T)=0.$$

This polynomial equation may have no solution if the scalars associated with X are real. However, if the scalars are complex, then $\sigma(T)$ will contain at least one point, and it may contain up to n distinct points. If $\lambda\in\sigma(T)$, then λ will be an *eigenvalue* of T; that is, there will exist a nonzero vector x, called an *eigenvector* associated with λ, such that $Tx=\lambda x$, and hence $(\lambda - T)x=0$. The null space of $\lambda - T$ is sometimes called the *eigenspace* corresponding to λ. In the infinite-dimensional case, $\sigma(T)$ may contain points that are not eigenvalues.

One of the goals in finite-dimensional spectral theory is to find a decomposition of the space X into a direct sum of subspaces, say, $X=M_1\oplus\cdots\oplus M_k$, such that each M_i is invariant under T; that is, $T(M_i)\subset M_i$. In this way the study of T is reduced to analyzing the behavior of T on each subspace M_i. When the scalars are complex, such a direct sum may be found with the property that if we restrict T to act only on a subspace M_i, its spectrum consists of only one point. If we choose a suitable basis for each M_i, the matrix representation of T takes on an especially simple form. When X is infinite dimensional, matters are usually much more complicated, but the pattern of developments in the finite-dimensional case still indicates a certain direction for investigations.

The most important application of the theory in this chapter will be to the study of compact linear operators. Such operators are the natural Banach space generalization of operators on a finite-dimensional space. They are interesting not only because their spectral theory is well developed, but also because they arise so frequently in integral and differential equations. In fact, we shall show how the spectral theory of compact operators includes important parts of the classical theory of Fredholm integral equations of the second kind.

V.1 ANALYTIC VECTOR-VALUED FUNCTIONS

If X is a complex Banach space and Δ is an open set in the complex plane, we shall say that a function f defined on Δ, with values in X, is *locally analytic* on Δ if for each $\lambda_0\in\Delta$ there is an element of X, denoted by $f'(\lambda_0)$, such that

$$\left\|\frac{f(\lambda)-f(\lambda_0)}{\lambda-\lambda_0}-f'(\lambda_0)\right\|\to 0 \qquad \text{as} \qquad \lambda\to\lambda_0.$$

(When Δ is connected, we say simply that f is analytic on Δ.) We call $f'(\lambda_0)$ the derivative of f at λ_0. If f is locally analytic on Δ, then for each $x' \in X'$, the linearity and continuity of x' imply that

$$\begin{aligned}(1\text{-}1) \qquad \lim_{\lambda\to\lambda_0} \frac{x'(f(\lambda)) - x'(f(\lambda_0))}{\lambda - \lambda_0} &= x'\left\{\lim_{\lambda\to\lambda_0} \frac{f(\lambda) - f(\lambda_0)}{\lambda - \lambda_0}\right\} \\ &= x'\{f'(\lambda_0)\}.\end{aligned}$$

Thus each complex-valued composite function $x' \circ f$ must be differentiable at each point of Δ. The remarkable fact is that this apparently weak necessary condition is also a sufficient condition for the local analyticity of f.

Theorem 1.1. *Let X be a complex Banach space, and let f be a function with values in X, defined on an open set Δ in the complex plane. Suppose that for each $x' \in X'$, the function $x' \circ f$ is differentiable at each point of Δ. Then f is locally analytic on Δ.*

Proof. Since X is complete, it suffices to prove that for each $\lambda_0 \in \Delta$, the expression

$$(1\text{-}2) \qquad \frac{f(\lambda) - f(\lambda_0)}{\lambda - \lambda_0} - \frac{f(\mu) - f(\lambda_0)}{\mu - \lambda_0}$$

approaches 0 as λ and μ independently approach λ_0. Let C be a counter-clockwise-oriented circle centered at λ_0, with a radius $r > 0$ such that C and its interior lie in Δ. For $x' \in X'$, the function $x' \circ f$ is continuous and hence bounded on C. By the uniform boundedness principle (Theorem III.9.2), there is a constant M such that

$$(1\text{-}3) \qquad \|f(\xi)\| \leq M, \qquad \xi \in C.$$

Take $x' \in X'$, $0 < |\lambda - \lambda_0| < \frac{1}{2}r$, and $0 < |\mu - \lambda_0| < \frac{1}{2}r$. By Cauchy's formula, we have

$$x'(f(\lambda)) = \frac{1}{2\pi i} \int_C \frac{x'(f(\xi))}{\xi - \lambda}\, d\xi,$$

with corresponding formulas for $x'(f(\mu))$ and $x'(f(\lambda_0))$. A straightforward calculation leads to the formula

$$(1\text{-}4) \qquad x'\left\{\frac{f(\lambda) - f(\lambda_0)}{\lambda - \lambda_0} - \frac{f(\mu) - f(\lambda_0)}{\mu - \lambda_0}\right\} = \frac{1}{2\pi i} \int_C x'(f(\xi)) \frac{\lambda - \mu}{(\xi - \lambda)(\xi - \mu)(\xi - \lambda_0)}\, d\xi.$$

Our choice of λ and μ ensures that $|\xi - \lambda| \geq \frac{1}{2}r$ and $|\xi - \mu| \geq \frac{1}{2}r$ for $\xi \in C$. From this and (1-3) we readily see that the absolute value of the left member of (1-4) does not exceed $4r^{-2}M\|x'\||\lambda - \mu|$. It follows (from Theorem III.3.2) that the norm of the expression in (1-2) does not exceed $4r^{-2}M|\lambda-\mu|$. The desired conclusion follows immediately. □

Another way of stating Theorem 1.1 is that, for each $\lambda_0 \in \Delta$, if the difference quotient $[f(\lambda)-f(\lambda_0)]/(\lambda-\lambda_0)$ converges *weakly* to an element $f'(\lambda_0)$, then the difference quotient also converges to $f'(\lambda_0)$ in the norm of X. This is indeed remarkable, for the weak convergence of a sequence does not in general imply its norm convergence.

Next, we consider an operator-valued analytic function, that is, a locally analytic mapping of Δ into the Banach space $L(X, Y)$, where X and Y are Banach spaces. It turns out that the notion of local analyticity of an operator-valued function is the same for the weak, strong, and uniform operator topologies on $L(X, Y)$.

Theorem 1.2. *Let X and Y be complex Banach spaces, and let Δ be an open set in the complex plane. For each $\lambda \in \Delta$, let A_λ be an element of $L(X, Y)$. The following statements are equivalent:*

(*a*) *For each $x \in X$, $y' \in Y'$, the mapping $\lambda \mapsto y'(A_\lambda x)$ is locally analytic on Δ.*

(*b*) *For each $x \in X$, the mapping $\lambda \mapsto A_\lambda x$ is locally analytic on Δ, with values in Y.*

(*c*) *The mapping $\lambda \mapsto A_\lambda$ is locally analytic on Δ, with values in $L(X, Y)$.*

The fact that (*c*) implies (*a*) involves a simple calculation like that given in (1-1). Theorem 1.1 shows that (*a*) implies (*b*). The proof that (*b*) implies (*c*) is similar to the proof of Theorem 1.1 and is left to the reader.

A great deal of the standard classical theory of analytic functions can be taken over to the vector-valued case, proofs and all, with absolute values replaced by norms. Cauchy's theorem, Liouville's theorem, Taylor's theorem, and many others retain their validity. An alternate method to prove many of these results is to use linear functionals to reduce a theorem to the classical case. See Theorem 1.4 below.

For a brief exposition of the history of analyticity as a concept in functional analysis, see Taylor [7]. An extensive bibliography is included.

Contour Integrals

Later in this chapter we shall make frequent use of line integrals of vector-valued functions. If f is a continuous function from an open set Δ in C into a Banach space X and if C is a simple rectifiable curve in Δ, then the (Stieltjes) integral $\int_C f(\lambda)\, d\lambda$ is defined exactly as in the classical theory of complex-valued functions. The defining sums converge to an element of X because f is continuous, C is rectifiable, and X is complete. The infinite dimensionality of X plays no role. (For reference, the reader may consult Taylor [5, Section 9-5].)

Theorem 1.3. *Let X, Y be Banach spaces, let f be continuous from an open set Δ in $\mathbf{C}$ into X, and let C be a simple rectifiable curve in Δ. Suppose that T is a closed linear operator from X into Y such that $f(\lambda) \in \mathscr{D}(T)$ for λ on C and $Tf(\cdot)$ is continuous from C into Y. Then the vector $\int_C f(\lambda)\,d\lambda$ belongs to $\mathscr{D}(T)$, and*

$$T\int_C f(\lambda)\,d\lambda = \int_C Tf(\lambda)\,d\lambda.$$

In particular,

$$x'\left\{\int_C f(\lambda)\,d\lambda\right\} = \int_C x'(f(\lambda))\,d\lambda$$

for $x' \in X'$.

The proof is left to the reader.

Definition. A set D in the complex plane is called a *Cauchy domain* if: (a) it is open; (b) it has a finite number of components; and (c) the boundary of D, ∂D, is composed of a finite positive number of simple closed rectifiable curves, no two of which intersect.

If C is one of the curves forming part of ∂D, the positive orientation of C is clockwise or counterclockwise according to whether the points of D near a point of C are outside or inside of C. The positively oriented boundary of D is denoted by $+\partial D$.

Theorem 1.4. *Let D be a bounded Cauchy domain, and let f be a continuous function from the closure $\bar{D}$ into a Banach space X such that f is locally analytic on D. Then*

$$\int_{+\partial D} f(\lambda)\,d\lambda = 0. \tag{1-5}$$

For $n = 0, 1, 2, \ldots$, the nth derivative $f^{(n)}$ is locally analytic on D, and

$$f^{(n)}(\lambda) = \frac{n!}{2\pi i}\int_{+\partial D} \frac{f(\xi)}{(\xi-\lambda)^{n+1}}\,d\xi, \qquad \lambda \in D. \tag{1-6}$$

Proof. We shall use the fact that the conjugate space X' separates the points of X; that is, if $x'(x) = x'(y)$ for all $x' \in X'$, then $x = y$. (See the remarks following Theorems III.2.8 and III.3.2.) We take $x' \in X'$ and apply Cauchy's theorem to $x' \circ f$. Using Theorem 1.3, we have

$$x'\left\{\int_{\epsilon+\partial D} f(\lambda)\, d\lambda\right\} = \int_{+\partial D} x'(f(\lambda))\, d\lambda = 0.$$

Since this is true for every $x' \in X'$, we obtain (1-5). The proof of (1-6) is similar. It is not difficult to show that f has derivatives of all orders on D (problem 1). Using (1-1), we find that

$$\text{(1-7)} \qquad \frac{d^n}{d\lambda^n} x' \circ f(\lambda) = x' \circ f^{(n)}(\lambda),$$

for $x' \in X'$, $\lambda \in D$, and $n = 1, 2, \ldots$. The formulas in (1-6) follow easily from this and Cauchy's integral formulas for $x' \circ f$ and the derivatives of $x' \circ f$. □

Power Series and Laurent Expansions

If f is a locally analytic function on a disc $|\lambda - \lambda_0| < r_0$, with values in X, then (1-6) leads to the usual Cauchy estimates on the norm of $f^{(n)}(\lambda)$ for $|\lambda - \lambda_0| = r$, where r satisfies $0 < r < r_0$. These estimates can be used to show that the Taylor expansion for f,

$$\text{(1-8)} \qquad \sum_{n=0}^{\infty} (\lambda - \lambda_0)^n \frac{f^{(n)}(\lambda_0)}{n!},$$

converges (in the norm of X) uniformly for $|\lambda - \lambda_0| \leq r$. From (1-7) it is clear that the series (1-8) converges weakly to $f(\lambda)$; hence the strong limit must also be $f(\lambda)$. On the other hand, if $\{a_n\}$ is a sequence in X such that the power series $\Sigma\, (\lambda - \lambda_0)^n a_n$ converges for λ in some disc $|\lambda - \lambda_0| < r_0$, then the series defines an analytic function f on that disc and $a_n = f^{(n)}(\lambda_0)/n!$. The classical proofs of these assertions carry over to the present situation.

Similar arguments may be given to show that if f is locally analytic on an annulus $0 \leq r_1 < |\lambda - \lambda_0| < r_2$, then f has a unique Laurent expansion,

$$f(\lambda) = \sum_{-\infty}^{\infty} (\lambda - \lambda_0)^n a_n, \qquad a_n = \frac{1}{2\pi i}\int_C (\lambda - \lambda_0)^{-n-1} f(\lambda)\, d\lambda,$$

where C is any counterclockwise circle $|\lambda - \lambda_0| = r$, with $r_1 < r < r_2$.

The Maximum Modulus Theorem

Theorem 1.5. *Let X be a complex Banach space, and let Δ be a connected open set in $\mathbf{C}$. Let F be analytic on Δ, with values in X, and suppose that $\|F(\lambda)\|$ is not constant on Δ. Then $\|F(\lambda)\|$ cannot attain an absolute maximum at any point of Δ.*

Proof. We could imitate one of the classical proofs that uses Cauchy's integral formula. Instead, we shall give a proof using linear functionals. Suppose the theorem is false, so that for some $\lambda_0 \in \Delta$ we have $\|F(\lambda)\| \leq \|F(\lambda_0)\|$ for each λ. By the Hahn–Banach theorem, there exists $x' \in X'$ such that $\|x'\| = 1$ and $x'(F(\lambda_0)) = \|F(\lambda_0)\|$. Then $x' \circ F$ is a complex-valued analytic function on Δ that attains its maximum modulus at λ_0. The classical maximum modulus theorem then implies that $x'(F(\lambda))$ is constant in Δ, its value being $\|F(\lambda_0)\|$. But $|x'(F(\lambda))| \leq \|F(\lambda)\|$ and, since $\|F(\lambda)\| < \|F(\lambda_0)\|$ at some points of Δ, we have a contradiction. $\square$

Theorem 1.5 is not a full generalization of the classical theorem. However, see problem 3 for the case when X is a strictly convex space. (All Hilbert spaces and the spaces ℓ^p and $L^p(0, 1)$ for $1 < p < \infty$ are strictly convex, for example. See Clarkson [1] for a proof of the stronger result that $L^p(0, 1)$ is uniformly convex.) An even wider class of Banach spaces is treated in Thorp and Whitley [1].

Means of Analytic Functions

We conclude this section with an interesting application of the theory of Banach-space-valued analytic functions to some classical theorems about means of complex-valued analytic functions. First, we need a corollary of Theorem 1.5 for the case when Δ is a disc $|\lambda| < R$, for some $R > 0$. For $0 \leq r < R$, let $M(r) = \max \{\|F(\lambda)\| : |\lambda| = r\}$. The next theorem follows from Theorem 1.5.

Theorem 1.6. *If F is analytic on the disc $|\lambda| < R$, then $M(r)$ is a nondecreasing function of r. Furthermore, if $0 < r_1 < r_2$ and $M(r_1) = M(r_2)$, then $M(r)$ is constant when $0 \leq r \leq r_2$.*

We also require a generalization of Hadamard's three-circles theorem.

Theorem 1.7. *If F is analytic and not identically 0 when $|\lambda| < R$, then* $\log M(r)$ *is a convex function of* $\log r$ *when $0 < r < R$.*

This theorem may be proved exactly as in one of the standard proofs for the complex-valued case (e.g., Titchmarsh [1, § 5.3 and § 5.32]), except that absolute values are replaced by norms.

Let us now turn to the definition of the means $\mathfrak{M}_p[f; r]$ given in Example 8 of § II.2 (page 60), where f is a complex-valued analytic function in the open unit disc of the complex plane. We are going to assume that $1 \leq p < \infty$ and prove:

1. $\mathfrak{M}_p[f; r]$ is, for $0 \leq r < 1$, a strictly increasing function of r, unless f is a constant.

2. If f is not identically 0, $\log \mathfrak{M}_p[f; r]$ is a convex function of $\log r$ when $0 < r < 1$.

These propositions go back to G. H. Hardy [1]. Very beautiful proofs of the propositions were given by F. Riesz [3], using the principle of subharmonic functions. The propositions are true for all $p > 0$, but our proofs require $p \geq 1$, because we depend on the fact that H^p is a Banach space if $p \geq 1$. See § II.2 and problem 1(d) in § II.4.

If f is analytic when $|z| < 1$ and if $|\lambda| < 1$, let f_λ be the function defined by $f_\lambda(z) = f(\lambda z)$. Then f_λ is analytic in a domain that includes the closed disc $|z| \leq 1$, so that $f_\lambda \in H^p$. Hence the function F defined by $F(\lambda) = f_\lambda$ is a function of λ with values in H^p. Furthermore, $F(\lambda)$ may be written as a power series in λ that converges in H^p for $|\lambda| < 1$. Thus F is analytic for $|\lambda| < 1$. See problems 4 to 6. If we compute $\|F(\lambda)\| = \|f_\lambda\|$ using the norm in H^p (see (2-11) on page 60, we find that $\|F(\lambda)\| = \|F(|\lambda|)\|$. Thus $\|F(\lambda)\|$ is constant on the circle $|\lambda| = r$; we denote it by $M(r)$. Theorems 1.6 and 1.7 apply to $M(r)$ for $r < 1$. The computation of $\|F(\lambda)\|$ also shows that

$$M(r) = \sup_{0 \leq t < r} \mathfrak{M}_p[f; t]. \tag{1-9}$$

From this we shall be able to show that

$$M(r) = \mathfrak{M}_p[f; r]. \tag{1-10}$$

First, we take note of the fact that

$$\mathfrak{M}_p[f; 0] = |f(0)| \leq \mathfrak{M}_p[f; r] \tag{1-11}$$

if $0 < r < 1$ and that the inequality here is strict unless f is a constant function. For, by Cauchy's formula,

$$f(0) = \frac{1}{2\pi} \int_0^{2\pi} f(re^{i\theta})\, d\theta,$$

whence $|f(0)| \leq \mathfrak{M}_1[f; r]$. The inequality here is easily seen to be strict unless f is constant. But then $\mathfrak{M}_1[f; r] \leq \mathfrak{M}_p[f; r]$ if $1 < p$ (see Hardy, Littlewood and Polya [1, page 143]), whence (1–11) follows.

Now let $\phi(r) = \mathfrak{M}_p[f; r]$; we know that ϕ is continuous. Evidently $M(0) = \phi(0)$ and $\phi(r) \leq M(r)$ because of (1-9). Suppose that for some r_0 we have $0 < r_0 < 1$ and $\phi(r_0) < M(r_0)$. Consider the maximum of $\phi(r)$ for $0 \leq r \leq r_0$. It occurs for some r_1 on the interval and is equal to $M(r_0)$, by (1-9). That is, $\phi(r_1) = M(r_0)$. Hence $r_1 < r_0$ because $\phi(r_1) \neq \phi(r_0)$. We see that $M(r_0) \leq M(r_1)$. But $M(r_1) \leq M(r_0)$, by the first conclusion in Theorem 1.6. Hence $M(r_1) = M(r_0)$, and it follows from the second conclusion in Theorem 1.6 that $M(r)$ is constant on $[0, r_0]$. Then $\phi(r_0) < M(r_0) = M(0) = \phi(0)$. This contradicts (1-11). Hence (1-10) is established.

Proposition 2 now follows at once from (1-10) and Theorem 1.7. From Theorem 1.6 we infer that $\mathfrak{M}_p[f; r]$ is a nondecreasing function of r. Actually, $\mathfrak{M}_p[f; r]$ must be strictly increasing, unless f is constant, for otherwise, by Theorem 1.6 we would have $M(r)$ constant on some interval $[0, r_1]$, whence $\mathfrak{M}_p[f; 0] = \mathfrak{M}_p[f; r_1]$, in contradiction to the statement made in connection with (1-11).

PROBLEMS

1. Let f be locally analytic on an open set Δ in C, with values in a Banach space X. Use the fact that a complex-valued locally analytic function has derivatives of all orders to show that f has derivatives of all orders.
2. Let X be a Hausdorff locally convex space, and let f be a function from C into X such that $x' \circ f$ is analytic on C for each $x' \in X'$. If $\{f(\lambda) : \lambda \in C\}$ is a weakly bounded set, then f is a constant function. In particular, this implies that the classical theorem of Liouville remains valid when the values of the analytic function lie in a complex Banach space. (Cf. problem 1(c) of § II.7.)
3. A Banach space X is *strictly convex* (or *rotund*) if every point on the surface of the unit ball B of X is an extremal point of B (cf. problem 5 of § III.11).
   a. If X is strictly convex, then each $x' \in X'$ assumes its maximum on B at most once; that is, $x'(x_0) = \|x'\|$ for at most one $x_0 \in B$.
   b. (The maximum modulus theorem) Let Δ be a connected open set in C, let X be a strictly convex Banach space, and let f be analytic on Δ with values in X. If $\|f(\lambda)\|$ attains an absolute maximum at some point in Δ, then $f(\lambda)$ is constant on Δ.
4. Let $u_n(z) = z^n$, $n = 0, 1, \ldots$. Show that u_n, as an element of H^p, has $\|u_n\| = 1$.
5. If $h \in H^p$ and $h(z) = \sum_0^\infty c_n z^n$, use Cauchy's formulas for c_n to show that $r^n|c_n| \leqslant \mathfrak{M}_p[h; r]$ if $0 < r < 1$, and hence $|c_n| \leqslant \|h\|$. Deduce that $|h(z)| \leqslant \|h\|(1 - |z|)^{-1}$ if $|z| < 1$.
6. If $f(z) = \sum_0^\infty a_n z^n$ converges when $|z| < 1$, show that $\sum_0^\infty a_n \lambda^n u_n$ is convergent to f_λ in the metric of H^p (with u_n as in problem 4). First, show that the series converges to some element $g \in H^p$. Then use problem 5 to show that $\sum_0^\infty a_n \lambda^n z^n = g(z)$, whence $g = f_\lambda$. The conclusion is that $F(\lambda)$ is given by a power series in λ with coefficients from H^p. Hence F is analytic. For this and more general results of the same type see Taylor [3, 4].

V.2 THE RESOLVENT OPERATOR

We consider a linear operator T with domain and range in a normed linear space X. As a convenience in stating some of the next few theorems, we shall write T_λ in place of $\lambda - T$. For $\lambda \in \rho(T)$, the resolvent operator will be denoted by either R_λ or T_λ^{-1}.

Our first fundamental result is that the resolvent set is an open set of scalars. When X is complete and T is in $L(X)$, this follows from Theorem IV.1.5. For T_μ has an inverse in $L(X)$ if $\mu \in \rho(T)$, and we know (from Theorem IV.1.5) that if T_λ is sufficiently close to T_μ, then T_λ also has an inverse in $L(X)$, that is, $\lambda \in \rho(T)$. But

$$T_\mu x - T_\lambda x = (\mu - \lambda)x \tag{2-1}$$

for $x \in \mathscr{D}(T)$. Hence $\|T_\mu - T_\lambda\| = |\mu - \lambda|$, and $\lambda \in \rho(T)$ when λ is sufficiently close to μ. The proof in the general case requires the following lemma.

Lemma 2.1. *Suppose μ is such that T_μ has a continuous inverse, and let $M(\mu)$ be the norm of T_μ^{-1} (as an operator from $\mathscr{R}(T_\mu)$ into X). Then T_λ has a continuous inverse if $|\lambda - \mu| < 1/M(\mu)$. Moreover, $\overline{\mathscr{R}(T_\lambda)}$ is not a proper subset of $\overline{\mathscr{R}(T_\mu)}$.*

Proof. If $x \in \mathscr{D}(T)$, then $\|x\| = \|T_\mu^{-1} T_\mu x\| \leq \|T_\mu^{-1}\| \|T_\mu x\|$, so that $\|T_\mu x\| \geq \|x\|/M(\mu)$. Therefore

$$\begin{aligned} \|T_\lambda x\| &= \|T_\mu x + (\lambda - \mu)x\| \geq \|T_\mu x\| - |\lambda - \mu| \|x\| \\ &\geq \left(\frac{1}{M(\mu)} - |\lambda - \mu|\right)\|x\|. \end{aligned} \tag{2-2}$$

It follows from Theorem II.1.2 that T_λ has a continuous inverse if $1/M(\mu) - |\lambda - \mu| > 0$. To prove that $\overline{\mathscr{R}(T_\lambda)}$ is not a proper subset of $\overline{\mathscr{R}(T_\mu)}$, assume the contrary. If θ is chosen so that $|\lambda - \mu| M(\mu) < \theta < 1$, there must exist an element $y_0 \in \overline{\mathscr{R}(T_\mu)}$ such that $\|y_0\| = 1$ and $\|y - y_0\| \geq \theta$ if $y \in \overline{\mathscr{R}(T_\lambda)}$ (by Riesz's lemma, Theorem II.3.5). Now choose $y_n \in \mathscr{R}(T_\mu)$ so that $y_n \to y_0$, and let $x_n = T_\mu^{-1} y_n$. Then $T_\lambda x_n \in \mathscr{R}(T_\lambda)$, and so

$$\begin{aligned} \theta \leq \|y_0 - T_\lambda x_n\| &\leq \|y_0 - T_\mu x_n\| + \|T_\mu x_n - T_\lambda x_n\| \\ &= \|y_0 - y_n\| + |\mu - \lambda| \|x_n\| \\ &\leq \|y_0 - y_n\| + |\mu - \lambda| M(\mu) \|y_n\|. \end{aligned}$$

If we now let $n \to \infty$, we obtain the contradiction $\theta \leq |\lambda - \mu| M(\mu)$. □

Theorem 2.2. *The resolvent set $\rho(T)$ is open, and hence the spectrum is closed.*

Proof. If $\mu \in \rho(T)$, then T_μ^{-1} exists and is continuous and $\overline{\mathscr{R}(T_\mu)} = X$. Lemma 2.1 shows that if λ is sufficiently near μ, then T_λ^{-1} also exists and is continuous and $\overline{\mathscr{R}(T_\lambda)}$ is not a proper subset of X. Hence $\overline{\mathscr{R}(T_\lambda)} = X$ and $\lambda \in \rho(T)$. □

The next theorem is also of fundamental importance. It establishes the fact that when T is closed and X is a complex Banach space, the resolvent operator R_λ is an analytic function of λ. As mentioned previously, this will permit us to use the theory of analytic functions in the general spectral theory developed in the last half of this chapter. Equation (2-3) below is often called the *resolvent equation*; it will play a crucial role in the construction of the operational calculus (§ 8).

It was noted at the beginning of this chapter that if X is complete and T is closed, the range of $\lambda - T$ is all of X when $\lambda \in \rho(T)$. In the next theorem we assume this directly, since this is all we need. (However, see problem 3).

Theorem 2.3. *Suppose T is such that $\mathcal{R}(T_\lambda) = X$ if $\lambda \in \rho(T)$. Then, if λ and μ are any two points in $\rho(T)$, R_λ and R_μ satisfy the relations*

$$R_\lambda - R_\mu = (\mu - \lambda) R_\lambda R_\mu, \tag{2-3}$$

$$R_\lambda R_\mu = R_\mu R_\lambda. \tag{2-4}$$

If $\mu \in \rho(T)$ and $|\mu - \lambda| \|R_\mu\| < 1$, then $\lambda \in \rho(T)$ and

$$R_\lambda = \sum_0^\infty (\mu - \lambda)^n R_\mu^{n+1}, \tag{2-5}$$

the series converging according to the metric in $L(X)$. As a function on $\rho(T)$ to $L(X)$, R_λ has derivatives of all orders, with

$$\frac{d^n}{d\lambda^n} R_\lambda = (-1)^n n! R_\lambda^{n+1}. \tag{2-6}$$

Proof. Given $y \in X$, we may write $y = T_\mu x$ where $x = R_\mu y \in \mathcal{D}(T)$. From (2-1) we obtain

$$y - T_\lambda R_\mu y = (\mu - \lambda) R_\mu y.$$

Hence, applying R_λ to both sides, we obtain

$$R_\lambda y - R_\mu y = (\mu - \lambda) R_\lambda R_\mu y.$$

This proves (2-3). By symmetry, the result holds with λ and μ exchanged; from this we conclude that (2-4) holds.

We know from Lemma 2.1 that $\mu \in \rho(T)$ and $|\lambda - \mu| \|R_\mu\| < 1$ imply $\lambda \in \rho(T)$. To prove (2-5) with this condition on λ and μ we use (2-3), (2-4), and induction to prove that

$$R_\lambda = \sum_{k=0}^{n} (\mu - \lambda)^k R_\mu^{k+1} + (\mu - \lambda)^{n+1} R_\mu^{n+1} R_\lambda$$

if $n \geqslant 0$. Thus (2-5) is equivalent to

$$\lim_{n \to \infty} |\mu - \lambda|^{n+1} \|R_\mu^{n+1} R_\lambda\| = 0. \tag{2-7}$$

Since $\|R_\mu^{n+1}R_\lambda\| \le \|R_\mu\|^{n+1}\|R_\lambda\|$ and $|\mu-\lambda|\,\|R_\mu\|<1$, we see that (2-7) is true.

To prove (2-6) when $n=1$, perhaps the easiest method is to start with (2-3) and show that

$$\|(R_\lambda - R_\mu)/(\lambda-\mu)+R_\mu^2\| = \|(\lambda-\mu)R_\lambda R_\mu^2\| \le |\lambda-\mu|\,\|R_\lambda\|\,\|R_\mu^2\|.$$

If we let $y = T_\lambda x$ and $x = R_\lambda y$, then it follows from (2-2) that when $|\lambda-\mu|$ is sufficiently small,

$$\|R_\lambda\| \le \|R_\mu\|(1-|\lambda-\mu|\,\|R_\mu\|)^{-1}.$$

Hence $(R_\lambda - R_\mu)/(\lambda-\mu)\to -R_\mu^2$ as $\lambda\to\mu$. For higher values of n, (2-6) may be proved by induction. See problem 1. □

When X is a Banach space, (2-5) is easily proved from Theorem IV.1.4. If $\|(\mu-\lambda)R_\mu\| = |\mu-\lambda|\,\|R_\mu\|<1$, then $I-(\mu-\lambda)R_\mu$ has an inverse in $L(X)$. From the resolvent equation, $R_\lambda(I-(\mu-\lambda)R_\mu)=R_\mu$. Hence

$$R_\lambda = R_\mu(I-(\mu-\lambda)R_\mu)^{-1}.$$

The series expansion of $(I-(\mu-\lambda)R_\mu)^{-1}$ given by Theorem IV.1.4 leads immediately to (2-5). It follows from (2-5) and our discussion of Taylor series in § 1 that when X is a complex Banach space, R_λ is an analytic function of λ whose derivatives are given by (2-6).

The Conjugate of R_λ

Now suppose that T is densely defined (but do not assume that X is complete). The conjugate of $\lambda - T$ is $\lambda I_{X'} - T'$, which we write simply as $\lambda - T'$. Because T' is closed and X' is complete, λ is in $\rho(T')$ if and only if $\mathscr{R}(\lambda - T') = X'$ and $(\lambda-T')^{-1}$ exists.

Theorem 2.4. *If $\overline{\mathscr{D}(T)} = X$, then T and T' have the same resolvent set (and hence the same spectrum). When $\lambda\in\rho(T)$, the conjugate of $(\lambda-T)^{-1}$ is $(\lambda-T')^{-1}$.*

Proof. Applying Theorems IV.8.4 and IV.9.2 to $\lambda - T$ and its conjugate, we see that $\overline{\mathscr{R}(\lambda-T)} = X$ and $\lambda - T$ has a continuous inverse if and only if $(\lambda-T')^{-1}$ exists and $\mathscr{R}(\lambda-T')=X'$. Thus $\rho(T)=\rho(T')$. Now take $\lambda\in\rho(T)$, $x\in\mathscr{D}(R_\lambda)$, and $x'\in X'$. Then $X'=\mathscr{R}(\lambda-T')$, and we have

$$\begin{aligned}\langle R_\lambda x, x'\rangle &= \langle R_\lambda x, (\lambda-T')(\lambda-T')^{-1}x'\rangle\\ &= \langle(\lambda-T)R_\lambda x, (\lambda-T')^{-1}x'\rangle\\ &= \langle x, (\lambda-T')^{-1}x'\rangle.\end{aligned}$$

This shows that the conjugate of R_λ has domain X' and equals $(\lambda-T')^{-1}$. □

PROBLEMS

1. To complete the proof of (2-6) by induction it suffices to prove that $(d/d\lambda)R_\lambda^n = -nR_\lambda^{n+1}$. Do this (at $\lambda = \mu$) by factoring $R_\lambda^n - R_\mu^n$.
2. Suppose, for each λ in a nonempty set S of scalars, A_λ is a linear operator on X into X such that $A_\lambda - A_\mu = (\mu - \lambda)A_\lambda A_\mu$ if $\lambda, \mu \in S$. Suppose also A_λ^{-1} exists for at least one $\lambda \in S$. Show that A_λ^{-1} exists for every $\lambda \in S$, that all the operators A_λ have the same range $\mathscr{D}$, and that there exists a linear operator T on $\mathscr{D}$ into X such that $\lambda - T = A_\lambda^{-1}$ for each $\lambda \in S$. Show further that if A_λ is closed for at least one λ, then T is closed and A_λ is closed for every $\lambda \in S$.
3. Suppose X is complete and T is a linear operator whose domain and range are in X. If $\mathscr{R}(T_\lambda) = X$ for some $\lambda \in \rho(T)$, then T is closed.
4. Consider the operator $T \in L(\ell^1)$ defined by $y = Tx$, where $y = \{\eta_k\}$, $x = \{\xi_k\}$, $\eta_1 = 0$, and $\eta_k = -\xi_{k-1}$ if $k \geqslant 2$. Show that $(\lambda - T)^{-1}$ exists for all λ, that $\rho(T)$ consists of all λ for which $|\lambda| > 1$, and that $\|R_\lambda\| = (|\lambda| - 1)^{-1}$. Use problem 4, § IV.6. If $|\lambda| \leqslant 1$ the range of $\lambda - T$ is not dense in ℓ^1. The inverse is continuous if $|\lambda| < 1$, but not if $|\lambda| = 1$.
5. Let $\hat{X}$ be the completion of X (X itself if X is complete). Assume $\rho(T)$ not empty and, for each $\lambda \in \rho(T)$, let A_λ be the unique extension of T_λ^{-1} to all of $\hat{X}$.
   a. Show that $A_\lambda - A_\mu = (\mu - \lambda)A_\lambda A_\mu$ for $\lambda, \mu \in \rho(T)$.
   b. Show that T is closable (as defined in problem 9, § IV.5) if and only if A_λ has an inverse for some $\lambda \in \rho(T)$ (and hence for all such λ, by problem 2). The next problem shows that T can fail to be closable.
6. Take $X = \ell^1$. Let $\mathscr{D}(T)$ be the set of $x = \{\xi_n\}$ such that $\xi_n = 0$ except for a finite number of indices. Define $Tx = y$ by $\eta_k = k^{-2}\sum_{i=k}^{\infty} i^2\xi_i$. Then T defines a one-to-one mapping of $\mathscr{D}(T)$ onto itself; T^{-1} is continuous and T is discontinuous. The extension A of T^{-1} to all of X is defined by $Ay = x$, where $k^2\xi_k = k^2\eta_k - (k+1)^2\eta_{k+1}$. But A has no inverse. Hence T is not closable (by problem 5).
7. Let X be an incomplete normed linear space, with $\hat{X}$ its completion. Suppose $T \in L(X)$, and let $\hat{T}$ be the unique linear extension of T to all of $\hat{X}$. Then $\rho(\hat{T}) = \rho(T)$, and hence $\sigma(\hat{T}) = \sigma(T)$. [For the proof, one does not need the result of the next problem.]
8. Let X be complete, let G be the graph of T, and let $\bar{G}$ be the closure of G in the product space $X \times X$. Suppose T is closable, so that $\bar{G}$ is the graph of a closed linear extension $\bar{T}$ of T. Show that $\bar{T}$ has the same resolvent set and spectrum as T.
9. Let X be a Banach space, and suppose $S, T \in L(X)$.
   a. If $\lambda \in \rho(S) \cap \rho(T)$, then the resolvents of S and T satisfy the equation
   $$(\lambda - S)^{-1} - (\lambda - T)^{-1} = (\lambda - S)^{-1}(S - T)(\lambda - T)^{-1}.$$
   (This is sometimes called "the second resolvent equation.")
   b. For a fixed λ_0, the set $\mathscr{S}$ of all $T \in L(X)$ such that $\lambda_0 \in \rho(T)$ is open. In fact, if $T \in \mathscr{S}$ and $\|S - T\| < 1/\|(\lambda_0 - T)^{-1}\|$, then $S \in \mathscr{S}$ and
   $$(\lambda_0 - S)^{-1} = (\lambda_0 - T)^{-1}\left\{I + \sum_{n=1}^{\infty} [(S - T)(\lambda_0 - T)^{-1}]^n\right\}.$$

c. (Newburgh) Given a nonempty open set Δ in C and given $T \in L(X)$ with $\sigma(T) \subset \Delta$, there exists an $\varepsilon > 0$ such that $\sigma(S) \subset \Delta$ whenever $S \in L(X)$ and $\|S - T\| < \varepsilon$. (We say that the spectrum of T is an upper semi-continuous function of T.)

V.3 THE SPECTRUM OF A BOUNDED LINEAR OPERATOR

In all of the theorems of this section we assume that $T \in L(X)$. The main results are concerned with the finding of the smallest constant r such that $|\lambda| \leq r$ if $\lambda \in \sigma(T)$ and with the expression of $(\lambda - T)^{-1}$ in terms of λ and T when $|\lambda| > r$.

Theorem 3.1. *If $T \in L(X)$ and $|\lambda| > \|T\|$, then $\lambda \in \rho(T)$, and*

$$R_\lambda y = \sum_1^\infty \lambda^{-n} T^{n-1} y \tag{3-1}$$

for each y in the (dense) range of $\lambda - T$. Thus $\sigma(T)$ is compact. If X is complete and $|\lambda| > \|T\|$,

$$R_\lambda = \sum_1^\infty \lambda^{-n} T^{n-1}, \tag{3-2}$$

the series converging in the space $L(X)$.

Proof. Assume first that X is complete. Since $\lambda - T = \lambda(I - \lambda^{-1}T)$, the assertion about (3-2) follows from Theorem IV.1.4 (the Neumann expansion), with $A = \lambda^{-1}T$. Thus $\lambda \in \rho(T)$ when $|\lambda| > \|T\|$. This is also true when X is not complete. For, if $\hat{T}$ is the unique linear extension of T to all of the completion $\hat{X}$, then $\|\hat{T}\| = \|T\|$ and $\rho(\hat{T}) = \rho(T)$ (problem 7 of § 2). Of course, $\sigma(T)$ is compact, since we have just proved it is bounded, and we already know it is closed. Finally, if $y = (\lambda - T)x$, we find by induction that

$$x = \lambda^{-1}y + \cdots + \lambda^{-n}T^{n-1}y + \lambda^{-n}T^n x.$$

Formula (3-1) follows from this because $\lambda^{-n}T^n x \to 0$ when $|\lambda| > \|T\|$. □

Before going further we note this useful fact: *If $T \in L(X)$ and if the series $\sum_1^\infty \lambda^{-n} T^{n-1}$ converges in $L(X)$ for some value of λ, then $\lambda \in \rho(T)$ and the operator defined by the series is R_λ.* For, denoting the series by A, it is easily seen that $(\lambda - T)A = A(\lambda - T) = I$. (Also, see problem 3.)

For the remainder of this section we must restrict our attention to complex Banach spaces. The reasons for this are not superficial. We need the theory of analytic functions discussed in § 1. We cannot relate the convergence of the series (3-2) to the extent of the set $\rho(T)$ unless we think of λ as a complex variable, for somewhat the same reason that the radius of convergence of the power series expansion of a real analytic function cannot be discovered merely by looking for singularities of the function on the real axis.

The key fact is that when T is closed and X is a complex Banach space, R_λ depends analytically on λ as λ varies in $\rho(T)$. This is true by Theorem 2.3.

Theorem 3.2. *If X is a complex Banach space and $T \in L(X)$, then $\sigma(T)$ is not empty.*

Proof. If $|\lambda| > \|T\|$, we have

$$\|\lambda x - Tx\| \geq |\lambda| \|x\| - \|Tx\| \geq (|\lambda| - \|T\|)\|x\|.$$

It follows that $\|R_\lambda\| \leq (|\lambda| - \|T\|)^{-1}$ if $|\lambda| > \|T\|$. Hence $\|R_\lambda\| \to 0$ as $|\lambda| \to \infty$. If $\sigma(T)$ were empty, it would follow that R_λ is analytic and bounded on the whole plane. But then it would be constant, by Liouville's theorem (cf. problem 2, § 1), and the constant would be the zero operator. This is impossible by the fact that R_λ sets up a one-to-one mapping of X onto itself, for X is assumed to have some nonzero elements (see the second sentence at the beginning of this chapter). □

Theorem 3.2 remains valid even if X is not complete, as we see from problem 7, § 2.

The Spectral Radius

Supposing that $\sigma(T)$ is nonempty and compact, we define

$$r_\sigma(T) = \sup_{\lambda \in \sigma(T)} |\lambda|, \tag{3-3}$$

and call $r_\sigma(T)$ the *spectral radius* of T.

Theorem 3.3. *If X is a complex Banach space and $T \in L(X)$, the resolvent is given by*

$$R_\lambda = \sum_1^\infty \lambda^{-n} T^{n-1} \tag{3-4}$$

if $|\lambda| > r_\sigma(T)$. This series also represents R_λ if the series converges and $|\lambda| = r_\sigma(T)$. The series diverges if $|\lambda| < r_\sigma(T)$.

Proof. We know that R_λ is analytic when $|\lambda| > r_\sigma(T)$. Hence it has a unique Laurent expansion in positive and negative powers of λ, convergent when $|\lambda| > r_\sigma(T)$. Now, we already know that (3-4) is valid when $|\lambda| > \|T\|$. By the uniqueness, then, this must be the Laurent expansion. The second assertion in the theorem follows from the italicized statement following the proof of Theorem 3.1. For the same reason, the series (3-4) cannot converge if $\lambda \in \sigma(T)$. Hence it cannot converge at λ_0 if $|\lambda_0| < r_\sigma(T)$ because, if it did, then, as in the general theory of power series, the series would converge when $|\lambda| > |\lambda_0|$, and so, in particular, for some $\lambda \in \sigma(T)$. □

Theorem 3.3 enables us to write a formula for the spectral radius of T. If we consider the series (3-4) as a power series in λ^{-1}, the standard formula for the radius of convergence of a power series tells us that

$$r_\sigma(T) = \limsup_{n \to \infty} \|T^n\|^{1/n}. \tag{3-5}$$

Actually, as we shall presently prove, $\{\|T^n\|^{1/n}\}$ is a convergent sequence, so that the limit superior in (3-5) is a limit. To prove this we first prove what is called the *spectral mapping theorem* for polynomials. Suppose

$$p(\lambda) = \alpha_n \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_0$$

is a polynomial with complex coefficients. If $T \in L(X)$, positive integral powers of T have a clear meaning, and we define

$$p(T) = \alpha_n T^n + \alpha_{n-1} T^{n-1} + \cdots + \alpha_0 I.$$

By the rules of algebra for operators, it is clear that if the polynomial $p(\lambda)$ is factored, there is a corresponding factored form of $p(T)$.

Now $p(T) \in L(X)$, and we can consider its spectrum.

Theorem 3.4. *Suppose $T \in L(X)$, where X is a complex Banach space. If p is a polynomial, the spectrum of $p(T)$ consists precisely of those points μ such that $p(\lambda) = \mu$ for some $\lambda \in \sigma(T)$. In symbolic form, $\sigma(p(T)) = p(\sigma(T))$.*

Proof. We can assume that $n \geqslant 1$ and $\alpha_n = 1$, leaving the case $n = 0$ to the reader. For a fixed μ let the zeros of $p(\lambda) - \mu$ be $\beta_1, \ldots, \beta_n$, so that

$$p(T) - \mu I = (T - \beta_1) \cdots (T - \beta_n). \tag{3-6}$$

If $T - \beta_1, \ldots, T - \beta_n$ have continuous inverses defined on all of X, so does $p(T) - \mu I$, the inverse of the latter being the product of the inverses of the former in the reverse order. Hence if $\mu \in \sigma(p(T))$, there must be some β_k such that $\beta_k \in \sigma(T)$. Since $p(\beta_k) = \mu$, this shows that $\sigma(p(T)) \subset p(\sigma(T))$. Suppose, on the other hand, that some β_k, say β_1, is in $\sigma(T)$. If $T - \beta_1$ has an inverse, the range of $T - \beta_1$ is not all of X, and (3-6) shows that the range of $p(T) - \mu I$ is likewise not all of X; hence $\mu \in \sigma(p(T))$. If $T - \beta_1$ has no inverse, we see by exchanging the positions of the factors $T - \beta_1$ and $T - \beta_n$ in (3-6) that $p(T) - \mu I$ also has no inverse, and again $\mu \in \sigma(p(T))$. This argument works just as well for any β_k as for β_1, and so the proof is complete. □

The conclusion of Theorem 3.4 can be restated as follows: *The equation $p(T)x = y$ has a unique solution for each $y \in X$ if and only if $p(\lambda) \neq 0$ when $\lambda \in \sigma(T)$.*

Theorem 3.5. *Suppose $T \in L(X)$, where X is a complex Banach space. Then $r_\sigma(T) \leq \|T^n\|^{1/n}$ for every positive integer n. Also, $\|T^n\|^{1/n}$ converges to $r_\sigma(T)$ as $n \to \infty$.*

Proof. Theorem 3.4 shows that $\sigma(T^n)$ consists of the nth powers of points of $\sigma(T)$. Hence $r_\sigma(T^n) = [r_\sigma(T)]^n$. We know (Theorem 3.1) that $r_\sigma(T^n) \leq \|T^n\|$. Hence $r_\sigma(T) \leq \|T^n\|^{1/n}$. It follows that

$$r_\sigma(T) \leq \liminf_{n\to\infty} \|T^n\|^{1/n}.$$

This, together with (3-5), leads to the final conclusion of the theorem. □

An operator T is said to be *nilpotent* if $T^n = 0$ for some n, and *quasinilpotent* if $\|T^n\|^{1/n} \to 0$ as $n \to \infty$. By Theorem 3.5, a quasinilpotent operator is one whose spectrum consists only of the complex number zero.

Example. Let T on ℓ^1 into ℓ^1 be defined by the infinite matrix

$$\begin{Vmatrix} 0 & 0 & 0 & 0 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & \\ 0 & \frac{1}{2} & 0 & 0 & 0 & \\ 0 & 0 & \frac{1}{3} & 0 & 0 & \\ 0 & 0 & 0 & \frac{1}{4} & 0 & \\ \vdots & & & & \ddots & \end{Vmatrix}$$

The norm of T as an element of $L(\ell^1)$ is the supremum of the ℓ^1 norms of the column vectors in its matrix representation. (See problem 4 of § IV.6.) By direct calculation we find that $\|T^n\| = 1/n!$. Hence $\|T^n\|^{1/n} \to 0$ and $\sigma(T) = \{0\}$.

PROBLEMS

1. Let S be a compact set in the plane. Let $\{\alpha_n\}$ be a sequence of points of S everywhere dense in S. Define $T \in L(\ell^2)$ by $Tx = y$, where $x = \{\xi_n\}$, $y = \{\alpha_n \xi_n\}$. Then $S = \sigma(T)$. Each α_n is an eigenvalue; for the other points $\lambda \in S$, the range of $\lambda - T$ is dense in ℓ^2 and the inverse is discontinuous.
2. The function $(\lambda - T)^{-1}$ cannot be continued analytically beyond $\rho(T)$. [Show that $\|R_\lambda\|^{-1} \to 0$ as λ approaches a point in $\sigma(T)$.]
3. Suppose X is a Banach space and $T \in L(X)$. If, for some fixed λ, the series $\Sigma_1^\infty \lambda^{-n} T^{n-1} x$ converges for each $x \in X$, then $\lambda \in \rho(T)$.

4. Suppose X is a Banach space and $T \in L(X)$. If $\lim_{n\to\infty} \|T^n x\|^{1/n} = 0$ for each $x \in X$, then T is quasinilpotent.
5. If $X = C[a, b]$ and T is a Volterra-type integral operator (see § IV.2), then T is quasinilpotent.
6. If T does not belong to $L(X)$, $\sigma(T)$ may not be compact. Example 3, § I.5 illustrates a situation in which the spectrum of a certain differential operator consists of the points $-n^2, n = 0, 1, 2, \ldots$. Example 4, § I.5 shows a differential operator whose spectrum is composed of all λ whose real part is $\leqslant 0$.
7. Let $X = C[0, 1]$, and let $\mathscr{D}$ be the set of continuously differentiable functions in X. Let T be the operator with domain $\mathscr{D}$ defined by $(Tx)(s) = x'(s)$. (See Example 1 of § IV.5.) This (closed) operator has an empty resolvent set. If T_1 is the restriction of T to $\mathscr{D}(T_1) = \{x \in \mathscr{D} : x(0) = 0\}$, then T_1 is closed and $\rho(T_1) = C$.
8. Let $A, B, C \in L(X)$, where X is complete.
   a. If $(I - AB)C = C(I - AB) = I$, then $I - BA$ has an inverse in $L(X)$. [Try $I + BCA$.]
   b. The nonzero points of $\rho(AB)$ and $\rho(BA)$ are the same. Hence $r_\sigma(AB) = r_\sigma(BA)$.
   c. There exist *no* operators $A, B \in L(X)$ such that $AB - BA = I$. (This fact has relevance in quantum physics.)
9. If X is a complex Banach space and $AB = BA$, where A and B are in $L(X)$, then $r_\sigma(AB) \leqslant r_\sigma(A) r_\sigma(B)$.
10. Suppose $A \in L(X)$, $A^n \neq 0$ for all n, and let $\|A^n\|^{1/n}$ be monotonic in n. Show that $\|A^{n+1}\|/\|A^n\| \leqslant \|A^n\|^{1/n}$, whence it can be inferred that $\overline{\lim}\, (\|A^{n+1}\|/\|A^n\|) = \lim \|A^n\|^{1/n}$. For a case in which $\|A^n\|^{1/n}$ is not monotone, let $A \in L(\ell^1)$ be defined by $Ax = y$, where $\eta_1 = 0$, $\eta_{2k} = \xi_{2k-1}$, $\eta_{2k+1} = 2\xi_{2k}$, $k = 1, 2, \ldots$.
11. Theorem IV.1.4 remains true when the condition that $\|A\| < 1$ is replaced by the condition that $r_\sigma(A) < 1$.
12. Suppose $T \in L(X)$, where X is a complex Banach space. Let r_0 be a real number larger than $r_\sigma(T)$, and suppose the series
$$f(\lambda) = \sum_0^\infty a_n \lambda^n$$
converges for $|\lambda| < r_0$. Then we may define
$$f(T) = \sum_0^\infty a_n T^n.$$
Show that this series converges to an operator in $L(X)$.
13. Let X be a complex Banach space. Suppose $A, B \in L(X)$, $AB = BA$, and A has an inverse in $L(X)$. If $r_\sigma(B) < 1/r_\sigma(A^{-1})$, then $(A + B)^{-1}$ exists and belongs to $L(X)$. Compare this result with Theorem IV.1.5.
14. Let X be a Banach space. If $T \in L(X)$, then $\lambda(\lambda - T)^{-1}$ converges in $L(X)$ to the identity I as $\lambda \to \infty$.

V.4 SUBDIVISIONS OF THE SPECTRUM

We can classify the various values of the parameter λ according to the state of the operator $\lambda - T$, using the definition of "states" as made in § IV.10. We say that λ is in one of the classes $\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{III}_2, \mathrm{III}_3$ if $\lambda - T$ is in the corresponding state, as an operator on $\mathscr{D}(T)$ into X. For this classification we do not insist that T be continuous or even closed. According to the definitions, $\lambda \in \rho(T)$ if and only if λ is in class I_1 or II_1. It has been customary to group the remaining classes as follows, thus dividing the spectrum into three mutually exclusive parts:

Classes I_2 and II_2 = the continuous spectrum, denoted by $C\sigma(T)$;

Classes III_1 and III_2 = the residual spectrum, denoted by $R\sigma(T)$;

Classes I_3, II_3, and III_3 = the point spectrum (eigenvalues), denoted by $P\sigma(T)$.

In § 2 we proved that $\sigma(T) = \sigma(T')$ when T is densely defined. However, if a point λ is in $P\sigma(T)$, it need not belong to $P\sigma(T')$; it may in fact be in $R\sigma(T')$. Also, if $\lambda \in R\sigma(T)$, then λ must be in $P\sigma(T')$. These facts and others similar to them may be deduced from the state diagram.

The Approximate Point Spectrum

If λ is an eigenvalue of T, then there exists an $x \in \mathscr{D}(T)$ such that $\|x\| = 1$ and $Tx = \lambda x$. More generally, we call λ an approximate eigenvalue if to each $\varepsilon > 0$ there corresponds some $x \in \mathscr{D}(T)$ such that $\|x\| = 1$ and $\|\lambda x - Tx\| < \varepsilon$. The set of all such λ is called the *approximate point spectrum* (of T). It is easy to see from Theorem II.1.2 that λ is an approximate eigenvalue if and only if $\lambda - T$ does not have a bounded inverse, that is, if and only if $\lambda - T$ is in a state whose subscript is 2 or 3. We thus conclude that the only points of $\sigma(T)$ that are *not* approximate eigenvalues are those for which $\lambda - T$ is in state III_1.

Theorem 4.1. *The approximate point spectrum of T is a closed subset of $\sigma(T)$ and includes all points on the boundary of $\sigma(T)$. If $T \in L(X)$, where X is complex, then the approximate point spectrum of T is nonempty.*

Proof. The key fact is that the set $\{\lambda : \lambda - T \text{ is in state } \mathrm{III}_1\}$ is open and hence is in the interior of $\sigma(T)$. The proof is left as an exercise (problem 1). Since $\sigma(T)$ is closed, the remarks above show that the approximate point spectrum is closed and includes all boundary points of $\sigma(T)$. If $T \in L(X)$, where X is complex, then $\sigma(T)$ is compact and nonempty and so must have a boundary point. $\square$

In contrast to the situation in Theorem 4.1, the point spectrum of a bounded linear operator on a Banach space may be empty or may fail to be closed. See problems 2 and 3.

The Fredholm Spectrum

Now suppose that X is complete and T is closed. The *Fredholm spectrum* of T, denoted by $\sigma_\Phi(T)$, is the set of λ for which $\lambda - T$ is *not* a Fredholm operator. The complement of $\sigma_\Phi(T)$ is called the Fredholm set of T, or the Fredholm resolvent of T. It is not difficult to show from Theorem IV.13.6 that the Fredholm set of T is an open set containing $\rho(T)$ (problem 4). Thus $\sigma_\Phi(T)$ is a closed subset of $\sigma(T)$. Furthermore, if T is a bounded linear operator on an infinite-dimensional complex Banach space, the Fredholm spectrum of T cannot be empty (problem 8, § VII. 3).

Examples

For illustrative purposes we shall discuss two operators, each of which may be considered as a bounded operator on ℓ^p into ℓ^p for any selected value of p, $1 \leqslant p \leqslant \infty$. We shall analyze the spectrum of each operator and see how the resulting classification of each spectral value λ depends on the value of p. Some of the details are left for the problems. We always write $x = \{\xi_n\}$, $y = \{\eta_n\}$, $n = 1, 2, \ldots$.

Example 1. Let T on ℓ^p into ℓ^p be defined by the infinite matrix

$$\left\| \begin{matrix} 0 & 1 & 0 & 0 & & \cdot \\ 0 & 0 & 1 & 0 & & \\ 0 & 0 & 0 & 1 & & \\ 0 & 0 & 0 & 0 & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \cdot \end{matrix} \right\|$$

so that $(\lambda - T)x = y$ means $\eta_k = \lambda\xi_k - \xi_{k+1}$, $k = 1, 2, \ldots$. It is easy to see that $\|T\| = 1$ for each value of p. Hence $|\lambda| > 1$ implies $\lambda \in \rho(T)$. It is readily seen that when $p = \infty$ and $|\lambda| \leqslant 1$, λ is an eigenvalue; the corresponding eigenmanifold is generated by the vector $(1, \lambda, \lambda^2, \ldots)$. When $1 \leqslant p < \infty$, we have this same eigenvalue and eigenmanifold if $|\lambda| < 1$. But $(\lambda - T)^{-1}$ exists if $|\lambda| = 1$ and $1 \leqslant p < \infty$, for $(1, \lambda, \lambda^2, \ldots)$ is not an element of ℓ^p in this case. Since $\sigma(T)$ is closed, we see that, for each value of p, $\sigma(T)$ is the set $\{\lambda : |\lambda| \leqslant 1\}$. When $|\lambda| = 1$ and $1 \leqslant p < \infty$, the inverse cannot be continuous, by Theorem 4.1.

To investigate the range of $\lambda - T$ we find by induction from $\eta_k = \lambda\xi_k - \xi_{k+1}$ that

$$\xi_{k+1} = \lambda^k\xi_1 - \lambda^{k-1}\eta_1 - \lambda^{k-2}\eta_2 - \cdots - \eta_k$$

and that

$$\xi_1 = \lambda^{-1}\eta_1 + \cdots + \lambda^{-k}\eta_k + \lambda^{-k}\xi_{k+1} \qquad \text{if } \lambda \neq 0.$$

Note that $\lambda^{-k}\xi_{k+1} \to 0$ as $k \to \infty$ if $|\lambda| \geq 1$ and $1 \leq p < \infty$ or if $|\lambda| > 1$ and $p = \infty$. This enables us to find x in terms of y when $y = (\lambda - T)x$. In particular, we get the formulas for the resolvent operator when $|\lambda| > 1$. When $|\lambda| = 1$ and $1 \leq p < \infty$, the range of $\lambda - T$ is dense in ℓ^p, for it is easily seen to contain y if the number of nonzero components of y is finite. The range cannot be all of ℓ^p in this case, however. Why not? We leave it until later to show that the range is *not* dense in ℓ^p for the case $|\lambda| = 1$ and $p = \infty$. When $\lambda = 0$, the range of $\lambda - T$ is obviously all of ℓ^p, for each p. When $0 < |\lambda| < 1$ and y is in the range of $\lambda - T$, any x such that $(\lambda - T)x = y$ is given by $x = \xi_1(1, \lambda, \lambda^2, \ldots) - \lambda^{-1}z$, where $z = \{\zeta_k\}$, $\zeta_1 = 0$, and

$$\zeta_{k+1} = \lambda^k\eta_1 + \lambda^{k-1}\eta_2 + \cdots + \lambda\eta_k \qquad \text{if } k \geq 1. \tag{4-1}$$

After discussing Example 2, we shall see that the range of $\lambda - T$ is all of ℓ^p in this case, for each p.

Example 2. Let A on ℓ^p into ℓ^p be defined by the infinite matrix

$$\left\|\begin{matrix} 0 & 0 & 0 & 0 & & \cdot \\ 1 & 0 & 0 & 0 & & \\ 0 & 1 & 0 & 0 & & \\ 0 & 0 & 1 & 0 & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \cdot \end{matrix}\right\|$$

This is the transpose of the matrix of Example 1. When $p = \infty$, A can be identified with the conjugate operator T', T being the operator of Example 1, with $p = 1$. If $1 < p < \infty$ and $p' = p/(p-1)$, A for p' can be identified with the conjugate T' of T for p. Finally, T for $p = \infty$ can be identified with the conjugate A' of A for $p = 1$. These remarks, and the state diagram of § IV.10, will help us in analyzing the spectrum of A. In particular, the relations between an operator and its conjugate assure us that $\|A\| = 1$ (which, of course, we can see directly) and that $\sigma(A) = \sigma(T) = \{\lambda : |\lambda| \leq 1\}$.

The equation $(\lambda - A)x = y$ is expressed by the equations $\eta_1 = \lambda\xi_1$, $\eta_{k+1} = -\xi_k + \lambda\xi_{k+1}$, $k \geq 1$. When $\lambda = 0$ we see that the inverse of $\lambda - A$ exists and is continuous, since $\|Ax\| = \|x\|$. The range is not dense in ℓ^p, however, because

$y=-Ax$ implies $\eta_1=0$; the range is a proper closed subspace of ℓ^p. When $\lambda\neq 0$ the inverse of $\lambda-A$ exists; the vector $x=(\lambda-A)^{-1}y$, for y in the range of $\lambda-A$, is expressed by

$$\xi_k=\lambda^{-k}\eta_1+\lambda^{1-k}\eta_2+\cdots+\lambda^{-1}\eta_k. \tag{4-2}$$

In particular, these equations define the resolvent of A when $|\lambda|>1$.

We can use equations (4-2) in discussing the range of $\lambda-T$. A scrutiny of (4-1) and (4-2) shows that the z of (4-1) can be written in the form $z=A(\lambda^{-1}-A)^{-1}y$. The solutions of $(\lambda-T)x=y$ are then $x=\xi_1(1,\lambda,\lambda^2,\ldots)-\lambda^{-1}A(\lambda^{-1}-A)^{-1}y$, where ξ_1 is arbitrary. If $0<|\lambda|<1$, this formula is applicable for every y in ℓ^p, for all p, and so we see that the range of $\lambda-T$ is all of ℓ^p in this case. The result is valid for $0\leq|\lambda|<1$ if we write it in the form $x=\xi_1(1,\lambda,\lambda^2,\ldots)+A(\lambda A-I)^{-1}y$.

Next, we show that, for $p=\infty$, the range of $\lambda-A$ is not dense in ℓ^∞ if $|\lambda|=1$. In fact, we show that y is not in the range if $\|y-w\|=\varepsilon<1$, where $w=(\lambda^{-1},\lambda^{-2},\lambda^{-3},\ldots)$. If we assume that $x\in\ell^\infty$ and $(\lambda-A)x=y$, we can write $y=w+(\theta_1,\theta_2,\ldots)$, where $|\theta_n|\leq\varepsilon$, and from (4-2) we have

$$\xi_n=n\lambda^{-n-1}+\lambda^{-n-1}(\lambda\theta_1+\lambda^2\theta_2+\cdots+\lambda^n\theta_n),$$

so that

$$|n^{-1}\xi_n-\lambda^{-n-1}|=n^{-1}|\lambda\theta_1+\cdots+\lambda^n\theta_n|\leq\varepsilon.$$

Letting $n\to\infty$, we obtain the contradiction $1\leq\varepsilon$.

An argument of a similar kind will show for the case $p=\infty$, $|\lambda|=1$, that the range of $\lambda-T$ is not dense in ℓ^∞. If $y=(\lambda-T)x$, it may be verified that $y=(\lambda\xi_1,0,0,\ldots)-(\lambda^{-1}-A)T(\lambda x)$, and from this it may be shown that y is not in the range of $\lambda-T$ if $\|y-w\|=\varepsilon<1$, where $w=(\lambda,\lambda^2,\lambda^3,\ldots)$.

It is now possible to make a complete classification of the points of the spectra of T and A for all values p. See the problems.

PROBLEMS

1. Let X be a normed linear space, and let T be a linear operator whose domain and range are in X.
   a. Suppose μ is such that $\mu-T$ has a continuous inverse, and suppose $|\lambda-\mu|M(\mu)<\frac{1}{2}$, using the notation of Lemma 2.1. Then $\overline{\mathcal{R}(T_\mu)}$ is not a proper subset of $\overline{\mathcal{R}(T_\lambda)}$. [*Hint.* Assume the contrary, and use Riesz's lemma, where θ is chosen so that $|\lambda-\mu|M(\mu)/(1-|\lambda-\mu|M(\mu))<\theta<1$.]
   b. $\{\lambda:\lambda-T$ is in state $\mathrm{III}_1\}$ is open in $\boldsymbol{C}$.
   c. If X is complete and T is closed, then $\{\lambda:\lambda-T$ is in state $\mathrm{I}_3\}$ is open in $\boldsymbol{C}$.
2. Let T be the operator in problem 1 of § 3. The approximate point spectrum and the Fredholm spectrum of T both coincide with $\sigma(T)$. $P\sigma(T)$ is dense in $\sigma(T)$.

Let T be the operator in problem 5 of § 3. The approximate point spectrum and the Fredholm spectrum of T both coincide with $\sigma(T)$, but $P\sigma(T)$ is empty.

Let T be a closed linear operator on a Banach space X.

a. The Fredholm spectrum of T contains the continuous spectrum, all points in classes II_3 and III_2, and some of the points in classes I_3, III_1, and III_3.
b. The Fredholm set of T is an open set containing $\rho(T)$.
c. If $\overline{\mathscr{D}(T)} = X$, then $\sigma_\Phi(T') = \sigma_\Phi(T)$.

Suppose $T \in L(X)$, where X is a complex Banach space. If $p(\lambda)$ is a polynomial, $\sigma_\Phi(p(T)) = p(\sigma_\Phi(T))$.

For the T of Example 1 there is the following classification of points of $\sigma(T)$. (*a*) $\lambda = 0$: I_3; (*b*) $0 < |\lambda| < 1$: I_3; (*c*) $|\lambda| = 1$: II_2 if $1 \leqslant p < \infty$, and III_3 if $p = \infty$; (*d*) the Fredholm spectrum consists of all $|\lambda| = 1$.

If $(\lambda - A)x = y$ in Example 2, show that $\eta_1 + \lambda\eta_2 + \cdots + \lambda^k \eta_{k+1} = \lambda^{k+1}\xi_{k+1}$. This may be used to show that the range of $\lambda - A$ is not dense in ℓ^p if $|\lambda| < 1$ and $p = \infty$, or if $|\lambda| \leqslant 1$ and $p = 1$.

For the A of Example 2 there is the following classification of points of $\sigma(A)$. (*a*) $|\lambda| < 1$: III_1; (*b*) $|\lambda| = 1$: II_2 if $1 < p < \infty$, and III_2 if $p = 1$ or ∞; (*c*) $|\lambda| = 1$: this is both the approximate point spectrum and the Fredholm spectrum. Some of these results may be obtained by using problem 6 and the state diagram.

If $|\lambda| = 1$ and $1 \leqslant p < \infty$ in Example 2, take $x_n = (1, \lambda^{-1}, \ldots, \lambda^{1-n}, 0, 0, \ldots)$, and show that $(\lambda - A)^{-1}$ is not continuous.

Let $\alpha_1, \alpha_2, \ldots$ be scalars such that $\sup_k |\alpha_k| < \infty$. Let $\beta_2, \beta_3, \ldots$ be scalars such that $\Sigma_2^\infty |\beta_k| < \infty$. Define $T \in L(\ell^1)$ by $\eta_1 = \Sigma_1^\infty \alpha_i \xi_i$, $\eta_k = \beta_k \xi_1$ if $k \geqslant 2$. Discuss $\sigma(T)$ and find the resolvent operator.

Let $T \in L(\ell^1)$ be defined by the infinite matrix

$$\left\| \begin{matrix} 0 & 1 & 1 & 1 & & \cdot \\ 1 & 0 & 0 & 0 & \cdot & \cdot \\ 0 & 1 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 1 & 0 & & \\ \cdot & & & & & \\ \cdot & & & & & \\ \cdot & & & & & \cdot \end{matrix} \right\|$$

Show that $\sigma(T)$ consists of $\lambda = (1+\sqrt{5})/2$ and all λ such that $|\lambda| \leqslant 1$. Classify the points of $\sigma(T)$. This operator has an interesting connection with the Fibonacci numbers, as may be seen by computing the matrix representation of T^n. The operator has been studied by Halberg [1, pages 21–29; and 2].

The operator $A \in L(\ell^1)$ defined by the matrix (α_{ij}) with $\alpha_{ij} = 1$ if $|i-j| = 1$, $\alpha_{ij} = 0$ otherwise, is interesting. Its spectrum consists of the interval $-2 \leqslant \lambda \leqslant 2$ of the real axis. The points $\lambda = \pm 2$ are classified II_2; the rest of $\sigma(A)$ is classified III_2 (see Halberg [1, pages 29–36]). We may also consider A as an element of $L(\ell^2)$. The spectrum is the same as before, but now all points are classified II_2 (Hellinger [1, pages 231–232]).

V.5 REDUCIBILITY

Let T be a linear operator with domain and range in the linear space X. For convenience we shall sometimes write $\mathscr{D}$ instead of $\mathscr{D}(T)$ for the domain of T. A subspace M of X is said to be *invariant* under T if $T(\mathscr{D} \cap M) \subset M$. We can then talk about the *restriction* of T to M, with M in place of X and $\mathscr{D} \cap M$ in place of $\mathscr{D}$.

Definition. Let M_1 and M_2 be linearly independent subspaces of X such that $X = M_1 \oplus M_2$, and let P_1, P_2 be the associated projections of X onto M_1 and M_2, respectively. The operator T is said to be *completely reduced* by the pair (M_1, M_2) if these subspaces are invariant under T and if $P_i\mathscr{D} \subset \mathscr{D}$, $i = 1, 2$.

This definition can be extended in an obvious manner to give meaning to the statement "T is completely reduced by the set of subspaces $M_1, \ldots, M_n$."

To illustrate the meaning of the foregoing concepts, let M_1, M_2, P_1, and P_2 be as above, and suppose only that T is an operator on X whose domain $\mathscr{D}$ satisfies $P_i\mathscr{D} \subset \mathscr{D}$, $i = 1, 2$. Then for $x \in \mathscr{D}$ we may write

$$Tx = P_1TP_1x + P_1TP_2x + P_2TP_1x + P_2TP_2x.$$

For $i, j = 1, 2$, let T_{ij} be the restriction of P_iTP_j to M_j considered as a mapping into M_i. If elements in X are represented as column vectors, $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, where $x_i = P_ix$, then T may be represented by the "operator matrix"

$$T \sim \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix}.$$

Now M_1 is invariant under T if and only if T_{21} is the zero operator, and T is completely reduced by (M_1, M_2) if and only if both T_{21} and T_{12} are zero.

If an operator is completely reduced by a pair of subspaces, the operator may be studied by studying the restrictions of it to these subspaces. If these restrictions can also be completely reduced, the study of the operator may be simplified further. Evidently, if X is n-dimensional and $\mathscr{D} = X$, the greatest simplification will occur if we can find a set of one-dimensional subspaces $M_1, \ldots, M_n$, which completely reduce the operator. This is the same as asking that X have a basis $u_1, \ldots, u_n$, each element of which is an eigenvector. The matrix representing the operator will then have eigenvalues down its main diagonal, and all other elements will be zero. This situation can occur only for operators of a very special sort.

Theorem 5.1. *Let P_1, P_2 be the projections determined by the direct sum $X = M_1 \oplus M_2$. Then an operator T with domain $\mathscr{D}$ is completely reduced by (M_1, M_2) if and only if*

(a) $P_1\mathscr{D} \subset \mathscr{D}$, *and*

(b) $P_1 Tx = TP_1 x, \qquad$ *for all* $x \in \mathscr{D}$.

Proof. Suppose (M_1, M_2) completely reduces T. Then (a) is true by definition, and if $x \in \mathscr{D}$, we may write $x = x_1 + x_2$, with $x_i \in \mathscr{D} \cap M_i$. Then $P_1 Tx = P_1 Tx_1 + P_1 Tx_2 = Tx_1$, since $Tx_1 \in M_1 = \mathscr{R}(P_1)$ and $Tx_2 \in M_2 = \mathscr{N}(P_1)$. But $TP_1 x = Tx_1$, so (b) is true.

Conversely, suppose (a) and (b) are true. If $x \in \mathscr{D} \cap M_1$, then $Tx = TP_1 x = P_1 Tx \in M_1$. If $x \in \mathscr{D} \cap M_2$, then $0 = P_1 x = TP_1 x = P_1 Tx$, and so $Tx \in \mathscr{N}(P_1) = M_2$. Also, (a) implies that P_2 $(= I - P_1)$ maps $\mathscr{D}$ into $\mathscr{D}$. Thus (M_1, M_2) completely reduces T. □

One method of studying an operator, then, is to look for projections that commute with T and map $\mathscr{D}(T)$ into itself. It turns out that such projections often arise as "functions" of T. We shall consider the problem of constructing these commuting projections in § 9, § 11 and throughout much of Chapter VI.

For formal reference we now state several elementary theorems. The proofs are left to the reader.

Theorem 5.2. *Let T be completely reduced by (M_1, M_2), and let T_i be the restriction of T to M_i. Then (a) $\mathscr{D}(T) = \mathscr{D}(T_1) \oplus \mathscr{D}(T_2)$; (b) $\mathscr{N}(T) = \mathscr{N}(T_1) \oplus \mathscr{N}(T_2)$; (c) $\mathscr{R}(T) = \mathscr{R}(T_1) \oplus \mathscr{R}(T_2)$; (d) T^{-1} exists if and only if T_1^{-1} and T_2^{-1} exist, in which case T^{-1} is completely reduced by (M_1, M_2) and the restriction of T^{-1} to M_i is T_i^{-1}; (e) $\mathscr{R}(T) = X$ if and only if $\mathscr{R}(T_1) = M_1$ and $\mathscr{R}(T_2) = M_2$.*

Theorem 5.3. *Consider the situation of Theorem* 5.2, *supposing now that X is a normed linear space and that the projections P_1, P_2 determined by the direct sum $X = M_1 \oplus M_2$ are continuous. Then $\mathscr{R}(T)$ is dense in X if and only if $\mathscr{R}(T_1)$ and $\mathscr{R}(T_2)$ are dense in M_1 and M_2, respectively. If T^{-1} exists, it is continuous if and only if T_1^{-1} and T_2^{-1} are continuous.*

We remark that P_1 and P_2 will certainly be continuous when X is complete and M_1 and M_2 are closed (Theorem IV.12.2).

When T is completely reduced by (M_1, M_2), so is $\lambda - T$ for each λ. Thus Theorems 5.2 and 5.3 lead easily to the following result.

Theorem 5.4. *Under the conditions of Theorem* 5.3, *we have*

(*a*) $\sigma(T)=\sigma(T_1)\cup\sigma(T_2)$.

(*b*) $P\sigma(T)=P\sigma(T_1)\cup P\sigma(T_2)$.

If it is further assumed that $\sigma(T_1)$ *and* $\sigma(T_2)$ *have no points in common, it follows that*

(*c*) $C\sigma(T)=C\sigma(T_1)\cup C\sigma(T_2)$.

(*d*) $R\sigma(T)=R\sigma(T_1)\cup R\sigma(T_2)$.

PROBLEMS

1. Write out the proofs of Theorems 5.2 and 5.3.
2. Write out the proof of Theorem 5.4.
3. Let $\pi(T)$ denote the approximate point spectrum of T. Let T, T_1, and T_2 be as in Theorem 5.3. (You may need to assume that X is complete and T is closed.)
   a. What can you say about $\pi(T)$, $\pi(T_1)$, and $\pi(T_2)$?
   b. What can you say about $\sigma_\Phi(T)$, $\sigma_\Phi(T_1)$, and $\sigma_\Phi(T_2)$?
4. If $T\in L(X)$, where X is a normed linear space, if T is completely reduced by (M_1, M_2) and if the corresponding projections P_1, P_2 are continuous, then T' is completely reduced by $(M_1^\perp, M_2^\perp)$. See problem 6, § IV.12.

V.6 THE ASCENT AND DESCENT OF AN OPERATOR

In this section, X denotes a linear space and T is a linear operator with domain and range in X. The considerations here are all algebraic. The main result is Theorem 6.2. It will be needed at a crucial point in the development of the spectral theory of a compact operator (in the next section). Later (in § 10), the results of this section will play a major role in the study of isolated points of the spectrum.

We define T^n by induction, with $T^0=I$ and $T^1=T$. Then $\mathscr{D}(T^0)=X$ and, if $n\geqslant 1$, $\mathscr{D}(T^n)$ is the set of all x in $\mathscr{D}(T^{n-1})$ such that $T^{n-1}x$ is in $\mathscr{D}(T)$. If $\mathscr{D}(T)\neq X$, $\mathscr{D}(T^n)$ is usually a proper subset of $\mathscr{D}(T^{n-1})$ (problem 3). The null space of T^n is the set $\mathcal{N}(T^n)$ of all x in $\mathscr{D}(T^n)$ such that $T^nx=0$. If $x\in\mathcal{N}(T^i)$, for some i, then $T^ix=0\in\mathscr{D}(T^j)$ for $j=0, 1, 2, \ldots$. It follows that $\mathcal{N}(T^i)\subset\mathscr{D}(T^j)$ for all $i, j=0, 1, 2, \ldots$. Clearly the null spaces of the iterates of T form an increasing chain of subspaces:

$$(0)=\mathcal{N}(T^0)\subset\mathcal{N}(T)\subset\mathcal{N}(T^2)\subset\cdots.$$

It is easy to see that

$$\mathcal{N}(T^{n+1})=\{x\in\mathscr{D}(T): Tx\in\mathcal{N}(T^n)\}.$$

From this it follows that if $\mathcal{N}(T^n)$ coincides with $\mathcal{N}(T^{n+1})$, it coincides with all the $\mathcal{N}(T^k)$ for $k>n$.

Definition. The smallest nonnegative integer p such that $\mathcal{N}(T^p) = \mathcal{N}(T^{p+1})$ is called the *ascent* of T and denoted by $\alpha(T)$. If no such integer exists, we set $\alpha(T) = \infty$. Note that $\alpha(T) = 0$ if and only if T is one-to-one.

Next, we consider the ranges $\mathcal{R}(T^n)$ of the iterates of T. They too form a nested chain of subspaces:

$$X = \mathcal{R}(T^0) \supset \mathcal{R}(T) \supset \mathcal{R}(T^2) \supset \cdots.$$

[If $x \in \mathcal{D}(T^{n+1})$, then $T^{n+1}x = T(T^n x) = T^2(T^{n-1}x) = \cdots = T^n(Tx) \in \mathcal{R}(T^n)$.] Evidently,

$$\mathcal{R}(T^{n+1}) = T\{\mathcal{R}(T^n) \cap \mathcal{D}(T)\}.$$

Hence, if $\mathcal{R}(T^n) = \mathcal{R}(T^{n+1})$, it follows that $\mathcal{R}(T^n) = \mathcal{R}(T^k)$ for $k > n$.

Definition. The smallest nonnegative integer q such that $\mathcal{R}(T^q) = \mathcal{R}(T^{q+1})$ is called the *descent* of T and denoted by $\delta(T)$. We set $\delta(T) = \infty$ if for each n, $\mathcal{R}(T^{n+1})$ is a proper subspace of $\mathcal{R}(T^n)$. Note that $\delta(T) = 0$ if and only if $\mathcal{R}(T) = X$.

We begin with a technical lemma. The condition of the lemma will be satisfied in our later work when X is a Banach space and T is a closed linear operator with a nonempty resolvent set. The lemma is unnecessary when $\mathcal{D}(T) = X$.

Lemma 6.1. *Suppose $\mathcal{R}(\lambda - T) = X$ for some λ. Then, given $i, j = 0, 1, 2, \ldots$, every $x \in X$ can be written in the form $x = u + v$, where $u \in \mathcal{D}(T^i)$ and $v \in \mathcal{R}(T^j)$; that is,*

$$X = \mathcal{D}(T^i) + \mathcal{R}(T^j). \tag{6-1}$$

Proof. We have $X = \mathcal{R}[(\lambda - T)^n]$ for $n = 0, 1, 2, \ldots$. (Thus (6-1) is obviously true when $\lambda = 0$.) Let $n = i + j$. Given $y \in X$, there is an $x \in \mathcal{D}[(\lambda - T)^n] = \mathcal{D}(T^n)$ such that

$$y = (\lambda - T)^n x = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \lambda^k T^{n-k} x.$$

For $0 \le k \le i$, $T^{n-k}x \in \mathcal{R}(T^{n-k}) \subset \mathcal{R}(T^{n-i}) = \mathcal{R}(T^j)$ and, for $i \le k \le n$, $T^{n-k}x \in \mathcal{D}(T^k) \subset \mathcal{D}(T^i)$; hence $y \in \mathcal{R}(T^j) + \mathcal{D}(T^i)$. □

Theorem 6.2. *If $\alpha(T) = p < \infty$ and $\delta(T) = q < \infty$, then $\alpha(T) \le \delta(T)$. If, in addition, $\mathcal{D}(T) = X$ or $\mathcal{R}(\lambda - T) = X$ for some λ, then $\alpha(T) = \delta(T) = p$ and*

$$X = \mathcal{R}(T^p) \oplus \mathcal{N}(T^p). \tag{6-2}$$

Proof. Fix $j \geq 1$. If $x \in \mathscr{R}(T^p) \cap \mathscr{N}(T^j)$, then $x = T^p v$ for some v, and $0 = T^j x = T^{p+j} v$. Thus $v \in \mathscr{N}(T^{p+j}) = \mathscr{N}(T^p)$, since $\alpha(T) = p$. This shows that $x = T^p v = 0$, and so

$$\mathscr{R}(T^p) \cap \mathscr{N}(T^j) = (0), \qquad j = 1, 2, \ldots. \tag{6-3}$$

Now, for $x \in \mathscr{D}(T^q)$, we have $T^q x \in \mathscr{R}(T^q) = \mathscr{R}(T^{q+j})$, since $\delta(T) = q$. Then $T^q x = T^{q+j} v$ for some v, and so $0 = T^q(x - T^j v)$. Thus $x = T^j v + (x - T^j v) \in \mathscr{R}(T^j) + \mathscr{N}(T^q)$. This proves that

$$\mathscr{D}(T^q) \subset \mathscr{R}(T^j) + \mathscr{N}(T^q), \qquad j = 1, 2, \ldots. \tag{6-4}$$

If $x \in \mathscr{N}(T^{q+1}) \subset \mathscr{D}(T^q)$, then from (6-4) we may write $x = x_1 + x_2$, where $x_1 \in \mathscr{R}(T^p)$ and $x_2 \in \mathscr{N}(T^q)$. Hence $x_1 = x - x_2 \in \mathscr{N}(T^{q+1}) + \mathscr{N}(T^q) = \mathscr{N}(T^{q+1})$. In view of (6-3), this implies that $x_1 = 0$ and $x = x_2 \in \mathscr{N}(T^q)$. Thus $\mathscr{N}(T^{q+1}) = \mathscr{N}(T^q)$, which shows that $p = \alpha(T) \leq q$.

If $\mathscr{D}(T) = X$ or $\mathscr{R}(\lambda - T) = X$ for some λ, then from (6-4) and Lemma 6.1 we find that

$$X = \mathscr{R}(T^j) + \mathscr{N}(T^q), \qquad j = 1, 2, \ldots. \tag{6-5}$$

Taking $j = q$ in (6-3) and $j = p$ in (6-5), we have

$$X = \mathscr{R}(T^p) \oplus \mathscr{N}(T^q).$$

From this and (6-5) it is clear that $\mathscr{R}(T^j)$ cannot be a proper subset of $\mathscr{R}(T^p)$ for any $j \geq 1$. Hence $q = \delta(T) \leq p$, which proves that $p = q$. □

Theorem 6.2 with $\mathscr{D}(T) = X$ has been known for many years. The general case together with the other results in this section come from the work of Taylor [6], Kaashoek [1], and Lay [1]. A shorter proof of Theorem 6.2 may be based on the following characterizations of finite ascent and finite descent.

Theorem 6.3. *For $i, j = 0, 1, 2, \ldots$, there is an algebraic isomorphism (denoted by $\cong$) such that*

$$\mathscr{N}(T^{i+j}) / \mathscr{N}(T^i) \cong \mathscr{R}(T^i) \cap \mathscr{N}(T^j). \tag{6-6}$$

Thus, for $j = 1, 2, \ldots$, $\alpha(T) \leq p$ if and only if $\mathscr{R}(T^p) \cap \mathscr{N}(T^j) = (0)$.

Proof. Fix i and j, and observe that T^i maps $\mathscr{N}(T^{i+j})$ into $\mathscr{R}(T^i) \cap \mathscr{N}(T^j)$. If $y \in \mathscr{R}(T^i) \cap \mathscr{N}(T^j)$, then $y = T^i x$ for some $x \in \mathscr{D}(T^i)$. In fact, $x \in \mathscr{N}(T^{i+j})$ because $0 = T^j y = T^j(T^i x)$. Thus T^i maps $\mathscr{N}(T^{i+j})$ onto $\mathscr{R}(T^i) \cap \mathscr{N}(T^j)$. If we define $S : \mathscr{N}(T^{i+j}) / \mathscr{N}(T^i) \to \mathscr{R}(T^i) \cap \mathscr{N}(T^j)$ by $S([x]) = T^i x$, then S is the desired isomorphism. The second statement of the theorem follows from (6-6) and the fact that, for any $j = 1, 2, \ldots$, $\alpha(T) \leq p$ if and only if $\mathscr{N}(T^{p+j}) / \mathscr{N}(T^p) = (0)$. □

Theorem 6.4. *Suppose $\mathscr{D}(T)=X$ or $\mathscr{R}(\lambda-T)=X$ for some λ. Then, for $i, j=0,1,2,\ldots$, there is an algebraic isomorphism $\cong$ such that*

$$\mathscr{R}(T^i)/\mathscr{R}(T^{i+j})\cong X/[\mathscr{R}(T^j)+\mathscr{N}(T^i)]. \tag{6-7}$$

Thus, for $j=1,2,\ldots$, $\delta(T)\leq q$ if and only if $X=\mathscr{R}(T^j)+\mathscr{N}(T^q)$.

Proof. Fix i and j, and consider the class of all linear manifolds in $\mathscr{R}(T^j)$ whose intersection with $\mathscr{D}(T^i)$ is (0). A standard argument using Zorn's lemma implies that there is a maximal element M in this class. Clearly

$$\mathscr{D}(T^i)\oplus M=\mathscr{D}(T^i)+\mathscr{R}(T^j)=X,$$

by Lemma 6.1. Let P be the projection of X onto $\mathscr{D}(T^i)$ along M. (If $\mathscr{D}(T)=X$, then $P=I$, and the rest of the proof may be shortened somewhat.) Let Q be the quotient mapping of $\mathscr{R}(T^i)$ onto $\mathscr{R}(T^i)/\mathscr{R}(T^{i+j})$, and let $J=QT^iP$.

$$\begin{tikzcd}
\mathscr{D}(T^i)\oplus M \arrow[d, "P"'] \arrow[drr, dashed, "J"] & & \mathscr{R}(T^i) \arrow[d, "Q"] \\
\mathscr{D}(T^i) \arrow[urr, "T^i"'] & & \mathscr{R}(T^i)/\mathscr{R}(T^{i+j})
\end{tikzcd}$$

Clearly $\mathscr{R}(J)=\mathscr{R}(T^i)/\mathscr{R}(T^{i+j})$.

We next determine $\mathscr{N}(J)$. If $Jx=0$, then $T^iPx\in\mathscr{R}(T^{i+j})$; that is, $T^iPx=T^{i+j}w$ for some $w\in\mathscr{D}(T^{i+j})$. Hence $T^i(Px-T^jw)=0$, and

$$x=((I-P)x+T^jw)+(Px-T^jw)\in\mathscr{R}(T^j)+\mathscr{N}(T^i),$$

because $(I-P)x\in M\subset\mathscr{R}(T^j)$. Now suppose $x\in\mathscr{R}(T^j)$. Then $Px=x-(I-P)x\in\mathscr{R}(T^j)+M\subset\mathscr{R}(T^j)$; hence $T^iPx\in\mathscr{R}(T^{i+j})$ and so $Jx=0$. Also, if $x\in\mathscr{N}(T^i)$, then $Jx=QT^iPx=QT^ix=Q(0)=0$. We conclude that the null space of J is precisely $\mathscr{R}(T^j)+\mathscr{N}(T^i)$. The one-to-one mapping

$$\hat{J}:X/[\mathscr{R}(T^j)+\mathscr{N}(T^i)]\to\mathscr{R}(T^i)/\mathscr{R}(T^{i+j})$$

induced by J is the desired isomorphism. The second statement of the theorem follows easily from (6-7). $\square$

PROBLEMS

1. It can happen that $\delta(T)=0$ and $\alpha(T)=\infty$. Consider $X=\ell^2$, $\mathscr{D}(T)=X$, and $x=(\xi_1,\xi_2,\ldots)$, $Tx=(\xi_2,\xi_3,\ldots)$.
2. If $\mathscr{D}(T)\neq X$, it can happen that $\alpha(T)<\delta(T)$ when both are finite. Take $\mathscr{D}(T)\neq X$, $Tx=x$ if $x\in\mathscr{D}(T)$. Then $\alpha(T)=0$, $\delta(T)=1$.
3. Suppose that $\mathscr{D}(T)\neq X$, but $\mathscr{R}(\lambda-T)=X$ for some λ. Then for $i=1,2,\ldots$,

$$\mathscr{D}(T^i)/\mathscr{D}(T^{i+1})\cong X/\mathscr{D}(T).$$

[Consider the mapping from $\mathscr{D}(T^i)$ onto $[\mathscr{R}(T^i)+\mathscr{D}(T)]/\mathscr{D}(T)$ given by $x \mapsto [T^i x]$.]

4. If (6-2) holds, then T is completely reduced by $(\mathscr{R}(T^p), \mathscr{N}(T^p))$, even when $\mathscr{D}(T) \neq X$. Furthermore, $\alpha(T) \leq p$ and $\delta(T) \leq p$, even if it is not assumed that $\mathscr{R}(\lambda - T) = X$ for some λ.
5. Assume $\mathscr{D}(T) = X$, and suppose T is nilpotent; that is, suppose $T^p = 0$ for some p. Then $\alpha(\lambda - T) = \delta(\lambda - T) = 0$ for every $\lambda \neq 0$.
6. Suppose T is completely reduced by a pair (M_1, M_2) of complementary subspaces of X. Let T_i be the operator T acting in the space M_i. Show that
   a. $\alpha(T) = \sup\{\alpha(T_1), \alpha(T_2)\}$ (∞ is allowed).
   b. $\delta(T) = \sup\{\delta(T_1), \delta(T_2)\}$ (∞ is allowed).
7. Let M be a subspace invariant under T and let T_1 be the restriction of T to M. Denote elements of X/M by $[x]$, and define T_2 on X/M by $\mathscr{D}(T_2) = \{[x] : x \in \mathscr{D}(T)\}$, $T_2[x] = [Tx]$. (Verify that T_2 is well defined!) Show that
   a. $\alpha(T_1) \leq \alpha(T) \leq \alpha(T_1) + \alpha(T_2)$.
   b. $\delta(T_2) \leq \delta(T) \leq \delta(T_1) + \delta(T_2)$.
   c. The inequalities in (a) and (b) cannot be sharpened.
8. Suppose X is a Banach space and $T \in L(X)$. If to each $x \in X$ there corresponds a positive integer k such that $T^k x = 0$, then T is in fact nilpotent.
9. Let X be a Banach space, and let T be a closed linear operator such that $\rho(T) \neq \varnothing$. Suppose that $0 \in \sigma(T)$ and (6-2) holds for some $p \geq 1$. Show that 0 is an isolated point of $\sigma(T)$. Use the fact (proved later in problem 8 of § V.8) that T^p is a closed operator. [*Hint.* First show that $\mathscr{R}(T^p)$ is closed.]
10. Let X be a Banach space, and let T be a closed linear operator such that $\rho(T) \neq \varnothing$ and $\mathscr{D}(T)$ is a proper dense subspace of X. Then each quotient space $\mathscr{D}(T^i)/\mathscr{D}(T^{i+1})$ is infinite dimensional, $i = 0, 1, 2, \ldots$.

V.7 COMPACT OPERATORS

In many ways the spectral theory of compact linear operators is a natural and direct generalization of parts of (finite-dimensional) linear algebra and of the elementary theory of integral equations. In this section we shall prove theorems that, when applied to Fredholm integral equations of the second kind, yield most of the fundamental information about such equations.

The essential notion of a compact mapping was introduced by David Hilbert in 1906; eleven years later F. Riesz [2] published a thorough study of compact operators. In 1930, J. Schauder added some refinements to the theory by proving that the conjugate of a compact operator is itself compact. A majority of the results in this section are due to Riesz, although our presentation is quite different.

Throughout this section X and Y will denote normed linear spaces.

Definition. Suppose T is a linear operator with domain X and range in Y. We say that T is *compact* if, for each bounded subset B of X, $T(B)$ is relatively compact in Y. Since Y is a metric space, T is compact if and only if, for each bounded sequence $\{x_n\}$ in X, the sequence $\{Tx_n\}$ contains a subsequence converging to some limit in Y.

A compact operator is also called *completely continuous* (in French, *complètement* or *totalement continue*; in German, *vollstetig*). We note at once that a compact linear operator is continuous. For, discontinuity of T would imply the existence of a sequence $\{x_n\}$ such that $\|x_n\| \leq 1$ and $\|Tx_n\| \to \infty$, and this cannot occur if T is compact. We shall denote the set of compact operators in $L(X, Y)$ by $\mathscr{K}(X, Y)$ and those in $L(X)$ by $\mathscr{K}(X)$.

When Y is complete there is a useful characterization of compact operators in terms of precompactness. A nonempty subset A of a metric space is *precompact*, or *totally bounded*, if for each $\varepsilon > 0$, A is contained in the union of a finite number of open balls of radius ε. It can be shown that a subset of a complete metric space is relatively compact if and only if it is precompact. (See Taylor [5, pages 122, 126].) Thus when Y is complete, T is a compact operator if and only if $T(B)$ is precompact for each bounded subset B of X.

Theorem 7.1. *The set $\mathscr{K}(X, Y)$ is a subspace of $L(X, Y)$. If Y is complete, this subspace is closed.*

Proof. The verification of the first assertion is left to the reader. Suppose, therefore, that Y is complete and that $\{T_n\}$ is a sequence in $\mathscr{K}(X, Y)$ converging to some $T \in L(X, Y)$. To show that T is compact, it suffices to show that $T(S)$ is precompact, where $S = \{x \in X : \|x\| < 1\}$. (Any bounded set in X is contained in a multiple of S.) Given $\varepsilon > 0$, we fix N such that

$$\|T_N - T\| < \frac{\varepsilon}{2}.$$

Since T_N is compact, $T_N(S)$ is precompact. Thus there exist $y_1, \ldots, y_m \in Y$ such that every element $T_N x$, for $x \in S$, is within $\varepsilon/2$ of some y_i, $1 \leq i \leq m$. It follows that, given $x \in S$, there is some y_i such that

$$\|Tx - y_i\| \leq \|T - T_N\| \|x\| + \|T_N x - y_i\| < \varepsilon.$$

Hence $T(S)$ is precompact in Y. $\square$

Before we consider some examples of compact operators, we note that the identity operator on an infinite-dimensional normed linear space X is *never* compact. For, if the identity were a compact operator, then the unit ball in X would be compact. By Theorem II.3.6, this is possible only if the space is finite-dimensional. More generally, suppose that $P \in L(X)$ and $P^2 = P$. If P is compact, then $\mathscr{R}(P)$ is finite dimensional, because P is the identity on $\mathscr{R}(P)$. The converse of this is true also, as we shall see in Example 1.

Examples of Compact Operators

Example 1. If $T \in L(X, Y)$ and $\dim \mathscr{R}(T) < \infty$, we say that T is of *finite rank*; the *rank* of T is $\dim \mathscr{R}(T)$. Such an operator is compact, because the image $T(S)$ of the unit ball in X is a bounded set in the finite-dimensional space $\mathscr{R}(T)$ and hence is relatively compact (by Theorem II.3.4). Thus a continuous projection on X is compact if and only if it is of finite rank.

Operators in $L(X, Y)$ of finite rank are sometimes said to be *degenerate*. This terminology stems from the fact that an integral operator with a degenerate kernel (see § IV.2) has a finite-dimensional range. For example, suppose both $\{\phi_1, \ldots, \phi_n\}$ and $\{\psi_1, \ldots, \psi_n\}$ are linearly independent subsets of $C[a, b]$, and let

$$k(s, t) = \sum_{i=1}^{n} \phi_i(s)\psi_i(t).$$

If K is the corresponding Fredholm-type integral operator on $C[a, b]$, then

$$(Kx)(s) = \int_a^b k(s, t)x(t)\,dt = \sum_{i=1}^{n} \left[\int_a^b x(t)\psi_i(t)\,dt\right]\phi_i(s).$$

Thus $\mathscr{R}(K)$ is contained in the linear span of $\{\phi_1, \ldots, \phi_n\}$, and K is compact.

Example 2. Let $k(s, t)$ be an $\mathscr{L}^2$ kernel (see § IV.3), and let K be the corresponding integral operator on $L^2(a, b)$. By Theorem IV.3.1, K is the limit of integral operators with degenerate kernels. It follows from Theorem 7.1 that K is compact.

In the next example and later in the section we need the Arzelà–Ascoli theorem. Let M be a compact metric space, with a metric d. A family $\mathscr{F}$ of continuous functions on M is said to be equicontinuous if to each $\varepsilon > 0$ corresponds a $\delta > 0$ such that $|f(t_1) - f(t_2)| < \varepsilon$ whenever $f \in \mathscr{F}$ and t_1, t_2 are points of M such that $d(t_1, t_2) < \delta$.

The Arzelà–Ascoli Theorem. *Let M be a compact metric space. If $\mathscr{F}$ is a uniformly bounded equicontinuous family in $C(M)$, then every sequence of functions in $\mathscr{F}$ contains a uniformly convergent subsequence.*

This is a standard theorem. See Dunford and Schwartz [1, page 266] or Taylor [5, page 167].

Example 3. Let $X = C[a, b]$, where $[a, b]$ is a finite interval, let $k(s, t)$ be a continuous function of s and t on $[a, b] \times [a, b]$, and let $K \in L(X)$ be the corresponding integral operator. Then K is compact. To prove this we let $\{x_n\}$ be any bounded sequence in $C[a, b]$ and let $y_n = Kx_n$. Clearly $\{y_n\}$ is bounded, for $\|y_n\| \leq M\|K\|$, where $M = \sup_n \|x_n\|$. We now show that the y_n's form an equicontinuous family. Since $[a, b] \times [a, b]$ is compact, $k(s, t)$ is uniformly continuous. Hence to each $\varepsilon > 0$ there corresponds some $\delta > 0$ such that $|k(s_1, t) - k(s_2, t)| < \varepsilon/(b-a)M$ for all t whenever $|s_1 - s_2| < \delta$. Then for each n we have

$$|y_n(s_1) - y_n(s_2)| = \left| \int_a^b [k(s_1, t) - k(s_2, t)] x_n(t)\, dt \right| \leq \varepsilon$$

if $|s_1 - s_2| < \delta$. Thus $\{y_n\}$ is a bounded equicontinuous family of functions in $C[a, b]$. It follows from the Arzelà–Ascoli theorem that $\{y_n\}$ contains a subsequence that is convergent in the topology of $C[a, b]$. Thus K is compact.

The operator K will still be compact with certain less severe restrictions on the kernel k. If $k(s, t)$ is of class $\mathscr{L}^2(a, b)$ as a function of t for each s, if

$$\int_a^b |k(s, t)|^2\, dt$$

is a bounded function of s, and if

$$\int_a^b |k(s_1, t) - k(s_2, t)|^2\, dt \to 0 \qquad \text{as} \qquad |s_1 - s_2| \to 0,$$

then K will be compact as an operator acting in $C[a, b]$. The proof that K maps a bounded sequence into an equicontinuous sequence uses the Schwarz inequality.

Example 4. If $1 \leq q < \infty$, a continuous linear operator A on ℓ^1 into ℓ^q is represented by an infinite matrix (α_{ij}), where the condition on the matrix is that

$$\|A\| = \sup_j \left(\sum_{i=1}^{\infty} |\alpha_{ij}|^q \right)^{1/q}$$

be finite (see problem 4, § IV.6). The operator will be compact if and only if, in addition,

$$\sum_{i=n}^{\infty} |\alpha_{ij}|^q \to 0 \qquad \text{as} \qquad n \to \infty, \text{ uniformly in } j. \tag{7-1}$$

That this is a sufficient condition for compactness of A may be seen by using the result of problem 5 for, if $y = Ax$, it follows from Minkowski's inequality that

$$\left(\sum_{i=n}^{\infty} |\eta_i|^q \right)^{1/q} \leq \|x\| \left(\sum_{i=n}^{\infty} |\alpha_{ij}|^q \right)^{1/q}.$$

The condition (7-1) is also necessary, again by problem 5, for the vectors $v_j = (\alpha_{1j}, \alpha_{2j}, \ldots)$ must form a conditionally sequentially compact set, owing to the compactness of A. (Note that $v_j = Au_j$, where u_j is the vector whose ith component is δ_{ij}.)

Examples 1, 3, and 4 all depended on knowing the compact sets in the underlying spaces. In general, it is desirable to have a characterization of the compact sets in a given normed linear space Y, since this leads to a characterization of all compact linear operators from a normed linear space X into Y.

Example 5. Let X and Y be Banach spaces. An operator $T \in L(X, Y)$ is *nuclear* if there exist sequences $\{x'_k\} \subset X'$, $\{y_k\} \subset Y$, $\{\lambda_k\} \in \ell^1$, with $\|x'_k\| \leq 1$ and $\|y_k\| \leq 1$ for all k, such that

$$Tx = \sum_{k=1}^{\infty} \lambda_k x'_k(x) y_k, \tag{7-2}$$

for all $x \in X$. Note that the series in (7-2) is absolutely convergent since

$$\|\lambda_k x'_k(x) y_k\| \leq |\lambda_k| \, \|x\| \tag{7-3}$$

and $\Sigma_1^\infty |\lambda_k| < \infty$. We shall show that a nuclear operator is compact.

Suppose that T is given by (7-2), and define T_n, $n = 1, 2, \ldots,$ by

$$T_n x = \sum_{k=1}^{n} \lambda_k x'_k(x) y_k.$$

Clearly $T_n \in L(X, Y)$ and $\dim \mathscr{R}(T_n) \leq n$. Furthermore, (7-3) implies that

$$\|Tx - T_n x\| \leq \left(\sum_{k=n+1}^{\infty} |\lambda_k| \right) \|x\|,$$

which shows that $\|T - T_n\| \to 0$ as $n \to \infty$. Since each T_n is compact, so is T, by Theorem 7.1.

In examples 2 and 5 we used the fact that (when Y is complete) the uniform limit of operators of finite rank is compact. If Y is a Hilbert space, every compact operator in $L(X, Y)$ is the uniform limit of operators of finite rank (problem 9). In 1932, Banach conjectured that this is true when Y is any Banach space. This remained an open question until 1973, when P. Enflo [1] constructed a separable reflexive Banach space X such that some compact operators in $L(X)$ are *not* the uniform limit of operators of finite rank. Several variations of Enflo's counterexample have since appeared, the simplest of which is probably the construction of A. M. Davie [1].

Some Properties of Compact Operators

Theorem 7.2. *Let X, Y, Z be normed linear spaces, and suppose $T \in L(X, Y)$, $S \in L(Y, Z)$. Then ST is compact whenever S or T is compact.*

Proof. This follows immediately from the fact that a continuous linear operator carries bounded sets into bounded sets and relatively compact sets into relatively compact sets. □

If T is a compact operator whose domain X is infinite dimensional, then T cannot have a bounded inverse. For, if T^{-1} were to exist and be continuous, then $T^{-1}T = I$ would be compact, which would imply $\dim X < \infty$.

Theorem 7.3. *If $T \in \mathcal{K}(X, Y)$, then $T' \in \mathcal{K}(Y', X')$.*

Proof. Let $S = \{x \in X : \|x\| \leq 1\}$, and let $\{y'_n\}$ be any bounded sequence in Y', say, $\|y'_n\| \leq M$ for all n. We now view $\{y'_n\}$ as a sequence in the space $C(\overline{T(S)})$ of continuous functions on the compact set $\overline{T(S)}$. For each n,

$$\sup_{y \in \overline{T(S)}} |y'_n(y)| = \sup_{y \in T(S)} |y'_n(y)| \leq M\|T\|.$$

Hence $\{y'_n\}$ is a bounded sequence in $C(\overline{T(S)})$. Furthermore, this sequence is equicontinuous, because

$$|y'_n(y_1) - y'_n(y_2)| \leq M\|y_1 - y_2\|$$

for all n. By the Arzelà–Ascoli theorem, there is some subsequence $\{y'_{n_k}\}$ of $\{y'_n\}$ that converges uniformly on $\overline{T(S)}$. Hence $y'_{n_k}(Tx) = T'y'_{n_k}(x)$ converges uniformly for $x \in S$; that is, $\{T'y'_{n_k}\}$ converges in the norm on X'. □

The converse of Theorem 7.3 is valid if Y is complete (problem 10).

Theorem 7.4. *If T is a compact operator in $L(X, Y)$ whose range is a complete subspace of Y, then* $\dim \mathcal{R}(T) < \infty$.

Proof. Let $Y_1 = \mathcal{R}(T)$, and define $T_1 \in L(X, Y_1)$ by $T_1x = Tx$, $x \in X$. It is easy to see that T_1 is compact. Hence T'_1 is compact. Since $\mathcal{R}(T_1) = Y_1$ and Y_1 is complete, T'_1 has a continuous inverse (Theorem IV.9.3). Then $\dim Y'_1 < \infty$ (by the remarks following Theorem 7.2), and hence $\dim Y_1 < \infty$. □

Theorem 7.5. *Let X, Z be Banach spaces, and let Y be a normed linear space. Suppose that $T \in \mathcal{K}(X, Y)$ and that A is an operator in $L(Z, Y)$ whose range is contained in $\mathcal{R}(T)$. Then A must be compact.*

Proof. Let $\hat{T}$ be the one-to-one operator induced by T; that is, $\hat{T}[x] = Tx$, for all $[x] \in X/\mathcal{N}(T)$. Then $\hat{T}$ is compact. For, if $S = \{x : \|x\| \leq 1\}$, then

$S+\mathcal{N}(T)$ is the unit ball in $X/\mathcal{N}(T)$ and $\hat{T}(S+\mathcal{N}(T))=T(S)$ is relatively compact. Since $\hat{T}$ is continuous, $\hat{T}^{-1}$ is closed. Also, $\mathcal{D}(\hat{T}^{-1})=\mathcal{R}(T)\supset\mathcal{R}(A)$, and it is easy to verify that $\hat{T}^{-1}A$ is a closed operator from Z onto $X/\mathcal{N}(T)$. By the closed graph theorem, $\hat{T}^{-1}A$ is continuous. Since $A=\hat{T}(\hat{T}^{-1}A)$, it follows from Theorem 7.2 that A is compact. □

Spectral Theory of Compact Operators

The spectrum of a compact operator T will be described in Theorem 7.10. However, in the process of obtaining this theorem, we shall prove some facts about the operator $\lambda - T$, which are often more useful than the actual statement of Theorem 7.10. To avoid unnecessary repetition in the statements of the remaining theorems, we shall assume henceforth in this section that T is a compact operator in $L(X)$ (X a normed linear space) and λ is any *nonzero* scalar. As usual, we often write T_λ for $\lambda - T$.

Theorem 7.6. *The null spaces $\mathcal{N}(T_\lambda^n)$, $n=1,2,\ldots$, are finite dimensional.*

Proof. We begin with $n=1$. If $x\in\mathcal{N}(T_\lambda)$, then $x=\lambda^{-1}Tx$, so that $\lambda^{-1}T$ is the identity on $\mathcal{N}(T_\lambda)$. Consequently, the unit ball in $\mathcal{N}(T_\lambda)$ must be compact (since $\lambda^{-1}T$ is compact). Hence $\dim\mathcal{N}(T_\lambda)<\infty$ (by Theorem II.3.6). For $n>1$ we write

$$T_\lambda^n=(\lambda-T)^n=\lambda^n-n\lambda^{n-1}T+\cdots+(-1)^nT^n=\lambda^n-A,$$

where A is compact by Theorems 7.1 and 7.2. The foregoing reasoning applied to λ^n-A shows that $\mathcal{N}(T_\lambda^n)$ is finite dimensional. □

Theorem 7.7. *Let M be any closed subspace of X such that $M\cap\mathcal{N}(T_\lambda)=(0)$. Then the restriction of T_λ to M has a bounded inverse and $T_\lambda(M)$ is closed in X.*

Proof. Suppose the restriction does not have a continuous inverse from $T_\lambda(M)$ back to M. Then there exists a sequence $\{x_n\}\subset M$ such that

$$\|x_n\|=1 \qquad \text{and} \qquad T_\lambda x_n\to 0.$$

(See Theorem II.1.2.) Since T is compact, there exists a subsequence $\{x_{n_k}\}$ such that $\{Tx_{n_k}\}$ converges. But

$$x_{n_k}=\lambda^{-1}(Tx_{n_k}+T_\lambda x_{n_k}), \tag{7-4}$$

and so $\{x_{n_k}\}$ converges. Its limit, x, is in M since M is closed. Then

$$0=\lim_{k\to\infty}T_\lambda x_{n_k}=T_\lambda x.$$

This is impossible because T_λ is clearly one-to-one on M and $\|x\|=1\neq 0$. Hence the restriction of T_λ to M has a bounded inverse.

Now suppose $y\in\overline{T_\lambda(M)}$ and $T_\lambda x_n\to y$ for some $\{x_n\}\subset M$. It follows from the first part of the theorem that $\{x_n\}$ must be bounded. Then there exists a subsequence $\{x_{n_k}\}$ such that $\{Tx_{n_k}\}$ converges. By (7-4), $\{x_{n_k}\}$ converges to some $x\in M$. Hence $y=\lim_{k\to\infty} T_\lambda x_{n_k}=T_\lambda x\in T_\lambda(M)$, which shows that $T_\lambda(M)$ is closed. □

Theorem 7.8. *The ranges $\mathcal{R}(T_\lambda^n)$ are closed.*

Proof. Just as in the proof of Theorem 7.6 it suffices to treat the case $n=1$. So take M in Theorem 7.7 such that $X=M\oplus\mathcal{N}(T_\lambda)$. ($M$ exists by Theorem IV.12.3, since $\dim\mathcal{N}(T_\lambda)<\infty$.) Then $\mathcal{R}(T_\lambda)=T_\lambda(M)$, which is closed. □

The next theorem contains the heart of the spectral theory of compact operators. The proof relies on Riesz's lemma, Theorem II.3.5. (F. Riesz first used his well-known lemma in his work on compact operators [2].)

Theorem 7.9. *The ascent and descent of T_λ are both finite (and hence equal). If $p=\alpha(T_\lambda)$, then*

$$X=\mathcal{R}(T_\lambda^p)\oplus\mathcal{N}(T_\lambda^p), \tag{7-5}$$

where both subspaces are closed.

Proof. Suppose $\alpha(\lambda-T)=\infty$. Then $\mathcal{N}(T_\lambda^{n-1})$ is a proper closed subspace of $\mathcal{N}(T_\lambda^n)$ for $n=1,2,\ldots$. By Riesz's lemma there exists $x_n\in\mathcal{N}(T_\lambda^n)$ such that $\|x_n\|=1$ and $\|x_n-x\|\geq\frac{1}{2}$ if $x\in\mathcal{N}(T_\lambda^{n-1})$. If $1\leq m<n$, then

$$\begin{aligned} Tx_n-Tx_m &= \lambda x_n-(\lambda x_m+T_\lambda x_n-T_\lambda x_m) \\ &= \lambda x_n - z, \end{aligned} \tag{7-6}$$

where z clearly belongs to $\mathcal{N}(T_\lambda^{n-1})$ (since $m\leq n-1$). Then

$$\|Tx_n-Tx_m\|=|\lambda|\,\|x_n-\lambda^{-1}z\|\geq\frac{|\lambda|}{2}>0. \tag{7-7}$$

This shows that $\{Tx_n\}$ can have no convergent subsequence, in contradiction to the fact that T is compact. Thus $\alpha(\lambda-T)$ must be finite.

The proof that $\delta(\lambda-T)<\infty$ is similar. If $\delta(\lambda-T)=\infty$, $\mathcal{R}(T_\lambda^{n+1})$ is a proper closed subspace of $\mathcal{R}(T_\lambda^n)$ for $n=1,2,\ldots$. We choose $x_n\in\mathcal{R}(T_\lambda^n)$ such that $\|x_n\|=1$ and $\|x_n-x\|\geq\frac{1}{2}$ if $x\in\mathcal{R}(T_\lambda^{n+1})$. Then for $1\leq n<m$, the element z in (7-6) belongs to $\mathcal{R}(T_\lambda^{n+1})$. Hence (7-7) holds, and we obtain a contradiction as before.

The final statement of the theorem follows from Theorems 6.2 and 7.8. □

Theorem 7.9 reveals a lot about the operator T_λ. First of all, T_λ is one-to-one if and only if its range is all of X. Put another way, solutions of the equation

$$Tx - \lambda x = y$$

will be unique if and only if at least one solution exists for every $y \in X$. Secondly, if λ is not an eigenvalue, that is, if T_λ is one-to-one (and $\mathscr{R}(T_\lambda) = X$), then Theorem 7.7 shows that $\lambda \in \rho(T)$. Thus *every nonzero point in $\sigma(T)$ must be an eigenvalue.*

Finally, observe that T is completely reduced by the subspaces $X_1 = \mathscr{R}(T_\lambda^p)$ and $X_2 = \mathscr{N}(T_\lambda^p)$, because X_1 and X_2 are invariant under T_λ and hence also under T. So let $T_k \in L(X_k)$ be the restriction of T to X_k, $k = 1, 2$. The image under T_1 of a bounded set in X_1 is relatively compact in X and hence in X_1, since X_1 is closed. Thus T_1 is compact. Clearly $\lambda - T_1$ is one-to-one, and so it follows from our remark above that $\lambda \in \rho(T_1)$. Then $\mu \in \rho(T_1)$ for all μ sufficiently close to λ (Theorem 2.2). Furthermore, $\lambda - T_2$ is nilpotent, and so $\sigma(T_2) = \{\lambda\}$. (Apply Theorem 3.5 to the operator $\lambda - T_2$ in the finite-dimensional space X_2. Also see problem 5 of § 6.) Thus $\mu - T_1$ and $\mu - T_2$ are both one-to-one if $\mu \neq \lambda$ and μ is sufficiently close to λ. By Theorem 5.2, the same is true for $\mu - T$. But if μ is not an eigenvalue, then $\mu \in \rho(T)$. Thus *each nonzero $\lambda \in \sigma(T)$ is an isolated point of the spectrum.*

Now the spectrum of T is compact (Theorem 3.1), and hence $\sigma(T) \cap \{\lambda : |\lambda| \geq r\}$ is compact for each $r > 0$. But this set can consist only of isolated eigenvalues and therefore must be a finite (or empty) set.

We combine these observations into the following theorem.

Theorem 7.10. *The spectrum of a compact operator in $L(X)$ contains at most a countable set of points, and these have no accumulation point except possibly the point $\lambda = 0$. Each nonzero point of the spectrum is an eigenvalue.*

By various examples it may be shown that when T is compact, the point $\lambda = 0$ can belong to $P\sigma(T)$, $C\sigma(T)$, or $R\sigma(T)$; it cannot belong to $\rho(T)$ if X is infinite dimensional (since a compact operator in this case cannot have a bounded inverse). See problems 13 to 15.

The decomposition in (7-5) also leads easily to the following theorem.

Theorem 7.11. *If T is compact and $\lambda \neq 0$, then $\lambda - T$ is a Fredholm operator of index zero.*

Proof. Let $X_1 = \mathscr{R}(T_\lambda^p)$, $X_2 = \mathscr{N}(T_\lambda^p)$, as in (7-5), and let T_2 be the restriction of T to X_2. Since $\alpha(T_\lambda) = \delta(T_\lambda) = p$, T_λ is a one-to-one mapping of X_1 onto itself. Hence, by Theorem 5.2, $\mathscr{N}(\lambda - T) = \mathscr{N}(\lambda - T_2)$ and $\mathscr{R}(\lambda - T) = X_1 \oplus \mathscr{R}(\lambda - T_2)$. Thus it is clear that $\lambda - T$ and $\lambda - T_2$ have the same nullity and defect. Since X_2 is finite dimensional, $\lambda - T_2$ is a Fredholm operator of index zero (see the remark preceding Theorem IV.13.4) and hence so is $\lambda - T$. $\square$

Theorem 7.11 has the following important application to perturbation theory.

Theorem 7.12. *Suppose X and Y are Banach spaces, A is a Fredholm operator in $L(X, Y)$ and $K \in \mathcal{K}(X, Y)$. Then $A+K$ is a Fredholm operator with the same index as A.*

Proof. Let S be any pseudoinverse of A (see page 251). Then SK is a compact operator, and so $I+SK$ is a Fredholm operator of index zero, by Theorem 7.11. The conclusions of the theorem are now given by Theorem IV.13.5. $\square$

There is a useful generalization of Theorem 7.12 to perturbations of an unbounded operator A. In this case one allows (possibly unbounded) perturbations by operators K that are A-*compact*; that is, $\mathcal{D}(K) \supset \mathcal{D}(A)$ and $K \in \mathcal{K}(\mathcal{D}(A), Y)$, when $\mathcal{D}(A)$ is given the graph norm. Further details may be found in Goldberg [2] and Schechter [1].

The final phase of our study of the spectrum of a compact operator T concerns the conjugate operator T'. We know that T' is compact (Theorem 7.3) and has the same spectrum as T (Theorem 2.4). Hence a nonzero λ is an eigenvalue of T if and only if it is an eigenvalue of T'. But more than this is true, as we shall see in Theorem 7.14. Let us first consider a useful characterization of the ranges of T_λ and T'_λ.

Theorem 7.13. *$\mathcal{R}(T_\lambda) = \mathcal{N}(T'_\lambda)^\perp$ and $\mathcal{R}(T'_\lambda) = \mathcal{N}(T_\lambda)^\perp$.*

Proof. If X is complete, both relations follow from the closed range theorem (Theorem IV.10.1), because we know that $\mathcal{R}(T_\lambda)$ is closed. If X is not complete, the first relation follows from Theorem IV.8.4; for the second relation we need only prove that $\mathcal{N}(T_\lambda)^\perp \subset \mathcal{R}(T'_\lambda)$, because of Theorem IV.8.5. We begin by considering $y' \in \mathcal{N}(T_\lambda)^\perp$. Let M be a closed subspace of X such that

$$X = \mathcal{N}(T_\lambda) \oplus M.$$

By Theorem 7.7, the restriction S_λ of T_λ to M has a continuous inverse. Thus the mapping $T_\lambda x \mapsto y'(S_\lambda^{-1} T_\lambda x) = y'(x)$, for $x \in M$, is a continuous linear functional on $T_\lambda(M)$. By the Hahn–Banach theorem, there exists an $x' \in X'$ such that $x'(T_\lambda x) = y'(x)$ for all $x \in M$. Since $y' \in \mathcal{N}(T_\lambda)^\perp$, it is evident that y' and $x' \circ T_\lambda$ agree on both M and $\mathcal{N}(T_\lambda)$, and thus on X itself. Hence $y' = T'_\lambda x' \in \mathcal{R}(T'_\lambda)$. $\square$

Theorem 7.14.
(*a*) $n(T_\lambda)=d(T_\lambda)=n(T'_\lambda)=d(T'_\lambda)<\infty$,
(*b*) $\alpha(T_\lambda)=\delta(T_\lambda)=\alpha(T'_\lambda)=\delta(T'_\lambda)<\infty$.

Proof. Since T' is also compact, the first and third equalities of (*a*) are given by Theorem 7.11. If we use Theorem 7.13 (cf. Theorem IV.8.4) and a congruence discussed in Theorem III.3.3, we have

$$\mathcal{N}(T'_\lambda)=\mathcal{N}(T'_\lambda)^{\perp\perp}=\mathcal{R}(T_\lambda)^{\perp}\cong(X/\mathcal{R}(T_\lambda))'.$$

Since these spaces are finite dimensional, we may conclude that

$$\begin{aligned} n(T'_\lambda)&=\dim\mathcal{N}(T'_\lambda)=\dim\,(X/\mathcal{R}(T_\lambda))'\\ &=\dim X/\mathcal{R}(T_\lambda)=d(T_\lambda). \end{aligned}$$

It follows from (*a*) that, for each n, $\mathcal{N}(T^n_\lambda)$ and $\mathcal{N}[(T'_\lambda)^n]$ have the same dimension; this is because $(T^n_\lambda)'=(T'_\lambda)^n$ and $T^n_\lambda=\lambda^n-A$, where A is compact. (See the proof of Theorem 7.6.) Clearly, then, $\lambda-T$ and $\lambda-T'$ must have the same ascent. The rest of (*b*) follows from Theorem 7.9. □

Invariant Subspaces

A famous problem that has challenged mathematicians for many years is whether every bounded linear operator on a Banach space has a nontrivial closed invariant subspace. Except for a few special classes of operators (such as the self-adjoint operators on a Hilbert space), nothing much was known for general Banach spaces until 1954, when N. Aronszajn and K. Smith [1] proved that every compact operator has a nontrivial closed invariant subspace. Twelve years later, A. Bernstein and A. Robinson [1] used "nonstandard analysis" to extend this result to operators T such that $p(T)$ is compact for some nonzero polynomial $p(\lambda)$. Although this new theorem stimulated several authors to find refinements and simplifications, it was clear that any substantial generalization would require new techniques. Then, in 1973, V. I. Lomonosov [1] published a remarkable generalization of the Bernstein–Robinson theorem.

Theorem 7.15 (Lomonosov). *Let X be a Banach space, and let K be an operator in $L(X)$ that is not a scalar multiple of the identity. If the class $\mathcal{A}$ of members of $L(X)$ that commute with K contains at least one nonzero compact operator, then there exists a nontrivial closed subspace of X that is invariant under each operator in $\mathcal{A}$.*

The proof of Lomonosov's theorem involves an elegant use of the Schauder fixed point theorem for nonlinear operators. See Pearcy and Shields [1] for an expository discussion of the proof and related results.

Observe that if K is itself a nonzero compact operator and if X is infinite dimensional, then K automatically satisfies the conditions of Theorem 7.15. This special case of Lomonosov's theorem is still an outstanding improvement on earlier results. Furthermore, it has a completely elementary proof (see below) due to M. Hilden, based on a clever modification of Lomonosov's argument.

Theorem 7.16. *If X is a Banach space and K is a nonzero compact operator in $L(X)$, then there is a nontrivial closed subspace of X that is invariant under each member of $L(X)$ that commutes with K.*

Proof. If $\sigma(K)$ contains a nonzero point λ, then λ is an eigenvalue of K and $\mathcal{N}(\lambda - K)$ is a nontrivial closed subspace of X. It is easy to see that $\mathcal{N}(\lambda - K)$ is invariant under every $A \in L(X)$ commuting with K. So it remains to consider the case when the spectral radius of K is zero. That is, we may suppose that

$$\lim_{n\to\infty} \|K^n\|^{1/n} = 0. \tag{7-8}$$

Let $\mathscr{A} = \{A \in L(X) : AK = KA\}$, and note that, for each nonzero y in X, the set

$$\mathscr{A}y = \{Ay : A \in \mathscr{A}\}$$

is a nonzero subspace of X. Since $BA \in \mathscr{A}$ for all A, $B \in \mathscr{A}$, the subspace $\mathscr{A}y$ is invariant under each $B \in \mathscr{A}$. By continuity of operator multiplication, the closure $\overline{\mathscr{A}y}$ must also be invariant. Thus if $\overline{\mathscr{A}y} \neq X$ for some nonzero y, then $\overline{\mathscr{A}y}$ is a suitable nontrivial closed invariant subspace. Therefore suppose that

$$\overline{\mathscr{A}y} = X \qquad \text{for all } y \neq 0. \tag{7-9}$$

We shall deduce a contradiction, which will complete the proof.

Take any $x_0 \in X$ such that $x_0 \neq 0$ and $Kx_0 \neq 0$. Since K is continuous at x_0, there exists an open ball B centered at x_0 such that

$$0 \notin \bar{B}, \qquad 0 \notin \overline{K(B)}.$$

Given y in $K(B)$, we have $x_0 \in \overline{\mathscr{A}y}$, by (7-9). Because B is a neighborhood of x_0, there exists an operator A in $\mathscr{A}$ (depending on y) such that $Ay \in B$. Since this A is continuous, there exists a neighborhood U_y of y such that $A(U_y) \subset B$. Now $\overline{K(B)}$ is a compact set because K is a compact operator. So a finite number of the U_y cover $\overline{K(B)}$. Let $A_1, \ldots, A_m$ be the corresponding operators. *Then for each y in $\overline{K(B)}$, we have $A_i y \in B$ for at least one of the A_i.* In particular, there exists A_{i_1}, $1 \leq i_1 \leq m$, such that $A_{i_1}(Kx_0) \in B$. Since $K(A_{i_1}Kx_0) \in K(B)$, there exists A_{i_2}, $1 \leq i_2 \leq m$, such that $A_{i_2}KA_{i_1}Kx_0 \in B$. Proceeding inductively, we obtain an infinite sequence $\{A_{i_k}\}$ of operators from the set $\{A_1, \ldots, A_m\}$ such that for each $n \geq 1$,

$$x_n = A_{i_n}K \cdots A_{i_1}Kx_0 \in B.$$

Let $M = \max\{\|A_1\|, \ldots, \|A_m\|\}$. Using the fact that the A_i commute with K, we find that

$$\|x_n\| \leq M^n \|K^n\| \|x_0\|.$$

Hence by (7-8), $\|x_n\|^{1/n} \to 0$ as $n \to \infty$. This implies that $0 \in \bar{B}$, contrary to our choice of B. □

The Fredholm Alternative

We conclude this section with a brief summary of the application of the theory of compact operators to the study of integral equations. We consider Fredholm equations of the second kind, in the form

$$y(s) = x(s) - \lambda \int_a^b k(s, t)x(t)\, dt, \tag{7-10}$$

where the kernel k is of such a sort that the integral operator K defined by it is compact when acting in a certain specified normed linear function space X. We think primarily of $X = C[a, b]$ or $X = L^2(a, b)$. We can write (7-10) in the form $y = x - \lambda Kx$. Note that the parameter λ here occupies a different position from the one it occupies in our previous work, where we have been studying the equation $y = \lambda x - Tx$. There is no interest in the equation (7-10) when $\lambda = 0$, so we can assume $\lambda \neq 0$ and rewrite the equation as

$$\lambda^{-1}y = \lambda^{-1}x - Kx.$$

Since $\lambda^{-1} \neq 0$, our theory of compact operators now gives us the following "Fredholm alternative."

For a given $\lambda \neq 0$, *either* (7-10) *has a unique solution* x *corresponding to each choice of* $y \in X$, *or else the homogeneous equation*

$$0 = x(s) - \lambda \int_a^b k(s, t)x(t)\, dt \tag{7-11}$$

has a solution $x \neq 0$. *In this latter case the number of linearly independent solutions of* (7-11) *is finite.* The number λ^{-1} is then an eigenvalue of the integral operator K. It is customary to call λ itself a *characteristic value* of K. *The set of characteristic values is at most countable. If there is an infinite sequence* $\{\lambda_n\}$ *of such values,* $|\lambda_n| \to \infty$.

If the function space in question is $C[a, b]$, the conjugate operator K' acts in a different space. But if we are dealing with $L^2(a, b)$ and if k is an $\mathscr{L}^2$ kernel, we can identify K' with the integral operator in $L^2(a, b)$ such that $K'x$

is represented by

$$\int_a^b k(t, s)x(t)\,dt.$$

In this case the theory goes on to tell us that (7-11) *and the equation*

$$0 = x(s) - \lambda \int_a^b k(t, s)x(t)\,dt \tag{7-12}$$

have the same number of linearly independent solutions. Furthermore, if λ *is a characteristic value,* (7-10) *has a solution in* L^2 *corresponding to a given* $y \in L^2$ *if and only if*

$$\int_a^b y(t)\overline{x(t)}\,dt = 0$$

whenever x *is a solution of* (7-12).

The classical example of an incomplete inner-product space is $C[a, b]$ when it is endowed with the L^2 inner product and norm. Important integral equations of mathematical physics and theoretical mechanics are studied in this setting (with $[a, b]$ sometimes replaced by certain subsets of $\boldsymbol{R}^n$). It is primarily for this reason that in most of the theorems of this section we have avoided the assumption that X is complete. Suppose that k is an $\mathscr{L}^2$ kernel such that Kx and $K'x$ are in $C[a, b]$ whenever $x \in L^2(a, b)$. This is the case, for instance, if $k(s, t)$ is continuous in both variables (we assume that $[a, b]$ is a finite interval). Then, if $y \in C[a, b]$ and if $x \in L^2(a, b)$ is a solution of (7-10), it follows that $x \in C[a, b]$. Hence, in particular, the solutions of (7-11) are continuous, and the same is true of (7-12). Therefore it follows that all of the foregoing italicized assertions remain valid if we regard the $\mathscr{L}^2$ kernel as defining an operator in $C[a, b]$ rather than in $L^2(a, b)$ and restrict x and y accordingly.

The theory of Fredholm integral equations of the second kind, in the form presented here, can be applied to obtain the existence of a solution of the Dirichlet problem in two dimensions (see Example 1, § I.5), provided the region in question satisfies some reasonably general conditions. (The essential restrictions are on the smoothness of the boundary.) In fact, Fredholm's solution in 1900 of the Dirichlet problem stimulated much of the subsequent mathematical interest in the theory of integral equations. He showed that any harmonic function u that satisfies the Dirichlet problem also satisfies an equation of the form (7-10). We pointed out in § I.5 that the *uniqueness* of the solution of a Dirichlet problem is easily demonstrated. Thus the Fredholm alternative guarantees the *existence* of u. Prior to Fredholm's achievement, the existence of u had been rigorously demonstrated only for a fairly restricted class of regions in the plane.

The same approach can be used in other boundary-value problems associated with the Laplace operator, for example, the Neumann problem. For the Dirichlet problem in spaces of higher dimension the procedure is slightly more complicated, because one has to deal with an integral operator that may not be compact. However, a certain power of the operator turns out to be compact, and this is sufficient to obtain the Fredholm alternative (see the remarks following the proof of Theorem 10.8). These are the decisive results for the application to the Dirichlet problem. For more details see Riesz and Sz.-Nagy [1, pages 190–193] and Kellogg [1, pages 311–315].

PROBLEMS

1. If $X = \ell^2$ and (α_{ij}) is an infinite matrix such that $\Sigma_{i,j=1}^{\infty} |\alpha_{ij}|^2 < \infty$, the equations $\eta_i = \Sigma_{j=1}^{\infty} \alpha_{ij}\xi_j$ define a compact linear mapping $y = Ax$ of X into itself, with $\|A\| \leqslant (\Sigma_{i,j=1}^{\infty} |\alpha_{ij}|^2)^{1/2}$.
2. Let X be a separable Hilbert space, and let $\{e_n\}$ be an orthonormal basis of X.
   a. If T is a compact linear mapping of X into a normed linear space Y, then $\lim_{n\to 0} Te_n = 0 \in Y$.
   b. If Y is a Banach space and $T \in L(X, Y)$ has the property that $\Sigma \|Te_n\|^2 < \infty$, then T is compact.
3. Given X, Y normed linear spaces and $T \in \mathscr{K}(X, Y)$, then $Tx_n \to Tx$ whenever $\{x_n\}$ is $\sigma(X, X')$-convergent to x. [*Hint.* Use an elementary theorem in topology to conclude that the norm topology and the $\sigma(Y, Y')$ topology coincide on the closure $\overline{T(B)}$, where B is a ball in X containing $\{x_n\}$ and x.]
4. a. Suppose $T \in L(X, Y)$, where X is reflexive and T has the property that $Tx_n \to Tx$ whenever $\{x_n\}$ is $\sigma(X, X')$-convergent to x. Then T is compact. For operators in Hilbert space, this is the form in which the concept of a *vollstetig* operator was originally introduced.
   b. Every continuous linear mapping from a reflexive Banach space into ℓ^1 is compact.
   c. The proposition in (a) does not remain true if the hypothesis that X is reflexive is dropped.
5. A set S in a normed linear space X is called conditionally sequentially compact if every sequence from S contains a convergent subsequence. If $X = \ell^p$, where $1 \leqslant p < \infty$, a set S in X has this property if and only if (a) S is bounded and (b) $\Sigma_{i=n}^{\infty} |\xi_i|^p \to 0$ as $n \to \infty$, uniformly for all $x = \{\xi_i\}$ in S. To prove the sufficiency of these conditions, use a diagonal procedure on the components of the given sequence $\{x_n\}$. Boundedness of S is clearly necessary; the necessity of (b) may be proved by the method of contradiction.
6. If $T \in \mathscr{K}(X, Y)$, where X and Y are infinite-dimensional Banach spaces and there exists a closed subspace M such that $Y = \mathscr{R}(T) \oplus M$, then T is degenerate.

7. If $T \in \mathcal{K}(X, Y)$, where X and Y are normed linear spaces, then $\mathcal{R}(T)$ is separable.
8. If Y is a separable Banach space, there exists a compact operator in $L(Y)$ whose range is dense in Y.
9. Let X be a normed linear space and let Y be a Hilbert space. Then every compact operator in $L(X, Y)$ is the uniform limit of operators of finite rank. [*Hint.* Show that the identity operator on Y can be approximated by operators in $L(Y)$ of finite rank for the topology of uniform convergence on compact subsets of Y.]
10. If $T \in L(X, Y)$, where Y is a Banach space and T' is compact, then T is compact.
11. Let X and Y be Banach spaces, and let $S = \{x \in X : \|x\| \leq 1\}$.
    a. If $T \in \mathcal{K}(X, Y)$, then given $\varepsilon > 0$ there exists a closed subspace M of X such that $\dim X/M < \infty$ and the restriction of T to M has norm not exceeding ε. [Let $y_1, \ldots, y_n$ be the centers of balls of radius $\varepsilon/2$ that cover $T(S)$. For $1 \leq k \leq n$, choose $y'_k \in Y'$ such that $\|y'_k\| = 1$ and $y'_k(y_k) = \|y_k\|$. Let $M = \{x : y'_k(Tx) = 0,\ 1 \leq k \leq n\}$. If $x \in M$ and $\|x\| \leq 1$, choose k such that $\|Tx - y_k\| < \varepsilon/2$. Show that $\|y_k\| < \varepsilon/2$ and hence that $\|Tx\| < \varepsilon$.]
    b. Suppose that $T \in L(X, Y)$ and that for each $\varepsilon > 0$ there exists a linear manifold $M \subset X$ such that $\dim X/M < \infty$ and the restriction of T to M has norm not exceeding ε. Then T is compact. [We may assume M is closed. Any projection P of X onto M is continuous. Then

    $$\|Tx\| \leq \varepsilon \|x\| + (\varepsilon + \|T\|)\|(I - P)x\|.$$

    For any $\delta > 0$, there exist $x_1, \ldots, x_n \in S$ such that $(I - P)(S)$ is covered by the balls of radius δ whose centers are $(I - P)x_1, \ldots, (I - P)x_n$. Then $Tx_1, \ldots, Tx_n$ are the centers of balls of radius $2\varepsilon + (\varepsilon + \|T\|)\delta$ that cover $T(S)$.]
12. If $T \in L(X)$ and $\mathcal{R}(T)$ is finite dimensional, then $\sigma(T)$ is a finite set.
13. Define $T \in L(\ell^1)$ by $Tx = \{k^{-1}\xi_k\}$, where $x = \{\xi_k\}$. Then T is compact, and $0 \in C\sigma(T)$. What are the eigenvalues?
14. Define $T \in L(\ell^1)$ by the matrix (t_{ij}) with $t_{i1} = 2^{1-i}$, $i \geq 2$, $t_{ii} = 2^{1-i}$, $i \geq 2$, and all other $t_{ij} = 0$. Then T is compact and $0 \in R\sigma(T)$. What are the eigenvalues?
15. Suppose $x_0 \in X$, $x'_0 \in X'$, $x'_0(x_0) \neq 0$. Define $Tx = x'_0(x)x_0$. Show that $\sigma(T)$ consists of the eigenvalues $\lambda = 0$, $\lambda = x'_0(x_0)$ and that $R_\lambda y = \lambda^{-1}y + \lambda^{-1}[\lambda - x'_0(x_0)]^{-1}Ty$. (Assume that X is not one-dimensional.)
16. Suppose X is not complete, and let $T \in L(X)$ be compact. Let $\hat{T}$ be the unique continuous extension of T to all of the completion $\hat{X}$. Then $\hat{T}$ is compact and $\mathcal{R}(\hat{T}) \subset X$. If $\lambda \neq 0$, T^n_λ and $\hat{T}^n_\lambda$ have the same null manifold, and so T_λ and $\hat{T}_\lambda$ have the same ascent; also, $\mathcal{R}(T^n_\lambda) = \mathcal{R}(\hat{T}^n_\lambda) \cap X$ and $\hat{T}^n_\lambda(\hat{X} \backslash X) \subset \hat{X} \backslash X$. Finally, $\sigma(T) = \sigma(\hat{T})$. (Also, see problem 7, § 2.)
17. Let X, Y be Banach spaces, and suppose $A \in L(X, Y)$, with $\mathcal{R}(A)$ closed, and $K \in \mathcal{K}(X, Y)$.
    a. If A^{-1} exists and is continuous, then $n(A + K) < \infty$ and $\mathcal{R}(A + K)$ is

closed. If $d(A)=\infty$, then $d(A+K)=\infty$. (X and Y need not be complete for this part.)

b. Suppose instead that $\mathscr{R}(A)=Y$. Then $\mathscr{R}(A+K)$ is closed and $d(A+K)<\infty$. If $n(A)=\infty$, then $n(A+K)=\infty$. [*Hint.* Use (*a*), but note that $\mathcal{N}(A)$ may not have a closed complementary subspace.]

c. If $n(A)<\infty$, then $n(A+K)<\infty$ and $\mathscr{R}(A+K)$ is closed. [Consider the restriction of $A+K$ to a closed subspace complementary to $\mathcal{N}(A)$.]

d. If $d(A)<\infty$, then $d(A+K)<\infty$ and $\mathscr{R}(A+K)$ is closed. [Use (*c*).]

18. Let T be a closed linear operator from a Banach space X into a Banach space Y, and suppose that T^{-1} exists and is a compact operator in $L(Y, X)$. Note that T is a Fredholm operator of index zero. Let B be any operator in $L(X, Y)$. Then $T+B$ is a Fredholm operator (with domain $\mathscr{D}(T)$) and has index zero.

19. Let X be a Banach space, and let $\{T_n\}$ be a sequence of operators in $L(X)$ that converge pointwise to an operator T. Show that if $K\in\mathscr{K}(X)$, then $\{T_nK\}$ converges to TK in the norm of $L(X)$.

20. Suppose $f\in C[a,b]$. Consider the inhomogeneous two-point problem

$$y''(s)+\lambda f(s)y(s)=x(s), \qquad y(a)=\alpha, \qquad y(b)=\beta,$$

and the corresponding homogeneous problem in which the function x is identically zero. Show that, for a given λ, either the inhomogeneous problem has a unique twice continuously differentiable solution corresponding to each $x\in C[a,b]$, or else the homogeneous problem has a nonzero twice continuously differentiable solution. Also, the λ's for which the latter situation occurs form an at most countable set with no finite point of accumulation in the extended complex plane. Use the discussion in § IV.4, with $\lambda f(s)$ replacing $a_2(s)$.

V.8 AN OPERATIONAL CALCULUS

If X is a complex Banach space and T is a closed operator with domain and range in X, the fact that the resolvent operator $R_\lambda\equiv(\lambda-T)^{-1}$ is an analytic function of λ enables us to obtain some important results using contour integrals in the complex plane. The completeness of X (and hence of $L(X)$) assures us of the existence of these integrals.

First, let us consider some aspects of the situation when $T\in L(X)$. In this case $\sigma(T)$ is bounded, and we have a power series formula for R_λ when $|\lambda|>r_\sigma(T)$; see (3-4). If C denotes a simple closed contour, oriented counterclockwise, enclosing the circle $|\lambda|=r_\sigma(T)$, we can integrate (3-4) term by term around the contour, and we obtain

$$I=\frac{1}{2\pi i}\int_C R_\lambda\, d\lambda, \tag{8-1}$$

because the integral of each term of the series is 0, except for the first one. In a similar way,

$$T^p = \frac{1}{2\pi i}\int_C \lambda^p R_\lambda \, d\lambda \qquad p = 0, 1, 2, \ldots. \tag{8-2}$$

Since R_λ is analytic except at the points of $\sigma(T)$, it is clear that the integral in (8-2) is unchanged in value if we deform the contour C in any manner, so long as it continues to enclose $\sigma(T)$. We may even replace the single contour C by several nonintersecting closed contours, provided that no one of them is inside any other and $\sigma(T)$ lies in the union of their interiors.

We now propose to define a certain class of complex-valued analytic functions and to associate with each such function f an element of the operator space $L(X)$. The operator associated with f will be denoted by $f(T)$. The feature of greatest importance in this association is that the correspondence between f and $f(T)$ preserves the basic algebraic operations. That is, the operators corresponding to $f+g$, αf, and fg, respectively, are $f(T)+g(T)$, $\alpha f(T)$, and $f(T)g(T)$. In particular, then, since $fg = gf$, *the operators* $f(T)$ *and* $g(T)$ *commute.*

In what follows we use $\Delta(f)$ to denote the domain of definition of f.

Definition. Suppose $T \in L(X)$. Let $\mathfrak{A}(T)$ be the class of all complex-valued functions f such that: (a) $\Delta(f)$ is an open set in $\mathbf{C}$ and it contains $\sigma(T)$; (b) f is differentiable at each point of $\Delta(f)$. We say that such an f is *locally analytic on* $\sigma(T)$. For $f \in \mathfrak{A}(T)$, we define

$$f(T) = \frac{1}{2\pi i}\int_{+\partial D} f(\lambda) R_\lambda \, d\lambda, \tag{8-3}$$

where D is any bounded Cauchy domain such that $\sigma(T) \subset D$ and $\bar{D} \subset \Delta(f)$.

Several comments are needed concerning this definition.

1. When f is given, there exists a Cauchy domain of the required sort. This is intuitively plausible. Let δ be the minimum (positive) distance from $\sigma(T)$ to the boundary of $\Delta(f)$, and cover the complex plane with a grid of hexagons of diameter $\delta/3$. It is not difficult to show that a suitable Cauchy domain is the interior of the union of all the closed hexagonal regions that contain points of $\sigma(T)$. (Also, see Theorem 3.3 in Taylor [2].)
2. The integral in (8-3) has a value independent of the particular choice of D. This is so by an application of Cauchy's theorem (Theorem 1.4). For, if D_1 and D_2 are two Cauchy domains of the sort considered, we have $\sigma(T) \subset D_1 \cap D_2$, and there exists a bounded Cauchy domain D such that $\sigma(T) \subset D$ and $\bar{D} \subset D_1 \cap D_2$. Now $D_1 \setminus \bar{D}$ is a

bounded Cauchy domain, and its oriented boundary consists of $+\partial D_1$ and $-\partial D$. Moreover, $f(\lambda)R_\lambda$ has no singularities in $D_1\backslash\bar{D}$ or on its boundary, and hence, by Cauchy's theorem, the integral over $+\partial D_1$ is equal to the integral over $+\partial D$. The same result holds with D_2 in place of D_1, and so our assertion is justified.

3. The integral in (8-3) is unchanged in value if we replace f by any other member of $\mathfrak{A}(T)$, say g, such that $f(\lambda)=g(\lambda)$ at each point λ of an open set containing $\sigma(T)$. This follows from (1) and (2).
4. Any $S \in L(X)$ that commutes with T also commutes with $f(T)$, because such an S must commute with $f(\lambda)R_\lambda$ for $\lambda \in \rho(T)$.

It is obvious from (8-3) that $(f+g)(T)=f(T)+g(T)$ and $(\alpha f)(T)=\alpha f(T)$. We now show that the operator corresponding to fg is $f(T)g(T)$. The proof depends heavily on the resolvent equation $R_\lambda - R_\mu = (\mu-\lambda)R_\lambda R_\mu$. Choose bounded Cauchy domains D_1, D_2 such that $\sigma(T)\subset D_1$, $\bar{D}_1 \subset D_2$ and $\bar{D}_2 \subset \Delta(f)\cap\Delta(g)$. Then we express $f(T)$ as an integral with respect to λ over $+\partial D_1$ and $g(T)$ as an integral with respect to μ over $+\partial D_2$. Then we can write

$$
\begin{aligned}
\text{(8-4)}\qquad f(T)g(T) &= \frac{1}{2\pi i}\int_{+\partial D_1} f(\lambda)R_\lambda g(T)\,d\lambda \\
&= \frac{1}{2\pi i}\int_{+\partial D_1} f(\lambda)\left\{\frac{1}{2\pi i}\int_{+\partial D_2} g(\mu)R_\lambda R_\mu\,d\mu\right\}d\lambda.
\end{aligned}
$$

(We can move operators such as $g(T)$ and R_λ, for fixed λ, in and out of integration signs because operator multiplication is a continuous operation.) We now replace the $R_\lambda R_\mu$ in (8-4) by

$$
\frac{R_\lambda - R_\mu}{\mu-\lambda} = \frac{R_\lambda}{\mu-\lambda} + \frac{R_\mu}{\lambda-\mu},
$$

and invert the order of the iterated integration where the second fraction is concerned. Since $\lambda \in D_2$ and μ is not in $\bar{D}_1$, we have

$$
\frac{1}{2\pi i}\int_{+\partial D_2}\frac{g(\mu)}{\mu-\lambda}\,d\mu = g(\lambda), \qquad \frac{1}{2\pi i}\int_{+\partial D_1}\frac{f(\lambda)}{\lambda-\mu}\,d\lambda = 0.
$$

Hence we obtain from (8-4) the desired result

$$
f(T)g(T) = \frac{1}{2\pi i}\int_{+\partial D_1} f(\lambda)g(\lambda)R_\lambda\,d\lambda = (fg)(T).
$$

It is instructive to regard the association of $f(T)$ with f as an algebraic homomorphism. But in order to be precise about this, it is first of all necessary to convert $\mathfrak{A}(T)$ into an algebra by an appropriate equivalence relation. We define two functions f, g as being equivalent (relative to T) if $f(\lambda)=g(\lambda)$ on some open set containing $\sigma(T)$. Then $\mathfrak{A}(T)$ is divided into equivalence classes, and these classes form a commutative algebra with a unit element if we define the algebraic operations in an obvious way, using representative functions. We have already noted that $f(T)$ is unchanged if f is replaced by an equivalent function.

Theorem 8.1. *The mapping $f \mapsto f(T)$ by the formula* (8-3) *is an algebraic homomorphism of the algebra of the equivalence classes of $\mathfrak{A}(T)$ into the algebra $L(X)$. This mapping carries the function $f(\lambda) \equiv 1$ into I and the function $f(\lambda) \equiv \lambda$ into T.*

The proof is covered by the preceding discussion. The last two assertions are justified by (8-1) and (8-2).

The reader may have noticed that the formula defining $f(T)$ has a striking appearance when for heuristic effect we write R_λ in the form $1/(\lambda - T)$, and

$$f(T) = \frac{1}{2\pi i} \int_{+\partial D} \frac{f(\lambda)}{\lambda - T}\, d\lambda.$$

In formal structure this is just Cauchy's formula, with T in place of a complex number. We refer to the use of the homomorphism $f \mapsto f(T)$ and the consequences flowing out of it as an operational calculus for T.

One use of the operational calculus is that it enables us to compute inverse operators in certain situations.

Theorem 8.2. *Suppose $T \in L(X)$ and $f \in \mathfrak{A}(T)$. Suppose $f(\lambda) \neq 0$ when $\lambda \in \sigma(T)$. Then $f(T)$ is a one-to-one mapping of X onto all of X, with inverse $g(T)$, where g is any member of $\mathfrak{A}(T)$ equivalent to the reciprocal of $f(\lambda)$.*

Proof. From $f(\lambda)g(\lambda) \equiv 1$ on a neighborhood of $\sigma(T)$ we infer $f(T)g(T) = g(T)f(T) = I$, and the conclusion follows. □

We now give an example that illustrates the operational calculus for a particular operator.

Example 1. Let $X = C[0, 1]$, and consider the special Volterra-type operator T, where $Tx = y$ means

$$y(s) = \int_0^s x(t)\, dt.$$

It is easily verified by induction that $T^{n+1}x = y$ means

$$y(s) = \frac{1}{n!} \int_0^s (s-t)^n x(t)\, dt.$$

In this case the series (3-4) for R_λ converges whenever $\lambda \neq 0$ (i.e., $\sigma(T)$ is the single point $\lambda = 0$). By using this series we find that $x = R_\lambda y$ means

$$x(s) = \frac{1}{\lambda} y(s) + \frac{1}{\lambda^2} \int_0^s e^{(s-t)/\lambda} y(t)\, dt. \tag{8-5}$$

In this case $\mathfrak{A}(T)$ consists of functions analytic in a neighborhood of $\lambda = 0$. If f is such a function, we can deduce the meaning of $f(T)$ from (8-3) and (8-5). The relation $y = f(T)x$ can be written

$$y(s) = \frac{1}{2\pi i} \oint f(\lambda) \left\{ \frac{1}{\lambda} x(s) + \frac{1}{\lambda^2} \int_0^s e^{(s-t)/\lambda} x(t)\, dt \right\} d\lambda,$$

or

$$y(s) = f(0)x(s) + \int_0^s x(t) \left\{ \frac{1}{2\pi i} \oint \frac{f(\lambda)}{\lambda^2} e^{(s-t)/\lambda}\, d\lambda \right\} dt, \tag{8-6}$$

where $\oint$ denotes integration counterclockwise around some sufficiently small circle $|\lambda| = r$. We also write this in the form

$$y(s) = f(0)x(s) + \int_0^s F(s-t)x(t)\, dt, \tag{8-7}$$

where

$$F(u) = \frac{1}{2\pi i} \oint \frac{f(\lambda)}{\lambda^2} e^{u/\lambda}\, d\lambda. \tag{8-8}$$

The function F turns out to be an entire function of exponential type. If $f(\lambda)$ is expressed as a power series in λ, with radius of convergence α, the type of F is exactly $1/\alpha$. See problem 6 for more on this subject.

We can use the operational calculus for this particular operator to solve the differential equation

$$y^{(n)}(s) + a_1 y^{(n-1)}(s) + \cdots + a_n y(s) = x(s)$$

with the initial conditions

$$y(0) = y'(0) = \cdots = y^{(n-1)}(0) = 0,$$

where $x \in C[0, 1]$ and the coefficients $a_1, \ldots, a_n$ are constants. It is easy to see that y is a solution of this differential equation and satisfies the initial condition if and only if

$$(I + a_1 T + \cdots + a_n T^n)y = T^n x.$$

Now let $g(\lambda) = 1 + a_1\lambda + \cdots + a_n\lambda^n$. The problem is to solve $g(T)y = T^n x$ for y. Since $g(0) \neq 0$, we know by Theorem 8.2 that $g(T)$ has an inverse. The operational calculus shows that $y = f(T)x$, where $f(\lambda) = \lambda^n/g(\lambda)$. Hence by (8-6), the solution of the problem is

$$y(s) = \int_0^s x(t)\left\{\frac{1}{2\pi i}\oint \frac{\lambda^{n-2}}{g(\lambda)} e^{(s-t)/\lambda}\, d\lambda\right\} dt,$$

where the contour encloses $\lambda = 0$, and all the zeros of $g(\lambda)$ are outside of the contour. By the change of variable $z = \lambda^{-1}$, this solution can be put in the form

$$y(s) = \int_0^s x(t)\left\{\frac{1}{2\pi i}\oint \frac{e^{(s-t)z}}{z^n + a_1 z^{n-1} + \cdots + a_n}\, dz\right\} dt,$$

where the contour is counterclockwise and encloses all the zeros of the polynomial in the denominator. The integral may be evaluated by computing residues.

A Generalization to Closed Operators

It is desirable to develop a generalization of the formula (8-3) so as to yield an operational calculus for T when T is any closed linear operator with domain and range in X. It turns out that such a development is possible provided that the resolvent set $\rho(T)$ is not empty. When we give up the condition that $T \in L(X)$, the spectrum of T need no longer be compact; it might be empty or it might be the whole plane. However, we assume explicitly that $\sigma(T)$ is *not* the whole plane; we also assume that T is closed. This permits us to utilize Theorem 2.3.

Definition. By $\mathfrak{A}_\infty(T)$ we mean the class of complex-valued functions f such that (a) $\Delta(f)$ is an open set in the complex plane that contains $\sigma(T)$ and is such that the complement of $\Delta(f)$ is compact; (b) f is differentiable in $\Delta(f)$ and $f(\lambda)$ is bounded as $|\lambda| \to \infty$.

We know from function theory that $f(\lambda)$ approaches a finite limit as $|\lambda| \to \infty$, and we denote this limit by $f(\infty)$. We may then say that f is locally analytic on $\sigma(T)$ and at ∞.

We may define an equivalence relation in $\mathfrak{A}_\infty(T)$: two functions are equivalent if they agree on a neighborhood of $\sigma(T)$ and also on a neighborhood of ∞. As before, the equivalence classes form a commutative algebra in an obvious way. The function $f(\lambda) \equiv 1$ determines a unit for the algebra.

In seeking the proper replacement for (8-3), we observe that Cauchy's formula for an element of $\mathfrak{A}_\infty(T)$ holds in the form

$$f(\xi) = f(\infty) + \frac{1}{2\pi i}\int_{+\partial D} \frac{f(\lambda)}{\lambda - \xi}\, d\lambda, \tag{8-9}$$

where D is an *unbounded* Cauchy domain such that $\bar{D} \subset \Delta(f)$ and $\xi \in D$. Note that the complement of an unbounded Cauchy domain is compact; in fact, there is just one unbounded component of such a domain, and its complement is compact.

The appropriate definition of $f(T)$ when $f \in \mathfrak{A}_\infty(T)$ is

$$f(T) = f(\infty)I + \frac{1}{2\pi i}\int_{+\partial D} f(\lambda)R_\lambda \, d\lambda, \tag{8-10}$$

where D is an unbounded Cauchy domain such that $\sigma(T) \subset D$ and $\bar{D} \subset \Delta(f)$. The four comments made after (8-3) have counterparts in the present situation. Observe that $f(T) \in L(X)$ even though T need not be in $L(X)$.

As an illustration of $\Delta(f)$ and D, suppose $\sigma(T)$ is the entire real axis. Then, for some positive r and ε, $\Delta(f)$ must include all points for which $|\lambda| > r$ and all points for which the imaginary part of λ is in absolute value less than ε. We could then take D to be the union of the sets $\{\lambda : |\lambda| > 2r\}$, $\{\lambda : |\lambda| \leq 2r$ and $|Im\,\lambda| < \varepsilon/2\}$.

As in the case of Theorem 8.1, the mapping $f \mapsto f(T)$ defined by (8-10) is an algebraic homomorphism of the algebra of equivalence classes of $\mathfrak{A}_\infty(T)$ into the algebra $L(X)$, and the algebra preserves the unit element; that is, $f(\lambda) \equiv 1$ maps into $f(T) = I$. To see the truth of this last assertion let D_1 be the complement of $\bar{D}$ (D as in (8-10)). Then D_1 is a bounded Cauchy domain, $\bar{D}_1 \subset \rho(T)$, and $+\partial D = -\partial D_1$. Since R_λ is analytic on $\rho(T)$, the integral in (8-10) vanishes if $f(\lambda) \equiv 1$. In this case $f(\infty) = 1$, and we get $f(T) = I$. The proof that $(fg)(T) = f(T)g(T)$ is similar to the corresponding proof based on (8-3), and we leave the argument to the reader. One must use (8-9).

If T is not in $L(X)$, we must use (8-10) instead of (8-3). But if $T \in L(X)$ and $f \in \mathfrak{A}_\infty(T)$, the operator $f(T)$ given by (8-10) is the same as that given by (8-3). To prove this, choose for the D in (8-10) the union of a bounded Cauchy domain D_1 and the exterior of a very large circle C that encloses D_1, where $\sigma(T) \subset D_1$ and $\bar{D}_1 \subset \Delta(f)$. Then the integral over $+\partial D$ in (8-10) becomes the integral over $+\partial D_1$ plus an integral around C. Since $\|R_\lambda\| \to 0$ as $|\lambda| \to \infty$ (when $T \in L(X)$), it is easy to prove by standard methods that the integral around C cancels the term $f(\infty)I$.

Corresponding to Theorem 8.2, we have the following result.

Theorem 8.3. *If T is closed, $f \in \mathfrak{A}_\infty(T)$ and f has no zeros on $\sigma(T)$ or at ∞, the operator $f(T)$ is invertible in $L(X)$. The inverse of $f(T)$ is $g(T)$, where g is any member of $\mathfrak{A}_\infty(T)$ equivalent to the reciprocal of $f(\lambda)$.*

The proof is the same as for Theorem 8.2.

Polynomials and $\mathfrak{A}_\infty(T)$

A polynomial of degree $n \geqslant 1$ does not belong to $\mathfrak{A}_\infty(T)$. Nevertheless, it is convenient to be able to deal with polynomials in connection with the operational calculus. We consider now how this is to be done.

If $p(\lambda)$ is a polynomial of degree n, we define $p(T)$ in the obvious way, by putting T^k in place of λ^k in the expression for $p(\lambda)$; the domain of $p(T)$ is $\mathscr{D}(T^n)$, as defined at the beginning of § 6. We have several lemmas that are useful in dealing with polynomials.

Lemma 8.4. *Suppose $f \in \mathfrak{A}_\infty(T)$, and suppose either (a) that f has a zero of order m at ∞ or (b) that f vanishes identically in a neighborhood of ∞. Let p be a polynomial of degree n, where $0 < n \leqslant m$ in case a and $0 < n$ in case b. Let $F(\lambda) = p(\lambda)f(\lambda)$. Then $F \in \mathfrak{A}_\infty(T)$, and the range of $f(T)$ lies in $\mathscr{D}(T^n)$, and $F(T) = p(T)f(T)$.*

For an indication of the proof see problem 11.

Lemma 8.5. *Suppose $f \in \mathfrak{A}_\infty(T)$, $\alpha \in \rho(T)$, and an integer $n \geqslant 0$ are given, and suppose that $g \in \mathfrak{A}_\infty(T)$, where $g(\lambda) = (\alpha - \lambda)^n f(\lambda)$. (Note that this hypothesis is essentially a condition on the behavior of f near $\lambda = \infty$.) Then $f(T)\mathscr{D}(T^k) \subset \mathscr{D}(T^{n+k})$ when $k \geqslant 0$.*

Proof. If $x \in \mathscr{D}(T^k)$, we can write $x = (R_\alpha)^k y$ for some y. Then $f(T)x = (R_\alpha)^n g(T)(R_\alpha)^k y = (R_\alpha)^{n+k} g(T)y \in \mathscr{D}(T^{n+k})$. ☐

Lemma 8.6. *If $f \in \mathfrak{A}_\infty(T)$ and $p(\lambda)$ is a polynomial of degree $n \geqslant 1$, then $f(T)p(T)x = p(T)f(T)x$ if $x \in \mathscr{D}(T^n)$.*

Proof. Let $G(\lambda) = (\alpha - \lambda)^{-n} p(\lambda)$. Then $G(T) = p(T)(R_\alpha)^n$ (see problem 11). If $x \in \mathscr{D}(T^n)$, we can write $x = (R_\alpha)^n y$. We know $f(T)x \in \mathscr{D}(T^n)$ (Lemma 8.5.). Then $p(T)f(T)x = p(T)f(T)(R_\alpha)^n y = G(T)f(T)y = f(T)p(T)(R_\alpha)^n y$, and the proof is complete. ☐

We see next how to express $p(T)x$ as an integral.

Theorem 8.7. *Suppose $\alpha \in \rho(T)$. Let D be any unbounded Cauchy domain such that $\sigma(T) \subset D$ and α is not in $\bar{D}$. Then, if $p(\lambda)$ is a polynomial of degree n and $x \in \mathscr{D}(T^n)$,*

$$p(T)x = \frac{1}{2\pi i}\int_{+\partial D} \frac{p(\lambda)}{(\lambda - \alpha)^{n+1}}(T - \alpha)^{n+1} R_\lambda x \, d\lambda. \tag{8-11}$$

If $f \in \mathfrak{A}_\infty(T)$, we can also compute $f(T)$ by this formula for any $n \geqslant 0$, by putting $f(\lambda)$ in place of $p(\lambda)$ and $f(T)$ in place of $p(T)$, provided we choose D so that $\sigma(T) \subset D$ and $\bar{D} \subset \Delta(f)$.

Proof. We start from the formula

$$(T-\alpha)^{n+1}R_\lambda x = (\lambda-\alpha)^{n+1}R_\lambda x - \sum_{k=0}^{n} (\lambda-\alpha)^{n-k}(T-\alpha)^k x,$$

which is easily established by induction. The evaluation of the integral in (8-11) then becomes a matter of evaluation of familiar integrals, from which the final results are easily obtained. We leave the details to the reader. □

The next theorem is a generalization of Theorem 8.3.

Theorem 8.8. *Suppose $f \in \mathfrak{A}_\infty(T)$ and $f(\lambda) \neq 0$ if $\lambda \in \sigma(T)$, but that $f(\infty)=0$, the zero at ∞ being of finite order m. Then $f(T)$ has an inverse, the range of $f(T)$ is $\mathscr{D}(T^m)$ and, for $x \in \mathscr{D}(T^m)$, we have*

$$[f(T)]^{-1}x = \frac{1}{2\pi i}\int_{+\partial D} \{f(\lambda)(\lambda-\alpha)^{m+1}\}^{-1}(T-\alpha)^{m+1}R_\lambda x\, d\lambda. \tag{8-12}$$

Here $\alpha \in \rho(T)$; α and the unbounded Cauchy domain D are to be chosen so that $\sigma(T) \subset D$, $\bar{D} \subset \Delta(f)$, α is not in $\bar{D}$, and $f(\lambda) \neq 0$ if $\lambda \in \bar{D}$.

Proof. Let $g(\lambda) = (\alpha-\lambda)^m f(\lambda)$. Then g has no zeros on $\sigma(T)$ or at ∞, so that $g(T)$ has an inverse belonging to $L(X)$. Now $f(T) = (R_\alpha)^m g(T)$. Hence $f(T)$ has the inverse $[g(T)]^{-1}(\alpha - T)^m$ with domain $\mathscr{D}(T^m)$. We calculate $[f(T)]^{-1}x = [g(T)]^{-1}(\alpha-T)^m x$ by using (8-11), putting $n=0$, $1/g(\lambda)$ in place of $p(\lambda)$, and $(\alpha-T)^m x$ in place of x. The result is (8-12). □

We now consider an example of a closed operator with unbounded spectrum.

Example 2. Let $X = C[0, 2\pi]$. Let $\mathscr{D}(T)$ be the set of continuously differentiable functions $x \in X$ such that $x(0) = x(2\pi)$, and let $Tx = y$ mean $y(s) = -ix'(s)$. To compute R_λ we solve the differential equation

$$x'(s) - i\lambda x(s) = -iy(s)$$

with the condition $x(0) = x(2\pi)$ on the solution. We find that the values $\lambda = 0, \pm 1, \pm 2, \ldots$ are in the point spectrum, the eigenfunctions corresponding to $\lambda = n$ being multiples of e^{ins}. All other values of λ are in $\rho(T)$, with $x = R_\lambda y$ expressed by

$$x(s) = \frac{e^{i\pi\lambda}}{2\sin\pi\lambda}\int_0^{2\pi} e^{i\lambda(s-t)}y(t)\,dt - i\int_0^s e^{i\lambda(s-t)}y(t)\,dt. \tag{8-13}$$

Now suppose $f \in \mathfrak{A}_\infty(T)$. This means that f is analytic at ∞ and also in some neighborhood of each of the points of $\sigma(T)$. For the purpose of computing $f(T)$ we may assume that $f(\lambda) = \Sigma_0^\infty a_n\lambda^{-n}$ when $|\lambda| > N + \frac{1}{3}$, where

N is some positive integer, and that near $\lambda = k$, $k = 0, \pm 1, \ldots, \pm N$, f is given by $f(\lambda) = g_k(\lambda)$, where g_k is analytic at $\lambda = k$. In computing $f(T)$ by formula (8-10), we can take D to consist of the exterior of the circle $C : |\lambda| = N + \frac{1}{2}$ and the union of the interiors of circles $C_k : |\lambda - k| < \frac{1}{4}$, $k = 0, \pm 1, \ldots, \pm N$. The contribution to $f(T)x$ from integration around C_k turns out to be

$$\frac{g_k(k)}{2\pi} \int_0^{2\pi} x(t) e^{ik(s-t)}\, dt = g_k(k) \xi_k e^{iks},$$

where ξ_k is the Fourier coefficient of $x(t)$ with respect to e^{ikt}. The general formula for $y = f(T)x$ is a bit lengthy, thus we forego writing it out here. However, in the special case where $f(\lambda) \equiv a_0$ when $|\lambda| > N + \frac{1}{3}$, things are much simpler, and $y = f(T)x$ is expressed by

$$y(s) = a_0 x(s) + \sum_{k=-N}^{N} (b_k - a_0) \xi_k\, e^{iks}, \tag{8-14}$$

where $b_k = g_k(k)$. The coefficients a_0 and b_k can be assigned arbitrarily. From Theorem 8.3 we get the interesting result that if a_0 and the b_k's are all different from zero, the solution of (8-14) for x is

$$x(s) = \frac{1}{a_0} y(s) + \sum_{k=-N}^{N} \left(\frac{1}{b_k} - \frac{1}{a_0} \right) \eta_k\, e^{iks},$$

where the η_k's are the Fourier coefficients of y.

If we impose special conditions on T, there may be various other ways besides that indicated by (8-10) for developing an operational calculus. The cases in which T is such that $\sigma(T)$ lies in a half plane (Hille and Phillips [1, Chapter XV] or in a strip (Bade [1]) are of importance, and in these cases very interesting and useful operational calculi have been developed.

PROBLEMS

1. If $f(\lambda) = \Sigma_0^\infty a_n \lambda^n$ in (8-3), show that $f(T) = \Sigma_0^\infty a_n T^n$, the series converging in $L(X)$. [This can be done directly, without using the more general result in problem 2.]
2. Let $T \in L(X)$, and let Δ be an open set in C containing $\sigma(T)$. Let $\{f_n\}$ be a sequence in $\mathfrak{A}(T)$ such that $\Delta(f_n) \supset \Delta$ for each n, and let f be in $\mathfrak{A}(T)$ with $\Delta(f) \supset \Delta$. If $\{f_n\}$ converges to f uniformly on compact subsets of Δ, then $\{f_n(T)\}$ converges to $f(T)$ in the uniform topology of $L(X)$.
3. Suppose $P \in L(X)$ and $P^2 = P$. Find an explicit representation for $(\lambda - P)^{-1}$, where $\lambda(\lambda - 1) \neq 0$. Use this to obtain a simple expression for $f(P)$ when $f \in \mathfrak{A}(P)$.
4. If $T \in L(X)$ and $f \in \mathfrak{A}(T)$, then $f \in \mathfrak{A}(T')$ and $f(T') = [f(T)]'$.
5. (The essential uniqueness of the operational calculus.) Let $T \in L(X)$, and let Δ be a Cauchy domain in C containing $\sigma(T)$ such that each component of Δ

contains at least one point of $\sigma(T)$. Let $\mathfrak{A}(T, \Delta)$ be the set of the restrictions to Δ of all $f \in \mathfrak{A}(T)$ such that $\Delta(f) \supset \Delta$. Then $\mathfrak{A}(T, \Delta)$ is an algebra. Let ϕ be an algebraic homomorphism of $\mathfrak{A}(T, \Delta)$ into $L(X)$ such that

(i) $f(\lambda) \equiv 1$ implies $\phi(f) = I$,

(ii) $f(\lambda) \equiv \lambda$ implies $\phi(f) = T$,

(iii) if $\{f_n\}$ converges uniformly to f on compact subsets of Δ, then $\phi(f_n)$ converges to $\phi(f)$ in $L(X)$.

Then, for all $f \in \mathfrak{A}(T, \Delta)$, $\phi(f)$ is the operator $f(T)$ defined in (8-3). (*Hint.* Use Runge's theorem. See Rudin [1, pages 288–290].)

6. If $f(\lambda) = \Sigma_0^\infty a_n \lambda^n$ in (8-6), show that $F(u) = \Sigma_0^\infty a_{n+1}(u^n/n!)$ in (8-8). If the radius of convergence of the f series is α, F is of exponential type $1/\alpha$. See R. P. Boas [1, page 839]. We see that f determines F; conversely, F determines f, except that a_0 is left arbitrary. Any Volterra-type integral operator with kernel $F(s-t)$, where F is an entire function of exponential type, is an operator $f(T)$, where f is analytic at $\lambda = 0$ and $f(0) = 0$.

7. Let $X = C[a, b]$ (a finite interval). Let T be the differentiation operator, $Tx = x'$ where $\mathscr{D}(T)$ is the set of those $x \in X$ such that $x(a) = 0$ and the derivative x' also belongs to X. This operator is closed and $\sigma(T)$ is empty (problem 7, § 3). Show that $x = R_\lambda y$ means

$$x(s) = -\int_a^s e^{\lambda(s-t)} y(t)\, dt.$$

If $f \in \mathfrak{A}_\infty(T)$ is defined by $f(\lambda) = \Sigma_0^\infty a_n \lambda^{-n}$, show that $f(T)x = y$ means

$$y(s) = a_0 x(s) + \int_a^s F(s-t) x(t)\, dt,$$

where $F(u) = \Sigma_0^\infty a_{n+1}(u^n/n!)$. The situation here is closely related to that in Example 1, because $y = T^{-1}x$ means $y(s) = \int_a^s x(t)\, dt$.

8. When T is closed and $\rho(T)$ is not empty, $p(T)$ is closed. Outline of proof: If $\alpha \in \rho(T)$, let $A = (T-\alpha)^{-1}$. Write $p(\lambda) = \Sigma_0^n b_k(\lambda - \alpha)^{n-k}$, $q(\mu) = \Sigma_0^n b_k \mu^k$, $b_0 \neq 0$. Then $p(T)x = q(A)(T-\alpha)^n x$ if $x \in \mathscr{D}(T^n)$. Now $(T-\alpha)^n$ is closed, for it is the inverse of A^n. From this one proves that $p(T)$ is closed. See Taylor [2, Theorem 6.1]. If $\rho(T)$ is void, it can occur that T is closed but T^2 is not. Here is an example. Let $X = \ell^2 \times \ell^2$. If $(x, y) \in X$, let $\|(x, y)\|^2 = \|x\|^2 + \|y\|^2$. If $x = \{\xi_k\}$ and $y = \{\eta_k\}$, define $T(x, y) = (\{k^{-2}\eta_k\}, \{k\xi_k\})$, with $\mathscr{D}(T)$ consisting of all (x, y) such that $y \in \ell^2$ and $\{k\xi_k\} \in \ell^2$.

9. If T is a closed operator with a dense domain and $\rho(T) \neq \varnothing$ and if $p(\lambda)$ is a polynomial of degree $n \geqslant 1$, then $p(T)$ is bounded on $\mathscr{D}(T^n)$ if and only if T is bounded. [*Hint.* Is $\mathscr{D}(T^n)$ dense?]

10. Suppose $\alpha \in \rho(T)$. Then the function $f(\lambda) = (\alpha - \lambda)^{-n}$, $n \geqslant 1$, belongs to $\mathfrak{A}_\infty(T)$, and $f(T) = (R_\alpha)^n$. [For $n = 1$, use (8-10) and the resolvent equation.]

11. To prove Lemma 8.4 let $g(\lambda) = (\alpha - \lambda)^n f(\lambda)$, where $\alpha \in \rho(T)$. Then $f(T) = (R_\alpha)^n g(T)$ (use problem 10). Let $G(\lambda) = (\alpha - \lambda)^{-n} p(\lambda)$, and show that $G(T) = p(T)(R_\alpha)^n$ by expressing $p(\lambda)$ as a sum of powers of $\alpha - \lambda$. The relation $F(T) = p(T)f(T)$ now follows at once.

12. Let $p(\lambda)$ be a polynomial of degree $n \geqslant 1$ all of whose zeros lie in $\rho(T)$. Then $p(T)$ has X for its range, and it has an inverse belonging to $L(X)$ and given by

$$[p(T)]^{-1} = \frac{1}{2\pi i} \int \frac{1}{p(\lambda)} R_\lambda \, d\lambda,$$

the integration being extended clockwise around a set of nonoverlapping circles (one centered at each zero of $p(\lambda)$), each circle and its interior lying in $\rho(T)$. Method of proof: Let $f(\lambda) = 1/p(\lambda)$, and apply Theorem 8.8. A comparison of (8-12) for this case with (8-11) shows that $[f(T)]^{-1}x = p(T)x$, whence $[p(T)]^{-1} = f(T)$. Then apply (8-10).

13. Let T be the operator of problem 7. Use the formula in problem 12 to solve $p(T)y = x$ for y, where $p(\lambda)$ is a polynomial of degree n. Compare with the last part of Example 1.

14. Let $X = H^2$ (see Example 8, § II.2). Define $Tx(t) = tx(t)$ ($x \in X$, t the complex variable). Then $T \in L(X)$, $\sigma(T) = \{\lambda : |\lambda| \leqslant 1\}$. If $f \in \mathfrak{A}(T)$, $f(T)x(t) = f(t)x(t)$.

15. Let $X = L(0, \infty)$, and let T be the differentiation operator $Tx = x'$, with $\mathscr{D}(T)$ consisting of those $x \in X$ such that x is absolutely continuous on $[0, a]$ for every finite $a > 0$ and x' is again in X. It can be seen from Example 4 in § I.5 that $\sigma(T)$ consists of all λ for which the real part of λ is $\leqslant 0$. Also, $x = R_\lambda y$ is expressed by

$$x(s) = \int_0^\infty y(t) e^{\lambda(s-t)} \, dt.$$

Discuss the nature of $\mathfrak{A}_\infty(T)$ and the form of $f(T)$ in this case.

16. (S. Caradus) Suppose $T \in L(X)$ and T has a pseudoinverse. If f is locally analytic on $\sigma(T) \cup \{0\}$, $f(0) = 0$, and if f is one-to-one, then $f(T)$ has a pseudoinverse. [*Hint.* Use Theorem IV.13.9.]

V.9 SPECTRAL SETS. THE SPECTRAL MAPPING THEOREM

The results of this section are related to the concept of reducibility (see § 5). As we shall see, we can obtain a pair of subspaces that reduce T completely if we can split the spectrum of T into two parts of a suitable nature. We assume throughout the section that X is a complex Banach space and that T is a closed linear operator with domain and range in X. We also assume as always that $\rho(T)$ is not empty.

In addition to the concept of the spectrum of T, we need the concept of the *extended spectrum* of T. This is a point set in the *extended* complex plane (i.e., the one-point compactification of the ordinary plane, by adjunction of the point ∞). We denote the extended spectrum of T by $\sigma_e(T)$. It is defined to be the same as $\sigma(T)$ if $T \in L(X)$ and to consist of $\sigma(T)$ and the point ∞ if T is not in $L(X)$. Observe that $\sigma_e(T)$ is always closed and nonempty. The basic reason for considering ∞ is that R_λ always has a singularity at ∞ when

$\mathscr{D}(T) \neq X$. (See problem 7 of § 10). In contrast, when $T \in L(X)$ the expansion of R_λ given by (3-4) shows that R_λ is analytic at ∞ and has a zero there of order one.

Definition. A subset σ of $\sigma_e(T)$ is called a *spectral set* of T if it is both open and closed in the relative topology of $\sigma_e(T)$ as a subset of the extended plane. This is the same as requiring that both σ and $\sigma_e(T)\backslash\sigma$ be closed in the extended plane.

An isolated point of $\sigma_e(T)$ is of course a spectral set. If σ is a spectral set and if one of the sets σ, $\sigma_e(T)\backslash\sigma$ contains ∞, then the other one is bounded as a subset of the ordinary plane.

If σ is a spectral set of T, there is an $f \in \mathfrak{A}_\infty(T)$ such that $f(\lambda) = 1$ on a neighborhood of σ while $f(\lambda) = 0$ on a neighborhood of $\sigma_e(T)\backslash\sigma$. We then denote the operator $f(T)$ by E_σ. Since $f(\lambda)f(\lambda) = f(\lambda)$, the operational calculus shows that $E_\sigma E_\sigma = E_\sigma$, so that E_σ is a (continuous) projection. We call it *the projection associated with* σ. (Note that E_σ does not depend on the particular choice of $f \in \mathfrak{A}_\infty(T)$; see remark (3) preceding Theorem 8.1.)

Let σ and τ be spectral sets of T. It is readily verified (using the operational calculus) that the associated projections have the following properties:

$$
\text{(9-1)} \qquad
\begin{aligned}
&(a) \quad E_\sigma = 0 \qquad \text{if } \sigma = \varnothing.\\
&(b) \quad E_\sigma = I \qquad \text{if } \sigma = \sigma_e(T).\\
&(c) \quad E_{\sigma\cap\tau} = E_\sigma E_\tau = E_\tau E_\sigma.\\
&(d) \quad E_{\sigma\cup\tau} = E_\sigma + E_\tau - E_\sigma E_\tau.
\end{aligned}
$$

If, in particular, τ is the complementary set $\sigma' = \sigma_e(T)\backslash\sigma$, then $E_\sigma E_{\sigma'} = E_{\sigma'} E_\sigma = 0$ and $E_\sigma + E_{\sigma'} = E_{\sigma\cup\sigma'} + E_\sigma E_{\sigma'} = I$. Letting X_σ and $X_{\sigma'}$ be the ranges of E_σ and $E_{\sigma'}$, respectively, we have

$$X = X_\sigma \oplus X_{\sigma'}.$$

To show that T is completely reduced by $(X_\sigma, X_{\sigma'})$ it suffices to show that $E_\sigma \mathscr{D}(T) \subset \mathscr{D}(T)$ and $E_\sigma Tx = TE_\sigma x$ if $x \in \mathscr{D}(T)$ (Theorem 5.1). These things follow from Lemmas 8.5 (with $n = 0$, $k = 1$) and 8.6. Clearly any polynomial in T is completely reduced by $(X_\sigma, X_{\sigma'})$. Furthermore, $f(T)$ is completely reduced by $(X_\sigma, X_{\sigma'})$ if $f \in \mathfrak{A}_\infty(T)$. This is because the operational calculus produces operators that commute with each other; that is $f(T)E_\sigma = E_\sigma f(T)$.

These considerations are easily generalized to a finite number of disjoint spectral sets. The proof of the following theorem is left to the reader.

Theorem 9.1. *Suppose $\sigma_e(T) = \sigma_1 \cup \cdots \cup \sigma_n$, where $\sigma_1, \ldots, \sigma_n$ are pairwise disjoint spectral sets of T. Let $E_{\sigma(i)}$ be the projection associated with*

σ_i, and let $X_{\sigma(i)}$ be the range of $E_{\sigma(i)}$. Then T is completely reduced by $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$, that is, $I = E_{\sigma(1)} + \cdots + E_{\sigma(n)}$, $E_{\sigma(i)}E_{\sigma(j)} = 0$ if $i \neq j$, $\mathscr{D}(T)$ is invariant under $E_{\sigma(i)}$, and $X_{\sigma(i)}$ is invariant under T.

In the next two results we let T_σ be the restriction of T to the invariant subspace X_σ. It is evident that T_σ is a closed operator, since T is closed and X_σ is the range of the continuous projection E_σ.

Theorem 9.2. *Let σ be a spectral set of T. Then*

(*a*) $\sigma = \sigma_e(T_\sigma)$.

(*b*) *If $f \in \mathfrak{A}_\infty(T)$, then $f \in \mathfrak{A}_\infty(T_\sigma)$ and the restriction of $f(T)$ to X_σ is $f(T_\sigma)$.*

(*c*) *If σ is a bounded set, then $X_\sigma \subset \mathscr{D}(T^n)$ for each n, and T_σ is continuous on X_σ.*

Proof. It is convenient to prove (*c*) first. When σ is bounded, we have $E_\sigma = f(T)$, where one property of f is that f vanishes on a neighborhood of ∞. We can then apply Lemma 8.5, with $k = 0$ and n arbitrary, to conclude that $X_\sigma \subset \mathscr{D}(T^n)$. By the closed graph theorem, T_σ is continuous.

Next, we prove (*a*). Let $\sigma' \doteq \sigma_e(T)\backslash\sigma$, and let $T_{\sigma'}$ be the restriction of T to $X_{\sigma'}$. Select any finite point μ not in σ. If f is the function used in obtaining $f(T) = E_\sigma$, let $g(\lambda) = (\mu - \lambda)^{-1} f(\lambda)$. Then $g \in \mathfrak{A}_\infty(T)$, and $(\mu - T)g(T) = g(T)(\mu - T) = E_\sigma$, from which it appears that $\mu - T$ and $g(T)$ are inverse to each other when restricted to X_σ. Consequently, $\mu \in \rho(T_\sigma)$. This shows that every point of $\sigma(T_\sigma)$ is a finite point of σ; likewise for $\sigma(T_{\sigma'})$ and σ'. Hence $\sigma(T_\sigma) \cap \sigma(T_{\sigma'}) = \varnothing$. Now $\sigma(T) = \sigma(T_\sigma) \cup \sigma(T_{\sigma'})$ (by Theorem 5.4). Therefore it follows that $\sigma(T_\sigma)$ is identical with the set of finite points of σ; likewise for $\sigma(T_{\sigma'})$ and σ'. In view of (*c*), this proves (*a*) for the case in which ∞ is not in σ. If ∞ *is* in σ, then $T \notin L(X)$. Also, ∞ is not in σ', so $T_{\sigma'}$ is continuous on $X_{\sigma'}$. It follows that T_σ cannot belong to $L(X_\sigma)$ for, otherwise, T would be in $L(X)$. Therefore $\infty \in \sigma_e(T_\sigma)$, and (*a*) is completely proved.

Now, if $f \in \mathfrak{A}_\infty(T)$, then $f \in \mathfrak{A}_\infty(T_\sigma)$ by part (*a*). In § 5 we observed that $(\lambda - T)^{-1}$ is completely reduced by $(X_\sigma, X_{\sigma'})$. Hence if $x \in X_\sigma$ and D is an unbounded Cauchy domain such that $\sigma(T_\sigma) \subset \sigma(T) \subset D$ and $\bar{D} \subset \Delta(f)$, then

$$f(T)x = f(\infty)x + \frac{1}{2\pi i}\int_{+\partial D} f(\lambda)(\lambda - T)^{-1}x \, d\lambda$$

$$= f(\infty)x + \frac{1}{2\pi i}\int_{+\partial D} f(\lambda)(\lambda - T_\sigma)^{-1}x \, d\lambda = f(T_\sigma)x.$$

This proves (*b*). □

From the proof of (*a*) of Theorem 9.2 and from Theorem 5.4, we immediately obtain the following result.

Corollary 9.3. *Let σ be a spectral set of T. Then*
(*a*) $\sigma \cap P\sigma(T) = P\sigma(T_\sigma)$.
(*b*) $\sigma \cap C\sigma(T) = C\sigma(T_\sigma)$.
(*c*) $\sigma \cap R\sigma(T) = R\sigma(T_\sigma)$.

We have seen that the correspondence between spectral sets σ and projections E_σ allows us to study T in terms of its behavior on the various subspaces X_σ. It is natural to wonder if the subspaces X_σ can themselves be decomposed into invariant subspaces. This is always possible, of course, if σ is not a connected spectral set. However, if σ is connected, then X_σ *may* not contain even *one* nontrivial closed invariant subspace. (Recall the discussion of invariant subspaces in § 7.)

In important special cases it is possible to find many more closed invariant subspaces than those arising from spectral sets. We shall return to this problem in Chapter VI (when T is self-adjoint) and again in Chapter VII (when T is normal).

Now suppose that σ is a spectral set of T such that $E_\sigma = 0$, and let $\sigma' = \sigma_e(T)\backslash\sigma$. Then $X_\sigma = (0)$ and $X_{\sigma'} = X$. If σ were unbounded, Theorem 9.2(*c*) would imply that $T = T_{\sigma'}$ is continuous on all of X. This is impossible since $\sigma_e(T)$ is bounded if $T \in L(X)$. Hence σ must be bounded. In this case, $\sigma = \sigma_e(T_\sigma) = \sigma(T_\sigma)$, by (*a*) and (*c*) of Theorem 9.2. But $\sigma(T_\sigma)$ is empty because X_σ is of dimension zero. We conclude that

$$E_\sigma = 0 \qquad \textit{if and only if} \qquad \sigma = \varnothing.$$

Combining this with (9-1), we also have that

$$E_\sigma = I \qquad \textit{if and only if} \qquad \sigma = \sigma_e(T).$$

These observations are needed in the next theorem. This theorem is sort of a companion to Theorem 8.8.

Theorem 9.4. *Suppose that $\sigma(T)$ is bounded and that f is an element of $\mathfrak{A}_\infty(T)$ that vanishes on a neighborhood of ∞ but has no zeros on $\sigma(T)$. Then $f(T)$ has the same range and null manifold as the projection associated with $\sigma(T)$. In particular, $f(T)$ has no inverse if $\mathscr{D}(T) \neq X$.*

Proof. Note that if $\mathscr{D}(T) \neq X$ and if $\sigma(T)$ is bounded, then $\sigma(T)$ is a spectral set, but the associated projection is not I. Thus the second conclusion follows from the first. Now define $g(\lambda) = 0$ and $h(\lambda) = 1$ on the component of $\Delta(f)$ containing ∞, and define $g(\lambda) = 1$ and $h(\lambda) = f(\lambda)$ on the rest of $\Delta(f)$. Then $g(T) = E_\sigma$, where $\sigma = \sigma(T)$; and $h(T)$ is invertible in $L(X)$, because h has no zeros on $\sigma_e(T)$. Also, $gh = f$, and hence $E_\sigma h(T) = h(T)E_\sigma = f(T)$. The truth of the theorem follows easily from these relations, since $h(T)$ is invertible in $L(X)$. ☐

Examples

Example 1. Let $X = \ell^1$ (ℓ^2 would do just as well). Let $\{\lambda_n\}$ be a sequence of distinct nonzero numbers such that $\lambda_n \to 0$, and define Tx by $T\{\xi_k\} = \{\lambda_k \xi_k\}$. Then $\sigma(T)$ consists of the points $\lambda_1, \lambda_2, \ldots$ and 0. The resolvent is defined by $R_\lambda\{\xi_k\} = \{(\lambda - \lambda_k)^{-1}\xi_k\}$. Let E_k be the projection associated with the spectral set formed by the single point λ_k, and let Q_N be the projection associated with the spectral set consisting of $\lambda_{N+1}, \lambda_{N+2}, \ldots$ and 0. Let $\{u_k\}$ be the standard countable basis for ℓ^1; that is, u_k has 1 in the kth place and zeros elsewhere. Then we have

$$x = \sum_1^\infty \xi_k u_k, \qquad E_k x = \xi_k u_k, \qquad Q_N x = \sum_{N+1}^\infty \xi_k u_k.$$

Note that $I = E_1 + \cdots + E_N + Q_N$. Also, $TE_k = \lambda_k E_k$, and so $T = \lambda_1 E_1 + \cdots + \lambda_N E_N + TQ_N$. It is interesting to see what happens as $N \to \infty$. We have $\|Q_N x\| \to 0$ for each x, and so, for each x,

$$x = \sum_1^\infty E_k x \qquad \text{(convergence in } X\text{)}.$$

But $\|Q_N\| = 1$, and so we *cannot* write $I = \Sigma_1^\infty E_k$ (with convergence in $L(X)$). However, $\|TQ_N\| \to 0$, and so we *do* have

$$T = \sum_1^\infty \lambda_k E_k \qquad \text{(convergence in } L(X)\text{)}.$$

We also have

$$R_\lambda x = \sum_1^\infty \frac{E_k x}{\lambda - \lambda_k}, \qquad \text{but not} \qquad R_\lambda = \sum_1^\infty \frac{E_k}{\lambda - \lambda_k}.$$

Example 2. Consider the operator T of Example 2, § 8, where the space X is $C[0, 2\pi]$. For this operator $\sigma_e(T)$ consists of the points $0, \pm 1, \pm 2, \ldots$ and ∞. If E_n is the projection associated with the spectral set consisting of the single point $\lambda = n$, we easily find from (8–13) that $E_n x(s) = \xi_n e^{ins}$, where ξ_n is the Fourier coefficient of $x(t)$ with respect to e^{int}. If Q_N is the projection associated with the spectral set consisting of $\pm(N+1), \pm(N+2), \ldots$ and ∞, we find that

$$Q_N x(s) = x(s) - \sum_{k=-N}^{N} \xi_k e^{iks}.$$

In this case it is not *always* true that $Q_N x \to 0$ as $N \to \infty$. It *is* true, however, if $x \in \mathscr{D}(T)$, by a standard theorem on Fourier series. Thus we can write

$$x = \sum_{-\infty}^{\infty} E_n x \tag{9-2}$$

when $x \in \mathscr{D}(T)$. We have $TE_n = nE_n$. In general, it is not true that

$$Tx = \sum_{-\infty}^{\infty} nE_n x \tag{9-3}$$

when $x \in \mathscr{D}(T)$. But this is true if $x \in \mathscr{D}(T^2)$, since $E_nTx = TE_nx = nE_nx$, and in this case we can use (9-2) with Tx in place of x.

It is interesting to see what happens in the foregoing example if we change the basic space X in which the operator T works. Suppose we take $X = L^2(0, 2\pi)$ and define $Tx = -ix'$, with $\mathscr{D}(T)$ the set of those x such that $x(s)$ is absolutely continuous on $[0, 2\pi]$, with $x(0) = x(2\pi)$, and such that x' is in X. The spectrum and the formula for the resolvent are just as before, and the formulas for E_n and Q_N are unchanged. But now (9-2) holds for all $x \in L^2(0, 2\pi)$. For, if we let $u_n(s) = e^{ins}/\sqrt{2\pi}$, the set $\{u_n\}$, $n = 0, \pm 1, \pm 2, \ldots$, is a complete orthonormal set (see § II.8), and (9-2) is just the standard expansion of x with respect to this orthonormal set. That is, $E_nx = (x, u_n)u_n$. Formula (9-3) is now true for every x in $\mathscr{D}(T)$, for $(Tx, u_n)u_n = (x, Tu_n)u_n = (x, nu_n)u_n = nE_nx$.

The Spectral Mapping Theorem

The following generalization of Theorem 3.4 will be needed in § 10.

Theorem 9.5. *If $f \in \mathfrak{A}_\infty(T)$, the spectrum of $f(T)$ is exactly the set of values assumed by $f(\lambda)$ as λ varies over the set $\sigma_e(T)$. In symbols, $\sigma(f(T)) = f(\sigma_e(T))$.*

Proof. Let us first consider the point $\mu = 0$, which may or may not be in $\sigma(f(T))$. Then we shall prove the following special case.

(9-4) *$f(T)$ is invertible in $L(X)$ if and only if f has no zeros on $\sigma_e(T)$.*

If f has no zeros on $\sigma_e(T)$, then $f(T)$ is invertible in $L(X)$, by Theorem 8.3. Now suppose $f(T)$ is invertible in $L(X)$. If $f(\zeta) = 0$, we may write $f(\lambda) = (\lambda - \zeta)g(\lambda)$ for some $g \in \mathfrak{A}_\infty(T)$. (Take $g(\lambda) = f(\lambda)/(\lambda - \zeta)$ when $\lambda \neq \zeta$, and $g(\zeta) = f'(\zeta)$.) Then $f(T)x = (T - \zeta I)g(T)x$ for every x and $f(T)x = g(T)(T - \zeta I)x$ if $x \in \mathscr{D}(T)$. (Lemmas 8.4 and 8.6.) This shows that $\zeta \notin \sigma(T)$ for, otherwise, either the range of $f(T)$ would not be all of X or else $f(T)$ would fail to be injective. So f cannot have a zero on $\sigma(T)$. If $\mathscr{D}(T) \neq X$, we must also consider $f(\infty)$. By Theorem 8.8, f cannot have a zero at ∞ of finite order, for in this case the range of $f(T)$ could not be all of X. Furthermore, f cannot vanish identically on a neighborhood of ∞, this being impossible by the foregoing if $\sigma(T)$ is not bounded and by Theorem 9.4 if $\sigma(T)$ is bounded. This concludes the proof of (9-4).

The spectral mapping theorem is now easily proved by applying (9-4) to the function $F(\lambda)=\mu-f(\lambda)$. For, $\mu \notin f(\sigma_e(T))$ if and only if $\mu-f(\lambda)$ has no zeros on $\sigma_e(T)$. By (9-4), this is true if and only if $\mu I-f(T)$ is invertible in $L(X)$, that is, if and only if $\mu \notin \sigma(f(T))$. □

The spectral mapping theorem for polynomials also holds when $\mathscr{D}(T) \neq X$ if T is closed and has a nonempty resolvent set.

Theorem 9.6. *If $p(\lambda)$ is a polynomial, $\sigma(p(T))=p(\sigma(T))$. If $\overline{\mathscr{D}(T)}=X$, then $\sigma_e(p(T))=p(\sigma_e(T))$.*

The proof can be patterned after that of Theorem 9.5. At one stage in the argument the result of problem 12, § 8, is needed.

It is possible to obtain spectral mapping theorems for certain subsets of the spectrum. For example, if $T \in L(X)$ and if $\pi(T)$ denotes the approximate point spectrum of T, then $\pi(f(T))=f(\pi(T))$ whenever $f \in \mathfrak{A}(T)$ and f is not constant on any component of $\Delta(f)$. (See Theorem 5.12.2 of Hille and Phillips [1].) As another example, let T be a closed operator with a nonempty resolvent set, and let $f \in \mathfrak{A}_\infty(T)$. Then the Fredholm spectrum of $f(T)$ is exactly the set of values assumed by $f(\lambda)$ as λ varies over $\sigma_\Phi(T) \cup \{\infty\}$. (See Theorem 7 of Gramsch and Lay [1].) Other instances of results similar to Theorem 9.5 occur in the theory of semigroups of linear operators. (See Hille and Phillips [1, pages 464–471], and Sz.-Nagy [1].)

Composition of Functions

The next theorem increases the scope of the operational calculus by enabling us to deal with composition of functions.

Theorem 9.7. *Suppose $f \in \mathfrak{A}_\infty(T)$, $S=f(T)$, and $g \in \mathfrak{A}_\infty(S)$. Suppose also that $f(\infty) \in \Delta(g)$ (if $T \in L(X)$, we may always suppose f modified near $\lambda=\infty$, if necessary, to make $f(\infty) \in \Delta(g)$). Define F by $F(\lambda)=g[f(\lambda)]$ if $f(\lambda) \in \Delta(g)$. Then $F \in \mathfrak{A}_\infty(T)$ and $F(T)=g(S)$.*

Proof. Using Theorem 9.5, we see that $\sigma(T) \subset \Delta(F)$. Choose a bounded Cauchy domain D such that $\sigma(S) \subset D$, $\bar{D} \subset \Delta(g)$, and $f(\infty) \in D$. Choose an unbounded Cauchy domain D_1 such that $\sigma(T) \subset D_1$, $\bar{D}_1 \subset \Delta(f)$, and $f(\bar{D}_1) \subset D$. Then

$$F(T)=F(\infty)I+\frac{1}{2\pi i}\int_{+\partial D_1} g[f(\xi)]R_\xi\,d\xi$$

and, if $\xi \in \partial D_1$,

$$g[f(\xi)]=\frac{1}{2\pi i}\int_{+\partial D} \frac{g(\lambda)}{\lambda-f(\xi)}\,d\lambda.$$

If $h(\xi)=[\lambda-f(\xi)]^{-1}$ (λ fixed on ∂D), we see that $h(T)=(\lambda-S)^{-1}$. Since

$$F(\infty)=\frac{1}{2\pi i}\int_{+\partial D}\frac{g(\lambda)}{\lambda-f(\infty)}\,d\lambda$$

and

$$g(S)=\frac{1}{2\pi i}\int_{+\partial D}g(\lambda)(\lambda-S)^{-1}\,d\lambda,$$

the conclusion of the theorem follows by easy calculations, which we leave to the reader. □

One important application of Theorem 9.7 is to prove the next result, which is used in proving Theorem 10.8.

Theorem 9.8. *Suppose $f\in\mathfrak{A}_\infty(T)$, $S=f(T)$, and let τ be a spectral set of S. Let $\sigma=\sigma_e(T)\cap f^{-1}(\tau)$. Then σ is a spectral set of T, and the projection E_σ associated with σ and T is the same as the projection F_τ associated with τ and S.*

Proof. Let us write $\sigma'=\sigma_e(T)\backslash\sigma$, $\tau'=\sigma(S)\backslash\tau$. From Theorem 9.5 we see that $f(\sigma\cup\sigma')=\tau\cup\tau'$. Hence $\sigma'=\sigma_e(T)\cap f^{-1}(\tau')$, and it follows that σ and σ' are complementary spectral sets of T. Since $E_\sigma+E_{\sigma'}=I$ and $F_\tau+F_{\tau'}=I$, the relation $E_\sigma=F_\tau$ is implied by $E_{\sigma'}=F_{\tau'}$. Hence, in proving the theorem, it is allowable to assume that σ is a bounded set in the ordinary plane and that $f(\infty)$ is not in τ for, if this is not true, we can deal instead with σ' and τ'. Now let U_1, U_2, U_3, U_4 be open sets with the following properties: $\bar{U}_1$ and $\bar{U}_2$ are disjoint and so are $\bar{U}_3$, $\bar{U}_4$; U_1 and U_3 are bounded neighborhoods of τ and σ, respectively; U_2 contains τ', $f(\infty)$, and a neighborhood of ∞; U_4 contains σ' and a neighborhood of ∞; $\bar{U}_3$ and $\bar{U}_4$ are in $\Delta(f)$ and $f(\bar{U}_3)\subset U_1$, $f(\bar{U}_4)\subset U_2$. Such sets do exist. Define $f_\tau=1$ on U_1, $f_\tau=0$ on U_2, $f_\sigma=1$ on U_3, $f_\sigma=0$ on U_4. Then $f_\tau[f(\lambda)]=f_\sigma(\lambda)$ on $U_3\cup U_4$. But $f_\tau(S)=F_\tau$, $f_\sigma(T)=E_\sigma$, and so $F_\tau=E_\sigma$, by Theorem 9.7. □

PROBLEMS

1. Suppose $\alpha\in\rho(T)$ and let $A=R_\alpha$. Use Theorem 9.5 to show that if $\mu(\alpha-\lambda)=1$, then $\mu\in\sigma(A)$ if and only if $\lambda\in\sigma(T)$. If $\mu\in\rho(A)$ and $\mu(\alpha-\beta)=1$, use Theorem 9.7 to show that $(\mu-A)^{-1}=\mu^{-1}+\mu^{-2}R_\beta$. Take $f(\lambda)=(\alpha-\lambda)^{-1}$, $g(\lambda)=(\mu-\lambda)^{-1}$. Show also that $R_\beta=\mu A(\mu-A)^{-1}$.
2. Suppose $\alpha\in\rho(T)$ and $A=R_\alpha$. Make the transformation $\mu=(\alpha-\lambda)^{-1}$ from the λ-plane to the μ-plane. If $f\in\mathfrak{A}_\infty(T)$ and g is defined by $g(\mu)=f(\lambda)$, $g(0)=f(\infty)$, show that $g(A)=f(T)$.

3. Suppose $A, B \in L(X)$, $A \neq B$, $AB = BA$. Then, if $\lambda_0 \in \sigma(A)$, there exists $\lambda_1 \in \sigma(B)$ with $|\lambda_1 - \lambda_0| \leq \|A - B\|$. [*Suggestion.* Suppose the proposition is false, and investigate $r_\sigma((\lambda_0 - B)^{-1})$].
4. Consider the operator A of Example 2, § 4, with X taken as any fixed ℓ^p. Suppose $\Sigma_0^\infty |\alpha_n| < \infty$, and let $f(\lambda) = \Sigma_0^\infty \alpha_n \lambda^n$, $|\lambda| \leq 1$. Also, let $p_n(\lambda) = \Sigma_0^n \alpha_k \lambda^k$. Define $S = \Sigma_0^\infty \alpha_n A^n$ (series convergent in $L(X)$). Show that $\sigma(S) = f(\sigma(A))$. [This is not a direct application of Theorem 9.5, in general. Why not? For $f(\sigma(A)) \subset \sigma(S)$ use Theorem IV.1.5, and for $\sigma(S) \subset f(\sigma(A))$ use problem 3 (above) on S and $p_n(A)$.]
5. Suppose $f(T)$ is a projection in $L(X)$, where $f \in \mathfrak{A}_\infty(T)$. Then there is a spectral subset σ of $\sigma_e(T)$ such that $f(T) = E_\sigma$. A stronger result (not using this problem) is in the next problem.
6. Let $T \in L(X)$, and let $\{f_n\}$ be a sequence in $\mathfrak{A}(T)$.
   a. If $\{f_n(T)\}$ converges in $L(X)$ to an operator $U \in L(X)$, then $\{f_n\}$ converges uniformly on $\sigma(T)$ to a function f that is continuous on $\sigma(T)$. Furthermore, $f(\sigma(T)) \subset \sigma(U)$.
   b. If $\{f_n(T)\}$ converges in $L(X)$ to a projection E, then there exists a spectral subset σ of $\sigma(T)$ such that $E_\sigma = E$.

V.10 ISOLATED POINTS OF THE SPECTRUM

As in several preceding sections, we assume here that X is a complex Banach space and that T is a closed linear operator with a nonempty resolvent set. Then R_λ is analytic as a function of λ on $\rho(T)$, and an isolated point λ_0 of $\sigma(T)$ is an isolated singular point of R_λ. Hence there is a Laurent expansion of R_λ in powers of $\lambda - \lambda_0$. We write this in the form

$$R_\lambda = \sum_0^\infty (\lambda - \lambda_0)^n A_n + \sum_1^\infty (\lambda - \lambda_0)^{-n} B_n. \tag{10-1}$$

The coefficients A_n and B_n are members of $L(X)$, and this series representation of R_λ is valid when $0 < |\lambda - \lambda_0| < \delta$ for any δ such that all of $\sigma(T)$ except λ_0 lies on or outside the circle $|\lambda - \lambda_0| = \delta$. These coefficient operators are given by the usual standard formulas:

$$\begin{aligned} A_n &= \frac{1}{2\pi i} \int_C (\lambda - \lambda_0)^{-n-1} R_\lambda \, d\lambda, \\ B_n &= \frac{1}{2\pi i} \int_C (\lambda - \lambda_0)^{n-1} R_\lambda \, d\lambda, \end{aligned} \tag{10-2}$$

where C is any counterclockwise circle $|\lambda - \lambda_0| = h$ with $0 < h < \delta$.

It turns out that there are several important relationships among these coefficient operators. The demonstration of these relationships can be made conveniently by using the operational calculus. Choose $r > 0$ so that $2r < \delta$.

Define functions f_n as follows:

$$n \geqslant 0: f_n(\lambda) = \begin{cases} 0 & \text{if } |\lambda - \lambda_0| < r \\ (\lambda - \lambda_0)^{-n-1} & \text{if } |\lambda - \lambda_0| > 2r \end{cases}$$

$$(10\text{-}3) \qquad n < 0: f_n(\lambda) = \begin{cases} (\lambda - \lambda_0)^{-n-1} & \text{if } |\lambda - \lambda_0| < r \\ 0 & \text{if } |\lambda - \lambda_0| > 2r. \end{cases}$$

These functions all belong to $\mathfrak{A}_\infty(T)$, and $f_n(\infty) = 0$. If we compare the definition of $f_n(T)$ with formulas (10-2), we see that

$$(10\text{-}4) \qquad \begin{aligned} A_n &= -f_n(T), & n &\geqslant 0, \\ B_n &= f_{-n}(T), & n &\geqslant 1. \end{aligned}$$

We note in particular that B_1 is the projection E_σ for the case in which σ is the spectral set consisting of the single point λ_0. Since σ is not empty, we know that $B_1 \neq 0$. The only case in which $B_1 = I$ is that in which $\mathscr{D}(T) = X$ and $\sigma(T)$ consists of the single point λ_0. (See the remarks preceding Theorem 9.4.)

Now let us observe that for λ in an open set containing $\sigma_e(T)$,

$$(\lambda - \lambda_0) f_0(\lambda) + f_{-1}(\lambda) = 1,$$

$$(\lambda - \lambda_0)^n f_n(\lambda) = f_0(\lambda), \qquad n \geqslant 0,$$

$$f_{-(n+1)}(\lambda) = (\lambda - \lambda_0)^n f_{-1}(\lambda), \qquad n \geqslant 1.$$

From these relations and Lemma 8.4 we obtain the formulas

$$(10\text{-}5) \qquad (T - \lambda_0) A_0 = B_1 - I,$$

$$(10\text{-}6) \qquad (T - \lambda_0)^n A_n = A_0,$$

$$(10\text{-}7) \qquad B_{n+1} = (T - \lambda_0)^n B_1.$$

Some further relations are indicated in problem 1.

Since $\{\lambda_0\}$ is a spectral set, T is completely reduced by the pair of closed subspaces $X_1 = \mathscr{N}(B_1)$ and $X_2 = \mathscr{R}(B_1)$. By Theorem 9.2, if T_1 and T_2 are the restrictions of T to X_1 and X_2, respectively, then $\sigma(T_1) = \sigma_e(T) \setminus \{\lambda_0\}$ and $\sigma(T_2) = \{\lambda_0\}$. Hence, for any $n \geqslant 1$, $(\lambda_0 - T)^n$ is a one-to-one mapping of X_1 onto itself. Using Theorem 5.2, we have for $n \geqslant 1$,

$$(10\text{-}8) \qquad \mathscr{R}[(\lambda_0 - T)^n] = X_1 \oplus \mathscr{R}[(\lambda_0 - T_2)^n] \supset X_1,$$

$$(10\text{-}9) \qquad \mathscr{N}[(\lambda_0 - T)^n] = (0) \oplus \mathscr{N}[(\lambda_0 - T_2)^n] \subset X_2.$$

An important situation arises if for some n the inclusion relations in (10-8) and (10-9) become equalities. It turns out that this happens if and only if λ_0 is a pole of R_λ.

Poles of the Resolvent Operator

As in the classical theory of functions, we shall say that λ_0 is a *pole of* R_λ *of order* p if and only if $p \geq 1$, $B_p \neq 0$, and $B_n = 0$ when $n > p$ (see (10-1)). From (10-7) we see that $B_{n+1} = 0$ if $B_n = 0$. Hence λ_0 is a pole of order p if and only if $B_p \neq 0$ and $B_{p+1} = 0$. In that case, $B_1, \ldots, B_p$ are all nonzero operators. If λ_0 is an isolated point of $\sigma(T)$ but not a pole of R_λ, we call it an *isolated essential singularity* of R_λ.

Theorem 10.1 *If λ_0 is a pole of R_λ of order p, then λ_0 is an eigenvalue of T. The ascent and descent of $\lambda_0 - T$ are both equal to p. The range of the projection B_1 is the null space of $(\lambda_0 - T)^p$, and the range of $I - B_1$ is the range of $(\lambda_0 - T)^p$, so that B_1 determines the same decomposition of X as that given by Theorem 6.2, namely,*

$$X = \mathscr{R}[(\lambda_0 - T)^p] \oplus \mathscr{N}[(\lambda_0 - T)^p]. \tag{10-10}$$

Proof. We shall use the notation established for (10-8) and (10-9). From (10-7) we have

$$\mathscr{R}(B_{n+1}) = (T - \lambda_0)^n B_1 X = (T - \lambda_0)^n X_2 = \mathscr{R}[(\lambda_0 - T_2)^n].$$

Thus $B_{n+1} = 0$ if and only if $\mathscr{R}[(\lambda_0 - T_2)^n] = (0)$. Clearly this happens if and only if $\mathscr{N}[(\lambda_0 - T_2)^n] = X_2$. By hypothesis, $B_{p+1} = 0$ and $B_p \neq 0$. It follows from (10-8) that $\mathscr{R}[(\lambda_0 - T)^p] = X_1$ and the descent of $\lambda_0 - T$ is p. Likewise, (10-9) shows that $\mathscr{N}[(\lambda_0 - T)^p] = X_2$ and the ascent of $\lambda_0 - T$ is p. Finally, λ_0 is an eigenvalue because $p \geq 1$. □

The next theorem provides with Theorem 10.1 a characterization of poles of the resolvent operator.

Theorem 10.2. *Suppose $\lambda_0 \in \sigma(T)$ and $\lambda_0 - T$ has finite ascent and descent. Then λ_0 is a pole of R_λ.*

Proof. Let $p = \alpha(\lambda_0 - T)$. Then $\delta(\lambda_0 - T) = p$ and $X = \mathscr{R}_p \oplus \mathscr{N}_p$, by Theorem 6.2, where $\mathscr{R}_p$ and $\mathscr{N}_p$ denote the range and null space of $(\lambda_0 - T)^p$, respectively. Note that $p \geq 1$, since $\lambda_0 \in \sigma(T)$. The operator $(\lambda_0 - T)^p$ is closed since T is closed and $\rho(T) \neq \varnothing$ (problem 8 of § 8); hence $\mathscr{N}_p$ is a closed subspace. Since the range of $(\lambda_0 - T)^p$ has a closed complementary subspace $\mathscr{N}_p$, it must be closed, by Theorem IV.5.10. Thus the projections of X onto $\mathscr{R}_p$ and $\mathscr{N}_p$ are continuous. It is easily verified that T is completely reduced by $(\mathscr{R}_p, \mathscr{N}_p)$ (problem 4 of § 6). If we let T_1 and T_2 denote the restrictions of T to $\mathscr{R}_p$ and $\mathscr{N}_p$, respectively, then

$$\rho(T) = \rho(T_1) \cap \rho(T_2), \tag{10-11}$$

by Theorem 5.4.

The next step is to show that λ_0 is an isolated point of $\sigma(T)$. Clearly $\lambda_0 - T_1$ is one-to-one and $(\lambda_0 - T_1)\mathscr{R}_p = (\lambda_0 - T)\mathscr{R}_p = \mathscr{R}_{p+1} = \mathscr{R}_p$. Hence $\delta(\lambda_0 - T_1) = 0$ and $\lambda_0 \in \rho(T_1)$. It follows that $\mu \in \rho(T_1)$ for all μ in some disc Δ centered at λ_0. On the other hand, $\lambda_0 - T_2$ is nilpotent (of order p), and hence $\sigma(T_2) = \{\lambda_0\}$. (Apply Theorem 3.5 to the operator $\lambda_0 - T_2$.) We conclude from (10-11) that $\rho(T) \subset \Delta \backslash \{\lambda_0\}$.

Now we can regard $\{\lambda_0\}$ as a spectral set of T and obtain the Laurent expansion (10-1) of R_λ. Let C be a counterclockwise circle inside Δ, with center at λ_0. For $x \in \mathscr{R}_p$,

$$B_1 x = \frac{1}{2\pi i}\int_C (\lambda - T)^{-1} x \, d\lambda = \frac{1}{2\pi i}\int_C (\lambda - T_1)^{-1} x \, d\lambda$$
$$= \left(\frac{1}{2\pi i}\int_C (\lambda - T_1)^{-1} \, d\lambda\right) x,$$

since R_λ is completely reduced by $(\mathscr{R}_p, \mathscr{N}_p)$. The last integral above defines an operator in $L(\mathscr{R}_p)$, but this operator is zero, by Cauchy's theorem, since $\Delta \subset \rho(T_1)$. Hence $B_{p+1}x = (T - \lambda_0)^p B_1 x = (T - \lambda_0)^p (0) = 0$ for $x \in \mathscr{R}_p$. Also, if $x \in \mathscr{N}_p \subset \mathscr{D}(T^p)$, then $0 = B_1(T - \lambda_0)^p x = (T - \lambda_0)^p B_1 x = B_{p+1}x$. Thus $B_{p+1}x = 0$ for all $x \in \mathscr{R}_p \oplus \mathscr{N}_p = X$, which proves that λ_0 is a pole of R_λ. $\square$

We say that a pole of R_λ is of *finite rank* if the projection B_1 has a finite-dimensional range. The following fact is an immediate consequence of Theorems 7.9, 7.6 and 10.2.

Corollary 10.3. *If X is a Banach space and T is a compact operator in $L(X)$, then each nonzero point of $\sigma(T)$ is a pole of R_λ of finite rank.*

Another characterization of poles of R_λ is given in Theorem 10.5. To obtain this result we first prove a theorem that says, together with Theorem 10.2, that if $\lambda_0 - T$ has finite descent, then $\delta(\lambda - T) = 0$ for all λ in a deleted neighborhood of λ_0. (See Lay [1].)

Theorem 10.4. *If $\alpha(\lambda_0 - T) = \infty$ and $\delta(\lambda_0 - T) = p < \infty$, then there is an $\varepsilon > 0$ such that $\lambda - T$ is in state I_3 whenever $0 < |\lambda - \lambda_0| < \varepsilon$.*

Proof. For simplicity, we let $A = T - \lambda_0$ and consider $\lambda - A$ for λ near 0. Since A^p is a closed operator (problem 8, § 8), the quotient space $\tilde{X} = X/\mathscr{N}(A^p)$ is a Banach space. Let B be the one-to-one (closed) linear operator from $\tilde{X}$ into X induced by A^p. Then $\mathscr{R}(A^p) = \mathscr{R}(B) = \mathscr{D}(B^{-1})$. Since B^{-1} is also a closed operator, $\mathscr{R}(A^p)$ becomes a Banach space under the "graph norm" for $\mathscr{D}(B^{-1})$ given by

$$|||x||| = \|x\| + \|B^{-1}x\|, \qquad x \in \mathscr{R}(A^p).$$

Let $X_0 = \mathscr{R}(A^p)$, and let $A_0 : X_0 \to X_0$ be the restriction of A to X_0. Endow X_0 with the norm just described. It is readily checked that A_0 is a closed operator on the Banach space X_0. Now $\delta(A) = p$, which implies that $A_0 X_0 = X_0$. Furthermore, $\alpha(A) = \infty$ and so $\mathcal{N}(A) \cap \mathscr{R}(A^p) \neq (0)$, by Theorem 6.3. Hence A_0 is not one-to-one and therefore is in state I_3. By problem 1 of § 4, there is an $\varepsilon > 0$ such that $\lambda - A_0$ is also in state I_3 for $|\lambda| < \varepsilon$.

Now fix λ such that $0 < |\lambda| < \varepsilon$. To prove that $\lambda - A$ is in state I_3 we use the fact that $\lambda - A_0$ is not one-to-one in $\mathscr{R}(A^p)$ and $(\lambda - A_0)\mathscr{R}(A^p) = \mathscr{R}(A^p)$. Certainly $\lambda - A$ is not one-to-one, since every $x \in \mathcal{N}(\lambda - A_0)$ also belongs to $\mathcal{N}(\lambda - A)$. Given $x \in X$, we may write $x = u + v$, where $u \in \mathscr{D}(A^p)$ and $v \in \mathscr{R}(A^p)$, by Lemma 6.1. Let $y_0 = \lambda^{-1}(\lambda - A)u$, $y_1 = \lambda^{-2}(\lambda - A)Au, \ldots, y_{p-1} = \lambda^{-p}(\lambda - A)A^{p-1}u$. It is easily seen that

$$u = y_0 + y_1 + \cdots + y_{p-1} + \lambda^{-p}A^p u,$$

and $y_0 + \cdots + y_{p-1} \in \mathscr{R}(\lambda - A)$. Now $\lambda^{-p}A^p u + v \in \mathscr{R}(A^p) = X_0 = \mathscr{R}(\lambda - A_0) \subset \mathscr{R}(\lambda - A)$. (Here we are thinking of X_0 as a subset of X, ignoring the other topology on X_0.) Hence

$$x = (y_0 + \cdots + y_{p-1}) + (\lambda^{-p}A^p u + v) \in \mathscr{R}(\lambda - A).$$

Thus $\lambda - A$ is in state I_3. □

Theorem 10.5. *Suppose that $\lambda_0 \in \sigma(T)$, and suppose that each neighborhood of λ_0 contains a point that is not an eigenvalue of T. Then λ_0 is a pole of R_λ if and only if $\lambda_0 - T$ has finite descent.*

Proof. Suppose $\delta(\lambda_0 - T) < \infty$. Then by Theorem 10.4, if $\alpha(\lambda_0 - T)$ were infinite, a whole neighborhood of λ_0 (including λ_0 itself) would be in the point spectrum. This is impossible, by hypothesis. Thus $\lambda_0 - T$ has finite ascent, and λ_0 is a pole of R_λ, by Theorem 10.2. Of course, if λ_0 is a pole, then $\delta(\lambda_0 - T) < \infty$, by Theorem 10.1. □

A point λ_0 will satisfy the conditions of Theorem 10.5 if $\delta(\lambda_0 - T) < \infty$ and if λ_0 is in the boundary of $P\sigma(T)$ or in the boundary of $\sigma(T)$.

Corollary 10.6. *If T is a quasinilpotent operator in $L(X)$ such that $\delta(T) = p < \infty$, then $T^p = 0$.*

Proof. Since $\sigma(T) = \{0\}$, it follows from Theorem 10.5 that 0 is a pole of R_λ. By Theorem 10.1, the order of the pole is p and the range of the corresponding projection B_1 is $\mathcal{N}(T^p)$. But then, since $\{0\}$ is all of $\sigma(T)$, $B_1 = I$ (as was noted just before the statement of Theorem 9.4). Therefore $X = \mathcal{N}(T^p)$, which means that $T^p = 0$. □

Applications to the Operational Calculus

The next theorem is concerned with conditions on $f \in \mathfrak{A}_\infty(T)$ such that $f(T)=0$. When this theorem is applied to the special case in which X is finite dimensional and $T \in L(X)$, it yields the conditions that determine the minimal polynomial associated with T, that is, the polynomial $p(\lambda)$ of lowest degree, with leading coefficient 1, such that $p(T)=0$.

Theorem 10.7. *Suppose $f \in \mathfrak{A}_\infty(T)$. Then $f(T)=0$ if and only if (a) $f(\lambda) \equiv 0$ on an open set containing all of $\sigma_e(T)$ with the possible exception of a finite set of poles of R_λ at the finite points $\lambda_1, \ldots, \lambda_p$, and (b) if these latter exceptional points exist and if the order of λ_i as a pole of R_λ is m_i, then f has a zero of order greater than or equal to m_i at λ_i.*

Proof. As a preliminary to the proof, let us suppose that λ_0 is an isolated point of $\sigma(T)$ and that there is no neighborhood of λ_0 in which $f(\lambda) \equiv 0$. There will then be some smallest $k \geq 0$ such that $f^{(k)}(\lambda_0) \neq 0$. Now choose an r suitable for (10-3), with the further property that f is analytic and $f(\lambda) \neq 0$ when $0 < |\lambda - \lambda_0| < r$. Define g_k as follows:

$$g_k(\lambda) = \begin{cases} (\lambda - \lambda_0)^{-k} f(\lambda) & \text{if } 0 < |\lambda - \lambda_0| < r \\ (\lambda - \lambda_0)^{-k-1} & \text{if } |\lambda - \lambda_0| > 2r. \end{cases}$$

Then $g_k \in \mathfrak{A}_\infty(T)$ and $[g_k(T)]^{-1}$ exists, by Theorem 8.8. Referring to (10-3), we see that $g_k(\lambda)(\lambda - \lambda_0)^k f_{-1}(\lambda) = f(\lambda) f_{-1}(\lambda)$. Hence $g_k(T)(T - \lambda_0)^k B_1 = f(T) B_1$. Using (10-7), we have

$$B_{k+1} = [g_k(T)]^{-1} f(T) B_1. \tag{10-12}$$

With these preliminaries established, let us now assume that $f(T)=0$. From the spectral mapping theorem we have $(0) = \sigma(f(T)) = f(\sigma_e(T))$. Thus $f(\lambda)=0$ at all points of $\sigma_e(T)$. In fact, $f(\lambda) \equiv 0$ in each component of $\Delta(f)$ that contains an accumulation point of $\sigma_e(T)$, since f is locally analytic on $\sigma_e(T)$. Also, if $\infty \in \sigma_e(T)$, then f is identically zero on a neighborhood of ∞. For, otherwise, there would exist an integer $k \geq 1$ such that $g(\lambda) = \lambda^k f(\lambda) \in \mathfrak{A}_\infty(T)$ and $g(\infty) \neq 0$. But $g(T) = T^k f(T) = 0$ and, therefore, $g(\infty)=0$, since $(0) = \sigma(g(T)) = g(\sigma_e(T))$. Thus any isolated zeros of f in $\sigma_e(T)$ must occur in the bounded components of $\Delta(f)$. Since $\sigma(T)$ is a closed subset of $\Delta(f)$, it follows that f vanishes identically on an open set containing $\sigma_e(T)$ except for at most a finite number of finite isolated points $\lambda_1, \ldots, \lambda_p$. Suppose $1 \leq i \leq p$. If k is the order of the zero of f at λ_i, then (10-12) shows that $B_{k+1}=0$. Hence λ_i is a pole of R_λ of order not greater than k.

Now, conversely, suppose that f satisfies the two conditions in the theorem. If there are no points λ_i as described, then $f(T)=0$ directly from the defining formula (8-10). Otherwise, (8-10) simplifies to the form

(10-13) $$f(T)=\sum_{j=1}^{p}\frac{1}{2\pi i}\int_{C_j} f(\lambda)R_\lambda\,d\lambda,$$

where $C_1,\ldots,C_p$ are suitably small nonoverlapping counterclockwise circles with centers at $\lambda_1,\ldots,\lambda_p$. By condition (*b*), $f(\lambda)R_\lambda$ has a removable singularity at each λ_j; hence each integral in (10-13) is zero, by Cauchy's theorem, and $f(T)=0$. ☐

We conclude this section with a theorem that extends the spectral theory for compact operators to an important class of closed linear operators (with nonempty resolvent sets).

Theorem 10.8. *Suppose that $f\in\mathfrak{A}_\infty(T)$ and $f(T)$ is compact. If $\lambda_0\in\sigma(T)$ and $f(\lambda_0)\neq 0$, then λ_0 is a pole of R_λ of finite rank.*

Proof. Let $S=f(T)$, $\mu=f(\lambda_0)\neq 0$. We know that $\mu\in\sigma(S)$, by the spectral mapping theorem. Since S is a compact operator, $\{\mu\}$ is a spectral set; if E is the corresponding projection, then $\mathscr{R}(E)=\mathscr{N}[(\mu-S)^m]$, for some m, by Corollary 10.3 and Theorem 10.1. Furthermore, this subspace is finite dimensional. Now let $\sigma=\sigma_e(T)\cap f^{-1}(\mu)$. Then σ is a spectral set of T and the projection associated with σ and T is just E, by Theorem 9.8. Let $X_\sigma=\mathscr{R}(E)$, and let T_σ be the restriction of T to X_σ. Then $p=\alpha(\lambda_0-T_\sigma)\leqslant\dim X_\sigma<\infty$ since, if $\mathscr{N}[(\lambda_0-T_\sigma)^k]$ (for some k) is a proper subspace of $\mathscr{N}[(\lambda_0-T_\sigma)^{k+1}]$, then the second subspace has a larger dimension than the first, and this cannot happen for infinitely many k. A similar argument shows that $\delta(\lambda_0-T_\sigma)\leqslant\dim X_\sigma$. Using problem 2 and Theorem 10.2, we conclude that $\alpha(\lambda_0-T)=\delta(\lambda_0-T)=p$ and λ_0 is a pole of R_λ. The range of the corresponding spectral projection B_1 is $\mathscr{N}[(\lambda_0-T)^p]$. By problem 2, this space is $\mathscr{N}[(\lambda_0-T_\sigma)^p]$, which is obviously finite dimensional since it is contained in X_σ. ☐

One important application of Theorem 10.8 is to the case in which $T\in L(X)$ and T^n is compact for some positive integer n. Clearly we can choose $f\in\mathfrak{A}_\infty(T)$ so that $f(\lambda)=\lambda^n$ on a neighborhood of $\sigma(T)$, and then $f(T)=T^n$. *Hence if T^n is compact for some $n\geqslant 1$, every nonzero point in $\sigma(T)$ is a pole of R_λ of finite rank.* In such a case, the decomposition (10-10) holds for each $\lambda_0\neq 0$, and $\mathscr{N}(\lambda_0-T)$ is finite dimensional. From (10-10) it follows that λ_0-T is one-to-one if and only if $\mathscr{R}(\lambda_0-T)=X$. That is, we have the Fredholm alternative for each nonzero λ. See the discussion of (7-11).

There are a number of important Banach spaces where every "weakly compact" operator T has the property that T^2 is compact, even though T itself may not be compact. These operators are discussed in Dunford and Schwartz [1, pages 489–511].

Example. Take $X = L(a, b)$, where (a, b) is a finite or infinite interval. Let $k(s, t)$ be a measurable function on $(a, b)\times(a, b)$ such that $|k(s, t)| \leq h(s)$ almost everywhere, where $h \in X$. Define $Kx = y$ to mean $y(s) = \int_a^b k(s, t)x(t)\, dt$. Then $K \in L(X)$; K need not be compact, but K^2 is compact. See Zaanen [1, pages 322–323]. Also Dunford and Pettis [1, page 370] and Phillips [1].

Another important application of Theorem 10.8 is to the case of an operator T such that R_λ is compact. In this case $\mathscr{D}(T)$ cannot be all of X unless X is finite dimensional, because the identity operator would be compact if T were in $L(X)$. We note that as a result of the resolvent equation (2-3), if R_λ is compact for one λ in $\sigma(T)$, it is compact for every such λ. Suppose $\alpha \in \rho(T)$ and that R_α is compact. Now $R_\alpha = f(T)$, where $f(\lambda) = (\alpha - \lambda)^{-1}$. *Theorem 10.8 then shows that in this case every point of $\sigma(T)$ is a pole of R_λ of finite rank.* This situation prevails in the case of many ordinary differential operators, that of Example 2, § 8, for instance. See Goldberg [2, pages 140–153].

PROBLEMS

1. Show that the coefficients in (10-1) satisfy the relations $A_m B_n = 0$, $A_n = (-1)^n A_0^{n+1}$, $B_{n+1} = B_2^n$, and B_2 is quasinilpotent. Show also that $\mathscr{R}(A_n) \subset \mathscr{D}(T^{n+1})$ if $n \geq 0$ and $\mathscr{R}(B_n) \subset \mathscr{D}(T^k)$ if $n \geq 1$ and $k \geq 0$.
2. Let σ be a spectral set of T, and let T_σ be the restriction of T to the invariant subspace X_σ associated with σ (see Theorem 9.2).
   a. If $\lambda \in \sigma$, then $\mathscr{N}[(\lambda - T_\sigma)^k] = \mathscr{N}[(\lambda - T)^k]$ for $k \geq 1$. Hence $\lambda - T_\sigma$ (acting in the space X_σ) and $\lambda - T$ have the same nullity and the same ascent.
   b. If $\lambda \in \sigma$, then $\lambda - T_\sigma$ and $\lambda - T$ have the same defect and the same descent. Also, $\mathscr{R}(\lambda - T_\sigma)$ is closed in X_σ if and only if $\mathscr{R}(\lambda - T)$ is closed in X.
   c. A point $\lambda_0 \in \sigma$ is a pole of $(\lambda - T)^{-1}$ if and only if λ_0 is a pole of $(\lambda - T_\sigma)^{-1}$.
3. If σ is a spectral set of T such that X_σ is finite dimensional, then σ is a finite set of poles of R_λ.
4. Suppose that $T \in L(X)$. A point λ_0 in $\sigma(T)$ is a pole of R_λ of finite rank if and only if there is a compact operator K such that $KT = TK$ and $\lambda_0 \in \rho(T + K)$. [*Hint.* Use the fact that $(\lambda_0 - T - K)^{-1}(\lambda_0 - T) = I + (\lambda_0 - T - K)^{-1}K$ when $\lambda_0 \in \rho(T + K)$.]
5. Suppose $\lambda = 0$ is a pole of R_λ, and let B_1 be the associated projection. Suppose that there exist positive constants M and ε such that $\|\lambda R_\lambda\| \leq M$ whenever $0 < |\lambda| < \varepsilon$. Then λR_λ converges to B_1 in $L(X)$ as $\lambda \to 0$.

6. If $\mathscr{D}(T) \neq X$ there exists no polynomial $p(\lambda) \not\equiv 0$ such that $p(T)x = 0$ for every $x \in \mathscr{D}(T^n)$ (n the degree of p). If such a polynomial p did exist, choose $\alpha \in \rho(T)$, $f(\lambda) = (\alpha - \lambda)^{-n-1} p(\lambda)$, and show that $f(T) = 0$. Then use Theorem 10.7.
7. Suppose that $\sigma(T)$ is bounded and that $\mathscr{D}(T) \neq X$. Then R_λ has an expansion in powers of λ valid when $|\lambda| > r_\sigma(T)$, say $R_\lambda = \Sigma_1^\infty \lambda^{-n} A_n + \Sigma_0^\infty \lambda^n B_n$. Show that $TB_0 + I = A_1$, $TB_n = B_{n-1}$, and $TA_n = A_{n+1}$. Show that $B_n \neq 0$ for each n and that A_1 is the projection associated with the spectral set $\sigma(T)$.
8. Take $X = \ell^1$, and define $A \in L(X)$ by the matrix (α_{ij}), where $\alpha_{ii} = \beta_i$, $\alpha_{12} = \alpha_{34} = \alpha_{56} = \cdots = 1$, and all other entries in the matrix are 0. Suppose $\beta_i \to 0$. Then A is not compact, but A^2 is compact. Discuss $\sigma(A)$ and the resolvent.

V.11 OPERATORS WITH A RATIONAL RESOLVENT

Let X be a Banach space, and let T be an operator in $L(X)$. We shall consider under what circumstances R_λ, the resolvent of T, may be represented in the form

$$R_\lambda = \frac{P(\lambda)}{q(\lambda)}, \tag{11-1}$$

where P is a polynomial in λ with coefficients in $L(X)$, q is a polynomial in λ with coefficients in C, and where, moreover, $P(\lambda)$ and $q(\lambda)$ are never both 0 for the same value of λ. When such a representation is possible, we say that T has a *rational resolvent.* The definition of a rational resolvent carries with it the implication that the resolvent set of T includes all finite points except those for which $q(\lambda) = 0$.

When $T \in L(X)$ and R_λ is rational, the degree of q must be at least 1, for $\sigma(T)$ is nonempty. Hence $\sigma(T)$ consists of poles of R_λ, with the order of each pole equal to its order as a zero of q. Suppose, on the other hand, we know only that $T \in L(X)$ and $\sigma(T)$ consists of a finite number of poles of R_λ. Then we may use Liouville's theorem and the fact that $\|R_\lambda\| \to 0$ as $\lambda \to \infty$ (see the proof of Theorem 3.2) to represent R_λ as the sum of the singular parts of its Laurent expansions at the various poles. From this we obtain a rational representation as in (11-1).

We need not consider the case when T is closed and $\mathscr{D}(T) \neq X$, for R_λ cannot be rational in this case. See problem 1.

An important class of operators with rational resolvents is described in the following theorem.

Theorem 11.1 *Suppose $T \in L(X)$ and $\mathscr{R}(T)$ is finite dimensional. Then R_λ is rational.*

Proof. Since T is compact, the nonzero points of $\sigma(T)$ are all poles of R_λ (Corollary 10.3), with zero the only possible accumulation point. Furthermore, the descent of T must be finite, because $\dim \mathcal{R}(T) < \infty$. Hence 0 is a pole of R_λ, by Theorem 10.5, and therefore cannot be an accumulation point of $\sigma(T)$. We conclude that because $\sigma(T)$ is compact, it must be finite. Hence R_λ is rational. □

In the proof above, one can avoid Theorem 10.5 and use Theorem 10.2 instead. Simply note that since $X/\mathcal{N}(T)$ is isomorphic to $\mathcal{R}(T)$ and hence is finite dimensional, the ascent (as well as the descent) of T must be finite.

The next theorem gives a useful characterization of an operator with a rational resolvent. Other characterizations are given in the problems.

Theorem 11.2. *An operator T in $L(X)$ has a rational resolvent if and only if there exists a nonzero scalar polynomial $p(\lambda)$ such that $p(T)=0$.*

Proof. Suppose $p(T)=0$ for some nonzero polynomial $p(\lambda)$. Since $p(\lambda)$ is not identically zero (and cannot be constant), Theorem 10.7 implies that $\sigma(T)$ is contained in the finite set of zeros of $p(\lambda)$ and each point in $\sigma(T)$ is a pole of R_λ. Hence R_λ is rational. Conversely, if R_λ is rational, it has at least one pole, because $\sigma(T)$ is not empty. Let $\lambda_1, \ldots, \lambda_k$ be its distinct poles, of orders $m_1, \ldots, m_k$, and let

$$p(\lambda) = (\lambda - \lambda_1)^{m_1} \cdots (\lambda - \lambda_k)^{m_k}. \tag{11-2}$$

By Theorem 10.7, $p(T)=0$. □

The polynomial $p(\lambda)$ in (11-2) is called the *minimal polynomial* associated with T. It follows from Theorem 10.7 that $p(\lambda)$ divides any other polynomial $p_1(\lambda)$ such that $p_1(T)=0$. We recall from Theorem 10.1 that m_i is the ascent of $\lambda_i - T$.

Suppose that T has a rational resolvent, suppose the minimal polynomial has degree n, and let M be the subspace of $L(X)$ spanned by I, T, $T^2, \ldots, T^{n-1}$. It is not difficult to show that every operator of the form $f(T)$ belongs to M, where $f \in \mathfrak{A}(T)$. (See problems 5 and 6.) In fact, A. F. Ruston has proved that if S is an operator that commutes with all elements of $L(X)$ commuting with T, then $S \in M$. Conversely, if $T \in L(X)$ and if every operator that commutes with all elements of $L(X)$ commuting with T is a polynomial in T, then T must have a rational resolvent. See Ruston [1, 2].

We now turn to a structure theorem for an operator with a rational resolvent. The result is an easy consequence of our earlier work. An elementary algebraic proof of the theorem is also possible. See problem 8.

Theorem 11.3. *Suppose R_λ is rational, and let $p(\lambda)=(\lambda-\lambda_1)^{m_1}\cdots(\lambda-\lambda_k)^{m_k}$ be the minimal polynomial of T. Then*

$$X=\mathcal{N}[(\lambda_1-T)^{m_1}]\oplus\cdots\oplus\mathcal{N}[(\lambda_k-T)^{m_k}],$$

and T is completely reduced by this direct sum decomposition.

Proof. If E_i denotes the spectral projection associated with the spectral set $\{\lambda_i\}$, $1\leq i\leq k$, then $\mathcal{R}(E_i)=\mathcal{N}[(\lambda_i-T)^{m_i}]$, by Theorem 10.1. The theorem now follows from Theorem 9.1, since $\sigma(T)=\{\lambda_1\}\cup\cdots\cup\{\lambda_k\}$. □

Let T be as in Theorem 11.3, let $E_1,\ldots,E_k$ be the associated spectral projections, and let

$$N=\sum_{i=1}^{k}(T-\lambda_i)E_i.$$

If $p=\max\{m_1,\ldots,m_k\}$, it is easily seen that

$$N^p=\sum_{i=1}^{k}(T-\lambda_i)^pE_i=0,$$

so that N is nilpotent. (Observe that $N=0$ if $m_1=\cdots=m_k=1$.) We can write

$$T=\sum_{i=1}^{k}\lambda_iE_i+N. \tag{11-3}$$

This representation of T is possible because R_λ is rational. The converse is given in the following theorem.

Theorem 11.4. *Let $E_1,\ldots,E_k$ be elements of $L(X)$ that are projections such that $E_iE_j=0$ if $i\neq j$ and such that $I=E_1+\cdots+E_k$. Let T be defined by (11-3), where $\lambda_1,\ldots,\lambda_k$ are complex constants, and N is a nilpotent member of $L(X)$ such that $NE_i=E_iN$ for each i. Then R_λ is rational.*

Proof. Suppose $N^p=0$. Now, $(T-\lambda_i)E_i=NE_i$, as we see from (11-3). Hence $(T-\lambda_i)^pE_i=N^pE_i=0$. It now follows from $I=E_1+\cdots+E_k$ that $(T-\lambda_1)^p\cdots(T-\lambda_k)^p=0$. Therefore, by Theorem 11.2, R_λ is rational. □

PROBLEMS

1. Let T be a closed linear operator with domain and range in a Banach space X and $\mathcal{D}(T)\neq X$. Use problem 7, § 10, to show that R_λ cannot be rational.
2. An operator T in $L(X)$ has a rational resolvent if and only if there exists a nonzero polynomial $q(\lambda)$ such that $q(T)$ has finite-dimensional range.
3. An operator T in $L(X)$ has a rational resolvent if and only if the operator $\lambda-T$ has finite descent for each $\lambda\in C$.

4. Suppose $T = F + N$, where $F, N \in L(X)$, $\dim \mathscr{R}(F) < \infty$, N is nilpotent, and $FN = NF = 0$. Then T has a rational resolvent.

5. a. Suppose that T has a rational resolvent and the minimal polynomial associated with T has degree n. Let M be the subspace of $L(X)$ generated by $I, T, T^2, \dots, T^{n-1}$. Then M is n-dimensional and $f(T)$ is in M for every $f \in \mathfrak{A}(T)$.
   b. An operator S in $L(X)$ has a rational resolvent if and only if for each $f \in \mathfrak{A}(S)$ the operator $f(S)$ is actually a polynomial in S.

6. Suppose that R_λ is rational. Using the notation of Theorem 11.3 and the paragraph that follows it, show that, if f is locally analytic on $\sigma(T)$,

$$f(T) = \sum_{i=1}^{k} \sum_{j=0}^{m_i - 1} \frac{f^{(j)}(\lambda_i)}{j!} (T - \lambda_i)^j E_i.$$

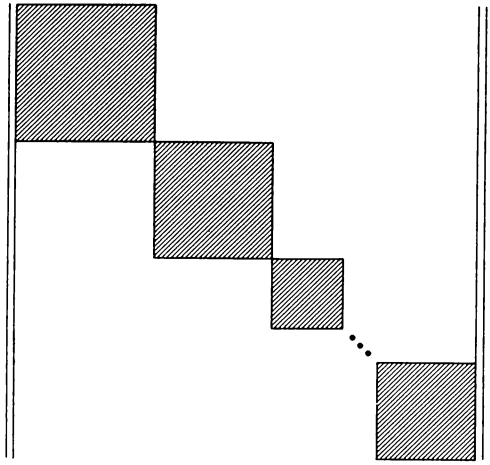
In particular, suppose m is fixed, $1 \leqslant m \leqslant k$, and let $f(\lambda)$ be a polynomial such that $f^{(j)}(\lambda_i) = 1$ if $i = m$, $j = 0$, while $f^{(j)}(\lambda_i) = 0$ otherwise as it occurs in the formula for $f(T)$. Then $f(T) = E_m$. Since a polynomial of this kind exists (of degree at most $m_1 + \cdots + m_k - 1$), each E_m is expressible as a polynomial in T. See Hamburger and Grimshaw [1, page 111].

7. Suppose $T \in L(X)$. If $p(T)$ has a rational resolvent for some nonconstant polynomial p, then R_λ is rational. Conversely, if R_λ is rational, then $p(T)$ has a rational resolvent for every polynomial p.

8. a. Let X be any complex linear space, and let T be a linear operator on X into X. Let $p_1(\lambda)$ and $p_2(\lambda)$ be scalar polynomials without common zeros. Let $p(\lambda) = p_1(\lambda)p_2(\lambda)$, and let M_1, M_2, M be the null spaces of $p_1(T), p_2(T), p(T)$, respectively. Then $M = M_1 \oplus M_2$. [*Hint.* Use the fact that since p_1 and p_2 are relatively prime, there exist polynomials q_1 and q_2 such that $q_1(\lambda)p_1(\lambda) + q_2(\lambda)p_2(\lambda) \equiv 1$.]
   b. Let $p(\lambda) = (\lambda - \lambda_1)^{m_1} \cdots (\lambda - \lambda_k)^{m_k}$, where the λ_i are distinct and the m_i are positive integers, and let $M_i = \mathscr{N}[(\lambda_i - T)^{m_i}]$, $1 \leqslant i \leqslant k$. If $p(T) = 0$, then $X = M_1 \oplus \cdots \oplus M_k$.

9. If X is finite dimensional, formula (11-3) plus certain well-known facts about nilpotent operators provide the basis for obtaining the *Jordan normal form* matrix representation of the operator T. Choose a basis $u_1, \dots, u_n$ for X in such a way that successive sets of u's form bases for $M_1, \dots, M_k$. Then the matrix representation of T takes the form shown in the diagram on page 340, where the only nonzero elements are inside the shaded blocks. The ith block, counting down the diagonal, is the matrix representation of the restriction of T to M_i. Owing to the nilpotency of N, whose restriction to M_i is the same as that of $T - \lambda_i$, the basis for M_i may be chosen in such a way that the matrix representation of the restriction of N to M_i has no nonzero elements except for the possibility of 1's in a certain arrangement along the diagonal directly below the main diagonal. There will be some such 1's if $m_i > 1$, but not if $m_i = 1$. Thus the matrix for T has

$$\underbrace{\lambda_1, \dots, \lambda_1}_{\nu_1}; \quad \underbrace{\lambda_2, \dots, \lambda_2}_{\nu_2}; \quad \dots; \quad \underbrace{\lambda_k, \dots, \lambda_k}_{\nu_k}$$



down the main diagonal, where ν_i is the dimension of M_i. The only other nonzero elements are the 1's already referred to, on the diagonal just below the main diagonal. For more details on the matrix representation of nilpotent operators see Halmos [1, pages 109–112].

VI SPECTRAL ANALYSIS IN HILBERT SPACE

In a Hilbert space, the spectral theory for certain classes of operators (self-adjoint, unitary, and normal) is more complete and definitive than that for operators in a Banach space. This richer theory is possible because the operators involved have special relationships to their adjoints. The theory of self-adjoint operators in Hilbert space has its origins in the algebraic study of symmetric and Hermitian matrices and the reduction of such matrices to diagonal form by orthogonal transformations. The matrix theory has applications in geometry (such as finding the principal axes of symmetry of figures defined by $Q(x_1, \ldots, x_n) = 1$, where Q is a symmetric real quadratic form) and to problems in mathematical physics (such as finding the normal modes of certain vibrating systems). The theory for infinite dimensions was launched by Hilbert in his study of integral equations in the first decade of the twentieth century, and this theory turned out to be important in quantum mechanics, differential equations, and a host of other areas.

The first section of this chapter contains important facts about bilinear forms and includes the Lax–Milgram theorem (Theorem 1.4). Elementary properties of symmetric, self-adjoint, and normal operators are explored in the next two sections. Then, in § 4, we obtain a spectral decomposition for compact symmetric operators in an inner-product space that need not be complete (Theorem 4.2). This material has applications to the Hilbert–Schmidt theory of integral equations with symmetric kernel and to symmetric differential operators with compact resolvent, in particular, to the classical Sturm–Liouville differential equation problems. The decomposition of a compact symmetric operator is used to motivate the spectral theorem for a bounded self-adjoint operator on a Hilbert space (Theorem 6.1). From this we derive an operational calculus (Theorem 6.3), with the contour integral of § V.8 replaced by a Riemann–Stieltjes integral over an interval on the real axis.

The spectral theorem for unitary operators is proved in § 7. After introducing Cayley transforms in § 8, we sketch how a spectral theorem for unbounded self-adjoint operators may be obtained from the theorem for unitary operators. The case of a bounded normal operator is postponed until the end of the next chapter, where the spectral theorem is deduced from the theory of B^*-algebras.

VI.1 BILINEAR AND QUADRATIC FORMS

Throughout this section, X denotes an inner-product space; completeness is assumed as needed. The scalar field may be either real or complex, except when we specify one or the other explicitly. If no such specification is made, a bar indicating the complex conjugate of a number is to be ignored if the scalars are real.

Suppose A is a linear operator with domain X and range in X. Then the inner product (Ax, y) is a linear functional of x for each fixed y. As a function of y, (Ax, y) is linear when the scalar field is real; in the complex case, however, we do not quite have linearity with respect to y, for $(Ax, \alpha y) = \bar{\alpha}(Ax, y)$, the scalar factor coming outside as $\bar{\alpha}$ rather than as α. However, for convenience in embracing both the real and complex cases in one terminology, we shall say that (Ax, y) is *bilinear* in x and y.

For some purposes it is desirable to study bilinearity directly, instead of through the medium of a linear operator. Hence we make the following definition: A scalar-valued function ϕ on $X \times X$ is called a *bilinear form* if $\phi(x, y)$ is linear in x for each y, while $\overline{\phi(x, y)}$ is linear in y for each x. With ϕ we associate the functional ψ on X defined by $\psi(x) = \phi(x, x)$. We call ψ the *quadratic form* corresponding to ϕ. Observe that $\psi(\alpha x) = |\alpha|^2 \psi(x)$.

One important relation between ϕ and ψ is expressed by the formula

$$\tag{1-1} \tfrac{1}{2}\{\phi(x, y) + \phi(y, x)\} = \psi\left(\frac{x+y}{2}\right) - \psi\left(\frac{x-y}{2}\right).$$

It may be verified by expanding the terms on the right, using the bilinearity. When the scalar field is complex we have

$$\tag{1-2} \phi(x, y) = \psi\left(\frac{x+y}{2}\right) - \psi\left(\frac{x-y}{2}\right) + i\psi\left(\frac{x+iy}{2}\right) - i\psi\left(\frac{x-iy}{2}\right).$$

In the complex case, therefore, ϕ can be expressed entirely in terms of ψ. This is not always true in the real case, however, for in this case the left member of (1-1) is a bilinear form, in general not the same as ϕ, whose corresponding quadratic form is the same as the one corresponding to ϕ.

A bilinear form ϕ is continuous jointly in its two variables if and only if $|\phi(x, y)|$ is bounded for all x and y such that $\|x\| \leq 1$, $\|y\| \leq 1$ (proof similar to that of Theorem II.1.1). When ϕ is continuous we define

$$\|\phi\| = \sup \frac{|\phi(x, y)|}{\|x\|\,\|y\|} \qquad \text{for } x \neq 0,\ y \neq 0,$$

or, equivalently, we may take the supremum merely for all x and y with $\|x\|=\|y\|=1$. To avoid trivialities we are assuming that the space X does not reduce to (0). For the corresponding quadratic form we define

$$\|\psi\|=\sup_{\|x\|=1}|\psi(x)|=\sup_{x\neq 0}\frac{|\psi(x)|}{\|x\|^2}.$$

Obviously $\|\psi\|\leq\|\phi\|$. When the field of scalars is complex it is easily seen with the aid of (1-2) and the parallelogram law (page 75) that $\|\phi\|\leq 2\|\psi\|$. Hence

$$\|\psi\|\leq\|\phi\|\leq 2\|\psi\| \tag{1-3}$$

in the complex case.

A bilinear form ϕ is called *symmetric* if $\phi(x, y)=\overline{\phi(y, x)}$ (in the complex case such a form is also called *Hermitian*). If ϕ is symmetric we can prove that

$$\|\phi\|=\|\psi\|. \tag{1-4}$$

We have only to prove that $\|\phi\|\leq\|\psi\|$. Owing to the symmetry, we have from (1-1) and the parallelogram law:

$$|\mathrm{Re}\,\phi(x, y)|\leq\tfrac{1}{2}\|\psi\|(\|x\|^2+\|y\|^2)$$

(with the Re symbol, for the "real part," superfluous in the real case). For fixed x and y with $\|x\|=\|y\|=1$ we can choose α so that $|\alpha|=1$ and $\alpha\phi(x, y)=|\phi(x, y)|$, whence

$$|\phi(x, y)|=\phi(\alpha x, y)=|\mathrm{Re}\,\phi(\alpha x, y)|\leq\|\psi\|,$$

and so $\|\phi\|\leq\|\psi\|$.

Theorem 1.1. *Suppose $\phi(x, y)=(Ax, y)$, where A is linear on X into X. Then ϕ is continuous if and only if A is continuous, and then $\|A\|=\|\phi\|$.*

Proof. If A is continuous, $|(Ax, y)|\leq\|Ax\|\,\|y\|\leq\|A\|\,\|x\|\,\|y\|$. If ϕ is continuous, $\|Ax\|^2=(Ax, Ax)\leq\|\phi\|\,\|x\|\,\|Ax\|$, or $\|Ax\|\leq\|\phi\|\,\|x\|$. The conclusions now follow. □

Theorem 1.2. *Suppose that X is complete and that ϕ is a continuous bilinear form. Then there exists $A\in L(X)$ such that $\phi(x, y)=(Ax, y)$.*

Proof. For fixed x, $\overline{\phi(x, y)}$ is a continuous linear functional of y. Hence, by the Fréchet–Riesz theorem (Theorem III.5.1), we can represent the functional in the form $\overline{\phi(x, y)}=(y, Ax)$, where Ax is some vector depending on x. This defines an operator A on X; the linearity of A is easily verified, since $(Ax, y)=\phi(x, y)$. That A is continuous follows from Theorem 1.1. □

Theorem 1.3. *Let ϕ and A be as in Theorem 1.2. Suppose there exists $m > 0$ such that*

$$|\phi(x, x)| \geq m\|x\|^2 \quad \text{for all } x \in X. \tag{1-5}$$

Then A is a one-to-one mapping of X onto X, and $A^{-1} \in L(X)$.

Proof. For $x \in X$,

$$m\|x\|^2 \leq |\phi(x, x)| = |(Ax, x)| \leq \|Ax\| \, \|x\|.$$

Hence $m\|x\| \leq \|Ax\|$ for all x, which implies that A has a bounded inverse and $\mathscr{R}(A)$ is closed, by Theorems II.1.2 and IV.5.8. Now if x is orthogonal to $\mathscr{R}(A)$, then $0 = (Ax, x) \geq m\|x\|^2$, which shows that $x = 0$. Since $\mathscr{R}(A)$ is closed, it follows from Theorem II.7.4 that $\mathscr{R}(A) = X$. Hence $A^{-1} \in L(X)$. □

The Fréchet–Riesz representation theorem concerns the symmetric bilinear form given by the inner product on X. The following generalization, known as the Lax–Milgram theorem, applies to any continuous bilinear form satisfying (1-5). This result is often used to establish the existence of solutions of certain linear partial differential equations.

Theorem 1.4 (Lax–Milgram). *Suppose that X is complete and that ϕ is a continuous bilinear form satisfying* (1-5) *for some $m > 0$. Then given any continuous linear functional x' on X, there exists a unique $y \in X$ such that $x'(x) = \phi(x, y)$ for every $x \in X$.*

Proof. By the Fréchet–Riesz theorem, there exists $z \in X$ such that $x'(x) = (x, z)$ for $x \in X$. Now by Theorem 1.3, the operator A associated with ϕ has a continuous inverse. Hence the range of the adjoint A^* is X, by problem 4, § IV.11. So $z = A^*y$ for some $y \in X$. Then from the way A is related to ϕ, we have

$$x'(x) = (x, z) = (x, A^*y) = (Ax, y) = \phi(x, y).$$

Now suppose y_1 satisfies $x'(x) = \phi(x, y_1)$ for $x \in X$. Then $0 = \phi(x, y_1) - \phi(x, y) = \phi(x, y_1 - y)$. Taking $x = y_1 - y$ in (1-5), we conclude that $y_1 = y$, which shows that y is unique. □

PROBLEMS

1. A bilinear form ϕ on a complex inner-product space X is symmetric if and only if the quadratic form ψ corresponding to ϕ is real valued. [*Hint.* If ψ is real valued, let $\phi_1(x, y) = \overline{\phi(y, x)}$, and observe that ψ is also the quadratic form corresponding to ϕ_1. From (1-2) it follows that $\phi = \phi_1$.]
2. a. If A is a linear operator on a complex inner-product space X such that $(Ax, x) = 0$ for each $x \in X$, then $A = 0$. Is this true for real inner-product spaces?

b. If A and B are linear operators on a complex inner-product space X such that $(Ax, x) = (Bx, x)$ for $x \in X$, then $A = B$.

VI.2 SYMMETRIC OPERATORS

A linear operator A with domain and range in the inner-product space X is said to be *symmetric* if

$$(Ax, y) = (x, Ay) \tag{2-1}$$

for each x and y in $\mathscr{D}(A)$. In case X is a complex space, a symmetric linear operator is also called a *Hermitian* operator. If we regard $\mathscr{D}(A)$ by itself as an inner-product space, $\phi(x, y) = (Ax, y)$ is a bilinear form on $\mathscr{D}(A) \times \mathscr{D}(A)$, and ϕ is symmetric (as defined in § 1) if and only if A is symmetric. The corresponding quadratic form, defined on $\mathscr{D}(A)$, is (Ax, x).

We assume that A is symmetric, with $\mathscr{D}(A) \neq (0)$. Then (Ax, x) is real, even when X is a complex space, and we define

$$m(A) = \inf_{\|x\|=1} (Ax, x), \qquad M(A) = \sup_{\|x\|=1} (Ax, x). \tag{2-2}$$

The possibilities $m(A) = -\infty$, $M(A) = +\infty$ are not excluded.

Theorem 2.1. *If A is symmetric and λ is an eigenvalue of A, then λ is real, and $m(A) \leqslant \lambda \leqslant M(A)$. Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

Proof. Suppose $\|x\| = 1$ and $Ax = \lambda x$. Then $(Ax, x) = (\lambda x, x) = \lambda$, so λ is real and $m(A) \leqslant \lambda \leqslant M(A)$. If $Ax = \lambda x$ and $Ay = \mu y$, where $\lambda \neq \mu$, we have $\lambda(x, y) = (\lambda x, y) = (Ax, y) = (x, Ay) = (x, \mu y) = \mu(x, y)$, or $(\lambda - \mu)(x, y) = 0$, whence $(x, y) = 0$. □

Next, we consider the question as to whether the bounds $m(A)$ and $M(A)$ can be attained by values of (Ax, x) when $\|x\| = 1$. If $\lambda = m(A)$ happens to be an eigenvalue of A, there is an x with $\|x\| = 1$ and $Ax = \lambda x$. Then $(Ax, x) = (\lambda x, x) = m(A)$. A similar thing is true if $M(A)$ is an eigenvalue. There is a valid converse if $\mathscr{D}(A)$ is dense in X.

Theorem 2.2. *Suppose that A is symmetric and that $\mathscr{D}(A)$ is dense in X. Suppose $y \in \mathscr{D}(A)$, $\|y\| = 1$, and $(Ay, y) = \lambda$, where λ is either $m(A)$ or $M(A)$. Then $Ay = \lambda y$, so that λ is an eigenvalue of A.*

Proof. Suppose for example that $\lambda = m(A)$. Let $B = A - \lambda$. Then B is symmetric, with $(By, y) = 0$ and $m(B) = 0$, so that $(Bx, x) \geqslant 0$ for every x in $\mathscr{D}(A)$. Let $\alpha = \theta(By, x)$, where x is arbitrary in $\mathscr{D}(A)$ and θ is an arbitrary real number. Then

$$0 \leqslant (B[y+\alpha x], y+\alpha x) = (By, y) + \alpha(Bx, y) + \bar{\alpha}(By, x) + |\alpha|^2(Bx, x),$$

or

$$0 \leqslant \theta |(x, By)|^2 \{2 + \theta(Bx, x)\}.$$

Since the right member of the inequality changes sign with θ when θ is small, unless $(x, By) = 0$, we conclude that $(x, By) = 0$. This is true for each $x \in \mathscr{D}(A)$ and, since $\mathscr{D}(A)$ is dense in X, we conclude that $By = 0$ or $Ay = \lambda y$. If $\lambda = M(A)$, we put $B = \lambda - A$, and the same argument applies. □

When X is complete and A is symmetric, with $\mathscr{D}(A) = X$, we can prove (see § 3) that A is continuous and that $\sigma(A)$ lies on the interval $m(A) \leqslant \lambda \leqslant M(A)$ of the real axis, with $m(A)$ and $M(A)$ actually belonging to $\sigma(A)$.

In the next theorem we consider the situation in which $\mathscr{D}(A) = X$, though X need not be complete.

Theorem 2.3. *If A is symmetric, with $\mathscr{D}(A) = X$, then A is continuous if and only if $m(A)$ and $M(A)$ are both finite, and in that case*

$$\|A\| = \sup_{\|x\|=1} |(Ax, x)| = \max\{|m(A)|, |M(A)|\}. \tag{2-3}$$

Proof. We know from Theorem 1.1 that A is continuous if and only if $|(Ax, y)|$ is bounded for all x and y for which $\|x\| = \|y\| = 1$; and, as we see in conjunction with (1-4), this is the same as demanding that $\sup_{\|x\|=1} |(Ax, x)|$ be finite, this supremum then being equal to $\|A\|$. Formula (2-3) now follows from the definitions (2-2). □

Example 1. Let X be the Hilbert space $L^2(0, 2\pi)$. Let $\mathscr{D}(A)$ be the set of those x such that $x(s)$ is absolutely continuous on $[0, 2\pi]$, such that $x(0) = x(2\pi)$ and such that the derivative x' is in X. For $x \in \mathscr{D}(A)$ define $Ax = y$ to mean $y(t) = -ix'(t)$. Then A is symmetric. For, if $x, y \in \mathscr{D}(A)$,

$$(x, Ay) - (Ax, y) = i\int_0^{2\pi} [x(t)\overline{y'(t)} + x'(t)\overline{y(t)}]\, dt$$

$$= ix(t)\overline{y(t)}\Big|_0^{2\pi} = 0.$$

Example 2. An important type of symmetric operator arises in connection with certain boundary-value problems for second-order ordinary differential equations. The symmetry of the operator in such problems depends both on the nature of the boundary conditions and on certain formal properties of the differential operator that occurs. The space X is taken to be either $L^2(a, b)$ or some subspace of $L^2(a, b)$, with the usual inner product of $L^2(a, b)$. When the domain $\mathscr{D}(A)$ is suitably defined, $Ax = u$ is expressed by

$$u(t) = -\frac{d}{dt}[p(t)x'(t)] + q(t)x(t)$$

where p and q are certain *real-valued* functions. This formula exhibits the essential formal structure of the differential operator. There must be conditions on p, q, and $\mathscr{D}(A)$ which ensure that Ax is a well-defined member of X when $x \in \mathscr{D}(A)$. For instance, if (a, b) is a finite interval and if X consists of those elements of $L^2(a, b)$ that correspond to functions continuous on $[a, b]$, we might require q to be continuous and p to be continuously differentiable. We could then take $\mathscr{D}(A)$ to consist of those functions that are twice continuously differentiable and satisfy certain boundary conditions. If X is taken to be all of $L^2(a, b)$, we might require that q be measurable and bounded and that p be absolutely continuous. Then $\mathscr{D}(A)$ could be taken to consist of those elements of $L^2(a, b)$ corresponding to functions $x(t)$ such that both $x(t)$ and $p(t)x'(t)$ are absolutely continuous, $p(t)x'(t)$ has a derivative in $L^2(a, b)$, and $x(t)$ satisfies certain boundary conditions.

If $Ax = u$ and $Ay = v$, it is easily verified that we have

$$u(t)\overline{y(t)} - x(t)\overline{v(t)} = \frac{d}{dt}\{p(t)[x(t)\overline{y'(t)} - x'(t)\overline{y(t)}]\}.$$

Thus

$$(Ax, y) - (x, Ay) = p(t)[x(t)\overline{y'(t)} - x'(t)\overline{y(t)}]\Big|_a^b.$$

The operator A is then seen to be symmetric, provided that the definition of $\mathscr{D}(A)$ ensures that

$$p(b)[x(b)\overline{y'(b)} - x'(b)\overline{y(b)}] = p(a)[x(a)\overline{y'(a)} - x'(a)\overline{y(a)}]. \tag{2-4}$$

In case the interval is infinite, this condition must be interpreted appropriately. For a finite interval the boundary conditions are of the form

$$\alpha_{11}x(a) + \alpha_{12}x'(a) + \beta_{11}x(b) + \beta_{12}x'(b) = 0$$

$$\alpha_{21}x(a) + \alpha_{22}x'(a) + \beta_{21}x(b) + \beta_{22}x'(b) = 0$$

where the α's and β's are certain real scalars. For these boundary conditions it can be shown that condition (2-4) is equivalent to the condition

$$p(a)\begin{vmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{vmatrix} = p(b)\begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{vmatrix}.$$

It is not our intent in this book to discuss in detail the spectral analysis of a symmetric operator A of the type just described. However, in § 5, as an application of the theory of compact symmetric operators, we do discuss the solution of boundary-value problems of this type in the special case in which A^{-1} is a compact operator belonging to $L(X)$.

We return now to the general consideration of symmetric operators defined on X. For such operators it is useful to introduce a partial-order relation. If A and B are symmetric and $(Ax, x) \leqslant (Bx, x)$ for $x \in X$; we write $A \leqslant B$. If $A \leqslant B$, then $\alpha A \leqslant \alpha B$ when $\alpha > 0$, and $A + C \leqslant B + C$ for every symmetric C. If $A \leqslant B$ and $B \leqslant A$, then one can show that $A = B$ (see problem 2). A symmetric operator A such that $0 \leqslant A$ (where 0 denotes the zero operator) is said to be *positive*.

Suppose that A is positive and write $\{x, y\}$ for (Ax, y). Then $\{x, y\}$ as a function on $X \times X$ has all the properties of an inner product except possibly the property that $\{x, x\} = 0$ implies $x = 0$. However, this last property is not needed in the derivation of the Cauchy–Schwarz inequality $|\{x, y\}|^2 \leqslant \{x, x\}\{y, y\}$ (see problem 3), and so we obtain the inequality

$$|(Ax, y)|^2 \leqslant (Ax, x)(Ay, y), \tag{2-5}$$

whenever A is a positive operator. This result is used in the proof of Theorem 3.2.

An important example of a positive operator is an orthogonal projection. Recall from Theorem IV.12.7 that a projection P is orthogonal if and only if it is symmetric. In this case

$$(Px, x) = (PPx, x) = (Px, Px) = \|Px\|^2. \tag{2-6}$$

Thus $(Px, x) \geqslant 0$. Also, if $P \neq 0$, then $\|P\| = 1$ by Theorem IV.12.6, and so $(Px, x) \leqslant \|Px\|\,\|x\| \leqslant \|x\|^2 = (x, x) = (Ix, x)$. Thus each orthogonal (and symmetric) projection P satisfies $0 \leqslant P \leqslant I$.

The following theorem will be needed in § 6.

Theorem 2.4. *If P_1 and P_2 are symmetric projections, then the following conditions are equivalent*:

(*a*) $P_1 \leqslant P_2$.
(*b*) $\|P_1 x\| \leqslant \|P_2 x\|$, *for* $x \in X$.
(*c*) $\mathcal{R}(P_1) \subset \mathcal{R}(P_2)$.
(*d*) $P_2 P_1 = P_1$.
(*e*) $P_1 P_2 = P_1$.

Proof. For $x \in X$, we use (2-6) to see that $\|P_1 x\|^2 = (P_1 x, x) \leqslant (P_2 x, x) = \|P_2 x\|^2$. Thus (*a*) implies (*b*). Now assuming (*b*) holds and taking $x \in \mathcal{R}(P_1)$, we have $\|x\| = \|P_1 x\| \leqslant \|P_2 x\| \leqslant \|x\|$, since P_2 is an orthogonal projection. Then $\|P_2 x\| = \|x\|$; while from the Pythagorean theorem we must have $\|x\|^2 = \|P_2 x\|^2 + \|x - P_2 x\|^2$. We conclude that $\|x - P_2 x\| = 0$; that is, $x = P_2 x \in \mathcal{R}(P_2)$.

Thus (b) implies (c). Given (c) and $x \in X$, we have $P_1x \in \mathscr{R}(P_1) \subset \mathscr{R}(P_2)$, and so $P_2(P_1x) = P_1x$. Thus (c) implies (d). It is easy to see that (d) and (e) are equivalent. For example, from $P_2P_1 = P_1$ we have

$$(P_1P_2x, y) = (P_2x, P_1y) = (x, P_2P_1y) = (x, P_1y) = (P_1x, y),$$

whence $P_1P_2 = P_1$. Finally, given (e), one obtains (d) and then it is easy to verify that $P_2 - P_1$ is a symmetric projection, whence $0 \leq P_2 - P_1$, or $P_1 \leq P_2$. □

PROBLEMS

1. Let A be linear on X into X, where X is a complex inner-product space. Suppose (Ax, x) is real for every x. Then A is symmetric.
2. Suppose A and B are linear and symmetric on X into X. If $(Ax, x) = (Bx, x)$ for $x \in X$, then $A = B$. Compare with problem 2, § 1.
3. The Cauchy–Schwarz inequality may be proved in a way that also applies to the generalization (2-5). Fill in the details of the following argument: Given x, y, let $\beta = (x, y)/|(x, y)|$ if $(x, y) \neq 0$ and $\beta = 1$ otherwise. Then for any real number t, examine $\|tx + \beta y\|^2$ to deduce the quadratic inequality

$$(x, x)t^2 + 2|(x, y)|t + (y, y) \geq 0$$

   for all real t. The Cauchy–Schwarz inequality follows from this.
4. If P_1 and P_2 are symmetric projections on X, show that $P_2 - P_1$ is a symmetric projection if and only if $P_1 \leq P_2$, and in this case characterize the range of $P_2 - P_1$.
5. Suppose A and B are linear on X and A is symmetric. Then $AB = 0$ if and only if $\mathscr{R}(A)$ and $\mathscr{R}(B)$ are orthogonal. (This result is often useful when A and B are symmetric projections on X.)

VI.3 NORMAL AND SELF-ADJOINT OPERATORS

Throughout this section, X denotes a Hilbert space, and all of the operators considered here belong to $L(X)$. Recall from § IV.11 that if $A \in L(X)$, then $A^* \in L(X)$ and the definition of A^* is fully expressed by the relation

$$(Ax, y) = (x, A^*y) \qquad x, y \in X.$$

The operator A is *normal* if $AA^* = A^*A$ and *self-adjoint* if $A = A^*$. Examples of self-adjoint operators were given in § IV.11.

Clearly a self-adjoint operator in $L(X)$ is symmetric. Conversely, a symmetric operator whose domain is all of X (with X complete) must be a self-adjoint operator in $L(X)$. The proof of this fact hinges upon the closed graph theorem or, alternatively, upon the principle of uniform boundedness (problem 1). Self-adjoint operators with domain a proper dense subspace of X

will be treated in § 8. There it will be seen that an unbounded symmetric operator need not be self-adjoint.

Our first theorem concerns the localization of the spectrum of any member of $L(X)$. For $A \in L(X)$, the *numerical range* of A is the set of values of (Ax, x) for all x with $\|x\| = 1$. We shall denote the closure of the numerical range of A by $V(A)$.

Theorem 3.1. *If $A \in L(X)$, then $\sigma(A) \subset V(A)$. If the distance d from λ to $V(A)$ is positive, then*

$$\|(\lambda - A)\| \leq d^{-1}. \tag{3-1}$$

Proof. Suppose λ is not in $V(A)$, the distance from λ to $V(A)$ being $d > 0$. Then, for $\|x\| = 1$,

$$d \leq |\lambda - (Ax, x)| = |(\lambda x - Ax, x)| \leq \|\lambda x - Ax\|$$

by the Cauchy–Schwarz inequality. Hence $d\|x\| \leq \|\lambda x - Ax\|$ for every x. This shows that $(\lambda - A)^{-1}$ exists and is continuous, with

$$\|(\lambda - A)^{-1} y\| \leq d^{-1}\|y\|$$

for every y in the range of $\lambda - A$. We know, then, that λ is in either $\rho(A)$ or $R\sigma(A)$. If $\lambda \in R\sigma(A)$, then $\mathscr{R}(\lambda - A)^{\perp} \neq (0)$; and, since $\mathscr{R}(\lambda - A)^{\perp} = \mathscr{N}(\bar{\lambda} - A^*)$ by (11-7) in § IV.11, we conclude that $\bar{\lambda}$ is an eigenvalue of A^*. Take x such that $A^* x = \bar{\lambda} x$ and $\|x\| = 1$. Then $(Ax, x) = (x, A^* x) = (x, \bar{\lambda} x) = \lambda$, so that $\lambda \in V(A)$. This contradiction shows that $\lambda \in \rho(A)$. □

It can be shown that the numerical range of an operator in $L(X)$ is a convex set (see Davis [1]). In the case when A is normal, $V(A)$ is the smallest closed convex set containing $\sigma(A)$ (see Halmos [3, page 321]). For finite-dimensional spaces these results go back to Toeplitz and Hausdorff (see Wintner [1, pages 33–38]).

Theorem 3.2. *Suppose A is a bounded self-adjoint operator. Then $\sigma(A)$ lies on the closed interval $[m(A), M(A)]$ of the real axis. The end points of this interval belong to $\sigma(A)$.*

Proof. The first assertion is a consequence of Theorem 3.1, for it follows from (2-2) that $V(A)$ lies on the closed interval in question. (In fact, $V(A)$ is the closed interval $[m(A), M(A)]$, as a result of the fact that $V(A)$ is convex; but we do not need this information.) To see that $m(A)$ is in $\sigma(A)$ let $\lambda = m(A)$ and observe that $([A - \lambda]x, x) \geq 0$ for each x. Therefore, by (2-5), with $A - \lambda$, x, and $(A - \lambda)x$ in place of A, x, and y, respectively, we have

$$\begin{aligned}\|(A - \lambda)x\|^4 &\leq ([A - \lambda]x, x)([A - \lambda]^2 x, [A - \lambda]x) \\ &\leq ([A - \lambda]x, x)\|A - \lambda\|^3 \|x\|^2.\end{aligned}$$

If $\|A-\lambda\|=0$, it is clear that $\lambda \in \sigma(A)$. Otherwise, we see from the foregoing that

$$\inf_{\|x\|=1} \|(A-\lambda)x\|=0,$$

whence $\lambda \in \sigma(A)$. This is because $\inf_{\|x\|=1} ([A-\lambda]x, x)=0$, by the definition of $m(A)$. The proof that $M(A)$ is in $\sigma(A)$ is similar. □

If λ is not on the interval $[m(A), M(A)]$ we can estimate the norm of $(\lambda - A)^{-1}$ in terms of the distance from λ to the interval, using (3-1). In particular, if X is a complex space and λ is not real, the distance in question is not less than the absolute value of the imaginary part of λ. Hence

$$\|(\lambda - A)^{-1}\| \leq 2/|\lambda - \bar{\lambda}|. \tag{3-2}$$

From Theorem 3.2 we also conclude that a bounded self-adjoint operator A is positive (i.e., $0 \leq A$) if and only if $\sigma(A)$ lies in the interval $[0, \infty)$.

Theorem 3.3. *If $A \in L(X)$ and $A = A^*$, the spectral radius of A is $\|A\|$.*

Proof. This is an immediate consequence of (2-3) and Theorem 3.2. □

This result, which is of crucial importance in our proof of the fundamental spectral theorem for self-adjoint operators (Theorem 6.1), can be proved in other ways (see problem 2). The property is true, more generally, of normal operators in complex spaces, as we shall see below.

Theorem 3.4. *If $A \in L(X)$, then A is normal if and only if $\|Ax\| = \|A^*x\|$ for every x.*

Proof. We have

$$\|Ax\|^2 = (Ax, Ax) = (x, A^*Ax),$$

and

$$\|A^*x\|^2 = (A^*x, A^*x) = (x, A^{**}A^*x) = (x, AA^*x).$$

Thus $\|Ax\| = \|A^*x\|$ if and only if $(x, A^*Ax) = (x, AA^*x)$. But this is true for every x if and only if $A^*A = AA^*$. See problem 2, § 2. □

Theorem 3.5. *If the space is complex and A is normal, the spectral radius of A is exactly $\|A\|$.*

Proof. We first prove that $\|A^2\| = \|A\|^2$. We have $\|A^2x\| = \|A^*Ax\|$ for every x, by Theorem 3.4. Thus $\|A^2\| = \|A^*A\|$. But we know that $\|A^*A\| = \|A\|^2$ (by Theorem IV.11.1). Since powers of A are normal, it now follows by induction that $\|A^p\| = \|A\|^p$ if p is a positive integer of the form 2^n. The fact that the spectral radius of A is $\|A\|$ now follows from Theorem V.3.5. □

Theorem 3.6. *If A is normal, then $\overline{\mathcal{R}(A)}$ and $\mathcal{N}(A)$ are orthogonal complements, so that $X = \overline{\mathcal{R}(A)} \oplus \mathcal{N}(A)$.*

Proof. Since $\mathcal{N}(A)$ is closed, $X = \mathcal{N}(A)^{\perp} \oplus \mathcal{N}(A)$, by Theorem II.7.4. But $\mathcal{N}(A) = \mathcal{N}(A^*)$ by Theorem 3.4, and so $\mathcal{N}(A)^{\perp} = \mathcal{N}(A^*)^{\perp} = \overline{\mathcal{R}(A)}$, by (11-7) of § IV.11. □

Theorem 3.7. *If A is normal, then the ascent of A is either* 0 *or* 1.

Proof. If $x \in \mathcal{N}(A^2)$, then $Ax \in \mathcal{R}(A) \cap \mathcal{N}(A)$. But $\mathcal{R}(A) \cap \mathcal{N}(A) = (0)$ by Theorem 3.6, and hence $Ax = 0$. We conclude that $\mathcal{N}(A^2) = \mathcal{N}(A)$. Thus the ascent of A does not exceed 1. (Cf. Theorem V.6.3.) □

Now suppose that X is a complex space. If A is normal, so is $\lambda - A$ for each λ. If λ is not an eigenvalue of A, then Theorem 3.6 shows that $X = \overline{\mathcal{R}(\lambda - A)}$. If $\mathcal{R}(\lambda - A)$ is closed, then $\lambda \in \rho(A)$; otherwise, $\lambda \in C\sigma(A)$. It follows that the residual spectrum of A is empty and $\sigma(A) = P\sigma(A) \cup C\sigma(A)$. In particular, the approximate point spectrum of A coincides with $\sigma(A)$.

Suppose X is complex and finite dimensional and A is normal. If $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of A, then the ascent of $\lambda_i - A$ is 1 for each i. If $M_i = \mathcal{N}(\lambda_i - A)$, it follows that $X = M_1 \oplus \cdots \oplus M_k$ and that $A = \lambda_1 E_1 + \cdots + \lambda_k E_k$ where E_i is the projection of X onto M_i determined by this particular direct sum decomposition of X. (We have here a special case of the situation portrayed in Theorem V.11.3 and the remarks following it.) The eigenspaces M_i are mutually orthogonal (see problem 8), and so the E_i are orthogonal projections. These facts constitute the spectral theorem for a normal operator on a finite-dimensional space. The generalization to an infinite-dimensional space will be given in § VII.7.

PROBLEMS

1. Suppose A and B are linear on X into X (with X a Hilbert space), and $(Ax, y) = (x, By)$ for all x and y. Find two different proofs that $A \in L(X)$ and $B = A^*$.
2. If $A \in L(X)$ and $A = A^*$, prove that $r_\sigma(A) = \|A\|$, starting as follows: Let $\alpha = \|A\|$. Show that
$$\|(\alpha^2 - A^2)x\|^2 = \alpha^4\|x\|^2 - 2\alpha^2\|Ax\|^2 + \|A^2x\|^2.$$
Choose x_n so that $\|x_n\| = 1$ and $\|Ax_n\| \to \|A\|$, and deduce that $\alpha^2 \in \sigma(A^2)$.
3. If $A \in L(X)$, where X is a complex space, we can write A in the form $A = H + iK$, where H and K are self-adjoint, in one and only one way, namely, with $H = (A + A^*)/2$, $K = (A - A^*)/2i$. Then A is normal if and only if $HK = KH$.

4. Suppose A is normal and $A = H + iK$, where H and K are self-adjoint. Then A^{-1} exists and belongs to $L(X)$ if and only if $(H^2+K^2)^{-1}$ exists and belongs to $L(X)$. In that case $A^{-1} = A^*(H^2+K^2)^{-1}$.
5. If A is normal, $\overline{\mathscr{R}(A^2)} = \overline{\mathscr{R}(A)}$.
6. If A is normal, its descent is finite if and only if $\mathscr{R}(A)$ is closed.
7. Suppose $A \in L(X)$. Then
   a. $\lambda \in \rho(A)$ if and only if $\bar{\lambda} \in \rho(A^*)$;
   b. $\lambda \in C\sigma(A)$ implies $\bar{\lambda} \in C\sigma(A^*)$;
   c. $\lambda \in R\sigma(A)$ implies $\bar{\lambda} \in P\sigma(A^*)$;
   d. $\lambda \in P\sigma(A)$ implies $\bar{\lambda} \in P\sigma(A^*) \cup R\sigma(A^*)$.
8. Suppose A is normal. If λ is an eigenvalue of A and M is the corresponding eigenspace, then M is also the eigenspace corresponding to $\bar{\lambda}$ as an eigenvalue of A^*. Eigenvectors corresponding to distinct eigenvalues of A are orthogonal.
9. Suppose A is normal and X is a complex space. If λ_0 is a pole of the resolvent R_λ of A, then the pole is of first order and $\mathscr{R}(\lambda_0 - A)$ is closed. Conversely, if $\mathscr{R}(\lambda_0 - A)$ is a closed proper subspace of X, then λ_0 is a pole of R_λ.
10. Let $\{A_n\}$ be a sequence of self-adjoint operators in $L(X)$ such that $A_n \leqslant A_{n+1}$ and $(A_n x, x) \leqslant \alpha \|x\|^2$ for each n, where α is a real constant. Then there exists a self-adjoint operator A such that $A_n x \to Ax$ for each x. A similar proposition is valid if all the inequalities are reversed. For the proof, first show that the sequence $\{\|A_n\|\}$ is bounded. Then apply (2-5) to $H = A_n - A_m$, where $n > m$, treating H like $A - \lambda$ in the proof of Theorem 3.2. The result is
$$\|A_n x - A_m x\|^4 \leqslant [(A_n x, x) - (A_m x, x)]k\|x\|^2,$$
where k is a constant. The rest of the argument is left to the reader.
11. Let $\{A_n\}$ be a sequence of positive operators in $L(X)$. Then $A_n x \to 0$ for each x if and only if $(A_n x, x) \to 0$ for each x.
12. Let $\{A_n\}$ be a sequence of self-adjoint operators in $L(X)$ such that $A_n \leqslant A_{n+1}$ for each n, and let A be an operator such that $(A_n x, x) \to (Ax, x)$ for each x. Then $A_n x \to Ax$. Instead of $A_n \leqslant A_{n+1}$, we may have $A_{n+1} \leqslant A_n$, and the conclusion is the same.

VI.4 COMPACT SYMMETRIC OPERATORS

For this section we assume that X is a real or complex inner-product space, with $X \neq (0)$. Completeness is assumed only as needed. We consider a compact symmetric operator A in $L(X)$. To avoid trivialities we assume $A \neq 0$. For the definition of a compact operator see § V.7. The present section is largely independent of § V.7.

The basic fact about the type of operator here considered is that it possesses a finite or countably infinite set of nonzero eigenvalues (and *at least one* such eigenvalue); moreover, the structure of the operator can be completely analyzed in terms of the eigenspaces corresponding to these eigenvalues. The exact story is told by Theorem 4.2. What we have here is a generalization of the fact that, if X is finite dimensional and A is a symmetric member of $L(X)$, a basis consisting of eigenvectors can be chosen for X in such a way that the matrix representing A is a diagonal matrix with each diagonal element an eigenvalue.

The initial step is that of proving the existence of at least one nonzero eigenvalue.

Theorem 4.1. *Suppose A is compact, symmetric, and $A \neq 0$. Then either $\|A\|$ or $-\|A\|$ is an eigenvalue of A, and there is a corresponding eigenvector x such that $\|x\| = 1$ and $|(Ax, x)| = \|A\|$.*

Proof. In view of (2-2) and (2-3) there exists a sequence $\{x_n\}$ such that $\|x_n\| = 1$ and $(Ax_n, x_n) \to \lambda$, where λ is real and $|\lambda| = \|A\|$. Now

$$\begin{aligned} 0 \leqslant \|Ax_n - \lambda x_n\|^2 &= \|Ax_n\|^2 - 2\lambda(Ax_n, x_n) + \lambda^2 \|x_n\|^2 \\ &\leqslant \|A\|^2 - 2\lambda(Ax_n, x_n) + \lambda^2. \end{aligned}$$

But then we see that $Ax_n - \lambda x_n \to 0$. Since A is compact, $\{Ax_n\}$ contains a convergent subsequence, which we denote by $\{Ay_k\}$, $\{y_k\}$ being a subsequence of $\{x_n\}$. The sequence $\{y_k\}$ is then convergent also because $\lambda \neq 0$. Suppose $y_k \to x$ as $k \to \infty$. Then $\|x\| = 1$ and $Ay_k \to Ax$, whence $Ax = \lambda x$. Evidently $|(Ax, x)| = |\lambda| \, \|x\|^2 = \|A\|$, so the proof is complete. □

We now apply Theorem 4.1 repeatedly. Denote the eigenvalue and eigenvector of Theorem 4.1 by λ_1 and x_1, respectively. Let $X = X_1$, and let $X_2 = \{x : (x, x_1) = 0\}$. Then X_2 is a subspace invariant under A, for $x \in X_2$ implies $(Ax, x_1) = (x, Ax_1) = (x, \lambda_1 x_1) = \lambda_1(x, x_1) = 0$. The restriction of A to X_2 is compact and symmetric. If the restriction is not the zero operator, we can assert the existence of λ_2 and x_2 such that $x_2 \in X_2$, $\|x_2\| = 1$, $Ax_2 = \lambda_2 x_2$, and λ_2 is the norm of the restriction of A to X_2. Evidently $|\lambda_2| \leqslant |\lambda_1|$. Continuing in this way, we obtain the nonzero eigenvalues $\lambda_1, \dots, \lambda_n$, with corresponding eigenvectors $x_1, \dots, x_n$ of unit norm. We also obtain $X_1 \supset X_2 \supset \cdots \supset X_{n+1}$, with X_{k+1} the set of elements of X_k that are orthogonal to $x_1, \dots, x_k$. At each step $x_k \in X_k$ and $|\lambda_k|$ is the norm of the restriction of A to X_k, so $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. The process stops with λ_n, x_n, and X_{n+1} if and only if the restriction of A to X_{n+1} is 0. In that case the range of A lies in the subspace generated by $x_1, \dots, x_n$. For if $x \in X$, let

$$y_n = x - \sum_{k=1}^{n} (x, x_k)x_k. \tag{4-1}$$

Then $(y_n, x_i) = 0$ if $i = 1, \dots, n$, so that $y_n \in X_{n+1}$ and therefore $Ay_n = 0$, or

$$Ax = \sum_{k=1}^{n} (x, x_k)Ax_k = \sum_{k=1}^{n} \lambda_k(x, x_k)x_k. \tag{4-2}$$

This situation may occur even if X is infinite dimensional. It will certainly occur eventually if X is finite dimensional, because $x_1, \ldots, x_n$ are linearly independent.

The foregoing considerations lead us to the statement of the fundamental theorem.

Theorem 4.2. *Suppose A is compact, symmetric, and $A \neq 0$. The procedure described in the foregoing discussion yields a possibly terminating sequence of nonzero eigenvalues $\lambda_1, \lambda_2, \ldots$ and a corresponding orthonormal set of eigenvectors $x_1, x_2, \ldots$. If the sequences do not terminate, then $|\lambda_n| \to 0$. The expansion*

$$Ax = \sum (Ax, x_k)x_k = \sum \lambda_k (x, x_k)x_k \tag{4-3}$$

is valid for each $x \in X$, the summation being extended over the entire sequence, whether finite or infinite. Each nonzero eigenvalue of A occurs in the sequence $\{\lambda_n\}$. The eigenmanifold corresponding to a particular λ_i is finite dimensional and its dimension is exactly the number of times this particular eigenvalue is repeated in the sequence $\{\lambda_n\}$.

Proof. Since $|\lambda_k| \geq |\lambda_{k+1}|$, we either have $\lambda_n \to 0$ or $|\lambda_n| \geq \varepsilon > 0$ for some ε and all n. Suppose the latter and that the sequence is infinite. Then $\{x_n/\lambda_n\}$ is a bounded sequence, and $A(x_n/\lambda_n) = x_n$, so that $\{x_n\}$ must contain a convergent subsequence. This is impossible, for the orthonormality yields $\|x_n - x_m\|^2 = 2$. Hence $\lambda_n \to 0$ when the sequence is infinite. If the sequence of λ_k's terminates with λ_n, (4-3) is equivalent to (4-2). In the nonterminating case we define y_n by (4-1) and obtain

$$\|y_n\|^2 = \|x\|^2 - \sum_{k=1}^{n} |(x, x_k)|^2 \leq \|x\|^2.$$

Since $y_n \in X_{n+1}$ and $|\lambda_{n+1}|$ is the norm of the restriction of A to X_{n+1}, we have

$$\|Ay_n\| \leq |\lambda_{n+1}| \, \|y_n\| \leq |\lambda_{n+1}| \, \|x\|.$$

Hence $Ay_n \to 0$. But

$$Ay_n = Ax - \sum_{k=1}^{n} (x, x_k)Ax_k,$$

and so we obtain (4-3) [note that $Ax_k = \lambda_k x_k$ and $(Ax, x_k) = (x, Ax_k) = \lambda_k(x, x_k)$].

If λ is a nonzero eigenvalue of A that is not in the sequence $\{\lambda_k\}$, there is a corresponding eigenvector x of unit norm, and it must be orthogonal to x_n for every n, by Theorem 2.1. Then $Ax = 0$, by (4-3). This contradicts $Ax = \lambda x \neq 0$. An eigenvalue cannot be repeated infinitely often in the sequence $\{\lambda_n\}$, because $\lambda_n \to 0$. Suppose that λ_k occurs p times. Then the corresponding eigenmanifold contains an orthonormal set of p eigenvectors, and therefore is at least p-dimensional. It cannot be of dimension greater than p, for this would entail the existence of an x such that $Ax = \lambda_k x$, $\|x\| = 1$, and $(x, x_n) = 0$ for every n. But such a thing is impossible, by an argument given at the beginning of this paragraph. □

The next theorem describes the inverse of $\lambda - A$.

Theorem 4.3. *Let A, $\{\lambda_n\}$, $\{x_n\}$ be as in Theorem 4.2. Then, if $\lambda \neq 0$ and if $\lambda \neq \lambda_k$ for each k, $\lambda - A$ has a continuous inverse defined on all of X and given by $x = (\lambda - A)^{-1}y$, where*

$$x = \frac{1}{\lambda} y + \frac{1}{\lambda} \sum \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} x_k. \tag{4-4}$$

Proof. We can discover the foregoing formula as follows. Suppose x and y given, such that $\lambda x - Ax = y$. Then $Ax = \lambda x - y$, and so from (4-3) we have

$$\lambda x - y = \sum \lambda_k (x, x_k) x_k.$$

We form the inner product with x_i and obtain

$$(\lambda x, x_i) - (y, x_i) = \lambda_i (x, x_i).$$

Thus

$$(x, x_i) = \frac{(y, x_i)}{\lambda - \lambda_i},$$

and so

$$\lambda x = y + \sum \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} x_k,$$

which gives (4-4). This shows that the solution of $(\lambda - A)x = y$ is unique, if it exists. On the other hand, if the series in (4-4) is convergent, the element x defined by (4-4) certainly satisfies $(\lambda - A)x = y$, for then

$$\lambda x - Ax = y + \sum \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} x_k - \frac{1}{\lambda} Ay - \frac{1}{\lambda} \sum \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} Ax_k.$$

We put $Ax_k = \lambda_k x_k$ in the last sum and use (4-3) with y in place of x; the result is $\lambda x - Ax = y$.

We now show that the series in (4-4) *does* converge, no matter how y is chosen. For this purpose let

$$\alpha = \sup_k \left| \frac{\lambda_k}{\lambda - \lambda_k} \right|, \qquad \beta = \sup_k \frac{1}{|\lambda - \lambda_k|}.$$

Also let

$$u_n = \sum_{k=1}^{n} \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} x_k, \qquad v_n = \sum_{k=1}^{n} \frac{(y, x_k)}{\lambda - \lambda_k} x_k.$$

Now, if $m < n$,

$$\|u_n - u_m\|^2 = \sum_{k=m+1}^{n} \left| \frac{\lambda_k}{\lambda - \lambda_k} \right|^2 |(y, x_k)|^2 \leq \alpha^2 \sum_{k=m+1}^{n} |(y, x_k)|^2.$$

Therefore $\{u_n\}$ is a Cauchy sequence, because $\Sigma\, |(y, x_k)|^2$ is convergent by Bessel's inequality (page 81). If X were complete, this would be enough for our purpose. If X is not complete, we continue the argument as follows:

$$\|v_n\|^2 = \sum_{k=1}^{n} \frac{|(y, x_k)|^2}{|\lambda - \lambda_k|^2} \leq \beta^2 \sum_{k=1}^{n} |(y, x_k)|^2 \leq \beta^2 \|y\|^2,$$

so that $\{v_n\}$ is bounded. Now $Av_n = u_n$. Hence the compactness of A shows that $\{u_n\}$ contains a convergent subsequence. Being a Cauchy sequence, $\{u_n\}$ must then be convergent to the same limit as the subsequence. Hence the series in (4-4) converges. We see from (4-4) that

$$\|x\| \leq \frac{1}{|\lambda|} \|y\| + \frac{1}{|\lambda|} \alpha \|y\|.$$

Thus we see that $(\lambda - A)^{-1}$ is continuous and defined on all of X, with

$$\|(\lambda - A)^{-1}\| \leq \frac{1}{|\lambda|} \left[1 + \sup_k \left| \frac{\lambda_k}{\lambda - \lambda_k} \right| \right]. \qquad \square \tag{4-5}$$

We round out the foregoing discussion by considering the null space and range of A. The situation is clearest if X is complete.

Theorem 4.4. *(a) Let A, $\{\lambda_n\}$ and $\{x_n\}$ be as in Theorem 4.2, and let M be the closed linear manifold generated by the eigenvectors $x_1, x_2, \ldots$. Then $M^\perp = \mathcal{N}(A)$. Hence the orthonormal set $\{x_n\}$ is complete if and only if 0 is not an eigenvalue of A. (b) When X is complete we have $X = M \oplus \mathcal{N}(A)$. Also, the range of A is composed of those elements y in M which are such that the series*

$$\sum \frac{(y, x_k)}{\lambda_k} x_k \tag{4-6}$$

is convergent.

Proof. (*a*) It follows from (4-3) that $M^\perp \subset \mathcal{N}(A)$. On the other hand, $x \in \mathcal{N}(A)$ implies $(x, x_k) = \lambda_k^{-1}(x, Ax_k) = \lambda_k^{-1}(Ax, x_k) = 0$, so that $x \in M^\perp$. Hence $M^\perp = \mathcal{N}(A)$. The orthonormal set $\{x_n\}$ is complete if and only if $M^\perp = (0)$, which then means that 0 is not an eigenvalue of A. (*b*) If X is complete, we have $X = M \oplus \mathcal{N}(A)$, by Theorem II.7.4. Now suppose $y = Ax$ for some x. Then from (4-3) we see that y is in M. From the orthonormality it follows that $(y, x_k) = \lambda_k(x, x_k)$. We can write $x = u + v$, $u \in M$, $v \in \mathcal{N}(A)$. Then $(x, x_k) = (u, x_k)$, since $v \perp M$, and so (see Theorem II.6.9)

$$u = \sum (u, x_k)x_k = \sum (x, x_k)x_k = \sum \frac{(y, x_k)}{\lambda_k} x_k,$$

the series necessarily being convergent if in fact it is infinite. Conversely, suppose $y \in M$ and that the series (4-6) is convergent, with u as its sum. Then

$$Au = \sum (y, x_k)x_k = y,$$

so that $y \in \mathcal{R}(A)$. □

We now examine formula (4-3) and express it in a somewhat different way, for the purpose of showing the relation between the formula and the general spectral representation theorem (of § 6) for self-adjoint operators. For the purpose of the following discussion it is assumed that X is complete.

The series (4-3) remains convergent, with the same sum, no matter how the terms are rearranged in order. (To see this, rewrite the series as $\Sigma (y, x_k)x_k$ where $y = Ax$, and then apply Theorem II.6.9.) It is convenient to rearrange the terms, if necessary, in such a way that all the terms for which λ_k has one particular value are brought so that they occur consecutively in the series. We shall now assume that the notation has been arranged so that this is true. For each λ_k let P_k be the operator defined by

$$P_k x = \sum_{\lambda_i = \lambda_k} (x, x_i)x_i.$$

Then $P_j = P_k$ if $\lambda_j = \lambda_k$. It is easy to verify that $P_j P_k = 0$ if $\lambda_j \neq \lambda_k$, that $P_k^2 = P_k$, and that P_k is symmetric. The series (4-3) can now be written

$$Ax = \sum{}' \lambda_k P_k x, \tag{4-7}$$

where the prime mark on the summation sign indicates that the sum is extended over the distinct values of λ_k. In like manner, (4-4) can be put in the form

$$(\lambda - A)^{-1} y = \frac{1}{\lambda} y + \frac{1}{\lambda} \sum{}' \frac{\lambda_k}{\lambda - \lambda_k} P_k y. \tag{4-8}$$

This equation shows that λ_k is a first-order pole of $(\lambda - A)^{-1} y$, the residue being $P_k y$. It is easy to verify that P_k is the projection associated with the spectral set (λ_k) (as in § V.10).

If we use the decomposition $X = M \oplus \mathcal{N}(A)$ (see Theorem 4.4), and write $x = u + v$, $u \in M$, $v \in \mathcal{N}(A)$, we have $P_k x = P_k u$ and

$$u = \sum{}' P_k x.$$

If we define P_0 by $P_0x = v$, we see that

$$x = \sum{}' P_k x + P_0 x. \tag{4-9}$$

The operator P_0 is a symmetric projection, and $P_0P_k = P_kP_0 = 0$ if $k \neq 0$.

Now let us define a one-parameter family of operators E_λ as follows (λ real, x arbitrary):

$$\begin{cases} E_\lambda x = \sum\limits_{\lambda_k \leqslant \lambda}{}' P_k x & \text{if } \lambda < 0 \\ E_\lambda x = x - \sum\limits_{\lambda_k > \lambda}{}' P_k x & \text{if } \lambda \geqslant 0. \end{cases} \tag{4-10}$$

It is understood that the meaning of a sum is 0 if there exist no points λ_k satisfying the indicated inequality. By separate consideration of the cases $\lambda \leqslant \mu < 0$, $\lambda < 0 \leqslant \mu$, $0 \leqslant \lambda \leqslant \mu$, it is easy to verify that $E_\mu E_\lambda = E_\lambda E_\mu = E_\lambda$ if $\lambda \leqslant \mu$. Hence in particular $E_\lambda^2 = E_\lambda$. Moreover, E_λ is symmetric. The operator E_λ is continuous from the right, in the sense that

$$\lim_{\lambda \to \mu^+} E_\lambda x = E_\mu x, \tag{4-11}$$

where $\lambda \to \mu^+$ means that we consider values $\lambda > \mu$. This is clear if we examine closely the definition of E_λ. The only possible point of accumulation of the points λ_k is at 0. As we move to the right from a given point μ, E_λ does not change in value except as λ reaches one of the points λ_k or the point 0. Thus (4-11) is evident if $\mu \neq 0$, and for $\mu = 0$ it is true as a result of the convergence of the series defining E_0. Since $\sigma(A)$ lies on $[m(A), M(A)]$, it is easy to see that $E_\lambda = 0$ if $\lambda < m(A)$ and $E_\lambda = I$ if $\lambda \geqslant M(A)$.

We shall denote $\lim\limits_{\lambda \to \mu^-} E_\lambda x$ by $E_{\mu-0}x$. It is easy to see from (4-10) that

$$E_\lambda - E_{\lambda-0} = P_k \qquad \text{if } \lambda = \lambda_k$$

and that

$$E_0 - E_{0-0} = P_0.$$

Formula (4-7) can now be written in the form

$$Ax = \int_\alpha^\beta \lambda \, dE_\lambda x, \tag{4-12}$$

where $[\alpha, \beta]$ is any closed interval such that $\alpha < m(A)$ and $M(A) \leqslant \beta$. The integral here is defined by the usual procedure for a Riemann–Stieltjes integral, involving sums of the type

$$\sum \mu_i [E(\mu_i) - E(\mu_{i-1})] x,$$

with $\alpha = \mu_0 < \mu_1 < \cdots < \mu_n = \beta$. (Here we have written $E(\mu_i)$ instead of E with subscript μ_i, to simplify printing.) See Taylor [5, pages 392–396] for a discussion of Riemann–Stieltjes integrals. Formula (4-8) can also be expressed in integral form. First, we modify (4-8) by use of (4-9) with y in place of x. In this way we find

$$(\lambda - A)^{-1}y = \sum{}' \frac{1}{\lambda - \lambda_k} P_k y + \frac{1}{\lambda - 0} P_0 y.$$

Then we have

$$(\lambda - A)^{-1}y = \int_\alpha^\beta \frac{1}{\lambda - \mu}\, dE_\mu y.$$

The results of this section have as their most important application the theory of integral equations with symmetric (or Hermitian) kernels, where the spaces and the kernels are such that the corresponding integral operators are compact. Compact integral operators were discussed in § V.7. For symmetry of the operator, the condition on the kernel is $\overline{k(s,t)} = k(t,s)$. For a classical exposition of the theory of symmetric integral equations see Courant and Hilbert [1, pages 122–134]. For an historical account, with many references, see Hellinger and Toeplitz [1, part III].

PROBLEMS

1. Let A, $\{\lambda_k\}$ and $\{x_k\}$ be as in Theorem 4.2. If $\lambda = \lambda_j$ for some j, show that the range of $\lambda - A$ consists of all vectors orthogonal to the eigenspace corresponding to λ_j. For such a vector y the general solution of $(\lambda - A)x = y$ is

$$x = \frac{1}{\lambda} y + \frac{1}{\lambda} \sum_{\lambda_k \neq \lambda} \lambda_k \frac{(y, x_k)}{\lambda - \lambda_k} x_k + w,$$

   where w is an arbitrary element of the eigenspace corresponding to λ_j.
2. Suppose X is complete. Let $\{x_n\}$ be an orthonormal set, and let $\{\lambda_n\}$ be any sequence of real numbers such that $\lambda_n \to 0$. Let A be defined by $Ax = \sum_{k=1}^\infty \lambda_k (x, x_k) x_k$. Then A is self-adjoint and compact. If $\lambda_n \geqslant 0$ for all n, then A is a positive operator. [*Hint.* For the compactness, show that A is the limit of a sequence of operators of finite rank.]
3. Suppose X is a complex inner-product space (not necessarily complete). Suppose S and T are compact members of $L(X)$ such that $ST = TS$ and $(Sx, y) = (x, Ty)$ for all x and y. Then there exist a finite or infinite orthonormal set $\{u_n\}$ and a corresponding sequence $\{\lambda_n\}$ such that $\lambda_n \neq 0$, $Su_n = \lambda_n u_n$, $Tu_n = \bar{\lambda}_n u_n$, $\|y\|^2 = \sum_n |(y, u_n)|^2$ for each $y \in \mathscr{R}(S)$, and $\lambda_n \to 0$ if the sequences are infinite. Method: Put $S + T = 2A$, $S - T = 2iB$, and consider A, B. They are symmetric and compact, $AB = BA$, and $S = A + iB$.

   The foregoing proposition can be used to prove the completeness of the orthonormal set $\{v_\lambda\}$, where $v_\lambda(t) = e^{i\lambda t}$ and λ takes on all real values, in the space of continuous almost-periodic functions, with

$$(x, y) = \lim_{h\to\infty} \frac{1}{h} \int_0^h x(t)\overline{y(t)}\, dt.$$

This is an important example of a nonseparable incomplete inner-product space. For this application of the abstract proposition $Sx = y$ is taken to mean

$$y(s) = \lim_{h\to\infty} \frac{1}{h} \int_0^h z(s-t)x(t)\, dt,$$

and $Tx = y$ means

$$y(s) = \lim_{h\to\infty} \frac{1}{h} \int_0^h \overline{z(s-t)}x(t)\, dt,$$

where z is a fixed continuous almost-periodic function. In this case the orthonormal set $\{u_n\}$ is a set $\{v_{\nu(n)}\}$, where $\nu(n)$ runs through all the values of λ for which $(z, v_\lambda) \neq 0$. The λ_n corresponding to u_n is $(z, v_{\nu(n)})$. For details see Rellich [1, pages 351–355].

VI.5 SYMMETRIC OPERATORS WITH COMPACT RESOLVENT

As in § 4, we assume that X is an inner-product space, not necessarily complete. We also assume that X is not finite dimensional.

Theorem 5.1. *Suppose that T is a symmetric linear operator with domain and range in X, and suppose that T^{-1} exists, belongs to $L(X)$, and is compact. Let $\{\lambda_n\}$, $\{x_n\}$ be the sequences of eigenvalues and eigenvectors associated with $A = T^{-1}$, as explained in connection with Theorem 4.2, and let $\mu_n = 1/\lambda_n$. The sequence $\{\mu_n\}$ is infinite, and $|\mu_n| \to \infty$. The orthonormal set $\{x_n\}$ is complete, and*

$$x = \sum_{k=1}^{\infty} (x, x_k)x_k \tag{5-1}$$

for each $x \in \mathscr{D}(T)$. A point μ is in $\sigma(T)$ if and only if it is one of the μ_k's. We have $Tx_n = \mu_n x_n$. If μ is not in $\sigma(T)$,

$$(\mu - T)^{-1}y = \sum_{k=1}^{\infty} \frac{(y, x_k)}{\mu - \mu_k} x_k \tag{5-2}$$

for each $y \in X$. This inverse operator is compact.

Proof. The symmetry of T implies that of A; hence we can apply the results of § 4. We observe that $Ax_n = \lambda_n x_n$ is equivalent to $x_n = \lambda_n Tx_n$, or $Tx_n = \mu_n x_n$. The orthonormal set $\{x_n\}$ is complete, by Theorem 4.4, because 0 is not an eigenvalue of A (since A^{-1} exists). The orthonormal set must therefore be infinite, since X is infinite dimensional. Hence $|\mu_n| \to \infty$, for we know that

$\lambda_n \to 0$. The range of A is the domain of T; since each $x \in \mathscr{D}(T)$ is of the form $x = Ay$, (5-1) is a consequence of (4-3). Next, we show that

$$(\mu - T)^{-1} = \mu^{-1}A\left(A - \frac{1}{\mu}\right)^{-1} \tag{5-3}$$

if μ is different from 0 and all of the μ_n's. First, let us suppose that $x \in \mathscr{D}(T)$ and $(\mu - T)x = y$. Then $\mu Ax - x = Ay$, $(A - \mu^{-1})x = \mu^{-1}Ay$, and $x = \mu^{-1}(A - \mu^{-1})^{-1}Ay$. On the other hand, if $y \in X$ and $x = \mu^{-1}A(A - \mu^{-1})^{-1}y$, then $x \in \mathscr{D}(T)$ and it is easily verified that $(\mu - T)x = (A - \mu^{-1})(A - \mu^{-1})^{-1}y = y$. Since A commutes with $(A - \mu^{-1})^{-1}$, this proves (5-3). The compactness of $(\mu - T)^{-1}$ follows from this formula and Theorem V.7.2. To obtain the formula (5-2) we use (4-4). First, we have

$$\left(\frac{1}{\mu} - A\right)^{-1} y = \mu y + \mu^2 \sum_{k=1}^{\infty} \frac{(y, x_k)}{\mu_k - \mu} x_k.$$

Then

$$\frac{1}{\mu}A\left(A - \frac{1}{\mu}\right)^{-1} y = -Ay + \mu \sum_{k=1}^{\infty} \frac{(y, x_k)}{\mu_k(\mu - \mu_k)} x_k.$$

If we express Ay by (4-3) and simplify, we obtain (5-2) from this result and (5-3). □

We observe that the operator T of Theorem 5.1 is not continuous, because $\|x_n\| = 1$ and $\|Tx_n\| = \|\mu_n x_n\| = |\mu_n| \to \infty$. However, T is closed, because $T = A^{-1}$ and A is closed.

Theorem 5.1 can be applied to certain kinds of differential equation problems. For instance, consider the type of second-order differential operator discussed in Example 2, § 2, with boundary conditions such that the operator is symmetric. We shall now denote the operator by T instead of A. Then $Tx = u$ means that $x \in \mathscr{D}(T)$ (which includes specification of the boundary conditions) and

$$-\frac{d}{dt}[p(t)x'(t)] + q(t)x(t) = u(t).$$

Also, $(\mu - T)x = y$ means $x \in \mathscr{D}(T)$ and

$$\frac{d}{dt}[p(t)x'(t)] + [\mu - q(t)]x(t) = y(t). \tag{5-4}$$

In order to be able to apply Theorem 5.1 to this operator T, we have to show that T^{-1} exists and is a compact operator defined on all of X. This can be shown to be true in certain cases by construction of a Green's function $k(s, t)$. The service provided by the Green's function is that of being the kernel for an integral operator which turns out to be T^{-1}. If one can show that the Green's function exists and that the corresponding integral operator is compact, then Theorem 5.1 is applicable. The Green's function is Hermitian (i.e., $\overline{k(t, s)} = k(s, t)$) as a consequence of the symmetry of the operator T.

There is not space in this book for a detailed treatment of boundary-value problems for ordinary differential equations. For an extensive discussion, using the full power of spectral-theory methods in Hilbert space, see Dunford and Schwartz [2, pages 1278–1628]. Another useful discussion of differential operators from the Hilbert space viewpoint is to be found in Akhiezer and Glazman [2, pages 162–215].

It is not always true that T has a compact resolvent. When it does not, there may be continuous spectrum in addition to or in place of point spectrum, and then the series expansions (5-1), (5-2) are replaced by integral representations.

PROBLEM

1. Suppose that H is a continuous symmetric operator such that H^{-1} exists and belongs to $L(X)$. Let T and $A = T^{-1}$ be as in Theorem 5.1. Let $B = HAH$ (a compact and symmetric operator), and let $\{\lambda_n\}$ be the nonzero eigenvalues and $\{x_n\}$ the corresponding eigenvectors associated with B, just as in § 4. Show that $(\mu_n H^2 - T)(H^{-1}x_n) = 0$, where $\mu_n = 1/\lambda_n$. The orthonormal set $\{x_n\}$ is complete, and $x = \Sigma_1^\infty (x, Hx_n)H^{-1}x_n$ for each $x \in \mathscr{D}(T)$. If μ is different from every μ_n, $\mu H^2 - T$ has a bounded inverse defined on all of X and given by

$$(\mu H^2 - T)^{-1}y = \sum_1^\infty \frac{(y, H^{-1}x_n)}{\mu - \mu_n} H^{-1}x_n.$$

To show this, show first that

$$(\mu H^2 - T)^{-1} = \frac{1}{\mu} H^{-1}\left(B - \frac{1}{\mu}\right)^{-1} HA = \frac{1}{\mu} AH\left(B - \frac{1}{\mu}\right)^{-1} H^{-1}$$

if $\mu \neq 0$.

This problem has an application to the boundary-value problem corresponding to the equation (5-4) with $\mu\rho(t)$ in place of μ, where $\rho(t)$ is a positive real function such that, if H maps $x(t)$ into $\sqrt{\rho(t)}x(t)$, then H and H^{-1} are bounded operators on X.

VI.6 THE SPECTRAL THEOREM FOR BOUNDED SELF-ADJOINT OPERATORS

In this section we consider an arbitrary bounded self-adjoint operator A on a complex Hilbert space X. These assumptions about A and X apply to all the theorems of this section. Our aim is to obtain generalizations of the formula (4-12) and to use the results to establish a useful operational calculus. The methods are quite different from those used in § 4, for now we do not assume that A is compact, and the spectrum can be much more complicated.

Before stating the first theorem we call the reader's attention to the definitions of $m(A)$ and $M(A)$ in (2-2). Since A is continuous, $m(A)$ and $M(A)$ are finite.

Theorem 6.1. *There exists a family of orthogonal projections E_λ defined for each real λ, with the following properties:*

(*a*) $E_\lambda E_\mu = E_\mu E_\lambda = E_\lambda$ *if* $\lambda \leqslant \mu$.

(*b*) $\lim_{\mu \to \lambda^+} E_\mu x = E_\lambda x$.

(*c*) $E_\lambda = 0$ *if* $\lambda < m(A)$, $E_\lambda = I$ *if* $M(A) \leqslant \lambda$.

(*d*) $E_\lambda A = AE_\lambda$.

For each x and y in X, $(E_\lambda x, y)$ is of bounded variation as a function of λ, and

$$(p(A)x, y) = \int_\alpha^\beta p(\lambda)\, d(E_\lambda x, y) \tag{6-1}$$

for each polynomial $p(\lambda)$ with real coefficients. The integral is an ordinary Stieltjes integral over any interval $[\alpha, \beta]$ such that $\alpha < m(A)$, $M(A) \leqslant \beta$.

Proof. Let $p(\lambda)$ be any polynomial with real coefficients. Then $p(A)$ is self-adjoint, and therefore the spectral radius of $p(A)$ is $\|p(A)\|$ (Theorem 3.3). Now $\sigma[p(A)]$ is the set of values of $p(\lambda)$ for $\lambda \in \sigma(A)$ (Theorem V.3.4); consequently,

$$\|p(A)\| = \sup_{\lambda \in \sigma(A)} |p(\lambda)|. \tag{6-2}$$

Now let C be the real Banach space of real-valued continuous functions f defined on the closed interval $m(A) \leqslant \lambda \leqslant M(A)$, with $\|f\|$ defined as the maximum of $|f(\lambda)|$ on the interval. If we consider the polynomial p as a member of C, we evidently have $\|p(A)\| \leqslant \|p\|$, by (6-2) and Theorem 3.2. We can consider the class P of polynomials p with real coefficients as a subspace of C. Since any member of C can be approximated uniformly as closely as we please by a member of P (the Weierstrass theorem), the subspace P is dense in C.

Now consider any two vectors x, y, and let

$$L(p) = (p(A)x, y).$$

It is readily evident that L is a complex-valued linear functional defined on P and that

$$|L(p)| \leqslant \|p\|\,\|x\|\,\|y\|. \tag{6-3}$$

Thus L is continuous on P and may be extended in a unique way by continuity to give a continuous linear functional defined on C. We denote the extension by L also. We now refer to § III.5 for the representation of continuous linear functionals on the space C. We write $L(f)=L_1(f)+iL_2(f)$, where L_1 and L_2 have real values. On applying Theorem III.5.5 to L_1 and L_2 and then combining results, we see that there exists a complex-valued function of bounded variation $V(\lambda; x, y)$, depending on x and y as parameters, such that

$$(p(A)x, y)=\int_{m(A)}^{M(A)} p(\lambda)\, dV(\lambda; x, y) \tag{6-4}$$

for each p in P. In order to have $V(\lambda; x, y)$ uniquely determined by the functional L, we agree to have $V(\lambda; x, y)$ normalized in the manner explained in § III.5; that is, $V[m(A); x, y]=0$ and $V(\lambda; x, y)=V(\lambda+0; x, y)$ if $m(A)<\lambda<M(A)$.

In several parts of subsequent arguments we shall need the following uniqueness principle: If v is a function of bounded variation on $[a, b]$, normalized as in § III.5, and if $\int_a^b t^n\, dv(t)=0$ for $n=0, 1, 2, \ldots$, it follows that $\int_a^b x(t)\, dv(t)=0$ for every continuous function. This depends on the Weierstrass theorem about approximating continuous functions uniformly by polynomials. Then $v\sim 0$, and hence $v(t)\equiv 0$ (see § III.5).

The next step is to show that for each λ, $V(\lambda; x, y)$ is a continuous symmetric bilinear form in x and y, in the sense of § 1. If α is any scalar and $n=0, 1, 2, \ldots$, we have

$$\begin{aligned}\int \lambda^n\, dV(\lambda; \alpha x, y)&=(A^n\alpha x, y)=\alpha(A^n x, y)=\alpha\int \lambda^n\, dV(\lambda; x, y)\\ &=\int \lambda^n\, d[\alpha V(\lambda; x, y)].\end{aligned}$$

We omit the limits on the integral signs, for convenience. The uniqueness principle mentioned above now shows that $V(\lambda; \alpha x, y)=\alpha V(\lambda; x, y)$. Similar arguments show that $V(\lambda; x_1+x_2, y)=V(\lambda; x_1, y)+V(\lambda; x_2, y)$ and $V(\lambda; y, x)=\overline{V(\lambda; x, y)}$. For this last relation we use the fact that A^n is symmetric. The inequality (6-3) enables us to assert that the total variation of $V(\lambda; x, y)$ does not exceed $2\|x\|\|y\|$, for the real and imaginary parts of V have variations equal to $\|L_1\|$ and $\|L_2\|$, respectively. Hence $|V(\lambda; x, y)|\leq 2\|x\|\|y\|$; this shows that the bilinear form is continuous. By Theorem 1.2, we now see that for each λ on $[m(A), M(A)]$ there exists a bounded linear operator $F(\lambda)$ such that $V(\lambda; x, y)=(F(\lambda)x, y)$. Since the bilinear form is symmetric, so is $F(\lambda)$. The fact that $V[m(A); x, y]=0$ implies that $F[m(A)]=0$. If we put $p(\lambda)\equiv 1$ in (6-4), we see that $V[M(A); x, y]=(x, y)$. Hence $F[M(A)]=I$. Next, we prove that

$$F(\lambda)F(\mu)=F(\mu)F(\lambda)=F(\lambda) \qquad \text{if } \lambda\leq\mu. \tag{6-5}$$

In particular, $[F(\lambda)]^2 = F(\lambda)$, so that $F(\lambda)$ is a symmetric (and orthogonal) projection (cf. Theorem IV.12.7). The proof involves use of the following easily proved theorem about Stieltjes integrals: *Suppose f and g are continuous and v is of bounded variation on* $[a, b]$. *Let*

$$u(t) = \int_a^t g(s)\, dv(s).$$

Then

$$\int_a^b f(t)\, du(t) = \int_a^b f(t)g(t)\, dv(t).$$

We also need to know that, if v is normalized, so is u. Now let

$$U(\lambda; x, y) = \int_a^\lambda \mu^m\, dV(\mu; x, y),$$

where we have set $a = m(A)$, $b = M(A)$. Then

$$\int_a^b \lambda^n\, dU(\lambda; x, y) = \int_a^b \lambda^{m+n}\, dV(\lambda; x, y) = (A^{m+n}x, y)$$

$$= (A^n x, A^m y) = \int_a^b \lambda^n\, dV(\lambda; x, A^m y).$$

By the uniqueness principle, we have

$$U(\lambda; x, y) = V(\lambda; x, A^m y).$$

Now

$$V(\lambda; x, A^m y) = (F(\lambda)x, A^m y) = (A^m F(\lambda)x, y) = \int_a^b \mu^m\, dV[\mu; F(\lambda)x, y].$$

Consequently,

$$\int_a^\lambda \mu^m\, d(F(\mu)x, y) = \int_a^b \mu^m\, d(F(\mu)F(\lambda)x, y).$$

The integral on the left can be written as

$$\int_a^b \mu^m\, dW(\mu; x, y),$$

where

$$W(\mu; x, y) = \begin{cases} (F(\mu)x, y) & \text{if } a \leqslant \mu \leqslant \lambda \\ (F(\lambda)x, y) & \text{if } \lambda \leqslant \mu \leqslant b. \end{cases}$$

Thus, using the uniqueness principle again, we find that $(F(\mu)F(\lambda)x, y) = (F(\nu)x, y)$, where $\nu = \min(\lambda, \mu)$. This implies the result (6-5).

Lastly, as regards $F(\lambda)$, we assert that $F(\mu)x \to F(\lambda)x$ as $\mu \to \lambda^+$ if $a < \lambda < b$. We write $F(\lambda + 0)x = \lim_{\mu \to \lambda^+} F(\mu)x$, by definition, so that our assertion is

$$F(\lambda + 0)x = F(\lambda)x, \qquad a < \lambda < b. \tag{6-6}$$

To prove this we observe from (6-5) that $F(\mu) - F(\lambda)$ is a symmetric projection if $\lambda < \mu$. Therefore

$$\|F(\mu)x - F(\lambda)x\|^2 = (F(\mu)x - F(\lambda)x, x) = (F(\mu)x, x) - (F(\lambda)x, x).$$

This expression approaches 0 as $\mu \to \lambda^+$ if $a < \lambda < b$, because $V(\lambda; x, y)$ is normalized. Thus (6-6) is proved. Likewise, if $a < \lambda_1 < \lambda_2$,

$$\|F(\lambda_2)x - F(\lambda_1)x\|^2 = (F(\lambda_2)x, x) - (F(\lambda_1)x, x) \to 0$$

as λ_1 and $\lambda_2 \to a^+$, because $(F(\lambda)x, x) = V(\lambda; x, x) \to V(a+0; x, x)$ as $\lambda \to a^+$. Therefore $\lim_{\lambda \to a^+} F(\lambda)x$ exists and defines an operator $F(a+0)$. We now know that

$$(p(A)x, y) = \int_a^b p(\lambda)\, d(F(\lambda)x, y). \tag{6-7}$$

We define $E_\lambda = 0$ if $\lambda < a$, $E_\lambda = I$ if $b \leqslant \lambda$, and $E_\lambda = F(\lambda + 0)$ if $a \leqslant \lambda < b$. This makes $E_\lambda = F(\lambda)$ if $a < \lambda \leqslant b$. It is easy to verify that conditions a, b, and c in Theorem 6.1 are fulfilled. Since E_λ may differ from $F(\lambda)$ at $\lambda = a$, we cannot replace $F(\lambda)$ by E_λ in (6-7). But, if $\alpha < a$ and $b \leqslant \beta$, simple calculations show that

$$\int_b^\beta p(\lambda)\, d(E_\lambda x, y) = 0$$

and that

$$\int_\alpha^a p(\lambda)\, d(E_\lambda x, y) + \int_a^b p(\lambda)\, d\{(E_\lambda x, y) - (F(\lambda)x, y)\} = 0.$$

Thus we obtain (6-1).

The fact that $AE_\mu = E_\mu A$ is easily deduced from (6-1). We have

$$(E_\mu A x, y) = (Ax, E_\mu y) = \int_\alpha^\beta \lambda\, d(E_\lambda x, E_\mu y),$$

$$(AE_\mu x, y) = \int_\alpha^\beta \lambda\, d(E_\lambda E_\mu x, y) = (E_\mu A x, y),$$

because $(E_\lambda E_\mu x, y) = (E_\mu E_\lambda x, y) = (E_\lambda x, E_\mu y)$. Thus $E_\mu A = AE_\mu$. $\square$

We observe that $(E_\lambda x, x)$ is a nondecreasing function of λ. This is equivalent to $E_\lambda \leqslant E_\mu$ if $\lambda < \mu$, which is equivalent to the known relation

$E_\lambda E_\mu = E_\lambda$, by Theorem 2.4. Just as in a previous argument about $F(\lambda)$, we can also conclude that $\lim_{\lambda \to \mu^-} E_\lambda x$ exists and defines an operator $E_{\mu-0}$, which is evidently a symmetric projection.

The next theorem shows that (6-1) is matched by a direct formula for $p(A)$, without intervention of x and y.

Theorem 6.2. *The formula*

$$p(A) = \int_\alpha^\beta p(\lambda)\, dE_\lambda \tag{6-8}$$

holds, the integral on the right being defined in the usual way as a limit of Riemann–Stieltjes sums, with convergence in the norm topology of $L(X)$.

Proof. We recall that $\alpha < m(A), M(A) \leq \beta$. Suppose that $\alpha = \lambda_0 < \lambda_1 < \cdots < \lambda_n = \beta$ and that $\lambda_{k-1} \leq \mu_k \leq \lambda_k$, $k = 1, 2, \ldots, n$. Let

$$B = \sum_{k=1}^n p(\mu_k)[E(\lambda_k) - E(\lambda_{k-1})].$$

Let $\varepsilon_k = \max |p(\lambda) - p(\mu)|$ as λ and μ vary over $[\lambda_{k-1}, \lambda_k]$, and let $\varepsilon = \max(\varepsilon_1, \ldots, \varepsilon_n)$. Now

$$(p(A)x, x) = \sum_{k=1}^n \int_{\lambda_{k-1}}^{\lambda_k} p(\lambda)\, d(E_\lambda x, x),$$

$$(Bx, x) = \sum_{k=1}^n \int_{\lambda_{k-1}}^{\lambda_k} p(\mu_k)\, d(E_\lambda x, x).$$

Since $(E_\lambda x, x)$ is nondecreasing, we see that

$$(p(A)x, x) - (Bx, x) \leq \sum_{k=1}^n \int_{\lambda_{k-1}}^{\lambda_k} \varepsilon\, d(E_\lambda x, x) = \varepsilon(x, x),$$

and hence that $M[p(A) - B] \leq \varepsilon$ [see (2-2)]. In a similar way we see that $-\varepsilon \leq m[p(A) - B]$. But then $\|p(A) - B\| \leq \varepsilon$, by (2-3). Now $\varepsilon \to 0$ as we carry out the usual limiting process in connection with a Stieltjes integral, because of the continuity of p. We see in this way that (6-8) is true in the sense asserted. □

The family of projections E_λ having the properties specified in Theorem 6.1 is unique. This follows as a result of the uniqueness principle for normalized functions of bounded variation which was mentioned in the course of the proof of Theorem 6.1. It can even be shown, though we do not give the details, that, if G_λ is a family of symmetric projections satisfying conditions a, b, and c of Theorem 6.1 and if the equation

$$(Ax, x) = \int_{\alpha}^{\beta} \lambda \, d(G_\lambda x, x)$$

holds for every x, then $G_\lambda = E_\lambda$ for every λ. Because of this uniqueness, the operator A fully determines the family E_λ (and is, of course, fully determined by the family). This family of projections is called the *resolution of the identity* corresponding to A.

Next, we consider the integral

$$\int_{\alpha}^{\beta} f(\lambda) \, dE_\lambda, \tag{6-9}$$

where f is an arbitrary complex-valued continuous function defined on the interval $m(A) \leqslant \lambda \leqslant M(A)$. We extend the definition of f by setting $f(\lambda) = f[m(A)]$ if $\alpha \leqslant \lambda \leqslant m(A)$ and $f(\lambda) = f[M(A)]$ if $M(A) \leqslant \lambda \leqslant \beta$. In order to show that the Stieltjes integral (6-9) exists (as a limit of sums in the topology of $L(X)$), we can proceed as follows: First, suppose that f has real values. Referring back to the proof of Theorem 6.1, we can now see that

$$L(f; x, y) = \int_{\alpha}^{\beta} f(\lambda) \, d(E_\lambda x, y),$$

and that this is a continuous symmetric bilinear form in x and y. Hence there exists a uniquely determined self-adjoint operator, which we denote by $f(A)$, such that

$$(f(A)x, y) = \int_{\alpha}^{\beta} f(\lambda) \, d(E_\lambda x, y).$$

It is now easy, just as in the proof of Theorem 6.2, to show that

$$f(A) = \int_{\alpha}^{\beta} f(\lambda) \, dE_\lambda. \tag{6-10}$$

If f is complex valued, the integral in (6-10) exists, as we see by separating f into its real and imaginary parts. In this case also we denote the value of the integral by $f(A)$.

Theorem 6.3. *The correspondence between the continuous function f and the operator $f(A)$ indicated in formula* (6-10) *has the properties:*

(*a*) $(f+g)(A) = f(A) + g(A)$.
(*b*) $(\alpha f)(A) = \alpha f(A)$.
(*c*) $(fg)(A) = f(A)g(A)$.
(*d*) $f(A)B = Bf(A)$ *if* $B \in L(X)$ *and* $BE_\lambda = E_\lambda B$ *for every* λ.
(*e*) $f(A)$ *is normal, with* $f(A)^*$ *corresponding to the function with values* $\overline{f(\lambda)}$.

(*f*) *$f(A)$ is self-adjoint if f is real-valued, and $f(A) \geqslant 0$ if $f(\lambda) \geqslant 0$ for every λ.*

(*g*) $\|f(A)\| \leqslant \max\{|f(\lambda)| : \lambda$ *varying over* $[m(A), M(A)]\}$.

(*h*) $\|f(A)x\|^2 = \int_\alpha^\beta |f(\lambda)|^2 \, d\|E_\lambda x\|^2$.

Proof. Properties (*a*) and (*b*) are obvious. To prove (*c*) we use an approximating sum for $f(A)$:

$$\sum_{k=1}^{n} f(\lambda_k)[E(\lambda_k) - E(\lambda_{k-1})],$$

and the exactly corresponding sum for $g(A)$. On multiplying the sums, the fact that

$$[E(\lambda_k) - E(\lambda_{k-1})][E(\lambda_j) - E(\lambda_{j-1})] = \begin{cases} E(\lambda_k) - E(\lambda_{k-1}) & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

yields us the sums

$$\sum f(\lambda_k) g(\lambda_k)[E(\lambda_k) - E(\lambda_{k-1})].$$

The limit of this sum is evidently $(fg)(A)$; but it is also $f(A)g(A)$. Thus (*c*) is proved. Property (*d*) may also be proved by using the approximating sums. We leave (*e*) and (*f*) to the reader. For (*g*) let $\|f\| = \max |f(\lambda)|$, and suppose $\alpha = \lambda_0 < \lambda_1 < \cdots < \lambda_n = \beta$. Then

$$x = [E(\lambda_n) - E(\lambda_0)]x = \sum_{k=1}^{n} [E(\lambda_k) - E(\lambda_{k-1})]x,$$

and by the orthogonality relations among the projections $E(\lambda_k) - E(\lambda_{k-1})$, we find that

$$\|x\|^2 = \sum_{k=1}^{n} \|[E(\lambda_k) - E(\lambda_{k-1})]x\|^2. \tag{6-11}$$

Also,

$$\begin{aligned} \|f(A)x\|^2 &= \lim \left(\sum_k f(\lambda_k)[E(\lambda_k) - E(\lambda_{k-1})]x, \sum_j f(\lambda_j)[E(\lambda_j) - E(\lambda_{j-1})]x \right) \\ &= \lim \sum_k |f(\lambda_k)|^2 \|[E(\lambda_k) - E(\lambda_{k-1})]x\|^2 \\ &\leqslant \|f\|^2 \|x\|^2, \end{aligned}$$

by (6-11). This proves (*g*) and at the same time proves (*h*), for

$$\|[E(\lambda_k) - E(\lambda_{k-1})]x\|^2 = ([E(\lambda_k) - E(\lambda_{k-1})]x, x) = \|E(\lambda_k)x\|^2 - \|E(\lambda_{k-1})x\|^2.$$

We can also obtain (*h*) from (*c*) and (*e*). □

Theorem 6.3 exhibits a homomorphism of the algebra of continuous functions on $[m(A), M(A)]$ into the algebra $L(X)$. As in the case of the homomorphism referred to in Theorem V.8.1, the present homomorphism furnishes what may be called an operational calculus.

Next, we shall see that the behaviour of E_λ as a function of λ is closely correlated with $\sigma(A)$.

Theorem 6.4. *If λ_0 is real, then $\lambda_0 \in \rho(A)$ if and only if there is some $\varepsilon > 0$ such that E_λ is constant when $\lambda_0 - \varepsilon \leqslant \lambda \leqslant \lambda_0 + \varepsilon$.*

Proof. We prove the "if" part first. Let $f(\lambda) = \lambda_0 - \lambda$, and define $g(\lambda) = (\lambda_0 - \lambda)^{-1}$ except when $\lambda_0 - \varepsilon \leqslant \lambda \leqslant \lambda_0 + \varepsilon$; the values of g on this latter interval can be assigned arbitrarily, except for the requirement that g be continuous. Then $f(\lambda)g(\lambda) = 1$ except on the interval $[\lambda_0 - \varepsilon, \lambda_0 + \varepsilon]$. Since E_λ is constant on this interval, it follows that

$$f(A)g(A) = \int_\alpha^\beta f(\lambda)g(\lambda)\, dE_\lambda = \int_\alpha^\beta dE_\lambda = I.$$

Therefore $g(A)$ is the inverse of $\lambda_0 - A = f(A)$. This implies that $\lambda_0 \in \rho(A)$.

For the "only if" part we can assume $m(A) < \lambda_0 < M(A)$, because of Theorem 3.2 and the known facts about E_λ for $\lambda < m(A)$ and $M(A) < \lambda$. Suppose, no matter how small ε is, that there are points λ_1 and λ_2 in the interval $[\lambda_0 - \varepsilon, \lambda_0 + \varepsilon]$ such that $\lambda_1 < \lambda_2$ and $E(\lambda_1) \neq E(\lambda_2)$. Since $E(\lambda_2)E(\lambda_1) = E(\lambda_1)$, the range M_1 of $E(\lambda_1)$ is properly contained in the range M_2 of $E(\lambda_2)$, and we can choose an element y in $M_2 \backslash M_1$ and orthogonal to M_1. Then $E(\lambda_2)y = y$ and $E(\lambda_1)y = 0$. If $\lambda \leqslant \lambda_1$, then $E_\lambda y = E_\lambda E(\lambda_1)y = 0$; if $\lambda_2 \leqslant \lambda$, then $E_\lambda y = E_\lambda E(\lambda_2)y = E(\lambda_2)y = y$. Thus, since $E_\lambda y$ is constant if $\lambda \leqslant \lambda_1$ and also if $\lambda_2 \leqslant \lambda$, part (h) of Theorem 6.3 shows that

$$\|(\lambda_0 - A)y\|^2 = \int_{\lambda_1}^{\lambda_2} (\lambda_0 - \lambda)^2\, d\|E_\lambda y\|^2 \leqslant \varepsilon^2 \|y\|^2.$$

The last inequality follows from the fact that $(E_\lambda y, y) = \|E_\lambda y\|^2$ is nondecreasing and never exceeds $\|y\|^2$. We now see that

$$\inf_{\|x\|=1} \|(\lambda_0 - A)x\| = 0,$$

which implies that $\lambda_0 \in \sigma(A)$. □

We see readily, somewhat as in the first part of the proof of Theorem 6.4, that the resolvent of A is given by

$$(\lambda_0 - A)^{-1} = \int_\alpha^\beta \frac{1}{\lambda_0 - \lambda}\, dE_\lambda \tag{6-12}$$

if λ_0 is not on the interval $[m(A), M(A)]$. If λ_0 is a point of $\rho(A)$ on the aforementioned interval, (6-12) is still valid; the singularity of $(\lambda_0-\lambda)^{-1}$ at λ_0 causes no trouble because of the fact that E_λ is constant on a neighborhood of λ_0.

Theorem 6.5. *The point spectrum $P\sigma(A)$ consists of those points μ for which $E_\mu \neq E_{\mu-0}$. The corresponding eigenspace is then the range of the projection $E_\mu - E_{\mu-0}$. The continuous spectrum $C\sigma(A)$ consists of those points μ for which $E_\mu = E_{\mu-0}$ but which are such that E_λ is not constant in any neighborhood of μ.*

Proof. (For the definition of $E_{\mu-0}$ see the remarks following the proof of Theorem 6.1.) From the fact that $E_\lambda E_\mu = E_\lambda$ if $\lambda \leqslant \mu$ we easily deduce that $E_\lambda E_{\mu-0} = E_{\mu-0}$ if $\mu \leqslant \lambda$ and $E_\lambda E_{\mu-0} = E_\lambda$ if $\lambda < \mu$. Hence also $E_{\mu-0}$ and $E_\mu - E_{\mu-0}$ are projections (see § 2, problem 4). Suppose $E_\mu \neq E_{\mu-0}$. Let $y = (E_\mu - E_{\mu-0})x$, with $y \neq 0$. The foregoing relations show that $E_\lambda y = 0$ if $\lambda < \mu$ and $E_\lambda y = y$ if $\mu \leqslant \lambda$. Then, using Theorem 6.3, we have

$$\|(\mu - A)y\|^2 = \int_\alpha^\beta (\mu-\lambda)^2\, d\|E_\lambda y\|^2 = \int_\alpha^\mu (\mu-\lambda)^2\, d\|E_\lambda y\|^2 = 0.$$

The vanishing of the last integral depends on the fact that $(\mu-\lambda)^2 = 0$ at $\lambda = \mu$. We see that μ is an eigenvalue of A.

Suppose, conversely, that $\mu \in P\sigma(A)$, $y \neq 0$, $Ay = \mu y$. Then

$$0 = \int_\alpha^\beta (\mu-\lambda)^2\, d\|E_\lambda y\|^2.$$

We may suppose $M(A) < \beta$, since the integral is independent of β as long as $M(A) \leqslant \beta$. Now $\alpha < m(A) \leqslant \mu < \beta$, so we may choose $\varepsilon > 0$ so that $\alpha < \mu - \varepsilon$ and $\mu + \varepsilon < \beta$. We use the facts that $(\mu-\lambda)^2 \geqslant 0$ and $\|E_\lambda y\|^2 = (E_\lambda y, y)$ is nondecreasing to conclude that

$$\int_\alpha^{\mu-\varepsilon} (\mu-\lambda)^2\, d\|E_\lambda y\|^2 = 0.$$

This integral, however, is not smaller than

$$\varepsilon^2(\|E_{\mu-\varepsilon} y\|^2 - \|E_\alpha y\|^2) = \varepsilon^2 \|E_{\mu-\varepsilon} y\|^2.$$

Therefore $E_{\mu-\varepsilon} y = 0$, whence $E_{\mu-0} y = 0$. We also conclude that

$$\int_{\mu+\varepsilon}^\beta (\mu-\lambda)^2\, d\|E_\lambda y\|^2 = 0,$$

and from this that

$$0 = \varepsilon^2(\|E_\beta y\|^2 - \|E_{\mu+\varepsilon} y\|^2) = \varepsilon^2(\|y\|^2 - \|E_{\mu+\varepsilon} y\|^2),$$

whence $E_{\mu+\varepsilon} y = y$ and, finally, $E_\mu y = y$. Therefore $E_\mu \neq E_{\mu-0}$. Also, $y =$

$(E_\mu - E_{\mu-0})y$, so that the assertion of the theorem is proved as far as $P\sigma(A)$ is concerned. The part about $C\sigma(A)$ follows with the aid of Theorem 6.4. $\square$

The actual determination of the resolution of the identity for a given operator A is not an easy matter, in general. In some comparatively simple cases it may be inferred or conjectured from (6-1) or (6-4). This is so with problems 7 and 8 at the end of this section. A methodical procedure is furnished by the following formula. Suppose $\alpha < m(A)$ and $\varepsilon > 0$. Then

$$\tag{6-13} \tfrac{1}{2}[(E_{\mu-0}x, y) + (E_\mu x, y)] = \lim_{\varepsilon\to 0} \frac{1}{2\pi i}\int_\Gamma ((\lambda - A)^{-1}x, y)\, d\lambda,$$

where Γ is the polygonal line joining $\mu + i\varepsilon$, $\alpha + i\varepsilon$, $\alpha - i\varepsilon$, and $\mu - i\varepsilon$ in that order. See Dunford [3, page 58]. The proof of (6-13) depends on the use of (6-12).

PROBLEMS

1. Let X, A and E_λ be as in Theorem 6.1. Then for each λ there exists a sequence p_n of polynomials such that $p_n(A)x \to E_\lambda x$ for each x. This is readily evident for $\lambda < m(A)$ or $M(A) \leqslant \lambda$. Suppose $\alpha < m(A) \leqslant \mu < M(A) \leqslant \beta$, $0 < h < \beta - \mu$, and define $f(\lambda) = 1$, $1 - h^{-1}(\lambda - \mu)$, or 0 according as $\alpha \leqslant \lambda \leqslant \mu$, $\mu \leqslant \lambda \leqslant \mu + h$, or $\mu + h \leqslant \lambda \leqslant \beta$. Then, using (6-10), we find

$$(f(A)x, x) = (E_\mu x, x) + \int_\mu^{\mu+h} f(\lambda)\, d(E_\lambda x, x).$$

It follows from this and problem 12 in § 3 that $f(A)x \to E_\mu x$ as $h \to 0$. If $h = 1/n$, let the corresponding f be f_n. Choose a polynomial p_n such that $|p_n(\lambda) - f_n(\lambda)| < 1/n$ if $\alpha \leqslant \lambda \leqslant \beta$. With the aid of Theorem 6.3 it can then be shown that $p_n(A)x \to E_\mu x$.
2. For $B \in L(X)$, show from problem 1 that if B commutes with A, that is, if $BA = AB$, then B commutes with each E_λ.
3. If A is a self-adjoint (and hence symmetric) operator in $L(X)$ such that $0 \leqslant A$ (see § 2), then there exists a positive operator B such that $B^2 = A$ and B commutes with every operator in $L(X)$ that commutes with A. Furthermore, if A has an inverse in $L(X)$, then so does B. We call B a *square root* of A. It will follow from Theorem VII.6.3 that each positive operator A has a unique positive square root.
4. If $0 \leqslant A$, $0 \leqslant B$ and $AB = BA$, then $0 \leqslant AB$.
5. Let A be an arbitrary operator in $L(X)$, and let B be the positive square root of the positive operator A^*A. Then $\|Bx\| = \|Ax\|$ for all x.
6. If A in $L(X)$ is self-adjoint, then A may be decomposed in the form $A = A^+ - A^-$, where A^+ and A^- are positive operators with the following properties: (*a*) A^+ and A^- commute with each operator in $L(X)$ that

commutes with A, (*b*) $A^+A^- = 0$, (*c*) $-A^- \leqslant A \leqslant A^+$, (*d*) $A = A^+$ if $0 \leqslant A$. [*Hint.* Consider the function $f(\lambda) = (|\lambda| + \lambda)/2$.]

7. With $X = l^2$, let $Ax = y$ be defined by $\eta_i = \alpha_i \xi_i$, where α_i is real and $\sup_i |\alpha_i| < \infty$. Then $\sigma(A)$ is the closure of the set of the α_i's, and the α_i's form $P\sigma(A)$. The resolution of the identity corresponding to A is defined by $(E_\lambda x, y) = \Sigma_{\alpha i \leqslant \lambda} \xi i \bar{\eta} i$. This means that the matrix representing E_λ is a diagonal matrix with 1 in the ith diagonal position if $\alpha_i \leqslant \lambda$, and 0 there otherwise.
8. Consider $X = L^2(a, b)$, where (a, b) is a finite interval. Let $Ax = y$ be defined by $y(t) = tx(t)$. Then $\sigma(A)$ is the closed interval $[a, b]$, and $P\sigma(A)$ is empty. The resolution of the identity corresponding to A is defined by $E_\lambda x = u$, where (for $a \leqslant \lambda \leqslant b$) $u(t) = x(t)$ if $a \leqslant t \leqslant \lambda$, $u(t) = 0$ if $\lambda < t \leqslant b$.
9. Consider $X = L^2(-\infty, \infty)$, and suppose $a(t) \in \mathscr{L}^1(-\infty, \infty)$. Suppose also that $a(t)$ is real and that $a(-t) =^0 a(t)$. Define $Ax = y$ by

$$y(s) = \int_{-\infty}^{\infty} a(s-t)x(t)\, dt.$$

Then A is self-adjoint and $\|A\| \leqslant \int_{-\infty}^{\infty} |a(t)|\, dt$. Let

$$b(s) = \int_{-\infty}^{\infty} e^{-ist} a(t)\, dt.$$

Then $\sigma(A)$ is the closure of the range of b, and $\lambda \in P\sigma(A)$ if and only if $b(s) = \lambda$ on a set of positive measure. The resolvent of A can be expressed in terms of Fourier transforms and then the resolution of the identity can be computed from (6-13). See Dunford [3, pages 60–64] and Pollard [1].

VI.7 UNITARY OPERATORS

Let X and Y be complex Hilbert spaces. We shall use the same notation for the inner product in both spaces. Recall from § IV.11 that if $A \in L(X, Y)$, then $A^* \in L(Y, X)$.

An operator $U \in L(X, Y)$ is said to be *unitary* if $U^*U = I_X$ (the identity on X) and $UU^* = I_Y$ (the identity on Y).

Theorem 7.1. *Given $U \in L(X, Y)$, the following statements are equivalent*:

(*a*) *U is unitary.*
(*b*) *$\mathscr{R}(U) = Y$ and $(Ux_1, Ux_2) = (x_1, x_2)$ for all $x_1, x_2 \in X$.*
(*c*) *U is an isometric mapping of X onto Y.*

Proof. If $U^*U = I_X$, then $(Ux_1, Ux_2) = (x_1, U^*Ux_2) = (x_1, x_2)$. Also, $\mathscr{R}(U) = Y$ when $UU^* = I_Y$. Thus (*a*) implies (*b*). Taking $x_1 = x_2$ in (*b*), we see that (*b*) implies (*c*). Finally, given (*c*), we note that $(U^*Ux, x) = (Ux, Ux) = \|Ux\|^2 = \|x\|^2 = (x, x)$ for $x \in X$. It follows from problem 2 in § 2 that

$U^*U = I_X$. Given $y \in Y$, we have $y = Ux$ for some x. Then $UU^*y = UU^*(Ux) = U(U^*U)x = Ux = y$. Thus $UU^* = I_Y$, which shows that (c) implies (a). ☐

Theorem 7.1 will be needed in § 8 and § VII.7. In the rest of this section, we assume $X = Y$ and $U \in L(X)$.

Theorem 7.2. *If U in $L(X)$ is unitary, $\sigma(U)$ lies on the circle $|\lambda| = 1$.*

Proof. Since U and U^* are isometric, we see that $\|U\| = \|U^*\| = 1$. Hence $\lambda \in \rho(U) \cap \rho(U^*)$ if $|\lambda| > 1$. We know $0 \in \rho(U)$, because $U^{-1} = U^*$. Suppose $0 < |\lambda| < 1$. Then $\lambda^{-1} \in \rho(U^*)$. Now $\lambda - U = \lambda U(U^* - \lambda^{-1})$, and it follows that $(\lambda - U)^{-1} = \lambda^{-1}(U^* - \lambda^{-1})^{-1}U^*$, so that $\lambda \in \rho(U)$. ☐

Example 1. With $X = L^2(a, b)$ and α real let $Ux = y$ mean $y(t) = e^{i\alpha t}x(t)$. Then U is unitary. The set $\sigma(U)$ consists of the closure of the set of values of $e^{i\alpha t}$, $a < t < b$. The interval may be infinite.

There is a spectral theorem for a unitary operator $U \in L(X)$, corresponding closely to Theorem 6.1. In order to state and prove the theorem, it is convenient to examine the operational calculus $f \mapsto f(U)$ of § V.8. Recall that if $j \geq 0$ and $f(\lambda) = \lambda^j$, then $f(U) = U^j$. Now suppose $j > 0$ and $f(\lambda) = \lambda^{-j}$. Then f is locally analytic on $\sigma(U)$ by Theorem 7.2. Since $\lambda^{-j}\lambda^j \equiv 1$ on a neighborhood of $\sigma(U)$, it follows that $f(U) = (U^j)^{-1}$. (Cf. Theorem V.8.2.) Thus, if

$$p(\lambda) = \sum_{k=-n}^{n} c_k \lambda^k,$$

where the coefficients c_k are complex scalars, then p is locally analytic on $\sigma(U)$ and

$$p(U) = \sum_{k=-n}^{n} c_k U^k,$$

where $U^k = (U^{-1})^{-k} = (U^*)^{-k}$ if $k < 0$.

Since U is unitary, it is clear that $p(U)$ is normal. Hence $\|p(U)\|$ equals the spectral radius of $p(U)$, by Theorem 3.5. Now the spectral mapping theorem (Theorem V.9.5) says that $\sigma(p(U))$ is the set of values assumed by $p(\lambda)$ for $\lambda \in \sigma(U)$. Writing each $\lambda \in \sigma(U)$ as e^{it} for some real t (cf. Theorem 7.2), we have

$$\|p(U)\| = \max\{|p(e^{it})| : e^{it} \in \sigma(U)\}. \tag{7-1}$$

Thus, in studying $\|p(U)\|$, we are led to consider the "trigonometric polynomial"

$$p(e^{it}) = \sum_{k=-n}^{n} c_k e^{ikt}.$$

The relationship between $p(e^{it})$ and the operator $p(U)$ is described in the next two theorems.

Theorem 7.3. *Corresponding to the unitary operator U there is a family of symmetric projections E_t such that*

(a) $E_sE_t = E_tE_s = E_s$ *if* $s \leqslant t$.

(b) $E_{t+0} = E_t$.

(c) $E_t = 0$ *if* $t \leqslant 0$, $E_t = I$ *if* $2\pi \leqslant t$.

(d) $E_tU = UE_t$.

For each x and y, (E_tx, y) is of bounded variation in t, and (E_tx, x) is a nondecreasing function of t. For each trigonometric polynomial $p(e^{it})$,

$$(p(U)x, y) = \int_0^{2\pi} p(e^{it})\, d(E_tx, y).$$

Proof. Let $P[0, 2\pi]$ be that subspace of the complex Banach space $C[0, 2\pi]$ that is obtained by selecting out those functions f such that $f(0) = f(2\pi)$. We need to know how to represent the normed conjugate of $P[0, 2\pi]$ as a subspace of the space $BV[0, 2\pi]$. Each continuous linear functional L on $P[0, 2\pi]$ may be extended with no increase in norm to a continuous linear functional on $C[0, 2\pi]$, by the Hahn–Banach theorem, and hence may be represented in the form

$$L(f) = \int_0^{2\pi} f(t)\, dv(t), \qquad f \in P[0, 2\pi], \tag{7-2}$$

where $v \in BV[0, 2\pi]$ and the total variation of v is $\|L\|$. (Cf. Theorem III.5.5.) Let us define an equivalence relation in $BV[0, 2\pi]$, $w_1 \sim w_2$ meaning that

$$\int_0^{2\pi} f(t)\, dw_1(t) = \int_0^{2\pi} f(t)\, dw_2(t)$$

for each $f \in P[0, 2\pi]$. It turns out that $w \sim 0$ means $w(0) = w(2\pi)$ and $w(t+0) = w(t-0) = w(0+0) = w(2\pi - 0)$ if $0 < t < 2\pi$. We shall now say that w is normalized (relative to $P[0, 2\pi]$) if $w(0) = 0$ and $w(t+0) = w(t)$ when $0 \leqslant t < 2\pi$. The normalized functions form a subspace of $BV[0, 2\pi]$ and each equivalence class contains exactly one normalized function. Thus, much as in § III.5, it can be shown that the conjugate of $P[0, 2\pi]$ is congruent to the subspace of normalized members of $BV[0, 2\pi]$, under the correspondence $L \leftrightarrow v$ exhibited in (7-2).

The proof now proceeds much as in the case of a self-adjoint operator. If $p(e^{it})$ is a trigonometric polynomial, it is an element of $P[0, 2\pi]$, and its norm is $\|p\| = \max\{|p(e^{it})| : 0 \leqslant t \leqslant 2\pi\}$. It is clear from (7-1) that $\|p(U)\| \leqslant \|p\|$. Then, since

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

the mapping $p(e^{it}) \mapsto (p(U)x, y)$ is a continuous linear functional on the trigonometric polynomials viewed as a subspace of $P[0, 2\pi]$. Since this subspace is dense in $P[0, 2\pi]$, $(p(U)x, y)$ determines uniquely a continuous linear functional $L(f; x, y)$ on $P[0, 2\pi]$. Let $V(t; x, y)$ be the normalized function of bounded variation corresponding to this functional. Then

$$(p(U)x, y) = \int_0^{2\pi} p(e^{it})\, dV(t; x, y).$$

The normalization ensures $V(0; x, y) = 0$, and $V(2\pi; x, y) = (x, y)$ follows by putting $p(\lambda) \equiv 1$. We leave to the reader the proof that $V(t; x, y)$ is a continuous bilinear form. Then there exists a family of symmetric operators E_t, defined when $0 \leqslant t \leqslant 2\pi$, such that $V(t; x, y) = (E_t x, y)$. Evidently $E_0 = 0$ and $E_{2\pi} = I$. We define $E_t = 0$ if $t < 0$ and $E_t = I$ if $2\pi < t$. The rest of the proof is similar to the argument in § 6, and we leave it to the reader. □

Next comes the counterpart of Theorem 6.2.

Theorem 7.4. *The formula*

$$p(U) = \int_0^{2\pi} p(e^{it})\, dE_t \tag{7-3}$$

holds, the integral being defined as a limit in the norm topology of $L(X)$.

Proof. It will suffice to prove (7-3) for the special case $p(e^{it}) = e^{int}$; the general case will then follow by linearity. Form a subdivision $0 = t_0 < t_1 < \cdots < t_m = 2\pi$ and choose arbitrary points s_k such that $t_{k-1} \leqslant s_k \leqslant t_k$. Let

$$B = \sum_{k=1}^{m} e^{ins_k}[E(t_k) - E(t_{k-1})],$$

and write $A = U^n - B$. Now $(U^n x, U^n x) = (x, x)$. The operator $P_k = E(t_k) - E(t_{k-1})$ is a projection, and $P_j P_k = 0$ if $j \neq k$. From this we see that

$$BB^* = B^*B = \sum_{k=1}^{m} |e^{ins_k}|^2 P_k = I,$$

so that B is unitary. Therefore

$$\|Ax\|^2 = 2(x, x) - (Bx, U^n x) - (U^n x, Bx). \tag{7-4}$$

Next,

$$(E_s x, U^n x) = \int_0^{s} e^{-int}\, d(E_t x, x). \tag{7-5}$$

This is because

$$(E_s x, U^n x) = (U^{-n} E_s x, x) = \int_0^{2\pi} e^{-int}\, d(E_t E_s x, x),$$

and we can use (*a*) of Theorem 7.3. We can now write

$$\begin{aligned}(Bx, U^n x) &= \sum_{k=1}^{m} e^{ins_k}(E(t_k)x - E(t_{k-1})x, U^n x)\\ &= \sum_{k=1}^{m} e^{ins_k}\int_{t_{k-1}}^{t_k} d(E_s x, U^n x)\\ &= \sum_{k=1}^{m}\int_{t_{k-1}}^{t_k} e^{in(s_k - t)}\, d(E_t x, x).\end{aligned}$$

We used (7-5) at the last step. We now have

$$(U^n x, Bx) + (Bx, U^n x) = 2\sum_{k=1}^{m}\int_{t_{k-1}}^{t_k} \cos n(s_k - t)\, d(E_t x, x).$$

Therefore from (7-4) we see that

$$\|Ax\|^2 = 2\sum_{k=1}^{m}\int_{t_{k-1}}^{t_k} [1 - \cos n(s_k - t)]\, d(E_t x, x).$$

If $\varepsilon > 0$, we can make all the intervals (t_{k-1}, t_k) so short that $1 - \cos n(s_k - t) < \varepsilon/2$ if $t_{k-1} \leqslant t \leqslant t_k$. Then

$$\|Ax\|^2 < \varepsilon \sum_{k=1}^{m}\int_{t_{k-1}}^{t_k} d(E_t x, x) = \varepsilon\|x\|^2.$$

This completes the proof. □

There are developments parallel to the later part of § 6. To each $f \in P[0, 2\pi]$ there corresponds a uniquely determined operator U_f such that

$$(U_f x, y) = \int_0^{2\pi} f(t)\, d(E_t x, y),$$

$$\|U_f x\|^2 = \int_0^{2\pi} |f(t)|^2\, d(E_t x, x),$$

and

$$U_f = \int_0^{2\pi} f(t)\, dE_t.$$

We write U_f rather than $f(U)$, because in a formal sense U_f results from putting U in place of e^{it}, not in place of t [e.g., $U_f = 1 - U^{-2}$ if $f(t) = 1 - e^{-2it}$]. We have $\|U_f\| \leqslant \|f\|$, and the mapping $f \mapsto U_f$ has the properties corresponding to (*a*) to (*d*) of Theorem 6.3. In particular, if $|\lambda| \neq 1$,

$$(\lambda - U)^{-1} = \int_0^{2\pi} \frac{1}{\lambda - e^{it}}\, dE_t.$$

A point e^{is} is in $\sigma(U)$ if and only if s is not interior to an interval of constancy of E_t; if $0<t\leq 2\pi$, $e^{it}\in P\sigma(U)$ if and only if $E_{t-0}\neq E_t$. We omit the details.

The family E_t is called the resolution of the identity for U.

Example 2. The Fourier–Plancherel transform defines a unitary operator F in $L^2(-\infty,\infty)$. The definition of F is $Fx=y$, where

$$y(t) =^0 \frac{1}{\sqrt{2\pi}}\frac{d}{dt}\int_{-\infty}^{\infty}\frac{e^{-ist}-1}{-is}x(s)\,ds.$$

An alternative formula is

$$y(t) =^0 \underset{\alpha\to\infty}{\text{l.i.m.}}\frac{1}{\sqrt{2\pi}}\int_{-\alpha}^{\alpha} e^{-ist}x(s)\,ds,$$

where l.i.m. means "limit in mean," that is, the limit in the metric of $L^2(-\infty,\infty)$. It turns out that $F^4=I$ and, with the aid of this, it is rather easy to show that $\sigma(F)$ consists of the four eigenvalues ± 1, $\pm i$. The resolution of the identity for F is $E_t=0$ if $t<\pi/2$, $E_t=P_1$ if $\pi/2\leq t<\pi$, $E_t=P_1+P_2$ if $\pi\leq t<3\pi/2$, $E_t=P_1+P_2+P_3$ if $3\pi/2\leq t<2\pi$, and $E_t=I$ if $2\pi\leq t$, where

$$P_1=\tfrac{1}{4}(I-iF-F^2+iF^3),$$
$$P_2=\tfrac{1}{4}(I-F+F^2-F^3),$$
$$P_3=\tfrac{1}{4}(I+iF-F^2-iF^3).$$

See Riesz and Sz.-Nagy [1, pages 293–295].

PROBLEMS

1. Let X, Y be Hilbert spaces. An operator A in $L(X,Y)$ is isometric if and only if $A^*A=I$. For an example of an isometric operator that is not unitary, consider $X=Y=\ell^2$ and $A(\xi_1,\xi_2,\ldots)=(0,\xi_1,\xi_2,\ldots)$.
2. If X is finite dimensional, every isometric operator in $L(X)$ is unitary.
3. Given $U\in L(X,Y)$, the following statements are equivalent:
   (*a*) U is unitary,
   (*b*) U^* is unitary, and
   (*c*) U and U^* are isometric linear transformations.
4. Each normal operator A in $L(X)$ may be written in the form $A=UP$, where U is unitary, P is positive, and U and P commute with each other and with A. This factorization of A is called a *polar decomposition* of A, by analogy with the factorization of a complex number λ as $\lambda=e^{i\theta}|\lambda|$, where $|e^{i\theta}|=1$. [*Hint.* Let P be the positive square root of A^*A. For each $y=Px\in\mathcal{R}(P)$, define $Uy=Ax$. Show that U is a well-defined isometric mapping from $\mathcal{R}(P)$ onto $\mathcal{R}(A)$. Then extend U in an appropriate manner to all of X.]

5. If A in $L(X)$ has an inverse in $L(X)$, then A admits a polar decomposition $A = UP$, where U is unitary and P is positive. If $UP = PU$, then A is normal. For an application of this polar decomposition, see problem 12 of § VII.7.
6. An operator V in $L(X)$ is said to be a *partial isometry* if the restriction of V to $\mathcal{N}(V)^{\perp}$ is an isometry. Each operator A in $L(X)$ admits a factorization $A = VP$, where V is a partial isometry and P is a positive operator. [*Hint.* Modify the construction of U suggested in problem 4.]

VI.8 UNBOUNDED SELF-ADJOINT OPERATORS

In this section we shall describe certain important properties of unbounded symmetric and self-adjoint operators on a complex Hilbert space X. The purpose of the section is to orient the reader who desires to pursue these matters elsewhere. There is not space in the book for an extensive treatment of these topics. The subject matter naturally leads into applications to symmetric differential operators, and there are important applications to quantum mechanics.

Throughout the section, A is a densely defined linear operator on X. Recall from § IV.11 that if y and z satisfy $(Ax, y) = (x, z)$ for all $x \in \mathcal{D}(A)$, then $y \in \mathcal{D}(A^*)$ and $A^*y = z$. This defines A^*. We showed in § IV.11 that A^* is closed. If $\mathcal{D}(A^*)$ is dense in X, then A^{**} is defined and is an extension of A (in symbols, $A \subset A^{**}$). This occurs if and only if A has a closed linear extension, and then A^{**} is the minimal such extension. (See problem 1.) Hence $A = A^{**}$ if A is closed.

If A is symmetric, that is, if $(Ax, y) = (x, Ay)$ for $x, y \in \mathcal{D}(A)$, then $y \in \mathcal{D}(A^*)$ and $A^*y = Ay$. Thus A^* is an extension of A. If $A^* = A$ we say that A is *self-adjoint*. In Example 2 below we shall examine a closed symmetric operator that is not self-adjoint.

If A is symmetric, the approximate point spectrum of A (defined in § V.4) is confined to the real axis, but there may be nonreal points in $R\sigma(A)$: such points, if any, must be in state III_1 (see § IV.10). Furthermore, if $\lambda \in \sigma(A)$ and $\operatorname{Im} \lambda \neq 0$, then all points on the same side of the real axis as λ are in $R\sigma(A)$ (problem 2). Thus, to determine if $\sigma(A)$ contains any nonreal points, it suffices to examine the points $\pm i$.

Theorem 8.1. *If A is symmetric, then $A + i$ and $A - i$ are both one-to-one operators with continuous inverses. If, in addition, A is a closed operator, then $\mathcal{R}(A+i)$ and $\mathcal{R}(A-i)$ are closed subspaces of X.*

Proof. (The proof does not use the fact that $\mathcal{D}(A)$ is dense.) For $x \in \mathcal{D}(A)$,

$$\begin{aligned}\|Ax + ix\|^2 &= \|Ax\|^2 + (Ax, ix) + (ix, Ax) + \|ix\|^2 \\ &= \|Ax\|^2 - i(Ax, x) + i(x, Ax) + \|x\|^2 \\ &= \|Ax\|^2 + \|x\|^2,\end{aligned}$$

because A is symmetric. A similar calculation holds with i replaced with $-i$. Thus

$$\|Ax \pm ix\|^2 = \|Ax\|^2 + \|x\|^2. \tag{8-1}$$

In particular, $\|(A \pm i)x\| \geq \|x\|$, which shows that $A + i$ and $A - i$ have continuous inverses. If A is closed, then so are $A + i$ and $A - i$, and hence $\mathscr{R}(A + i)$ and $\mathscr{R}(A - i)$ are closed, by Theorem IV.5.8. □

From (8-1) we note that $\|(A - i)x\| = \|(A + i)x\|$, and so $\|(A - i)(A + i)^{-1}y\| = \|y\|$, where $y = (A + i)x$. Thus the linear mapping V defined by

$$V = (A - i)(A + i)^{-1}$$

is an isometry from its domain $\mathscr{R}(A + i)$ onto $\mathscr{R}(A - i)$. This operator V is called the *Cayley transform* of the symmetric operator A.

Theorem 8.2. *If A is self-adjoint, then its Cayley transform V is a unitary operator on X.*

Proof. Since V is an isometry, it suffices, by Theorem 7.1, to show that both the domain and range of V equal X, that is, $\mathscr{R}(A + i) = \mathscr{R}(A - i) = X$. Now the range of a closed operator is all of X if and only if its adjoint has a continuous inverse. (This was proved for conjugates of operators on Banach spaces in § IV.9. For adjoints, see problem 4, § IV.11.) Since $A = A^*$, both A and $A + i$ are closed operators, and $(A + i)^* = A^* - i = A - i$. (See problem 1, § IV.11.) Hence $\mathscr{R}(A + i) = X$ because the adjoint $A - i$ has a continuous inverse, by Theorem 8.1. Similarly, $\mathscr{R}(A - i) = X$ because $A + i$ has a continuous inverse. □

We observe from the proof of Theorem 8.2 that both i and $-i$ belong to $\rho(A)$. It follows from the remarks preceding Theorem 8.1 that the spectrum of a self-adjoint operator must be a subset of the real axis.

When A is self-adjoint, we can use the spectral representation of its Cayley transform V to obtain a spectral representation of A and a generalization of the results of § 6. Applying Theorem 7.3 to V, we can write

$$(Vx, y) = \int_0^{2\pi} e^{i\theta}\, d(F_\theta x, y),$$

where F_θ is the resolution of the identity for V. The relation between V and A then permits us to deduce (after some details that we omit) the relation

$$(Ax, y) = -\int_0^{2\pi} \cot\frac{\theta}{2}\, d(F_\theta x, y), \qquad x \in \mathscr{D}(A),\ y \in X. \tag{8-2}$$

This integral is improper at 0 and 2π. We then define $E_\lambda = F_\theta$, where

$\lambda = -\cot(\theta/2)$. Then (8-2) becomes

$$(Ax, y) = \int_{-\infty}^{\infty} \lambda \, d(E_\lambda x, y).$$

The family of orthogonal projections E_λ has properties much as in Theorem 6.1, but we have $E_\lambda x \to 0$ as $\lambda \to -\infty$, and $E_\lambda x \to x$ as $\lambda \to +\infty$, in place of property (c) in this earlier theorem. The domain of A consists exactly of those x for which

$$\int_{-\infty}^{\infty} \lambda^2 \, d(E_\lambda x, x)$$

is convergent; the value of the integral is then $\|Ax\|^2$. The behavior of E_λ near a particular λ_0 indicates the classification of this point in $\rho(T)$ or $\sigma(T)$ just as in the bounded case. Finally, an operational calculus can be developed, generalizing Theorem 6.3.

Our final theorem shows that every symmetric (and hence every self-adjoint) extension of A may be described as a restriction of A^* to some appropriate subspace of $\mathscr{D}(A^*)$.

Theorem 8.3. *If A is symmetric and if A_1 is a symmetric extension of A, then $A \subset A_1 \subset A_1^* \subset A^*$.*

Proof. By hypothesis, we have $A \subset A_1 \subset A_1^*$. Given $y \in \mathscr{D}(A_1^*)$, we have $(Ax, y) = (A_1 x, y) = (x, A_1^* y)$ for all $x \in \mathscr{D}(A)$. Hence $y \in \mathscr{D}(A^*)$ and $A^* y = A_1^* y$. Thus $A_1^* \subset A^*$. $\square$

An immediate corollary of Theorem 8.3 is that a self-adjoint operator has no proper symmetric extensions.

It is possible to make a thorough study of the subspaces of $\mathscr{D}(A^*)$ that correspond to symmetric extensions of a symmetric operator A. For example, see Dunford and Schwartz [2, pages 1224–1239]. Crucial in the investigation of whether A has self-adjoint extensions are the subspaces $\mathscr{R}(A-i)^\perp$ and $\mathscr{R}(A+i)^\perp$. The dimensions of these subspaces are called *the deficiency indices of A*. It follows from the proof of Theorem 8.2 that when A is self-adjoint, the deficiency indices are both zero. Conversely, if the indices are zero, then the Cayley transform of A is unitary, and one can show that A is self-adjoint (problem 3). It can also be shown that a (densely defined) symmetric operator A possesses at least one self-adjoint extension if and only if the deficiency indices of A are equal.

We conclude this section with several examples of symmetric and self-adjoint operators.

Example 1. The operator A of Example 1, § 2 is self-adjoint. Its spectrum is made up of the points $\lambda_n = n$, $n = 0, \pm 1, \pm 2, \ldots$; λ_n is an eigenvalue corresponding to the eigenvector u_n, where $u_n(s)=(2\pi)^{-1/2}e^{ins}$. In this case the resolution of the identity corresponding to A is given by

$$[E(\lambda_n)-E(\lambda_n-0)]x=(x, u_n)u_n.$$

For $x\in\mathscr{D}(A)$ the formula

$$Ax=\sum_{-\infty}^{\infty}(Ax, u_n)u_n=\sum_{-\infty}^{\infty}n(x, u_n)u_n$$

is just the same as the formula

$$Ax=\int_{-\infty}^{\infty}\lambda\, dE_\lambda x,$$

and $x\in\mathscr{D}(A)$ is equivalent to the convergence of

$$\int_{-\infty}^{\infty}\lambda^2\, d(E_\lambda x, x)=\sum_{-\infty}^{\infty}n^2|(x, u_n)|^2.$$

The reader should refer to the discussion of Example 2, § V.9; there, however, E_n is the operator now denoted by $E(\lambda_n)-E(\lambda_n-0)$. See also Stone [1, pages 428–435].

Example 2. Let $X=L^2(0, 1)$, and let D be the set of those x such that $x(t)$ is absolutely continuous on $[0, 1]$ and x' is in X. Let $\mathscr{D}(A)=\{x\in D: x(0)=x(1)=0\}$ and, for $x\in\mathscr{D}(A)$, define $(Ax)(t)=-ix'(t)$. It is easily seen that A is densely defined. We shall show that the adjoint operator A^* is defined by $\mathscr{D}(A^*)=D$ and $A^*y=-iy'$ when $y\in D$. (The calculations will be similar to those in § IV.8.) For $x\in\mathscr{D}(A)$ and $y\in D$, an integration by parts yields

$$(Ax, y)=\int_0^1 -ix'(t)\overline{y(t)}\, dt=i\int_0^1 x(t)\overline{y'(t)}\, dt=(x, -iy'). \tag{8-3}$$

Hence $y\in\mathscr{D}(A^*)$ and $A^*y=-iy'$. Also, (8-3) proves that A is symmetric. It remains only to show that $\mathscr{D}(A^*)\subset D$.

Suppose that $y\in\mathscr{D}(A^*)$. Let us define $z\in X$ by

$$z(s)=\int_0^s (A^*y)(t)\, dt+\alpha \tag{8-4}$$

where α is a constant to be determined. Now if $x\in\mathscr{D}(A)$,

$$\begin{aligned}\int_0^1 -ix'(t)\overline{y(t)}\, dt=(Ax, y)=(x, A^*y)&=\int_0^1 x(t)\overline{(A^*y)(t)}\, dt\\ &=x(1)\overline{z(1)}-x(0)\overline{z(0)}-\int_0^1 x'(t)\overline{z(t)}\, dt\\ &=-\int_0^1 ix'(t)\overline{[iz(t)]}\, dt.\end{aligned}$$

Thus

$$\int_0^1 x'(t)\overline{[y(t)-iz(t)]}\,dt=0 \tag{8-5}$$

for every $x \in \mathscr{D}(A)$. In particular, (8-5) holds when x is defined by

$$x(s)=\int_0^s [y(t)-iz(t)]\,dt, \qquad 0\leqslant s\leqslant 1,$$

if we now choose α in (8-4) so that $x(1)=0$. In this case (8-5) becomes

$$\int_0^1 |y(t)-iz(t)|^2\,dt=0,$$

which implies that $y(s)={}^0iz(s)$. Since $A^*y \in X$, it follows from (8-4) that $y \in D$. Thus A^* is a proper extension of A, with $\mathscr{D}(A^*)=D$. If A_1 denotes the restriction of A^* to $\{x \in D : x(0)=x(1)\}$, it turns out that A_1 is a self-adjoint extension of A. See problem 5.

Example 3. Let $X=L^2(-\infty,\infty)$, and let D be the set of those x in X such that $x(t)$ is absolutely continuous on every finite interval and x' is in X. Then the operator A defined for $x \in D$ by $(Ax)(t)=-ix(t)$ is self-adjoint. A proof of this fact may be constructed along the lines of Example 2, by replacing the interval $[0,1]$ in that example by an arbitrary finite interval $[a,b]$ with variable end points and by using the fact that, for $x \in D$,

$$\lim_{a\to-\infty} x(a)=\lim_{b\to\infty} x(b)=0.$$

The operator A is closely related to the operator T defined by $(Tx)(t)=tx(t)$ for all $x \in L^2(-\infty,\infty)$ with the property that $tx(t)$ also determines an element of $L^2(-\infty,\infty)$. It can be shown that $A=FTF^{-1}$, where F is the Fourier–Plancherel operator of Example 2, § 7. The entire real axis belongs to $\sigma(T)$, and it is all continuous spectrum. Hence the same is true for A. The resolution of the identity $\{E_\lambda\}$ for T is given by $E_\lambda x=y$, where $y(t)=x(t)$ if $t\leqslant\lambda$, and $y(t)=0$ if $t>\lambda$. For A the resolution of the identity $\{E_\lambda\}$ is given by

$$(E_\lambda-E_\mu)x(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{-i\lambda(s-t)}-e^{-i\mu(s-t)}}{i(s-t)}\,x(s)\,ds.$$

The operational calculus for A may be described using the transformation F. Suppose f is a bounded function in $\mathscr{L}^2(-\infty,\infty)$, and let g be the Fourier–Plancherel transform of f. Then the operator $f(A)$ corresponding to f is given by $f(A)x=y$, where

$$y(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(t-s)x(s)\,ds.$$

For details see Akhiezer and Glazman [1, pages 103–113, and 2, pages 84–86] and Stone [1, pages 441–446].

PROBLEMS

1. Let A be a densely defined linear operator in X.
   a. If $\mathscr{D}(A^*)$ is dense in X, then A^{**} is a closed extension of A.
   b. If A has a closed linear extension, then $\mathscr{D}(A^*)$ is dense in X and A^{**} is the minimal such extension. [*Hint.* See Theorem IV.8.1, problem 9, § IV.5, and problem 2, § IV.11.]
2. Let A be a symmetric operator.
   a. The approximate point spectrum of A is a subset of the real axis. [*Hint.* Show that if λ is nonreal, then $\lambda - A$ has a continuous inverse defined on $\mathscr{R}(\lambda - A)$.]
   b. If λ is nonreal and $\lambda \in \sigma(A)$, then all points on the same side of the real axis as λ belong to $R\sigma(A)$.
3. Let A be a symmetric operator whose Cayley transform V is a unitary operator in $L(X)$.
   a. $\mathscr{D}(A)$ is dense in X. [*Hint.* For $x \in \mathscr{D}(A)$, show that $(I - V)(A + i)x = 2ix$ and study the operator $I - V$.]
   b. $A^* = A$. [*Hint.* Given $y \in \mathscr{D}(A^*)$, show that there exists $x \in \mathscr{D}(A)$ such that $(A^* + i)y = (A + i)x = (A^* + i)x$. Then deduce that $y = x$.]
4. A closed symmetric operator A is densely defined and self-adjoint if and only if $\sigma(A)$ is confined to the real axis.
5. a. If A is the operator in Example 2, then $A^{**} = A$.
   b. If A_1 is the operator mentioned at the end of Example 2, then A_1 is self-adjoint. [*Hint.* For $x \in \mathscr{D}(A_1)$ and $y \in \mathscr{D}(A_1^*)$, examine $(A_1 x, y) - (x, A_1^* y)$.]
6. Consider the Hilbert space H^2 (cf. problem 1, § II.7), and define an operator A by

$$(Af)(z) = i\frac{1+z}{1-z} f(z)$$

for every $f \in H^2$ such that $Af \in H^2$. Then A is symmetric and possesses no self-adjoint extension. Also, the Cayley transform of A is the operator V defined by $(Vf)(z) = zf(z)$ for all $f \in H^2$.

VII || BANACH ALGEBRAS

The spectral theory of the preceding two chapters arose historically out of problems associated with specific linear operators. Today it is common to view spectral theory of *bounded* linear operators in the context of an abstract Banach algebra. This approach is not only very elegant but it also provides new and powerful tools for use in other parts of mathematics such as harmonic analysis.

The theory of Banach algebras has developed extensively since the pioneering work of I. M. Gelfand in 1939 and the early 1940s. We shall discuss only the more elementary parts of the theory. Our choice of material is partially influenced by our desire to prove the spectral theorem for a normal operator. Along the way to this result, we hope to provide a good introduction to the theory of commutative Banach algebras and to convey the beauty and power of the subject.

A *Banach algebra* is a complex Banach space A in which an operation of multiplication is defined, subject to the following axioms (for all $x, y, z \in A, \lambda \in C$):

1. $(xy)z = x(yz)$.
2. $x(y+z) = xy + xz$ and $(y+z)x = yx + zx$.
3. $\lambda(xy) = (\lambda x)y = x(\lambda y)$.
4. $\|xy\| \leq \|x\| \, \|y\|$.

If, in addition, $xy = yx$ for every pair x, y in A, then A is said to be *commutative*.

A *subalgebra* of A is a linear manifold A_1 in A such that xy is in A_1 whenever x and y are both in A_1. If the subalgebra A_1 has the stronger property that zy and yz are in A_1 whenever $y \in A_1$ and $z \in A$, then A_1 is called an *ideal*.

To simplify the discussion in this chapter, we shall tacitly assume that a Banach algebra has a *unit* (i.e., a multiplicative identity) e such that $ex = xe = x$ for all x and such that $\|e\| = 1$. However, since some important Banach algebras lack a unit (see Example 4, § 1), we shall indicate in § 3 and § 5 and in some exercises how the general theory would proceed in the absence of a unit. Often the theory we develop may be carried over to a Banach algebra B that has no unit by the simple device of *adjoining a unit* to B as follows: Let $A = \{(x, \lambda) : x \in B,\ \lambda \in C\}$, define the linear structure of A componentwise, define multiplication by

$$(x, \lambda)(y, \mu) = (xy + \lambda y + \mu x, \lambda\mu),$$

and define a norm on A by

$$\|(x, \lambda)\| = \|x\| + |\lambda|.$$

It is readily checked that A is a Banach algebra with a unit $e = (0, 1)$, and the mapping $x \mapsto (x, 0)$ is an isometric isomorphism of B onto a closed subalgebra of A.

Many of the important Banach spaces of classical analysis are also Banach algebras, and yet it was not until the middle 1930s that this richer structure began to be systematically exploited. Some of these Banach algebras are described in § 1. The examples there deserve careful study because they are the prototypes for all of the Banach algebras considered in this chapter. The final example in § 1 shows that, in one sense, all Banach algebras "are" algebras of operators. This result is useful, for it allows us in § 2 to carry over some key facts about operators to the spectral theory for an abstract Banach algebra. However, this "representation" of a Banach algebra as an algebra of operators on a Banach space has the serious disadvantage that the underlying Banach space is far too large. Much of this chapter is devoted to the search for more useful representations of Banach algebras.

The theory begins in § 2 with the Gelfand–Mazur theorem (Theorem 2.1). This result enables us to identify each maximal ideal in a commutative Banach algebra A with a nonzero multiplicative linear functional on A (Theorem 4.3). Using this, we establish the fundamental Gelfand representation theorem (Theorem 4.6) which characterizes those commutative Banach algebras that are isomorphic to an algebra of continuous functions on a compact Hausdorff space. Several applications of the representation theorem are discussed in § 5. The main application, however, is to the Gelfand–Naimark theorem (Theorem 6.2), which states that a commutative B^*-algebra A is isometrically isomorphic to $C(\mathfrak{M})$, where $\mathfrak{M}$ is a compact Hausdorff space. If A is the B^*-algebra generated by a normal operator T on a Hilbert space, $\mathfrak{M}$ can be taken to be the spectrum of T (Theorem 7.1). Finally, we show how this form of the Gelfand–Naimark theorem leads to two versions of the spectral theorem for a normal operator (Theorems 7.3 and 7.5).

VII.1 EXAMPLES OF BANACH ALGEBRAS

Concrete Banach algebras commonly consist either of complex-valued functions (or possibly equivalence classes of functions) or of linear operators. The operations of addition and scalar multiplication are defined pointwise, but there are three different types of multiplication—pointwise multiplication of functions, convolution of functions, and composition of operators. The corresponding algebras are often referred to as function algebras, group algebras, and operator algebras.

Example 1. The most important example of a commutative Banach algebra is the Banach space $C(X)$ of all bounded continuous complex-valued functions on a Hausdorff topological space X, with multiplication defined pointwise. The constant function $f(t) \equiv 1$ is the unit of the algebra. We shall have more to say later about the situation when X is compact.

Example 2. When X is the closed unit disc Δ in the complex plane, an important subalgebra of $C(\Delta)$ is the disc algebra $A(\Delta)$ of all f in $C(\Delta)$ that are analytic in the open disc. As a normed space, $A(\Delta)$ was introduced in Example 9, § II.2. It is complete in the supremum norm (cf. problem 1, § II.4). The study of the disc algebra provides a particularly good opportunity to observe the interplay of classical analysis (of analytic functions) and abstract analysis (in this case the general theory of commutative Banach algebras). See Examples 3 and 4 in § 4 and the ensuing discussion.

Example 3. Let W be the set of all continuous complex-valued functions x on $[0, 2\pi)$ with absolutely convergent Fourier series:

$$x(t) = \sum_{-\infty}^{\infty} a_n e^{int}, \qquad \sum_{-\infty}^{\infty} |a_n| < \infty. \tag{1-1}$$

W is a normed linear space when addition and scalar multiplication are defined pointwise and the second sum in (1-1) is used as a norm. Since the representation in (1-1) is unique, there is an obvious isometric isomorphism between W and the space $\ell^1(Z)$ of absolutely convergent "doubly infinite" sequences $\{a_n\}_{-\infty}^{+\infty}$. It is easily checked that $\ell^1(Z)$ is complete; therefore W is a Banach space. (Here Z denotes the set of all integers.)

We can make W into a Banach algebra by defining multiplication pointwise. It will be instructive to compute the Fourier series for a product xy, given the series $\Sigma_m a_m e^{imt}$ and $\Sigma_n b_n e^{int}$ for x and y, respectively. We have

$$\begin{aligned}
(xy)(t) &= x(t)y(t) \\
&= \left(\sum_m a_m e^{imt}\right)\left(\sum_n b_n e^{int}\right) = \sum_n \left(\sum_m a_m e^{imt}\right) b_n e^{int} \\
&= \sum_n \left(\sum_m a_{m-n} e^{i(m-n)t}\right) b_n e^{int} = \sum_n \sum_m a_{m-n} b_n e^{imt},
\end{aligned}$$

because the series are absolutely convergent and because the interior sum on

m is from $-\infty$ to $+\infty$. Interchanging the order of summation, we obtain

$$(xy)(t)=\sum_m c_m e^{imt}, \quad \text{where} \quad c_m=\sum_n a_{m-n}b_n. \tag{1-2}$$

It follows that xy has an absolutely convergent Fourier series, and

$$\begin{aligned}\|xy\|=\sum_m |c_m| &\leq \sum_m \sum_n |a_{m-n}|\,|b_n| \\ &= \sum_n \sum_m |a_{m-n}|\,|b_n| = \sum_n \left(\sum_m |a_m|\right)|b_n| = \|x\|\,\|y\|.\end{aligned}$$

Thus the norm satisfies the inequality required of a Banach algebra.

The algebra W is often called the *Wiener algebra* because of the important work on absolutely convergent Fourier series done by N. Wiener in the early 1930s. We shall say more about this in § 5.

The space $\ell^1(Z)$ mentioned above is also a Banach algebra. In fact, two different multiplicative structures may be imposed on it (but not at the same time, of course). The elements of $\ell^1(Z)$ may be viewed as functions defined on the integers. Hence it is easy to see that $\ell^1(Z)$ would become a Banach algebra under pointwise multiplication of these functions. However, this structure is seldom used. The usual definition of multiplication is derived from the congruence between $\ell^1(Z)$ and the Wiener algebra: if $f, g \in \ell^1(Z)$, we define $f * g$ to be the function on Z such that

$$(f * g)(m)=\sum_n f(m-n)g(n), \qquad m \in Z. \tag{1-3}$$

Comparing this definition with (1-2), we see that the isometric isomorphism between $\ell^1(Z)$ and W is actually an isometric *algebra* isomorphism (i.e., it also preserves multiplication), and hence $\ell^1(Z)$ is a Banach algebra. The unit for $\ell^1(Z)$ is the function whose value at 0 is 1 and whose value at the other integers is 0.

The product in (1-3) is called the *convolution* of f and g, and $\ell^1(Z)$ is called a convolution algebra or a *group algebra.* The "group" in this case is the additive group Z of the integers. (The term "group algebra" originated in the study of finite groups, where an associated algebra or "group ring" was used to investigate the structure of the group.) The next example presents another group algebra, one that is more typical of many group algebras currently studied in abstract harmonic analysis.

Example 4. Let $L^1(R)$ be the Banach space of (equivalence classes of) complex-valued Lebesgue integrable functions on the real line (cf. Example 5, § II.2). It is easily seen that the pointwise product of two integrable functions need not be integrable. The operation of convolution, however, does provide a suitable multiplicative structure for $L^1(R)$. For $f, g \in L^1(R)$ we define

$$(f * g)(t) = \int_{-\infty}^{\infty} f(t-s)g(s)\,ds. \tag{1-4}$$

This is the continuous analogue of the multiplication defined in $\ell^1(\boldsymbol{Z})$. However, there is no unit in $L^1(\boldsymbol{R})$ under this multiplication. The proof that the convolution product has the desired properties involves calculations similar to those for $\ell^1(\boldsymbol{Z})$ and W. For example, integrating over $\boldsymbol{R}$, we have

$$\begin{aligned}\|f\|\|g\| &= \int |f(t)|\,dt \cdot \int |g(s)|\,ds \\ &= \int \left[\int |f(t-s)|\,dt\right]|g(s)|\,ds = \int \left[\int |f(t-s)g(s)|\,dt\right] ds.\end{aligned}$$

Since this iterated integral is finite, by Tonelli's theorem its value is unchanged when the order of integration is reversed. But

$$\int \left[\int |f(t-s)g(s)|\,ds\right] dt \geqslant \int |(f * g)(t)|\,dt = \|f * g\|.$$

Thus $f * g$ is in $L^1(\boldsymbol{R})$ and $\|f * g\| \leqslant \|f\|\,\|g\|$. (One step omitted in this argument was the verification that the mapping $(s, t) \mapsto f(t-s)g(s)$ is a measurable function on $\boldsymbol{R} \times \boldsymbol{R}$. See Taylor [5, page 333], for example.) Associativity of convolution is proved similarly using Tonelli's theorem. Finally, it is apparent that convolution obeys the distributive laws. Thus $L^1(\boldsymbol{R})$ is a Banach algebra. (It is the group algebra of the additive group $\boldsymbol{R}$ of real numbers.) Furthermore, it is commutative. This follows from the simple change of variable $u = t - s$ in (1-4) and the fact that, for Lebesgue measure, $du = ds$.

Example 5. Let H be a Hilbert space. Much of the importance of the familiar algebra $L(H)$ lies in its rich supply of interesting and useful subalgebras, both commutative and noncommutative. Chief among these are the self-adjoint subalgebras. A subalgebra of $L(H)$ is *self-adjoint* if it contains the identity and the adjoint of each of its elements; if the algebra is also closed in the norm of $L(H)$, then it is called a *C^*-algebra*. Obviously, $L(H)$ is a C^*-algebra. Another C^*-algebra is the algebra of all compact linear operators in $L(H)$. (This fact follows immediately from the first three theorems of § V.7.)

If M is an arbitrary subset of $L(H)$, the *commutant* of M is the set M' of all operators in $L(H)$ that commute with each operator in M. It is easy to see that M' is a Banach algebra (under the operator norm) containing the identity operator. If M is self-adjoint, then M' is a C^*-algebra. The set $M'' = (M')'$ is called the *bicommutant* of M. Clearly $M \subset M''$, since everything in M commutes with the elements in M'. If M is a self-adjoint subalgebra of $L(H)$

such that $M = M''$, then M is called a *von Neumann algebra* or a *W^*-algebra.* The theory of W^*-algebras is extensively developed, but it lies beyond the scope of this chapter.

Example 6. This example will be connected with the study of normal operators in § 7. It is the prototype for all commutative C^*-algebras. Let $(X, \mathcal{S}, \mu)$ be a measure space, where X is a locally compact Hausdorff space, $\mathcal{S}$ is the σ-field of all Borel subsets of X, and μ is a finite regular Borel measure on X. (See page 151 for a definition of such a measure.) Let $B(X)$ denote the set of all complex-valued bounded Borel measurable functions on X. Then $B(X)$ is a Banach algebra under pointwise operations and the supremum norm, $\|\cdot\|_\infty$. Let $L^2(X, \mu)$ be the Hilbert space of (equivalence classes of) measurable functions f such that $\int_X |f|^2\, d\mu < \infty$. Then for $h \in B(X)$, we can define a multiplication operator M_h on $L^2(X, \mu)$ by $M_h f = hf$, $f \in L^2(X, \mu)$. Clearly M_h is a linear operator, and the inequality

$$\|M_h f\|_2 = \left(\int_X |hf|^2\, d\mu\right)^{1/2} \leq \left(\int_X \|h\|_\infty^2 |f|^2\, d\mu\right)^{1/2} = \|h\|_\infty \|f\|_2$$

shows that M_h is continuous, with $\|M_h\| \leq \|h\|_\infty$. It is readily seen that $\{M_h : h \in B(X)\}$ is a commutative subalgebra of the algebra of all bounded linear operators on $L^2(X, \mu)$. Further, $\{M_h : h \in B(X)\}$ is a self-adjoint algebra. To see this, take $f, g \in L^2(X, \mu)$ and $h \in B(X)$. Then

$$(M_h f, g) = \int_X (hf)\bar{g}\, d\mu = \int_X f(\overline{\bar{h}g})\, d\mu = (f, M_{\bar{h}} g).$$

Since this holds for all $f, g \in L^2(X, \mu)$, we conclude that $M_h^* = M_{\bar{h}}$. Finally, the algebra of multiplication operators is closed in the operator norm and so is a C^*-algebra. The first step in the proof is to replace $B(X)$ by the Banach algebra $L^\infty(X, \mu)$ of all equivalence classes of μ-essentially bounded measurable functions on X, under the essential supremum norm. (See Example 6, § II.2.) One may view $L^\infty(X, \mu)$ as the quotient space of $B(X)$ modulo the subspace of functions that vanish μ-almost everywhere. Clearly $M_h = M_k$ if $h, k \in B(X)$ and $h = k$ μ-almost everywhere. Then it can be shown that $\{M_h : h \in L^\infty(X, \mu)\}$ is isometrically isomorphic to $L^\infty(X, \mu)$. See problem 2.

Example 7. Let A be a Banach algebra, and for each $a \in A$, define a mapping L_a by

$$L_a(x) = ax, \qquad x \in A.$$

Clearly L_a is a linear operator from A (considered as a Banach space) into itself. Also, $\|L_a(x)\| = \|ax\| \leq \|a\| \|x\|$, which shows that L_a is a bounded operator with

$$\|L_a\| \leq \|a\|. \tag{1-5}$$

It is easily seen that the correspondence $a \mapsto L_a$ is a linear mapping that also preserves the operation of multiplication. For example, given $a, b, x \in A$, we have $L_{ab}(x) = (ab)x = a \cdot L_b(x) = L_a L_b(x)$; hence $L_{ab} = L_a L_b$. The mapping $a \mapsto L_a$ is called the *left regular representation of A*.

Now suppose that A has a unit e with $\|e\| = 1$. Then $\|L_a\| = \sup\{\|L_a(x)\| : \|x\| = 1\} \geq \|L_a(e)\| = \|a\|$. Hence, by (1-5),

$$\|L_a\| = \|a\|, \qquad a \in A.$$

Thus the left regular representation of A is an isometric algebra isomorphism of A onto a (closed) subalgebra of $L(A)$. Furthermore, this representation preserves the property of invertibility. An element a in A is said to be *invertible in A* if there exists an inverse element a^{-1} in A such that $a \cdot a^{-1} = a^{-1} \cdot a = e$. If a is invertible, then the operator L_a is obviously invertible in the algebra $L(A)$, because $L_a L_{a^{-1}} = L_{a^{-1}} L_a = L_e = I$. Thus

$$(L_a)^{-1} = L_{a^{-1}}. \tag{1-6}$$

Conversely, suppose that L_a has an inverse L_a^{-1} in $L(A)$. Let $b = L_a^{-1}(e)$, and observe that $e = L_a \cdot L_a^{-1}(e) = L_a(b) = ab$. And from this, $a \cdot e = ea = (ab)a = a \cdot (ba)$. Since L_a is invertible, we conclude that $e = ba$. Thus a is invertible in A.

Sometimes when a Banach algebra B has no unit, it is technically advantageous to first adjoin a unit to B (as described on page 387); let $A = \{(x, \lambda) : x \in B, \lambda \in C\}$, and then apply the left regular representation to A. The composition of the mapping $x \mapsto (x, 0)$ with the left regular representation of A provides an isometric isomorphism of B onto a closed subalgebra of $L(A)$. In this sense, then, *every* Banach algebra "is" an algebra of bounded linear operators on a suitable Banach space.

PROBLEMS

1. Let H be a Hilbert space, and let M be a subset of $L(H)$.
   a. Verify that M' is a Banach algebra with a unit, and M' is closed in the weak operator topology.
   b. Prove that M is a von Neumann algebra if and only if $M = N'$ for some subset N of $L(H)$ that contains the adjoint of each of its elements.
   c. Show that M'' is a commutative algebra if the elements of M commute.
2. Let $(X, \mathscr{S}, \mu)$ be a measure space as in Example 6. Show that the mapping $h \mapsto M_h$ is an isometric isomorphism from $L^\infty(X, \mu)$ into $L(L^2(X, \mu))$. [*Hint.* Given $h \in L^\infty(X, \mu)$ take $\varepsilon > 0$, and let $\delta = \{t \in X : |h(t)| \geq \|h\|_\infty - \varepsilon\}$. Examine $\|M_h f\|$ where f is $\mu(\delta)^{-1/2}$ times the characteristic function of δ.]

3. a. Show that multiplication is continuous in a Banach algebra A. That is, show that if $\{x_n\}$ and $\{y_n\}$ are sequences in A such that $x_n \to x$ and $y_n \to y$, then $x_n y_n \to xy$.
   b. Let A be a nonzero complex Banach space in which an operation of multiplication is defined, satisfying the first three axioms listed on page 386. Suppose that A has a unit e whose norm need not be equal to 1, and suppose that the multiplication operation $(x, y) \mapsto xy$ on $A \times A$ is continuous in x for each fixed y and continuous in y for each fixed x. Construct an equivalent norm $\|\cdot\|_1$ on A such that $\|e\|_1 = 1$ and $\|xy\|_1 \leq \|x\|_1 \|y\|_1$, thus making A into a Banach algebra. [*Hint.* Consider the left regular representation of A.]
4. Let A be a Banach algebra with a unit e, and let $\Phi(A)$ denote the image of A in $L(A)$ under the left regular representation $\Phi: A \to L(A)$. What can you say about an operator $T \in L(A)$ that commutes with every operator $L_a \in \Phi(A)$?

VII.2 SPECTRAL THEORY IN A BANACH ALGEBRA

We assume throughout this section that A is a Banach algebra with a unit e (with $\|e\| = 1$). Our first few results reflect the ideas presented in the beginning sections of Chapters IV and V. In general, two types of proof are available. One method is to use the left regular representation to carry over to A the results already known for $L(A)$; the other is simply to repeat the proofs of the earlier theorems, replacing $T \in L(X)$ by $x \in A$. Such arguments are straightforward, and we shall usually suppress most of the details.

Invertible elements of A were introduced in Example 7 of § 1. It is important for us to know that the open unit ball centered at e consists of invertible elements. That is,

$$\|x - e\| < 1 \quad \text{implies} \quad x \text{ is invertible in } A. \tag{2-1}$$

It is surprising how many important results in this section and the next depend either directly or indirectly on (2-1). Since the left regular representation is an isometric mapping that preserves the property of invertibility, (2-1) follows immediately from Theorem IV.1.4. Also, from Theorem IV.1.5, the set G of invertible elements is an open subset of A and $x \mapsto x^{-1}$ is a continuous mapping of G onto itself. Note that G is a group under multiplication, since the inverse of xy is $y^{-1}x^{-1}$, whenever $x, y \in G$. The elements in $A \backslash G$ are said to be *singular*.

The Spectrum

The *spectrum* of $x \in A$ is the set $\sigma(x)$ of all complex numbers λ such that $\lambda e - x$ is not invertible (i.e., is singular). The *resolvent set* of x is the set $\rho(x)$ of all λ such that $\lambda e - x$ is invertible.

Since the left regular representation of $\lambda e - x$ is the operator $\lambda I - L_x$, it follows that

(2-2) $$\sigma(x) = \sigma(L_x).$$

From our knowledge of the spectrum of a bounded linear operator on a Banach space (Theorems V.3.1 and V.3.2), we conclude that $\sigma(x)$ must be a nonempty compact subset of $\mathbf{C}$.

The fact that $\sigma(x)$ cannot be empty leads immediately to one of the key theorems in the structure theory of commutative Banach algebras. The result is usually called the Gelfand–Mazur theorem. Gelfand made the first (and most spectacular) use of the theorem in his 1939 paper [1], although Mazur had proved it a year earlier by different methods. It was also obtained in 1941 by E. R. Lorch, who was unaware of Gelfand's work.

Theorem 2.1 (The Gelfand–Mazur Theorem). *If A is a (complex) Banach algebra with a unit and if every nonzero element in A is invertible, then A is isometrically isomorphic to the algebra of complex numbers.*

Proof. Given $x \in A$, there is some $\lambda \in \sigma(x)$, so that $\lambda e - x$ is not invertible. Then, by hypothesis, $\lambda e - x = 0$. Hence $x = \lambda e$, and we see that A is just the set of scalar multiples of e. The correspondence $\lambda e \mapsto \lambda$ is obviously an isometric isomorphism, since $\|\lambda e\| = |\lambda| \, \|e\| = |\lambda|$. $\square$

We note that commutativity of A is not a hypothesis of this theorem but rather is part of the conclusion.

The Spectral Radius

The spectral radius $r_\sigma(x)$ of x is defined by

$$r_\sigma(x) = \sup\{|\lambda| : \lambda \in \sigma(x)\}.$$

From the usual formula for the spectral radius of an operator (Theorem V.3.5) we obtain

$$r_\sigma(x) = r_\sigma(L_x) = \lim_{n\to\infty} \|(L_x)^n\|^{1/n} = \lim_{n\to\infty} \|L_{x^n}\|^{1/n}.$$

Hence

(2-3) $$r_\sigma(x) = \lim_{n\to\infty} \|x^n\|^{1/n}.$$

Since $\|x^n\| \leq \|x\|^n$, (2-3) implies that $r_\sigma(x) \leq \|x\|$. In some algebras, $r_\sigma(x) = \|x\|$ for all x. It will be useful later to have a characterization of a slightly more general situation.

Theorem 2.2. *There exists a constant $c>0$ such that $c\|x\|^2 \leq \|x^2\|$ for all x if and only if there exists a constant $d>0$ such that $d\|x\| \leq r_\sigma(x)$ for all x. Furthermore, $\|x\|^2 = \|x^2\|$ for all x if and only if $\|x\| = r_\sigma(x)$ for all x.*

Proof. From the identity $\lambda^2 e - x^2 = (\lambda e - x)(\lambda e + x)$ it is easy to see that $\sigma(x^2) = \{\lambda^2 : \lambda \in \sigma(x)\}$, and hence $r_\sigma(x^2) = [r_\sigma(x)]^2$. Now, if $d\|x\| \leq r_\sigma(x)$ for all x, then

$$d^2\|x\|^2 \leq [r_\sigma(x)]^2 = r_\sigma(x^2) \leq \|x^2\|,$$

and we may take $c = d^2$. Conversely, suppose that $c\|x\|^2 \leq \|x^2\|$ for all x. Applying this inequality successively to $x^2, x^4, x^8, \ldots$, we find that

$$c^{n-1}\|x\|^n \leq \|x^n\|$$

for $n = 2^k$, $k = 1, 2, \ldots$, and

$$c^{(n-1)/n}\|x\| \leq \|x^n\|^{1/n}.$$

Since the limit in (2-3) exists, we have

$$c\|x\| \leq r_\sigma(x).$$

Thus in this case we may take $d = c$. It is clear from the preceding arguments that if c may be 1, then d may be 1, and conversely. Since we always have $\|x^2\| \leq \|x\|^2$ and $r_\sigma(x) \leq \|x\|$, this proves the second assertion of the theorem. □

An unexpected consequence of the conditions described in this theorem is that A must be commutative. This was proved in 1967 by C. LePage [1] and independently in 1968 by Hirschfeld and Zelasko [1]. Both proofs involved a clever use of the exponential function.

For each $x \in A$, let $x^0 = e$ and let

$$\exp(x) = \sum_0^\infty \frac{x^n}{n!}.$$

This series is absolutely convergent (by comparison with the usual exponential series for $e^{\|x\|}$) and converges to an element of A since A is complete. An important property of the exponential is that $\exp(x+y) = \exp(x)\exp(y)$ whenever x and y commute. Consequently, $\exp(x)$ is invertible, with $\exp(-x)$ as its inverse. See problem 5.

Theorem 2.3. *Given $x, y \in A$, suppose there is a constant M (depending on x and y) such that*

$$\|\exp(\lambda x) \cdot y \cdot \exp(-\lambda x)\| \leq M \tag{2-4}$$

for all $\lambda \in \mathbf{C}$. Then $xy = yx$.

Proof. Let $F(\lambda) = \exp(\lambda x) \cdot [y \cdot \exp(-\lambda x)]$. Since multiplication in A is continuous, for each λ the function $F(\lambda)$ is the product of two absolutely convergent series, and we may multiply them term-by-term and collect the coefficients of each power of λ, thus obtaining an absolutely convergent power series in λ:

$$F(\lambda)=\left[e+\lambda x+\frac{(\lambda x)^2}{2}+\cdots\right]\left[y+y(-\lambda x)+\frac{y(-\lambda x)^2}{2}+\cdots\right]$$
$$=y+\lambda(xy-yx)+\lambda^2\left(\frac{x^2y}{2}-xyx+\frac{yx^2}{2}\right)+\cdots.$$

Since the series converges for all λ, $F(\lambda)$ is an entire function. But (2-4) says that $F(\lambda)$ is bounded. Hence $F(\lambda)$ is a constant, by Liouville's theorem. From the uniqueness of the Taylor series expansion, we conclude that the coefficients of the positive powers of λ must be zero. Thus $xy=yx$. □

Theorem 2.4. *Suppose that there exists a constant $c>0$ such that $c\|x\|^2\leqslant\|x^2\|$ for all $x\in A$. Then A is commutative.*

Proof. Given $x, y\in A$ and $\lambda\in C$, let $z=\exp(\lambda x)\cdot y\cdot\exp(-\lambda x)$. Then, for each $\mu\in C$, $\mu e=\exp(\lambda x)\cdot\mu e\cdot\exp(-\lambda x)$ and

$$\mu e-z=\exp(\lambda x)\cdot(\mu e-y)\cdot\exp(-\lambda x).$$

It follows that $\mu e-z$ is invertible if and only if $\mu e-y$ is invertible; that is, $\sigma(z)=\sigma(y)$. Consequently, $r_\sigma(z)=r_\sigma(y)$. Now the hypothesis of the theorem implies (by Theorem 2.2) that there exists $d>0$ such that $d\|x\|\leqslant r_\sigma(x)$ for all x. Hence

$$\|z\|\leqslant\frac{1}{d}r_\sigma(z)=\frac{1}{d}r_\sigma(y).$$

Applying Theorem 2.3, we conclude that $xy=yx$. □

Theorem 2.4 provides a striking example of an analytic condition that leads to an algebraic conclusion.

The Resolvent

When λ belongs to the resolvent set of an $x\in A$, the inverse of $\lambda e-x$ is called the *resolvent* of x and is denoted by $R_\lambda(x)$. The resolvent of x is a continuous function of λ and x because it is the composition of two continuous mappings: $(\lambda, x)\mapsto(\lambda e-x)\mapsto(\lambda e-x)^{-1}$. Now if $\lambda, \mu\in\rho(x)$, then

$$(\mu-\lambda)R_\lambda(x)=R_\lambda(x)[(\mu e-x)-(\lambda e-x)]$$
$$=R_\lambda(x)(\mu e-x)-e.$$

Multiplying both sides by $R_\mu(x)$ on the right, we obtain

$$(\mu-\lambda)R_\lambda(x)R_\mu(x)=R_\lambda(x)-R_\mu(x).$$

This is the *resolvent equation.* From this we immediately obtain the important fact that $R_\lambda(x)$ is an analytic function of x in $\rho(x)$, since

$$\frac{R_\lambda(x)-R_\mu(x)}{\lambda-\mu}=-R_\lambda(x)R_\mu(x)\to -R_\mu(x)^2$$

as $\lambda\to\mu$.

Topological Divisors of Zero

An element $x\in A$ is called a *left* (resp., *right*) *topological divisor of zero* if there exists a sequence $\{z_n\}$ in A such that $\|z_n\|=1$ and $xz_n\to 0$ (resp., $z_nx\to 0$). A (two-sided) *topological divisor of zero* is an element that is both a left and a right topological divisor of zero. A left topological divisor of zero cannot be invertible. For, if $yx=e$ and $xz_n\to 0$, then $z_n=y(xz_n)\to 0$, which shows that $\|z_n\|=1$ for all n is impossible. Similarly, a right topological divisor of zero cannot be invertible. For an example of a topological divisor of zero, take the function $x(t)=t$ in $C[0,1]$. (See Figure 1.)

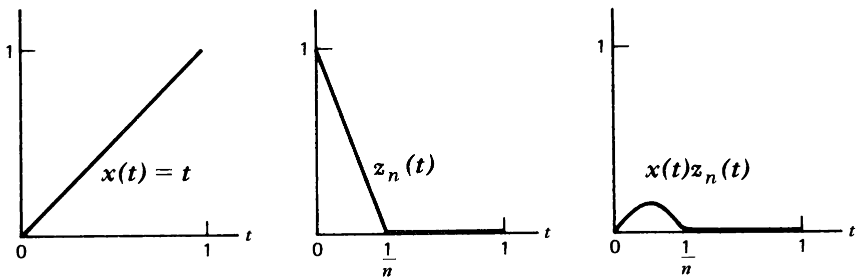


Figure 1. A topological divisor of zero in $C[0, 1]$.

Theorem 2.5. *Every boundary point of the group G of invertible elements of A is a topological divisor of zero.*

Proof. Suppose that x is in the boundary of G. Then $x\notin G$ since G is open, but there exists a sequence $\{x_n\}$ in G such that $x_n\to x$. Note that $x_n^{-1}x\notin G$, because $x_n\in G$. But then, by (2-1),

$$1\leq\|x_n^{-1}x-e\|\leq\|x_n^{-1}\|\,\|x-x_n\|,$$

which shows that $1/\|x_n^{-1}\|\to 0$. If we let $z_n=x_n^{-1}/\|x_n^{-1}\|$, then $\|z_n\|=1$ and

$$xz_n=(x-x_n)z_n+e/\|x_n^{-1}\|\to 0.$$

Similarly, $z_nx\to 0$, and so x is a topological divisor of zero. $\square$

Spectral Permanence

Our interest in topological divisors of zero lies mainly in the fact that they are *permanently singular*. To explain this we consider a closed subalgebra B of A that contains the unit e. If an element $x \in B$ is not invertible in B (i.e., is singular in B), then x may have an inverse in the larger algebra A. But this cannot happen if x is a topological divisor of zero in B. For a sequence $\{z_n\}$ in B that satisfies $\|z_n\| = 1$ and $xz_n \to 0$ is automatically a sequence in A with the same properties. Hence x is a topological divisor of zero in A and must be singular in A. Thus we cannot hope to make a topological divisor of zero invertible by enlarging the algebra.

Now consider any $x \in B$, let $\sigma_B(x)$ be the spectrum of x in B; that is, let $\sigma_B(x) = \{\lambda : \lambda e - x$ is not invertible in $B\}$, and let $\sigma_A(x)$ be the spectrum of x in A. The foregoing observations enable us to draw some useful conclusions about the relation between $\sigma_B(x)$ and $\sigma_A(x)$. If $\lambda e - x$ has no inverse in A, then it certainly has no inverse in B. Consequently, $\sigma_B(x) \supset \sigma_A(x)$; that is, the spectrum of x may shrink when we consider the larger algebra A. Suppose, however, that μ is a boundary point of $\sigma_B(x)$. Then $\mu e - x$ is evidently on the boundary of the group of invertible elements of B. By Theorem 2.5, $\mu e - x$ is a topological divisor of zero in B, and hence in A. Thus $\mu \in \sigma_A(x)$. In fact, μ is in the boundary of $\sigma_A(x)$, because a boundary point of $\sigma_B(x)$ is the limit of a sequence of points in $\rho_B(x)$ and this sequence automatically lies in $\rho_A(x)$. Thus the shrinking of $\sigma_B(x)$ is achieved by a hollowing out process, for no boundary points are lost. For reference, we state this in a theorem.

Theorem 2.6. *Let B be a closed subalgebra of A containing e. Then for all $x \in B$, $\sigma_A(x) \subset \sigma_B(x)$, and the boundary of $\sigma_B(x)$ is contained in the boundary of $\sigma_A(x)$.*

Corollary 2.7. *Let B be a closed subalgebra of A containing x and e. If $\rho_A(x)$ is connected, then $\sigma_A(x) = \sigma_B(x)$.*

Proof. The set $\sigma_B(x) \backslash \sigma_A(x)$ consists of the "holes" formed in $\sigma_B(x)$ when the algebra B is enlarged to A. This set is open in C because the boundary points of $\sigma_B(x)$ are all in $\sigma_A(x)$. Also, the resolvent set of x in A, $\rho_A(x)$, is a disjoint union of open sets, namely,

$$\rho_A(x) = \rho_B(x) \cup [\sigma_B(x) \backslash \sigma_A(x)].$$

Since $\rho_A(x)$ is connected, it follows that $\sigma_B(x) \backslash \sigma_A(x) = \varnothing$. Thus $\rho_A(x) = \rho_B(x)$, and so $\sigma_A(x) = \sigma_B(x)$. $\square$

One useful application of this corollary is to an operator $T \in L(X)$ whose spectrum is known to be real. In this case we may determine the spectrum of T by working in any closed subalgebra of $L(X)$ containing T and I.

PROBLEMS

In these problems, A denotes a Banach algebra having a unit e (with $\|e\| = 1$).

1. Suppose that to each nonzero element $x \in A$ there corresponds $y \in A$ such that $xy = e$. Then A is isometrically isomorphic to C.

2. Suppose that there exists a positive constant C such that $\|xy\| \leq C\|yx\|$ for all $x, y \in A$. Then A is commutative. If, in fact, $\|x\|\,\|y\| \leq C\|yx\|$ for all $x, y \in A$, then A is isomorphic to C.

3. Let $\Phi: A \to L(A)$ be the left regular representation of A. Let f be locally analytic on the spectrum of an element $a \in A$, and define $f(a)$ by an integral like that in (8-3) of Chapter V. Note that f is also locally analytic on the spectrum of $\Phi(a)$, and so $f(\Phi(a))$ may be defined by a similar integral.
   a. Show that the operator $\Phi(f(a))$ is $f(\Phi(a))$.
   b. Use (a) and Theorem V.9.5 to prove the spectral mapping theorem in A.

4. Suppose that $\|x\|^2 = \|x^2\|$ for all $x \in A$. Then $\|x\|^n = \|x^n\|$ for all x and for $n \geq 1$.

5. a. Given $a \in A$, let $f(\lambda) = \exp(\lambda a)$, $\lambda \in C$. Then f is an entire function and $f'(\lambda) = a \cdot \exp(\lambda a) = \exp(\lambda a) \cdot a$.
   b. Let y represent an analytic function with values in A. Show that the most general solution of the differential equation

$$y' = a \cdot y$$

   is of the form $f(\lambda) = \exp(\lambda a) \cdot b$, for some $b \in A$.
   c. If $a, b \in A$ and $ab = ba$, then $\exp(a + b) = \exp(a) \cdot \exp(b)$.

6. Let X be a compact Hausdorff topological space. Prove that a function f in $C(X)$ is either invertible or a topological divisor of zero. What can you say about $\sigma(f)$ for $f \in C(X)$?

7. The set of all left (right) topological divisors of zero is a closed subset of A.

8. For $x \in A$, let L_x be its left regular representation in $L(A)$. Show that $\lambda e - x$ is a left topological divisor of zero in A if and only if λ is in the approximate point spectrum of L_x.

9. Let A be the algebra of all continuous functions on the circle $|z| = 1$. Let B be the set of all $f \in A$ that can be extended to a function on the unit disc that is analytic for $|z| < 1$. (Note that B is isometrically isomorphic to the disc algebra. Why?) Let $f \in B$ be defined by $f(e^{it}) = e^{it}$. Show that $\sigma_B(f)$ is the closed unit disc and $\sigma_A(f)$ is the unit circle.

10. Let M be a subset of A, and let B be the bicommutant M'' of M. (The bicommutant is defined as in Example 5, § 1.) Then B is a closed subalgebra of A containing e, and $\sigma_B(x) = \sigma_A(x)$ for $x \in B$.

11. Let j be an element in A distinct from 0 and e such that $j^2 = j$. Let $B = jAj \equiv \{jxj: x \in A\}$. Then B is a Banach algebra whose unit is j. For $x \in B$, what can you say about $\sigma_B(x)$ and $\sigma_A(x)$?

VII.3 IDEALS AND HOMOMORPHISMS

In our study of Banach spaces we looked at subspaces and linear mappings. In the theory of Banach algebras, the appropriate objects for study are ideals and homomorphisms. In particular, homomorphisms into the algebra of complex numbers will be singled out for special attention, just as we did with linear functionals in the case of a Banach space.

As before, we let A denote a Banach algebra with a unit e. Recall that an *ideal* (i.e., a "two-sided" ideal) in A is a linear subspace J such that $xJ \equiv \{xy : y \in J\} \subset J$ and $Jx \equiv \{yx : y \in J\} \subset J$ for each $x \in A$. The *trivial* ideals are the zero ideal $\{0\}$ and A itself.

The set $\mathcal{K}(X)$ of all compact linear operators on a Banach space X is an ideal in $L(X)$ (cf. Theorems V.7.1 and V.7.2). The set of all functions in $C[0, 1]$ that vanish on a fixed subset of the interval $[0, 1]$ is an ideal in $C[0, 1]$.

A *homomorphism* from A into another Banach algebra B is a linear mapping $\Phi: A \to B$ such that $\Phi(xy) = \Phi(x)\Phi(y)$ for $x, y \in A$. The left regular representation of A is an example of a homomorphism of A into $L(A)$.

The *kernel* (or null space) of a homomorphism $\Phi: A \to B$ is the set $\{x \in A : \Phi(x) = 0 \in B\}$. This set is an ideal in A. For we know that it is a subspace (since Φ is a linear mapping) and, if x is in the kernel of Φ and $y \in A$, then $\Phi(xy) = \Phi(x)\Phi(y) = 0 \cdot \Phi(y) = 0$, which shows that xy is in the kernel of Φ. Similarly, yx is in the kernel of Φ.

Given an ideal J in A, we may consider A and J as linear spaces and form the quotient space A/J (cf. § I.6). It is readily checked that A/J becomes an algebra, the *quotient algebra*, when multiplication is defined by $[x][y] = [xy]$ for $[x], [y] \in A/J$. With this definition of multiplication, the canonical quotient mapping $q : A \to A/J$ becomes a homomorphism, because $q(x)q(y) = [x][y] = [xy] = q(xy)$.

When J is a closed ideal, the quotient space A/J is a Banach space under the quotient norm (Theorem II.5.1). In fact, A/J is a Banach algebra. In order to verify the necessary norm inequality in A/J, we use the easily established fact that $[x][y] = \{uv : u \in [x], v \in [y]\}$. Then

$$\begin{aligned}
\|[x][y]\| &= \inf \{\|uv\| : u \in [x], v \in [y]\} \\
&\leq \inf \{\|u\| \, \|v\| : u \in [x], v \in [y]\} \\
&= \inf \{\|u\| : u \in [x]\} \cdot \inf \{\|v\| : v \in [y]\} \\
&= \|[x]\| \cdot \|[y]\|.
\end{aligned}$$

The element $[e]$ is the unit for A/J, and one can show that $\|[e]\| = 1$ (problem 1).

Maximal Ideals

We say that an ideal J is *proper* if $J \neq A$. A *maximal ideal* is a proper ideal that is not properly contained in any proper ideal.

An important fact about proper ideals is that they contain no invertible elements. Indeed, if $x \in J$ and if $x^{-1}x = e$, then e must be in J, and hence $y = ye \in J$ for all $y \in A$, so that J is not a proper ideal. Thus the group of invertible elements in A lies in the complement of each proper ideal. Since this group is an open set, it follows that a proper ideal cannot be dense in A.

Theorem 3.1 (*a*) *Every maximal ideal in A is closed.* (*b*) *Every proper ideal is contained in a maximal ideal.*

Proof. (*a*) Let J be a maximal ideal. From the continuity of the algebraic operations, it is evident that the closure $\bar{J}$ is an ideal containing J. Since J cannot be dense in A, $\bar{J}$ is a proper ideal containing J. Hence $\bar{J}$ must coincide with J.

(*b*) Let J be a proper ideal, and let $\mathscr{J}$ be the family of all proper ideals that contain J. Partially order $\mathscr{J}$ by inclusion. If $\mathscr{I}$ is a completely ordered subfamily of $\mathscr{J}$, then $\bigcup_{I \in \mathscr{I}} I$ is clearly an ideal, and it is proper since it contains no invertible elements. Hence this union is an upper bound for $\mathscr{I}$. By Zorn's lemma, $\mathscr{J}$ contains a maximal element. □

A difficulty arises when an algebra lacks a unit, for in this case proper ideals can be dense. However, Theorem 3.1 is true in any Banach algebra B provided "ideal" is replaced by "modular ideal." An ideal M in B is said to be *modular* if there exists $u \in B$ such that $B - Bu \subset M$ and $B - uB \subset M$. We say that u is a unit for B modulo M. (See problem 7.) If B has a unit, then all ideals in B are modular.

Example 1. Let X be a compact Hausdorff space and let E be a closed subset of X. Then $\{f \in C(X) : f(t) = 0 \text{ for all } t \in E\}$ is a closed ideal in $C(X)$. It can be demonstrated that all closed ideals in $C(X)$ arise in this fashion. See Simmons [1, pages 328–330]. A maximal ideal is obtained by taking E to be a single point. We shall verify this in Theorem 3.4.

Example 2. In the disc algebra the closed ideals have been completely characterized by both A. Beurling and W. Rudin, but their description is too involved to reproduce here. See Hoffman [1, pages 83–87]. Later we shall show that each maximal ideal in the disc algebra is of the form $\{f : f(\lambda) = 0\}$, for some $|\lambda| \leq 1$.

Example 3. If H is a separable infinite-dimensional Hilbert space, there is only one closed ideal in $L(H)$, namely, the ideal $\mathcal{K}(H)$ of compact operators. Hence $\mathcal{K}(H)$ must be maximal. See Naimark [1, pages 296–298].

In general, it is extremely difficult to characterize the closed ideals in a Banach algebra. The task of describing the maximal ideals is sometimes more manageable. We shall consider some examples in § 4 and § 5.

Homomorphisms into C

A homomorphism ϕ of A into the algebra of complex numbers is often called a *multiplicative linear functional* because it must be a linear functional with the added property that $\phi(xy) = \phi(x)\phi(y)$ for $x, y \in A$. Note that, if ϕ is not identically zero, then $\phi(e) = 1$, because there exists an x such that $\phi(x) \neq 0$, and $\phi(x) = \phi(ex) = \phi(e)\phi(x)$.

We recall that the kernel of a nonzero linear functional is a hyperplane (i.e., a maximal subspace). The kernel M of a nonzero multiplicative linear functional ϕ is, in addition, an ideal. As such, M must be a maximal ideal, for there is no room for M to be properly contained in a proper subspace, much less in a proper ideal. (Cf. problem 2.) But then M must be closed, by Theorem 3.1. This implies that ϕ must be continuous! (Recall from § III.1 that a linear functional is continuous if and only if its null space is closed.) This proves the first half of the following theorem.

Theorem 3.2. *Each homomorphism $\phi : A \to C$ is necessarily continuous, and $\|\phi\| = 1$ when $\phi \neq 0$.*

Proof. Suppose that for some $x \in A$ with $\|x\| \leq 1$ we have $|\phi(x)| > 1$. Then $\|x^n\| \leq 1$ and $|\phi(x^n)| = |\phi(x)|^n \to \infty$ as $n \to \infty$. This is impossible since a continuous linear functional is bounded. It follows that $\|\phi\| \leq 1$. If $\phi \neq 0$, then $\phi(e) = 1 = \|e\|$, which implies that $\|\phi\| = 1$. □

The following theorem gives another important property of homomorphisms onto C.

Theorem 3.3. *If $\phi : A \to C$ is a nonzero homomorphism, then $\phi(x) \in \sigma(x)$, for each $x \in A$.*

Proof. For $x \in A$, the element $\phi(x)e - x$ obviously lies in the kernel of ϕ. As we have seen, this kernel is a proper ideal; hence it contains no invertible elements. Thus $\phi(x)$ must be in the spectrum of x. □

It is interesting that the converse of Theorem 3.3 is also true: if ϕ is a linear functional on A such that $\phi(x) \in \sigma(x)$ for all $x \in A$, then ϕ is a nonzero homomorphism onto $\mathbf{C}$. See Zelasko [1]. (Also, see Rudin [2, page 379], for references to other contributions to the converse of Theorem 3.3.) We shall not need this result.

Maximal Ideals in $C(X)$

We conclude this section by characterizing the maximal ideals in the Banach algebra $C(X)$, where X is a compact Hausdorff space. For each $t \in X$, let δ_t be the evaluation functional given by $\delta_t(f) = f(t)$ for $f \in C(X)$. (We called δ_t a "Dirac measure" in § III.5.) Then δ_t is obviously a homomorphism of $C(X)$ onto $\mathbf{C}$. The set $M_t = \{f : f(t) = 0\}$ is the kernel of δ_t; hence M_t is a maximal ideal (cf. the remarks preceding Theorem 3.2).

Theorem 3.4. *Every maximal ideal in $C(X)$ is of the form M_t for some $t \in X$. Every nonzero homomorphism of $C(X)$ onto $\mathbf{C}$ is an evaluation functional δ_t. Distinct points in X correspond to distinct maximal ideals and to distinct nonzero homomorphisms.*

Proof. Let M be a proper ideal, and suppose that $M \not\subset M_t$ for all $t \in X$. Then for each t there exists $f_t \in M$ with $f_t(t) \neq 0$ and, consequently, $|f_t|^2$ is (strictly) positive on a neighborhood of t. The compactness of X implies that there exist $t_1, \ldots, t_n \in X$ such that $f = |f_{t_1}|^2 + \cdots + |f_{t_n}|^2$ is positive on X. Then $1/f \in C(X)$; that is, f is invertible in $C(X)$. But $f = \sum_{i=1}^n f_{t_i} \bar{f}_{t_i} \in M$, since M is an ideal. This contradicts the fact that a proper ideal contains no invertible elements. Thus $M \subset M_t$ for some $t \in X$. If M is maximal, then $M = M_t$.

Now let ϕ be a nonzero homomorphism of $C(X)$ onto $\mathbf{C}$. Its kernel is some maximal ideal, say M_t. But the kernel of δ_t is also M_t. Since linear functionals that have the same null space must be proportional (Theorem III.1.1) and since ϕ and δ_t both equal 1 at the unit of $C(X)$, it follows that $\phi = \delta_t$. Finally, if $t_1, t_2 \in X$ with $t_1 \neq t_2$, then by Urysohn's lemma there exists $f \in C(X)$ such that $f(t_1) = 0$ and $f(t_2) \neq 0$. Hence $\delta_{t_1} \neq \delta_{t_2}$ and $M_{t_1} \neq M_{t_2}$. □

PROBLEMS

In these problems A denotes a Banach algebra with a unit e.

1. Let J be a closed ideal in A. Verify that multiplication is well defined in the quotient algebra A/J, and show that $\|[e]\| = 1$.
2. Suppose that M is both a maximal ideal and a hyperplane in A. Show that there exists a unique nonzero multiplicative linear functional ϕ_M whose kernel is M. [*Hint.* By Theorem III.1.1, M determines a family of linear functionals, all proportional and all having M as their null space. Let ϕ_M be the element in this family satisfying $\phi_M(e) = 1$.]

3. Let H be a separable infinite-dimensional Hilbert space. Explain why there exist no nonzero homomorphisms of $L(H)$ onto $\mathbf{C}$.
4. Let P be the normed algebra of all polynomial functions on the unit disc $|z| \leq 1$, with the supremum norm. Fix $\alpha \in \mathbf{C}$, with $|\alpha| > 1$. The mapping $p \mapsto p(\alpha)$ is an unbounded multiplicative linear functional on P.
5. If A is commutative, then a linear functional ϕ is multiplicative if and only if $\phi(x^2) = \phi(x)^2$ for all $x \in A$.
6. Use the result of Zelasko mentioned after Theorem 3.3 to prove:
   a. A linear functional ϕ is a homomorphism if and only if $\phi(e) = 1$ and the kernel of ϕ contains no invertible elements;
   b. A hyperplane in A is a maximal ideal if and only if it contains no invertible elements.
7. Let B be a Banach algebra that has no unit. Prove:
   a. If ϕ is a homomorphism of B into $\mathbf{C}$, then ϕ is continuous and $\|\phi\| \leq 1$.
   b. Let M be a modular ideal, and let u be a unit for B modulo M. (See page 401.) If $x \in M$, then $\|x - u\| \geq 1$. The closure of a proper modular ideal is a proper modular ideal. In particular, every maximal modular ideal is closed. Every proper modular ideal is contained in a maximal modular ideal. If M is a closed modular ideal and if u is a unit for B modulo M such that $\|u\| = 1$, then $[u]$ has norm 1 in B/M. Hence B/M is a Banach algebra with a unit.
   c. If ϕ is a nonzero homomorphism of B onto $\mathbf{C}$, then the kernel of ϕ is a maximal modular ideal.
   d. If M is both a hyperplane and a modular ideal in B, then M is the kernel of a homomorphism of B onto $\mathbf{C}$.
8. Let X be an infinite-dimensional complex Banach space, and let $\sigma_\Phi(T)$ denote the Fredholm spectrum of $T \in L(X)$ (cf. § V.4).
   a. Let $q : L(X) \to L(X)/\mathcal{K}(X)$ be the canonical quotient mapping, where $\mathcal{K}(X)$ is the ideal of all compact operators in $L(X)$. Show that the spectrum of $q(T)$ in $L(X)/\mathcal{K}(X)$ is $\sigma_\Phi(T)$, and deduce that $\sigma_\Phi(T)$ is not empty.
   b. Prove a spectral mapping theorem for the Fredholm spectrum: $\sigma_\Phi(f(T)) = f(\sigma_\Phi(T))$ for suitable functions f.

VII.4 COMMUTATIVE BANACH ALGEBRAS

The main purpose of this section is to determine when a Banach algebra is isomorphic to an algebra of continuous functions on a compact Hausdorff space. Further refinements will involve studying when this isomorphism is a homeomorphism and when it is an isometric isomorphism. Throughout this section A will denote a commutative Banach algebra with a unit e.

Definition. The *carrier space* of A is the set $\mathfrak{M}$ (or $\mathfrak{M}_A$) of all nonzero multiplicative linear functionals on A, endowed with the topology of pointwise convergence on A. For $x \in A$, the *Gelfand transform* of x is the function $\hat{x}$ (read x-hat) defined on $\mathfrak{M}$ by

$$\hat{x}(\phi) = \phi(x), \qquad \phi \in \mathfrak{M}.$$

Since each nonzero multiplicative linear functional ϕ is automatically continuous on A, with $\|\phi\| = 1$ (Theorem 3.2), $\mathfrak{M}$ is a subset of the unit ball in the conjugate space A' (where A is regarded as a Banach space). And the topology of pointwise convergence on A is just the weak* topology $\mathfrak{M}$ inherits as a subset of A'. Consequently, each function $\hat{x}$ is continuous and bounded on $\mathfrak{M}$, that is, $\hat{x} \in C(\mathfrak{M})$. In fact, from Example 1, § II.11, we see that $\mathfrak{M}$ has the *weakest* topology for which the mappings $\phi \mapsto \phi(x) = \hat{x}(\phi)$ are continuous.

Theorem 4.1. *Let A be a commutative Banach algebra with a unit. Then the carrier space $\mathfrak{M}$ of A is a compact Hausdorff space.*

Proof. The relative weak* topology on $\mathfrak{M}$ is Hausdorff (cf. § III.10). By Alaoglu's theorem, the unit ball S in A' is weak*-compact. Since $\mathfrak{M} \subset S$, it suffices to show that $\mathfrak{M}$ is weak*-closed in S. Suppose that $x' \in S$ is in the weak*-closure of $\mathfrak{M}$. Given $x, y \in A$ and $\varepsilon > 0$, the set

$$U = \{y' \in A' : |(x' - y')(x)| < \varepsilon,\ |(x' - y')(y)| < \varepsilon,\ |(x' - y')(xy)| < \varepsilon\}$$

is a weak*-neighborhood of x'. Hence there exists $\phi \in \mathfrak{M} \cap U$, and we find that

$$x'(xy) - x'(x)x'(y) = [x'(xy) - \phi(xy)] + \phi(x)[\phi(y) - x'(y)] + [\phi(x) - x'(x)]x'(y)$$

and

$$\begin{aligned} |x'(xy) - x'(x)x'(y)| &\leq \varepsilon + |\phi(x)| \cdot \varepsilon + \varepsilon \cdot |x'(y)| \\ &< \varepsilon(1 + \|x\| + \|y\|) \end{aligned}$$

(using the facts that $\|\phi\| = 1, \|x'\| \leq 1$). Since $\varepsilon > 0$ was arbitrary, $x'(xy) = x'(x)x'(y)$. Similar reasoning shows that $x'(e) = 1$. Thus $x' \neq 0$ and $x' \in \mathfrak{M}$. □

The correspondence $x \mapsto \hat{x}$ is called the *Gelfand representation of A*. The mapping is obviously linear. It is also multiplicative because, if $x, y \in A$ and $\phi \in \mathfrak{M}$,

$$\widehat{xy}(\phi) = \phi(xy) = \phi(x)\phi(y) = \hat{x}(\phi)\hat{y}(\phi).$$

Thus the Gelfand representation is a homomorphism of A onto a subalgebra $\hat{A}$ of $C(\mathfrak{M})$. Denoting the norm in $C(\mathfrak{M})$ by $\|\cdot\|_\infty$, we have

$$\|\hat{x}\|_\infty = \sup_{\phi \in \mathfrak{M}} |\hat{x}(\phi)| = \sup_{\phi \in \mathfrak{M}} |\phi(x)|.$$

Since $\phi(x) \in \sigma(x)$ for each ϕ by Theorem 3.3, it follows that

$$\|\hat{x}\|_\infty \leq r_\sigma(x) \leq \|x\|, \qquad x \in A.$$

So the Gelfand representation is norm-decreasing and hence continuous.

Our next step is to show that $\mathfrak{M}$ contains a good supply of elements. This will follow from Theorem 4.4. Theorems 4.3 and 4.4 contain the machinery that makes the Gelfand theory useful. Their proofs require the fact that A is commutative, something we have not used thus far. However, commutativity enters these proofs in only one essential way, through the following simple lemma.

Lemma 4.2. *Let A be a commutative Banach algebra with a unit. If an element x in A is not invertible, then the set*

$$Ax = \{wx : w \in A\}$$

is a proper ideal containing x.

Proof. Clearly $y(Ax) \subset Ax$ for each $y \in A$. And since A is commutative, $(Ax)y \subset Ax$. Thus Ax is always an ideal. Also, $x = ex \in Ax$. If Ax is not a *proper* ideal then $e \in Ax$; that is, $e = wx$ for some $w \in A$. Hence $xw = wx = e$, and x must be invertible. □

Theorem 4.3. *An ideal M in A is maximal if and only if it is the kernel of a nonzero multiplicative linear functional.*

Proof. From § 3 we know that the kernel of each $\phi \in \mathfrak{M}$ is a maximal ideal. Conversely, suppose M is a maximal ideal. Then A/M is a Banach algebra since M is closed (Theorem 3.1). Let $q : A \to A/M$ be the canonical quotient mapping. If J were a nontrivial ideal in A/M, then it is readily checked that $q^{-1}(J)$ would be a proper ideal in A that properly contains M. This is impossible since M is maximal. Consequently, A/M contains no nontrivial ideals. From Lemma 4.2 we conclude that every nonzero element of A/M must be invertible in A/M. Hence, by the Gelfand–Mazur theorem (Theorem 2.1), there exists an algebraic isomorphism Φ of A/M onto $\boldsymbol{C}$. Clearly the composition $\phi = \Phi \circ q$ is a nonzero homomorphism of A onto $\boldsymbol{C}$ whose kernel is M. □

Theorem 4.4. *For each $x \in A$,*

$$\sigma(x) = \{\hat{x}(\phi) : \phi \in \mathfrak{M}\}.$$

Hence,

$$r_\sigma(x) = \sup_{\phi \in \mathfrak{M}} |\hat{x}(\phi)| = \|\hat{x}\|_\infty.$$

Proof. If $\lambda \in \sigma(x)$, then $\lambda e - x$ is contained in a proper ideal (Lemma 4.2), which in turn is in some maximal ideal (Theorem 3.1). It follows from Theorem 4.3 that $\lambda e - x$ is in the kernel of some $\phi \in \mathfrak{M}$; that is, $0 = \phi(\lambda e - x) = \lambda - \phi(x) = \lambda - \hat{x}(\phi)$. This shows that $\sigma(x) \subset \{\hat{x}(\phi) : \phi \in \mathfrak{M}\}$. Containment in the other direction was established earlier in Theorem 3.3. □

This theorem produces for each x enough elements in $\mathfrak{M}$ to make the range of $\hat{x}$ equal to $\sigma(x)$. In particular, $\mathfrak{M}$ is not empty. The theorem also provides a means of identifying those elements in the kernel of the Gelfand mapping; that is, $\hat{x}=0$ if and only if $r_\sigma(x)=0$. From the formula (2-3) for the spectral radius, this is equivalent to having $\lim_{n\to\infty} \|x^n\|^{1/n}=0$. In any Banach algebra, an element for which this limit is zero is said to be *quasinilpotent*. Thus the kernel of the Gelfand mapping is the set of quasinilpotent elements. For technical reasons it is useful to describe this set in still other terms.

The Radical of A

The *radical* of A is the intersection of all the maximal ideals of A. If the radical of A is $\{0\}$, then A is said to be *semisimple*.

Theorem 4.5. *The following statements are equivalent:*

(*a*) *x is in the radical of A.*

(*b*) $x \in \bigcap_{\phi \in \mathfrak{M}} \{y : \phi(y)=0\}$.

(*c*) $\hat{x}(\phi)=0$ *for all* $\phi \in \mathfrak{M}$; *that is,* $\hat{x}=0 \in C(\mathfrak{M})$.

(*d*) $r_\sigma(x)=0$.

(*e*) $\lim_{n\to\infty} \|x^n\|^{1/n}=0$; *that is, x is quasinilpotent.*

Proof. Use Theorem 4.3, the definition of $\hat{x}$, Theorem 4.4 and the discussion above. □

Part (*b*) of this theorem shows that A is semisimple if and only if there are enough elements in $\mathfrak{M}$ to separate the points of A. For, if A is semisimple and $\phi(x)=\phi(y)$ for all $\phi \in \mathfrak{M}$, then $x-y$ is in the radical of A, and hence $x=y$. Conversely, if $\mathfrak{M}$ separates the points of A and if $x \neq 0$, then there exists $\phi \in \mathfrak{M}$ such that $\phi(x) \neq \phi(0)=0$, which shows that the radical of A is $\{0\}$. An important application of this property of semisimplicity is given in problem 2. The reader is asked there to prove that the (norm) topology in a semisimple commutative Banach algebra is unique. This was proved by Gelfand in 1941. Later C. E. Rickart proved a similar result for a certain class of (possibly) noncommutative Banach algebras, and he conjectured that each semisimple Banach algebra has a unique norm topology. In 1967, B. E. Johnson [1] verified this conjecture. (The general definition of a semisimple algebra need not concern us here.)

The Gelfand Representation Theorem

The main features of the Gelfand representation are summarized in the following theorem.

Theorem 4.6. *Let A be a commutative Banach algebra with a unit e, and let $\mathfrak{M}$ be its carrier space. The Gelfand representation $x \mapsto \hat{x}$ is a norm-decreasing homomorphism of A onto an algebra $\hat{A}$ of continuous functions on the compact Hausdorff space $\mathfrak{M}$, with the following properties:*

(*a*) $\hat{e}(\phi) = 1$ *for all* $\phi \in \mathfrak{M}$.
(*b*) $\hat{A}$ *contains the constant functions and separates the points of* $\mathfrak{M}$.
(*c*) $\hat{x}$ *is invertible in* $C(\mathfrak{M})$ *if and only if* x *is invertible in* A.
(*d*) $\|\hat{x}\|_\infty = \lim_{n\to\infty} \|x^n\|^{1/n}$.
(*e*) $\hat{A}$ *is isomorphic to* A *if and only if* A *is semisimple.*

Proof. Part (*a*) follows from our observation in § 3 that $\phi(e) = 1$ for each nonzero multiplicative linear functional ϕ. For $\lambda \in \mathbf{C}$, $\widehat{\lambda e}(\phi) = \phi(\lambda e) = \lambda$ for all $\phi \in \mathfrak{M}$, and so $\hat{A}$ contains the constant functions. Also, if $\hat{x}(\phi_1) = \hat{x}(\phi_2)$ for all $x \in A$, then $\phi_1(x) = \phi_2(x)$ for all x, and hence $\phi_1 = \phi_2$. This proves (*b*). Now a function $\hat{x} \in C(\mathfrak{M})$ is invertible in $C(\mathfrak{M})$ if and only if $\hat{x}(\phi) \neq 0$ for all $\phi \in \mathfrak{M}$. By Theorem 4.4, this happens if and only if $0 \notin \sigma(x)$, that is, if and only if x is invertible in A. This proves (*c*). Theorem 4.4 also gives (*d*) when used with the spectral radius formula. Theorem 4.5 shows that the kernel of the Gelfand mapping is the radical of A, which proves (*e*). □

The algebra $\hat{A}$ need not be complete in the supremum norm. For example, the Wiener algebra W has the property that $\hat{W}$ is a proper dense subalgebra of $C(\mathfrak{M})$. (We shall discuss this in § 5, following the proof of Theorem 5.3.) However, in many important cases the Gelfand representation is at least a topological isomorphism, thus forcing $\hat{A}$ to be complete. We consider this situation in the next two theorems and in the examples that follow.

Theorem 4.7. *Let A be a commutative Banach algebra with a unit. Then $\hat{A}$ is topologically isomorphic to A if and only if there is a constant $c > 0$ such that $c\|x\|^2 \leq \|x^2\|$ for all $x \in A$. And $\hat{A}$ is isometrically isomorphic to A if and only if $\|x\|^2 = \|x^2\|$ for all x.*

Proof. Since $r_\sigma(x) = \|\hat{x}\|_\infty \leq \|x\|$, $\hat{A}$ is topologically isomorphic to A if and only if there is a constant $d > 0$ such that $d\|x\| \leq r_\sigma(x)$ for all x. By Theorem 2.2, this happens if and only if there exists $c > 0$ such that $c\|x\|^2 \leq \|x^2\|$ for all x. By the same theorem, $\|x\|^2 = \|x^2\|$ for all x if and only if $\|x\| = r_\sigma(x) = \|\hat{x}\|_\infty$ for all x. □

Uniform Algebras

A *uniform algebra* is a closed subalgebra of $C(X)$ that contains the constant functions and separates the points of X, where X is a compact Hausdorff space. Theorems 4.7 and 2.4 together give a nice characterization of this important class of algebras.

Theorem 4.8. *Let B be a Banach algebra (not a priori commutative) with a unit. Then B is topologically isomorphic to a uniform algebra if and only if there exists $c>0$ such that $c\|x\|^2 \leqslant \|x^2\|$ for all $x \in B$. And B is isometrically isomorphic to a uniform algebra if and only if $\|x\|^2 = \|x^2\|$ for all x.*

Proof. If $c\|x\|^2 \leqslant \|x^2\|$ for all x, then B must be commutative by Theorem 2.4, and Theorem 4.7 applies. Thus B is topologically isomorphic to $\hat{B}$. If $\|x\|^2 = \|x^2\|$ for all x, then B is isometrically isomorphic to $\hat{B}$. In both cases, $\hat{B}$ must be complete since B is complete. It follows from this and Theorem 4.6(b) that $\hat{B}$ is a uniform algebra. Now suppose there is a topological isomorphism T from B onto a uniform algebra. Then there exist positive constants m, M such that $m\|x\| \leqslant \|Tx\| \leqslant M\|x\|$ for all x. Hence $m^2\|x\|^2 \leqslant \|Tx\|^2$. But Tx is in a uniform algebra, and an obvious property of the supremum norm is that $\|Tx\|^2 = \|(Tx)^2\|$. Since T is a homomorphism, $(Tx)^2 = T(x^2)$. Thus

$$m^2\|x\|^2 \leqslant \|T(x^2)\| \leqslant M\|x^2\|,$$

and

$$\frac{m^2}{M}\|x\|^2 \leqslant \|x^2\|$$

for all x. If T is an isometric isomorphism, then $\|x^2\| = \|T(x^2)\| = \|(Tx)^2\| = \|Tx\|^2 = \|x\|^2$ for all x. $\square$

Because the Gelfand representation replaces abstract algebras by certain algebras of functions, one would hope that if A is already such an algebra of functions, then $\hat{A}$ should not be much different from A itself. Of course, the first example that comes to mind is $C(X)$.

Example 1. Let X be a compact Hausdorff space and let $\mathfrak{M}$ be the carrier space of $C(X)$. Recall from Theorem 3.4 that each element of $\mathfrak{M}$ is an evaluation functional δ_t and the correspondence $t \leftrightarrow \delta_t$ is bijective. Note that for $f \in C(X)$ and $\delta_t \in \mathfrak{M}$, we have

$$\hat{f}(\delta_t) = \delta_t(f) = f(t). \tag{4-1}$$

Let us identify the sets X and $\mathfrak{M}$ via the correspondence $t \leftrightarrow \delta_t$. Then $\hat{f}(t)$ means $\hat{f}(\delta_t)$, and hence by (4-1) the Gelfand transform of f is identified with f itself. Now there are two possible topologies on X—the original compact Hausdorff topology τ_1 and the "Gelfand" topology τ_2 that X inherits from $\mathfrak{M}$. Since τ_2 is the weakest topology for which the functions $\hat{f}$ are continuous (cf. the remark before Theorem 4.1), it follows that τ_2 is weaker than τ_1. However,

X is a compact Hausdorff space under both topologies, and so τ_1 and τ_2 must coincide, by a well-known theorem of topology. In this sense, *X "is" the carrier space of $C(X)$*. In summary, the Gelfand representation $f \mapsto \hat{f}$ of $C(X)$ may be viewed as the identity mapping of $C(X)$ onto itself.

Example 2. Let A be a uniform algebra on X. The supremum norm satisfies the condition in Theorem 4.7 that $\|f\|^2 = \|f^2\|$, and so A is semisimple and isometrically isomorphic to $\hat{A}$. Let $\mathfrak{M}$ be the carrier space of A. Once again, the evaluation functionals δ_t are in $\mathfrak{M}$. The correspondence $t \mapsto \delta_t$ is one-to-one because A separates the points of X. Thus we may identify X with a subset of $\mathfrak{M}$. For $f \in A$, the restriction of $\hat{f}$ to the points in X satisfies $\hat{f}(t) = \hat{f}(\delta_t) = \delta_t(f) = f(t)$; that is, $\hat{f}$ agrees with f on the points in X. Thus $\hat{f}$ *becomes an extension of f*. In particular, the restriction of $\hat{f}$ to X is continuous with respect to the initial topology on X. The relative topology that X inherits as a subset of $\mathfrak{M}$ is the weakest topology for which the $\hat{f}$ are continuous. Hence, as before the two topologies coincide on X. Thus *X is imbedded homeomorphically in $\mathfrak{M}$*. Furthermore, since the Gelfand representation $f \mapsto \hat{f}$ is an isometry and since X is compact, the extension $\hat{f}$ of f already assumes its maximum modulus on X.

Example 3. As a concrete illustration of Example 2, let X be the unit circle $\{\lambda \in \boldsymbol{C} : |\lambda| = 1\}$, let Δ be the closed disc $\{\lambda : |\lambda| \leq 1\}$, and let A be the algebra of all functions in $C(X)$ that can be approximated uniformly on X by polynomials in λ. By the maximum modulus theorem, any sequence of polynomials that converges uniformly on X to f in A also converges uniformly on Δ to a unique continuous extension $\tilde{f}$ of f. We shall prove that Δ is (homeomorphic to) the carrier space $\mathfrak{M}$ of A and $\tilde{f}$ is the Gelfand transform of f. First of all, each $\lambda \in \Delta$ gives rise to a homomorphism ϕ_λ if we let $\phi_\lambda(f)$ be the value of the extended function $\tilde{f}$ at λ, that is,

$$\phi_\lambda(f) = \tilde{f}(\lambda).$$

Secondly, given $\phi \in \mathfrak{M}$, let $\lambda_0 = \phi(f_0)$, where $f_0(\lambda) = \lambda$ for $|\lambda| = 1$. Then $|\lambda_0| \leq \|\phi\| \|f_0\| = 1$, and $\lambda_0 \in \Delta$. A polynomial function p defined on X is just a linear combination of "powers" of f_0, and since ϕ is a multiplicative linear functional, $\phi(p) = p(\phi(f_0)) = p(\lambda_0)$. Thus ϕ agrees with ϕ_{λ_0} on a dense set in A. By continuity, $\phi = \phi_{\lambda_0}$, and hence $\mathfrak{M} = \{\phi_\lambda : \lambda \in \Delta\}$. Also, each homomorphism is determined by its value at f_0, so $\phi_\lambda \mapsto \lambda$ is a one-to-one mapping of $\mathfrak{M}$ onto Δ. Furthermore, this mapping is continuous, for it is just the Gelfand transform of f_0,

$$\hat{f}_0(\phi_\lambda) = \phi_\lambda(f_0) = \lambda.$$

Thus $\hat{f}_0$ is a homeomorphism of $\mathfrak{M}$ onto Δ, because $\mathfrak{M}$ is compact. Now, for $f \in A$.

$$\hat{f}(\phi_\lambda) = \phi_\lambda(f) = \tilde{f}(\lambda).$$

Under the identification $\phi_\lambda \leftrightarrow \lambda$, the Gelfand transform $\hat{f}$ is the continuous extension $\tilde{f}$ of f to Δ. The algebra $\hat{A}$ consists of all uniform limits of polynomials on Δ, and hence $\hat{A}$ is the disc algebra.

Example 4. Let $A = A(\Delta)$ be the disc algebra and let $f_0(\lambda) = \lambda$ for $|\lambda| \leq 1$. The argument in Example 3 shows that every homomorphism of A onto $\boldsymbol{C}$ is an evaluation functional ϕ_λ for some $\lambda \in \Delta$ and that the Gelfand transform of f_0 is a homeomorphism of $\mathfrak{M}_A$ onto Δ. For $f \in A(\Delta)$, we have $\hat{f}(\phi_\lambda) = \phi_\lambda(f) = f(\lambda)$. Hence under the identification $\phi_\lambda \leftrightarrow \lambda$, the Gelfand transform of f is f itself. The Gelfand representation does not extend the domain of the functions in $A(\Delta)$.

In § 3 we stated that each maximal ideal in the disc algebra consists of the set of all functions that vanish at some point of Δ. This assertion now follows from Example 4 and from the fact that the maximal ideals are precisely the kernels of the complex-valued homomorphisms (Theorem 4.3). This technique of determining the maximal ideals indirectly, by finding the homomorphisms onto $\boldsymbol{C}$, illustrates another use of the Gelfand theory.

Examples 3 and 4 taken together suggest that when A is a uniform algebra on a compact Hausdorff space X, the carrier space $\mathfrak{M}_A$ of A may be regarded as the "natural" domain for the functions in A. To verify this, suppose that X_1 is a compact Hausdorff space containing X (and inducing the same topology on X), with the property that each $f \in A$ is extendable with no increase in norm to a continuous function on X_1, and suppose that the set of such extensions forms a uniform algebra A_1 on X_1. Then A_1 and A are isometrically isomorphic and their carrier spaces may be identified (as topological spaces). To each $t \in X_1$ corresponds the homomorphism $\delta_t \in \mathfrak{M}_{A_1} = \mathfrak{M}_A$. Since A_1 is a uniform algebra, the mapping $t \mapsto \delta_t$ imbeds X_1 homeomorphically in $\mathfrak{M}_A$. In this sense, then, *$\mathfrak{M}_A$ is the largest compact set on which A may be realized as a uniform algebra.* Once A is sitting on $\mathfrak{M}_A$, the Gelfand representation becomes the identity mapping and produces no further extension of the underlying domain of the functions in A.

The ease with which we were able to determine the carrier space in Example 3 depended mainly on the fact that the elements of the carrier space were determined by their value at one fixed function f_0. This situation is clarified in the next theorem.

An Algebra Generated by a Single Element

Given an element a in A, the *algebra generated by a in A* is the smallest closed subalgebra B of A that contains a and the unit e. If p is a complex polynomial, then the element $p(a)$ (defined in the obvious way) is in B. In fact, the set of all polynomials in a is an algebra contained in B and containing a and e. Hence

B must be the closure of this set. Note that B is always commutative even if A is not. An interesting case arises when A itself is generated by one element.

Theorem 4.9. *Suppose that A is generated by the element a. Then $\hat{a}$ is a homeomorphism of the carrier space $\mathfrak{M}$ of A onto $\sigma(a)$.*

Proof. By Theorem 4.4, $\hat{a}$ maps $\mathfrak{M}$ onto $\sigma(a)$. Suppose $\hat{a}(\phi_1) = \hat{a}(\phi_2)$ for some $\phi_1, \phi_2 \in \mathfrak{M}$. Then $\phi_1(a) = \phi_2(a)$, and since the ϕ_i are linear and multiplicative, $\phi_1(p(a)) = \phi_2(p(a))$ for all polynomials p. By continuity, $\phi_1 = \phi_2$. Thus $\hat{a}$ is a one-to-one mapping. Since $\mathfrak{M}$ is compact and $\hat{a}$ is continuous on $\mathfrak{M}$, $\hat{a}$ is a homeomorphism. □

When A is generated by a, the algebra $\hat{A}$ may be described explicitly. We shall use $\hat{a}$ to identify $\mathfrak{M}$ and $\sigma(a)$ and thus regard $\hat{A}$ as an algebra of functions on $\sigma(a)$. More precisely, if f is the inverse of the homeomorphism $\hat{a}$, then each function $\hat{x} \circ f$ is defined on $\sigma(a)$. We shall still call $\hat{x} \circ f$ the Gelfand transform of x, and write $\hat{x}$ in place of $\hat{x} \circ f$. With this (standard) abuse of notation, $\hat{a}$ itself becomes the identity function on $\sigma(a)$, $\hat{a}(\lambda) = \lambda$, and the transform of a polynomial in a becomes a polynomial function of λ. Now, given $x \in A$, let $\{p_n\}$ be a sequence of polynomials in a that converge in A to x. Then $\|\hat{x} - \hat{p}_n\|_\infty \leq \|x - p_n\| \to 0$, which shows that $\hat{x}$ is the uniform limit on $\sigma(a)$ of polynomials in λ. It follows that each $\hat{x}$ is analytic on the interior of $\sigma(a)$ (when the interior is nonempty). More can be said, but we need a deep theorem of Mergelyan:

A nonempty compact subset K of $\mathbf{C}$ has a connected complement if and only if every function continuous on K and analytic in the interior of K can be approximated uniformly on K by polynomials.

(A proof of this would require a lengthy digression. See Rudin [1, pages 423–427].) In the case we are considering, $\mathbf{C} \backslash \sigma(a)$ must be connected (problem 10). Hence $\hat{A}$ consists of all functions continuous on $\sigma(a)$ and analytic in the interior of $\sigma(a)$. In particular, $\hat{A} = C(\sigma(a))$ if and only if $\sigma(a)$ has an empty interior.

In the theory of uniform algebras it is also important to consider algebras that are generated by a finite set of elements. If A is generated by $x_1, \ldots, x_n$, then the set of all polynomials $p(x_1, \ldots, x_n)$ in these n elements is dense in A. Therefore each $\phi \in \mathfrak{M}$ is determined by its values $\phi(x_1), \ldots, \phi(x_n)$, and the mapping $\phi \mapsto (\phi(x_1), \ldots, \phi(x_n))$ is one-to-one from $\mathfrak{M}$ onto the set

$$\Sigma = \{(\hat{x}_1(\phi), \ldots, \hat{x}_n(\phi)) : \phi \in \mathfrak{M}\} \subset \mathbf{C}^n.$$

The mapping is continuous because the $\hat{x}_i$ are continuous, and hence $\mathfrak{M}$ is homeomorphic to Σ. This latter set is called the *joint spectrum* of $x_1, \ldots, x_n$.

By identifying $\mathfrak{M}$ with Σ, $\hat{A}$ becomes the algebra of all uniform limits on Σ of polynomials $p(\lambda_1, \ldots, \lambda_n)$ in n variables. A further description of A is possible, but it involves parts of the theory of analytic functions of several complex variables. The reader is referred to Gamelin [1, Chapter III].

Some authors may have Theorem 4.9 and its generalizations in mind when they call the carrier space the "spectrum" of the algebra, or when they sometimes write $\sigma(A)$ in place of $\mathfrak{M}$.

We conclude this section with a note on another common name for the carrier space—the *maximal ideal space.*

The Set of Maximal Ideals

Historically, the symbol $\mathfrak{M}$ was used for the set of all maximal ideals in A. Of course, there is a natural identification of this set with our carrier space, via Theorem 4.3. In fact, some authors define $\mathfrak{M}$ as the set of nonzero multiplicative linear functionals but still refer to $\mathfrak{M}$ as the maximal ideal space. In Gelfand's original work, the transform of $x \in A$ was a function that assigned to each maximal ideal M a complex number $x(M)$. This number was defined as the image of x under the composite mapping $A \to A/M \to C$ described in the proof of Theorem 4.3. A geometric interpretation of $x(M)$ may be given by recalling that since M is a hyperplane in A not containing e, we have $A = M \oplus \{\lambda e : \lambda \in C\}$. Then $x(M)$ is the unique scalar such that $x = m + x(M)e$ for some $m \in M$. See Figure 2. Thus the correspondence $x \mapsto x(M)$ is the

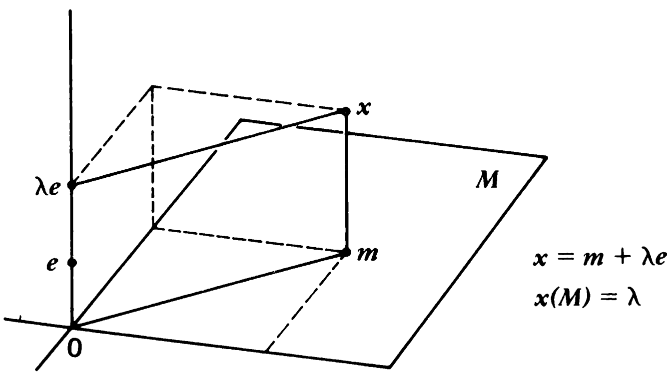


Figure 2

homomorphism ϕ whose kernel is M, so that $x(M) = \phi(x) = \hat{x}(\phi)$. However, Gelfand chose to emphasize the maximal ideals rather than the corresponding homomorphisms. Some recent presentations still use this approach.

PROBLEMS

1. If A is a semisimple commutative Banach algebra with a unit and if $\hat{A}$ is complete, then $\hat{A}$ is topologically isomorphic to A.
2. Let A be a semisimple commutative Banach algebra with a unit.
   a. Let B be a Banach algebra with a unit. Then each homomorphism Ψ from B into A is continuous. [*Hint.* Show that Ψ is a closed linear operator.]
   b. Suppose that A is also a Banach algebra under a second norm. Then the two norms on A are equivalent.
   c. Let B be a Banach algebra with a unit. If B is isomorphic to A, then B is topologically isomorphic to A.
3. Let A be a commutative Banach algebra with a unit, and let A_1 and A_2 be nontrivial closed subalgebras of A such that $A = A_1 \oplus A_2$ and such that A_1 and A_2 annihilate each other, that is, $ab = 0$ whenever $a \in A_1$, $b \in A_2$. (The latter condition is satisfied when A_1 and A_2 are ideals.) Show that the carrier space $\mathfrak{M}$ is the union of two nonempty disjoint open sets $\mathfrak{M}_1$ and $\mathfrak{M}_2$. [The converse is true too, but the proof requires some fairly deep results from the theory of analytic functions of several complex variables.]
4. Let B be a Banach algebra with a unit. If $x, y \in B$ and $xy = yx$, then $\sigma(x+y) \subset \sigma(x) + \sigma(y)$ and $\sigma(xy) \subset \sigma(x)\sigma(y)$. [*Hint.* See problem 10 of § 2.]
5. Suppose that $f_1, \ldots, f_k$ are functions in the disc algebra $A(\Delta)$, with no common zero in the closed disc Δ. Then there exist $g_1, \ldots, g_k$ in $A(\Delta)$ such that $\sum_{i=1}^{k} f_i(\lambda)g_i(\lambda) \equiv 1$ for $\lambda \in \Delta$. [*Hint.* Consider the set of elements of the form $\sum_{i=1}^{k} f_i g_i$, for $g_1, \ldots, g_k \in A(\Delta)$.]
6. Let A be the algebra of 2×2 matrices with complex entries a_{ij} such that $a_{21} = 0$. Determine the carrier space of A, and show that the radical is the set of matrices whose entries satisfy $a_{11} = a_{21} = a_{22} = 0$.
7. Let X be a compact Hausdorff space, and let A be a closed subalgebra of $C(X)$ that contains the unit $f(t) \equiv 1$ and that contains the complex conjugate of each of its elements. Let $\mathfrak{M}$ be the carrier space of A and, for $t \in X$, let δ_t be the evaluation functional on A.
   a. Show that if $f \in A$ is real valued, then $\hat{f}$ is real valued on $\mathfrak{M}$.
   b. Show that A is isometrically isomorphic to $C(\mathfrak{M})$.
   c. Show that $t \mapsto \delta_t$ is a continuous and open mapping from X onto $\mathfrak{M}$.
   d. Construct a topological quotient space of X that is homeomorphic to $\mathfrak{M}$.
8. Use the result of Zelasko mentioned after Theorem 3.3 to prove the following generalizations (due to Kahane and Zelasko).
   a. Let A_1 and A_2 be Banach algebras, each with a unit, and suppose that A_2 is commutative and semisimple. If T is a linear mapping from A_1 into A_2 such that $\sigma(Tx) \subset \sigma(x)$ for $x \in A_1$, then T is a homomorphism.
   b. Let A_1 and A_2 be Banach algebras with units e_1 and e_2, respectively. Let T be a homomorphism of A_1 into A_2 such that $Te_1 = e_2$. Then $\sigma(Tx) \subset \sigma(x)$, $x \in A_1$.
9. If A and B are commutative Banach algebras, each with a unit, and if A and B are isomorphic as algebras, then their carrier spaces are homeomorphic.

10. Suppose that a Banach algebra A with a unit is generated by the element a. Then the complement of $\sigma(a)$ in C is connected. [*Hint.* Let K be the union of $\sigma(a)$ and the bounded components of $C\backslash\sigma(a)$. Given $\lambda_0 \in K$, consider the mapping on the set of all polynomials in a defined by $p(a) \mapsto p(\lambda_0)$. Show that $\lambda_0 \in \sigma(a)$.]
11. In the algebra of all bounded linear operators on $C[0, 1]$, let A be the closed subalgebra generated by the Volterra integral operator (see § IV.2). Show that the carrier space of A consists of only one point.
12. Let A_0 be the Banach algebra $L^1(0, 1)$, where multiplication is defined for $f, g \in L^1(0, 1)$ by

$$(f * g)(t) = \int_0^t f(t-s)g(s)\, ds.$$

Then let A be the Banach algebra obtained by adjoining a unit to A_0. Show that the radical of A is A_0. [*Hint.* Show that A is generated by the function $f(t) \equiv 1, 0 \leq t \leq 1$.]
13. Let A be a commutative Banach algebra with a unit. Given $x_1, \ldots, x_k \in A$, show that the joint spectrum Σ of $x_1, \ldots, x_k$ is the set of all $(\lambda_1, \ldots, \lambda_k) \in C^k$ such that for any collection $y_1, \ldots, y_k \in A$ the element $\Sigma_{i=1}^k y_i(x_i - \lambda_i e)$ is not invertible. [*Hint.* Given $(\lambda_1, \ldots, \lambda_k) \in C^k$, the set of elements of the form $\Sigma_{i=1}^k y_i(x_i - \lambda_i e)$ is an ideal in A.] Use this characterization of the joint spectrum to prove the statement in problem 5.
14. Let B be a commutative Banach algebra without a unit. If M is a maximal modular ideal in B, then B/M is isometrically isomorphic as an algebra to C. The converse assertion was given in problem 7, § 3.

VII.5 APPLICATIONS AND EXTENSIONS OF THE GELFAND THEORY

The applications in this section have historical roots in problems that antedate the study of Banach algebras. Furthermore, early work on these problems may well have influenced the genesis of the Gelfand theory. The results we mention are not only of historical interest, however. They are still of value in mathematical research today—sometimes as tools, sometimes as models for further abstraction.

The first two topics in this section are related to the work of M. H. Stone. The next two are based on the work of N. Wiener. Near the end of the section we describe how the Gelfand theory extends to algebras that have no unit, and then we discuss the group algebra $L^1(\boldsymbol{R})$ and its generalizations.

The Stone–Čech Compactification

A topological space T is said to be *completely regular* if it is a T_1-space such that if t_0 is any point in T and if U is any open set containing t_0, then there is a continuous function f on T satisfying $0 \leq f(t) \leq 1$, $f(t_0) = 1$ and $f(t) = 0$ for $t \in T \backslash U$. The concept of complete regularity was introduced in 1930 by A. Tychonoff, who showed that a space is completely regular if and only if it can be imbedded homeomorphically in a compact Hausdorff space. In fact, he proved that if T is a completely regular space, then there exists a compact Hausdorff space $\beta(T)$ such that (*a*) T is dense in $\beta(T)$ and (*b*) every bounded continuous real-valued function on T has a continuous extension to $\beta(T)$. In 1937, E. Čech and M. H. Stone independently published detailed examinations of $\beta(T)$ and its relation to T, both proving that $\beta(T)$ is characterized up to a homeomorphism by (*a*) and (*b*). The space $\beta(T)$ is commonly referred to now as the *Stone–Čech compactification* of T.

This compactification has an interesting connection with the theory of Banach algebras. If T is completely regular and if one assumes Tychonoff's result about the existence of $\beta(T)$, then one can show that the carrier space of $C(T)$ is homeomorphic to $\beta(T)$ and the Gelfand transform of an f in $C(T)$ is essentially the extension of f to $\beta(T)$. On the other hand, rather than assuming the existence of $\beta(T)$, it is possible to demonstrate that the carrier space of $C(T)$ has exactly the properties required of $\beta(T)$. (See problem 1.) In fact, Stone's construction of $\beta(T)$ basically involved this approach. But in his case this was not just an application of the theory of commutative Banach algebras, for Stone's work [2] was done in 1936! The "points" in his compactification were essentially the maximal ideals in the algebra $C(T)$, although their formal construction involved certain collections of prime ideals in a Boolean ring of subsets of $C(T)$. Also, Stone's definition of the extension $\tilde{f}$ of a function $f \in C(T)$ amounted to the Gelfand transform of f, although he considered only real-valued functions. To do this, Stone first proved that an ideal M is maximal in the real algebra $C_{\boldsymbol{R}}(T)$ if and only if the quotient algebra $C_{\boldsymbol{R}}(T)/M$ is isomorphic to $\boldsymbol{R}$. (We let $C_{\boldsymbol{R}}(T)$ denote the real parts of the functions in $C(T)$.) Then, letting $Z(M)$ be the point in $\beta(T)$ corresponding to a maximal ideal M, he defined $\tilde{f}(Z(M))$ to be the image of f under the homomorphism $C_{\boldsymbol{R}}(T) \to C_{\boldsymbol{R}}(T)/M \to \boldsymbol{R}$.

Stone's 1937 paper contained other ideas that also prefigured the development of abstract Banach algebras. For a topology on $\beta(T)$, Stone used the "hull-kernel" topology, already known to algebraic topologists. In the case at hand, it is equivalent to our weak* topology on the carrier space of $C(T)$. This work later stimulated Gelfand and Shilov [1] to study this topology on the space of maximal ideals. The hull-kernel topology is of importance today in parts of harmonic analysis (cf. Rickart [1]). In preparation for his discussion of $\beta(T)$, Stone listed the properties of $C(T)$ that make it what he called a "topological ring." These properties correspond exactly to those used now to define an abstract commutative Banach algebra. Near the end of his paper, Stone pointed out the need for characterizations of the algebra of continuous functions on a compact Hausdorff space. A preprint of his paper was sent to Moscow in the early part of 1936. Two and one-half years later, Gelfand and Kolmogoroff [1] published a greatly simplified construction of $\beta(T)$ using the space of maximal ideals of $C(T)$. Soon after that, Gelfand [1] announced his general representation theory for commutative Banach algebras, which provided, among other things, a solution to the characterization problem mentioned by Stone.

The Conjugate of a Homomorphism

In § 4 we used the rather obvious fact that if two commutative Banach algebras with a unit are isometrically isomorphic, then their carrier spaces must be homeomorphic. Actually, the conclusion holds if we merely know that the two algebras are isomorphic, for the algebraic structure of a commutative Banach algebra determines its carrier space, including the topology of the latter. Applying this observation to Example 1 in § 4, we have the following theorem due to Gelfand and Kolmogoroff.

Theorem 5.1. *Two compact Hausdorff spaces X_1 and X_2 are homeomorphic if and only if $C(X_1)$ and $C(X_2)$ are isomorphic as algebras.*

Proof. If $C(X_1)$ and $C(X_2)$ are isomorphic, then X_1 and X_2 are homeomorphic since they are homeomorphic to the carrier spaces of $C(X_1)$ and $C(X_2)$, respectively. Conversely, if h is a homeomorphism of X_2 onto X_1, then the mapping Φ defined by

$$(\Phi f)(t)=f(h(t)), \qquad f\in C(X_1),\ t\in X_2, \tag{5-1}$$

is easily seen to be an algebraic (and isometric) isomorphism of $C(X_1)$ onto $C(X_2)$. $\square$

It is not difficult to show that every algebra isomorphism between $C(X_1)$ and $C(X_2)$ arises as in (5-1) from a homeomorphism of the underlying spaces. This fact is just a special instance of a general phenomenon involving the conjugate of a homomorphism.

Let A and B be commutative Banach algebras, each with a unit, let $\mathfrak{M}_A$ and $\mathfrak{M}_B$ be their respective carrier spaces, and let T be a homomorphism of A into B. If $\psi\in\mathfrak{M}_B$, the composition $\psi\circ T$ is a homomorphism of A into $\mathbf{C}$, and we denote it by $T'\psi$. Either $T'\psi\in\mathfrak{M}_A$ or $T'\psi$ is the zero homomorphism. The mapping $\psi\mapsto T'\psi$ is just the restriction to $\mathfrak{M}_B$ of the conjugate of T, when T is considered as a linear operator. We still refer to this restriction as the *conjugate* of T. This conjugate mapping T' is always continuous. For, if U is an open neighborhood in $\mathfrak{M}_A\cup\{0\}$ of some point $T'\psi_0$, say

$$U=\{\phi\in\mathfrak{M}_A\cup\{0\}:|\phi(x_i)-T'\psi_0(x_i)|<\varepsilon,\ 1\leq i\leq n\},$$

for some $x_1, \ldots, x_n \in A$, then its inverse image under T' is

$$\{\psi \in \mathfrak{M}_B : |T'\psi(x_i) - T'\psi_0(x_i)| < \varepsilon,\ \ 1 \leq i \leq n\}$$
$$= \{\psi \in \mathfrak{M}_B : |\psi(Tx_i) - \psi_0(Tx_i)| < \varepsilon,\ \ 1 \leq i \leq n\},$$

and this set is open in $\mathfrak{M}_B$.

It will be convenient in the next theorem to denote the kernel of a homomorphism T by $\ker(T)$.

Theorem 5.2. *Let A and B be commutative Banach algebras, each with a unit, and let T be a homomorphism of A onto B. Then the conjugate T′ of T is a homeomorphism of $\mathfrak{M}_B$ onto the closed set*

$$\{\phi \in \mathfrak{M}_A : \phi(x) = 0 \textit{ for all } x \in \ker(T)\}.$$

Proof. Since the range of T is B, $\psi \circ T \neq 0$ for $\psi \in \mathfrak{M}_B$, so that $T'(\mathfrak{M}_B) \subset \mathfrak{M}_A$. Also, T' is one-to-one, because the equality $\psi_1 \circ T = \psi_2 \circ T$ implies $\psi_1 = \psi_2$. Since $\mathfrak{M}_B$ is compact and T' is continuous, T' is a homeomorphism of $\mathfrak{M}_B$ onto a closed subset of $\mathfrak{M}_A$. In fact, $T'(\mathfrak{M}_B) \subset \{\phi \in \mathfrak{M}_A : \phi(x) = 0$ for all $x \in \ker(T)\}$, since $T'\psi(x) = \psi(Tx) = 0$ whenever $x \in \ker(T)$. Now if $\phi \in \mathfrak{M}_A$ vanishes on $\ker(T)$, then $\phi(x) = \phi(y)$ whenever $Tx = Ty$. Hence the relation

$$\psi(Tx) = \phi(x), \qquad x \in A,$$

defines a homomorphism ψ on the range of T, that is, on B. Thus $\psi \in \mathfrak{M}_B$ and $T'\psi = \phi$, which proves that $T'(\mathfrak{M}_B) = \{\phi \in \mathfrak{M}_A : \phi(\ker(T)) = 0\}$. □

To tie in this result with Theorem 5.1, assume that T is an isomorphism of $C(X_1)$ onto $C(X_2)$, and let δ_t denote the evaluation mapping. Then, for $t \in X_2$, $T'\delta_t$ is a homomorphism of $C(X_1)$ onto $\mathbf{C}$. Hence $T'\delta_t = \delta_{h(t)}$ for some point $h(t) \in X_1$, by Theorem 3.4. The function h in (5-1) is consequently the composition of three homeomorphisms:

$$t \mapsto \delta_t \overset{T'}{\mapsto} \delta_{h(t)} \mapsto h(t).$$

Thus the isomorphism T is implemented by a homeomorphism of the underlying spaces.

The Wiener Algebra

In 1939, six months after announcing his theory of commutative Banach algebras, Gelfand [2] showed how his work provides an elegant proof of a well-known lemma due to N. Wiener (1932). This lemma was of crucial importance in Wiener's proof of his renowned Tauberian theorem, but Wiener's proof of the lemma was quite technical and complicated. Gelfand's short proof attracted widespread interest in Banach algebras, for it raised the possibility of applying Banach algebra techniques to other related problems in harmonic analysis. During the next decade this approach met with considerable success.

Theorem 5.3 (Wiener). *If f is a continuous complex-valued function on $[0, 2\pi)$ with an absolutely convergent Fourier series, and if $f(t)$ is never zero, then $1/f$ also has an absolutely convergent Fourier series.*

Recall from Example 3, § 1, that the Wiener algebra W consists of all continuous functions on $[0, 2\pi)$ with an absolutely convergent Fourier series. Thus the theorem claims to give a sufficient condition for a function f in W to be invertible in W. Gelfand's proof rests on two key ideas. The first is that an element in W is invertible if and only if it belongs to no maximal ideal. (This follows, of course, from Lemma 4.2 and Theorem 3.1.) The second key idea is to look at homomorphisms instead of maximal ideals (cf. Theorem 4.3).

Proof. Let ϕ be a homomorphism of W onto C. Then $|\phi(e^{it})| \le \|e^{it}\| = 1$ and $|1/\phi(e^{it})| = \|\phi(e^{-it})\| \le \|e^{-it}\| = 1$. Hence $|\phi(e^{it})| = 1$; that is,

$$\phi(e^{it}) = e^{it_0}, \tag{5-2}$$

for some $t_0 \in [0, 2\pi)$. Clearly $\phi(e^{-it}) = e^{-it_0}$ and $\phi(e^{int}) = e^{int_0}$ for all n, since ϕ is a homomorphism. Then ϕ maps a trigonometric polynomial $\Sigma_{-N}^{N} a_n e^{int}$ into the complex number $\Sigma_{-N}^{N} a_n e^{int_0}$. By continuity, $\phi(f) = f(t_0)$ for every $f \in W$. The kernel of ϕ is $\{f \in W : f(t_0) = 0\}$. Thus if f never vanishes, it is not in the kernel of any nonzero homomorphism onto C; that is, f belongs to no maximal ideal and hence is invertible. □

In the proof of Theorem 5.3 we practically determined the carrier space $\mathfrak{M}$ of W. From (5-2) we see that the Gelfand transform of e^{it} is a continuous mapping from $\mathfrak{M}$ into the unit circle T. In fact, it is bijective and hence is a homeomorphism, because each $\phi \in \mathfrak{M}$ is determined by its value at e^{it} and because $e^{it} \mapsto e^{it_0}$ determines a homomorphism of W onto C for each $t_0 \in [0, 2\pi)$. Thus $\mathfrak{M}$ is homeomorphic to T.

If $\phi \in \mathfrak{M}$ is determined by (5-2) for some $t_0 \in [0, 2\pi)$, then the proof of Theorem 5.3 shows that $\phi(f) = f(t_0)$, that is, $\hat{f}(\phi) = f(t_0)$ for $f \in W$. When $\mathfrak{M}$ is identified with T, ϕ is identified with e^{it_0}, and we write $\hat{f}(e^{it_0}) \equiv \hat{f}(\phi) = f(t_0)$. Since this holds for each $t_0 \in [0, 2\pi)$, we have

$$\hat{f}(e^{it}) = \sum_{-\infty}^{\infty} a_n e^{int}, \qquad \text{when } f(t) = \sum_{-\infty}^{\infty} a_n e^{int}.$$

Clearly $\hat{f} = \hat{g}$ if and only if $f = g$. Thus W is semisimple and the algebra $\hat{W}$ of transforms is isomorphic to W (cf. Theorem 4.6).

Although W and $\hat{W}$ are isomorphic as algebras, they are not topologically isomorphic. We recall that the norm in W is given by

$$\|f\| = \sum_{-\infty}^{\infty} |a_n|, \qquad \text{when } f(t) = \sum_{-\infty}^{\infty} a_n e^{int},$$

while the norm in $\hat{W}$ is defined by

$$\|\hat{f}\|_\infty = \sup_{0 \le t < 2\pi} |\hat{f}(e^{it})| = \sup_{0 \le t < 2\pi} \left| \sum_{-\infty}^{\infty} a_n e^{int} \right|.$$

As a normed algebra, $\hat{W}$ is a proper dense (and hence incomplete) subalgebra of the Banach algebra $C(T)$ (see problem 5).

Functions that "Operate" on an Algebra

In 1934, P. Lévy proved that if F is an analytic function on an open domain D in C and if f is any function in the Wiener algebra whose range lies in D, then the composite function $F \circ f$ belongs to W. The special case where $F(\lambda) = 1/\lambda$ is Wiener's lemma, proved above. In modern terminology, we would say that W is closed under the action of analytic functions, or simply that analytic functions *operate* on W. The problem of determining the class of functions that operate on a given algebra is of particular interest in harmonic analysis (see Rudin [3, Chapter 6]).

In order to generalize the Wiener–Lévy theorem to an arbitrary commutative Banach algebra, we must shift our attention to the algebra of Gelfand transforms, so that "$F \circ f$" will make sense. Since W is so easily identified with $\hat{W}$, it should be clear that the following theorem includes the results of Wiener and Lévy.

Theorem 5.4. *Let A be a commutative Banach algebra with a unit e. Then analytic functions operate on $\hat{A}$. That is, if $a \in A$ and if a complex function F is locally analytic on a neighborhood of the range of $\hat{a}$, then there exists $b \in A$ such that $\hat{b} = F \circ \hat{a}$.*

Proof. Given a and F, we recall that the range of $\hat{a}$ is the spectrum of a. Thus there exists a bounded (open) Cauchy domain D containing $\sigma(a)$ such that $\bar{D}$ is contained in the domain of analyticity of F. Then it is a routine matter to demonstrate that the following contour integral exists and defines an element $F(a)$ in A:

$$F(a) = \frac{1}{2\pi i} \int_{+\partial D} F(\lambda)(\lambda e - a)^{-1} \, d\lambda. \tag{5-3}$$

(This was done in § V.8 for the algebra $L(X)$.) For each $\phi \in \mathfrak{M}$ we have, using the continuity of ϕ and Cauchy's integral formula,

$$\begin{aligned}\phi(F(a)) &= \frac{1}{2\pi i}\int_{+\partial D} F(\lambda)\phi[(\lambda e - a)^{-1}]\, d\lambda \\ &= \frac{1}{2\pi i}\int_{+\partial D} F(\lambda)(\lambda - \phi(a))^{-1}\, d\lambda = \frac{1}{2\pi i}\int_{+\partial D} \frac{F(\lambda)}{\lambda - \hat{a}(\phi)}\, d\lambda \\ &= F(\hat{a}(\phi)).\end{aligned}$$

If we let $b = F(a)$, then $\hat{b} = F \circ \hat{a}$. □

Corollary 5.5 (The Spectral Mapping Theorem). *Let A be a Banach algebra with a unit. If a function F is locally analytic on the spectrum of an element $a \in A$, then*

$$\sigma(F(a)) = F(\sigma(a)),$$

where $F(a)$ is defined as in (5-3).

Proof. Let $b = F(a)$. If A is commutative, we have

$$\sigma(b) = \{\hat{b}(\phi) : \phi \in \mathfrak{M}\} = \{F(\hat{a}(\phi)) : \phi \in \mathfrak{M}\} = F(\sigma(a)).$$

The extension to the noncommutative case follows easily by considering the bicommutant of a. See problem 7. □

Using the arguments of § V.8, one can show that, for each fixed $a \in A$, the correspondence $F \mapsto F(a)$ is a homomorphism (an "operational calculus") of the algebra of functions locally analytic on $\sigma(a)$ into the algebra A. The idea of using integrals and residue calculations for operators on an infinite-dimensional vector space goes back at least to F. Riesz [1, pages 117–121]. But the construction of a full operational calculus in an infinite dimensional space was not discussed until Gelfand [3] developed the operational calculus for a commutative Banach algebra and used it to prove Theorem 5.4 (for the case of a simply connected domain).

An important generalization of Theorem 5.4 can be proved when A is semisimple and F is a function of n complex variables that is analytic on a neighborhood of the joint spectrum of elements $a_1, \ldots, a_n \in A$. In this case there exists $b \in A$ such that $\hat{b}(\phi) = F(\hat{a}_1(\phi), \ldots, \hat{a}_n(\phi))$ for all $\phi \in \mathfrak{M}$. This was proved by Shilov in 1953 for finitely generated algebras and then was extended to the general case by Arens and Calderón in 1955. No simple proof of this result is known. The reader is referred to Rickart [1, pages 159–162] and Hörmander [1, pages 68–70].

An Algebra Without a Unit

Important applications of the Gelfand theory have been made in harmonic analysis to algebras that lack a unit. We shall describe here how the results of § 4 extend to such algebras.

Let B be a commutative Banach algebra that has no unit. We define the carrier space of B to be the set $\mathfrak{M}$ of all nonzero multiplicative linear functionals on B. The elements of $\mathfrak{M}$ are all continuous and lie in the closed unit ball of the conjugate space B' (see problem 7, § 3). Thus $\mathfrak{M}$ is relatively compact in the weak* topology of B'. Examining the proof of Theorem 4.1, we see that if an element $x' \in B'$ is in the weak*-closure of $\mathfrak{M}$, then $x'(xy) = x'(x)x'(y)$ for all $x, y \in B$. Thus either $x' \in \mathfrak{M}$ or $x' \equiv 0$. Consequently, there are two possibilities: either $\mathfrak{M}$ is closed (and hence compact), or $\overline{\mathfrak{M}} = \mathfrak{M} \cup \{0\}$ (in which case $\mathfrak{M}$ is locally compact). Instances of both situations occur. In extreme cases (the "radical algebras"), $\mathfrak{M}$ is void. The kernels of the homomorphisms in $\mathfrak{M}$ are precisely the maximal modular ideals of B (see problem 14, § 4).

The Gelfand transform $\hat{x}$ of $x \in B$ is defined as before by $\hat{x}(\phi) = \phi(x)$, for $\phi \in \mathfrak{M}$. Of course, $\hat{x} \in C(\mathfrak{M})$ and $\|\hat{x}\|_\infty = \sup\{|\phi(x)| : \phi \in \mathfrak{M}\} \leq \|x\|$. When $\mathfrak{M}$ is not compact, it is useful to observe that $\hat{x}$ has the following property. Given $\varepsilon > 0$, there exists a compact set K in $\mathfrak{M}$ such that $|\hat{x}(\phi)| < \varepsilon$ if $\phi \in \mathfrak{M}\backslash K$. (A suitable choice for K is $\{\phi \in \mathfrak{M} : \varepsilon \leq |\hat{x}(\phi)|\}$. This set is compact because it is the inverse image under $|\hat{x}|$ of the compact interval $[\varepsilon, \|x\|]$.) This property of $\hat{x}$ is described by saying that $\hat{x}$ *vanishes at infinity*. The class of all continuous functions on $\mathfrak{M}$ that vanish at infinity is denoted by $C_\infty(\mathfrak{M})$. It is evident that $C_\infty(\mathfrak{M})$ is a Banach algebra under the supremum norm. When $\mathfrak{M}$ is compact, we shall understand $C_\infty(\mathfrak{M})$ to be the same as $C(\mathfrak{M})$.

In order to define the spectrum of an element $x \in B$, we first adjoin a unit e to B (see page 387), and then define $\sigma(x)$ to be the set of λ for which $\lambda e - x$ is not invertible in the larger algebra with the unit. When $\mathfrak{M}$ is not empty, it can be shown that the range of $\hat{x}$ is either $\sigma(x)$ or $\sigma(x)\backslash\{0\}$; in either case the spectral radius equals $\|\hat{x}\|_\infty$, and this radius is given by $\lim_{n\to\infty} \|x^n\|^{1/n}$. We define the radical of B to be the set of all quasinilpotent elements (in B); this is the same as the intersection of all the maximal modular ideals in B. If B is semisimple, that is, if the radical of B is $\{0\}$, then B is isomorphic to a subalgebra of $C_\infty(\mathfrak{M})$.

The Algebra $L^1(R)$

Let us apply the above discussion to $L^1(\boldsymbol{R})$. Given $x \in \boldsymbol{R}$, consider the linear functional ϕ_x on $L^1(\boldsymbol{R})$ defined by

$$\phi_x(f) = \int_{-\infty}^{\infty} f(s)e^{-ixs}\,ds, \qquad f \in L^1(\boldsymbol{R}). \tag{5-4}$$

Using standard arguments (Fubini's theorem, changes of variable, etc.), one can show that ϕ_x is a homomorphism onto $\boldsymbol{C}$. Thus $x \mapsto \phi_x$ is a mapping of $\boldsymbol{R}$ into the carrier space of $L^1(\boldsymbol{R})$. If f is the characteristic function of an interval $[a, b]$, then

$$\phi_x(f) - \phi_y(f) = \int_a^b (e^{-ixs} - e^{-iys})\, ds.$$

If $x \neq y$, one can easily find a and b (with $a < b$) so that the above integral is nonzero; this shows that $\phi_x \neq \phi_y$. Now let ϕ be any homomorphism of $L^1(\boldsymbol{R})$ onto $\boldsymbol{C}$. We shall show that $\phi = \phi_x$ for some $x \in \boldsymbol{R}$. Since ϕ is a continuous linear functional, there exists $h \in L^\infty(\boldsymbol{R})$ such that

$$\phi(f) = \int_{-\infty}^{\infty} f(s)h(s)\, ds, \qquad f \in L^1(\boldsymbol{R}). \tag{5-5}$$

Using the facts that ϕ is a homomorphism and Lebesgue measure is translation invariant, we obtain, for $f, g \in L^1(\boldsymbol{R})$,

$$\begin{aligned}
\iint f(s)g(t)h(s+t)\, ds\, dt &= \iint f(s-t)g(t)h(s)\, ds\, dt \\
&= \iint f(s-t)g(t)h(s)\, dt\, ds \\
&= \phi(f * g) = \phi(f) \cdot \phi(g) \\
&= \left[\int f(s)h(s)\, ds\right]\left[\int g(t)h(t)\, dt\right] \\
&= \iint f(s)g(t)h(s)h(t)\, ds\, dt.
\end{aligned}$$

It follows that

$$h(s+t) \overset{0}{=} h(s)h(t). \tag{5-6}$$

Since $\phi \not\equiv 0$, we know from (5-5) that h cannot vanish almost everywhere on $\boldsymbol{R}$, and so there must exist some finite interval $[\alpha, \beta]$ such that $C = \int_\alpha^\beta h(t)\, dt \neq 0$. (See Lemma 9-8V of Taylor [5] for a proof.) Then

$$\int_\alpha^\beta h(s+t)\, dt \overset{0}{=} h(s) \int_\alpha^\beta h(t)\, dt = Ch(s)$$

and

$$h(s) \overset{0}{=} \frac{1}{C} \int_{\alpha+s}^{\beta+s} h(t)\, dt. \tag{5-7}$$

Since a definite integral is a continuous function of its limits of integration, h is equal almost everywhere to a continuous function. We may assume, therefore, that the function h representing ϕ in (5-5) is continuous. Thus we may read (5-6) and (5-7) as true for all values of the variables involved. From this we conclude that h is in fact differentiable everywhere and

$$h'(s) = \frac{1}{C}[h(\beta+s) - h(\alpha+s)] = \frac{h(\beta)-h(\alpha)}{C} \cdot h(s).$$

This differential equation implies that h has the form $h(s) = k \cdot e^{\lambda s}$, where $\lambda, k \in \boldsymbol{C}$; (5-6) then implies that $k = 1$. Now

$$|e^{\lambda s}| \leq \|h\|_\infty = \|\phi\| \leq 1, \qquad \text{for all } s \in \boldsymbol{R}.$$

This is possible only if λ is a pure imaginary number, say, $\lambda = -ix$, for some $x \in \boldsymbol{R}$. Then $h(s) = e^{-ixs}$, and $\phi(f)$ is given by the integral in (5-4).

Thus there is a one-to-one correspondence between $\boldsymbol{R}$ and the carrier space $\mathfrak{M}$ of the algebra $L^1(\boldsymbol{R})$. It is easy to show that $\boldsymbol{R}$ and $\mathfrak{M}$ are homeomorphic (problem 9). When these spaces are identified, the Gelfand transform of f is given by

$$\hat{f}(x) = \int_{-\infty}^{\infty} f(s) e^{-ixs}\, ds. \tag{5-8}$$

This is precisely the classical *Fourier transform* of f.

Generalizations of the algebra $L^1(\boldsymbol{R})$ were being studied at about the same time that Gelfand was developing his theory of Banach algebras. In 1938, A. Weil published a treatise that developed harmonic analysis on an arbitrary locally compact abelian group G. A central concern was (and still is) a "Fourier transform" of functions in the group algebra $L^1(G)$. Then in 1940, Gelfand and D. Raikov [1] studied the carrier space of $L^1(G)$ and showed that the general Fourier transform of a function is in fact its Gelfand transform. This result forged an important link between Banach algebra theory and abstract harmonic analysis (cf. Gelfand, Raikov, and Shilov [1, Chapter IV]).

PROBLEMS

1. Let T be a completely regular space, let $\mathfrak{M}$ be the carrier space of $C(T)$, and let h denote the mapping $t \mapsto \delta_t$. Then h is a homeomorphism from T into $\mathfrak{M}$, and $h(T)$ is dense in $\mathfrak{M}$. [*Hint.* A basic open set U in $\mathfrak{M}$ has the form $U = \{\phi : |\phi(f_i) - \phi_0(f_i)| < \varepsilon, 1 \leq i \leq n\}$ for some $f_i \in C(T)$, $\phi_0 \in \mathfrak{M}$ and $\varepsilon > 0$. The function $g(t) = \Sigma_1^n |f_i(t) - \phi_0(f_i)|^2$ is not invertible in $C(T)$ because $\phi_0(g) = 0$. Hence $\inf_{t \in T} g(t) = 0$.] Also, if T and $h(T)$ are identified, then each f in $C(T)$ has a unique continuous extension to $\mathfrak{M}$.
2. Let A be a commutative Banach algebra with a unit e, let $\mathfrak{M}$ be its carrier space, and let X be a compact Hausdorff space.

a. If T is a homomorphism of A into $C(X)$ such that Te is the unit of $C(X)$, then T may be factored through the algebra $\hat{A}$, that is, $T = T_1 \circ \Gamma$, where Γ is the Gelfand mapping and T_1 is a homomorphism of $C(\mathfrak{M})$ into $C(X)$.
b. Without using problem 2 of § 4, show that T_1 (and hence T) in part (a) is continuous.
c. Generalize part (a) to any continuous homomorphism T of A into $C(X)$. [*Hint.* Consider the set $X_1 = \{x \in X : Te(x) = 1\}$.]

3. Let A and B be commutative Banach algebras, each with a unit, and let T be a homomorphism of A onto a dense subalgebra of B. Then the conjugate T' of T is a homeomorphism of $\mathfrak{M}_B$ onto a compact subset of $\mathfrak{M}_A$.
4. Let B be the subalgebra of the Wiener algebra generated by the function $f(t) = e^{it}$. Determine the carrier space of B.
5. If W is the Wiener algebra, then $\hat{W}$ is a proper dense subalgebra of $C(T)$. [*Hint.* If $f \in C[0, 2\pi]$ and $f(0) = f(2\pi)$, then the Fourier series of f need not converge at every point in $[0, 2\pi]$, but f may be approximated uniformly on $[0, 2\pi]$ by trigonometric polynomials. See Katznelson [1, pages 13–15, 48–51].]
6. Let f be an analytic function in the open unit disc whose Taylor series is absolutely convergent for $|z| \leq 1$. If f does not vanish on the closed disc $|z| \leq 1$, then $1/f$ has an absolutely convergent Taylor series.
7. Complete the proof of Corollary 5.5 for the case when A is not commutative. [Show that $F(a)$ belongs to the bicommutant of a, and then use problem 10, § 2.] A different proof of Corollary 5.5 was outlined in problem 3, § 2.
8. Let B be a commutative Banach algebra without a unit, and let A be the algebra obtained by adjoining a unit to B; that is, $A = \{(a, \lambda) : a \in B, \lambda \in C\}$. Identify B with $\{(a, 0) : a \in B\}$.
   a. $\phi((a, \lambda)) = \phi((a, 0)) + \lambda$ for each $\phi \in \mathfrak{M}_A$.
   b. $\mathfrak{M}_A$ is homeomorphic to $\mathfrak{M}_B \cup \{0\}$, the one-point compactification of $\mathfrak{M}_B$.
   c. An ideal M in B is a maximal modular ideal if and only if there is a maximal ideal M_1 in A such that $M_1 \neq B$ and $M = M_1 \cap B$.
   d. For each $x \in B$, one has $0 \in \sigma(x)$ and $\sigma(x)\backslash\{0\} \subset \{\hat{x}(\psi) : \psi \in \mathfrak{M}_B\} \subset \sigma(x)$.
   e. $\|\hat{x}\|_\infty = \lim_{n\to\infty} \|x^n\|^{1/n}$, when $\mathfrak{M}_B \neq \varnothing$.
9. Let $R \cup \{\infty\}$ be the one-point compactification of R, with the usual topology τ. Let σ be the topology on $R \cup \{\infty\}$ induced by identifying R with the carrier space of $L^1(R)$ and ∞ with the zero homomorphism on $L^1(R)$. For $f \in L^1(R)$, let $\hat{f}$ be the Gelfand transform defined by (5-8), and define $\hat{f}(\infty) = 0$.
   a. Each $\hat{f}$ is τ-continuous.
   b. The topologies σ and τ coincide, and thus R is homeomorphic to the carrier space of $L^1(R)$.
10. Given $x = \{a_n\}_{-\infty}^{\infty}$ in the group algebra $\ell^1(Z)$, the *Fourier transform* of x is the continuous function $\tilde{x}$ on the unit circle T defined by

$$\tilde{x}(e^{it}) = \sum_{-\infty}^{\infty} a_n e^{int}.$$

Show that the carrier space of $\ell^1(\boldsymbol{Z})$ may be identified with $\boldsymbol{T}$, and under this identification the Gelfand transform of x in $\ell^1(\boldsymbol{Z})$ is the Fourier transform $\tilde{x}$.

VII.6 B^*-ALGEBRAS

The Banach algebra of all bounded linear operators on a Hilbert space H has a certain structural symmetry found also in many other Banach algebras. The distinguishing feature of $L(H)$ is the operation $T \mapsto T^*$ that sends a linear operator into its adjoint. In this section we consider Banach algebras on which a similar operation is defined.

A mapping $x \mapsto x^*$ of an algebra A into itself is called an *involution* if it has the following properties:

1. $(x^*)^* = x$.
2. $(x+y)^* = x^* + y^*$.
3. $(\alpha x)^* = \bar{\alpha} x^*$, α complex.
4. $(xy)^* = y^* x^*$.

In addition to $L(H)$, an important example of a Banach algebra with an involution is $C(X)$, with complex conjugation $f \mapsto \bar{f}$ as its involution. Another example is given by the disc algebra, where f^* for $f \in A(\Delta)$ is defined by $f^*(z) = \overline{f(\bar{z})}$.

Because an involution is conjugate linear (properties (2) and (3)), it will be continuous if and only if there exists $C > 0$ such that $\|x^*\| \leqslant C\|x\|$ for all x (cf. problem 1). The involutions mentioned above are all continuous. In fact they are isometries, $\|x^*\| = \|x\|$. We now consider a stronger condition on an involution.

Definition. An algebra A is called a *B^*-algebra* if it is a Banach algebra with an involution * satisfying

$$\|x^* x\| = \|x\|^2 \tag{6-1}$$

for every $x \in A$. We shall assume, moreover, that a B^*-algebra has a unit e, although for some results this requirement could be dropped. A *B^*-subalgebra* of a B^*-algebra A is a closed subalgebra B such that $e \in B$ and $x^* \in B$ whenever $x \in B$.

The condition (6-1) is readily verified for a number of Banach algebras. It is obviously satisfied by functions in $C(X)$, and Theorem IV.11.1 shows that (6-1) is also satisfied by operators in $L(H)$. Recall from § 1 (Example 5) that a C^*-algebra is a closed subalgebra of $L(H)$ that contains the identity and the adjoint of each of its elements. Obviously, (6-1) remains true in such a subalgebra. Hence *every C^*-algebra is a B^*-algebra.*

Our first theorem involves terminology adopted from operator theory. If A is an algebra with an involution, then an element $x \in A$ is said to be *normal* if $x^*x = xx^*$, and *self-adjoint* if $x^* = x$. Any element of the form x^*x is self-adjoint, because properties (1) and (4) of the involution show that $(x^*x)^* = x^*x^{**} = x^*x$. The unit e is self-adjoint, because $e = (e^*)^* = (e^*e)^* = e^*e = e^*$.

Theorem 6.1. *Let A be a B^*-algebra. Then*

(*a*) $\|x^*\| = \|x\|$, $x \in A$.

(*b*) *x^* is invertible if and only if x is invertible, in which case* $(x^*)^{-1} = (x^{-1})^*$.

(*c*) *If x is normal, then $\|x^2\| = \|x\|^2$ and $r_\sigma(x) = \|x\|$.*

(*d*) *If x is self-adjoint, then $\sigma(x)$ is real.*

Proof. The inequality $\|x\|^2 = \|x^*x\| \leq \|x^*\|\,\|x\|$ implies $\|x\| \leq \|x^*\|$. But then $\|x^*\| \leq \|x^{**}\| = \|x\|$, which proves that $\|x^*\| = \|x\|$. Now if x is invertible, then the identity $xx^{-1} = e$ implies $(x^{-1})^*x^* = e^* = e$. Similarly, $x^*(x^{-1})^* = e$. Thus x^* is invertible and $(x^*)^{-1} = (x^{-1})^*$. From this it follows that if x^* is invertible, so is $x^{**} = x$. This proves (*b*). If x is normal, then

$$\begin{aligned}\|x^2\|^2 &= \|(x^2)^*(x^2)\| = \|(x^*)^2x^2\| = \|(x^*x)(x^*x)\| \\ &= \|(x^*x)^*(x^*x)\| = \|x^*x\|^2 = \|x\|^4\end{aligned}$$

Hence $\|x^2\| = \|x\|^2$. By induction, for n in the form 2^k, $k = 1, 2, \ldots$, we have $\|x^n\| = \|x\|^n$ and $\|x^n\|^{1/n} = \|x\|$. Hence $r_\sigma(x) = \|x\|$ by the spectral radius formula (2-3).

Now suppose $x^* = x$. Let B be the commutative subalgebra generated by x, and let $\mathfrak{M}$ be its carrier space. Take $\phi \in \mathfrak{M}$ and write $\phi(x) = r + is$ for real r, s. Set $y = x + ite$, where t is real, and note that $y^*y = (x^* - ite)(x + ite) = x^2 + t^2e$, since $x^* = x$. Also, $y \in B$ and $\phi(y) = \phi(x) + it = r + i(s + t)$. Using the fact that $\|\phi\| \leq 1$, we have

$$r^2 + (s+t)^2 = |\phi(y)|^2 \leq \|y\|^2 = \|y^*y\| \leq \|x^2\| + t^2.$$

Hence

$$r^2 + s^2 + 2st \leq \|x^2\|.$$

This inequality cannot hold for all real t unless $s = 0$, that is, unless $\phi(x)$ is real. It follows (from Theorem 4.4) that $\sigma_B(x) \subset \boldsymbol{R}$. Since $B \subset A$, $\sigma_A(x) \subset \sigma_B(x) \subset \boldsymbol{R}$. This proves (*d*). □

Given two B^*-algebras A and B, a homomorphism Φ of A into B is called a **-homomorphism* if it preserves involutions; that is $\Phi(x^*) = \Phi(x)^*$. If Φ is also one-to-one, it is called a **-isomorphism*.

Theorem 6.2 (The Gelfand–Naimark Theorem). *Let A be a commutative B^*-algebra, with carrier space $\mathfrak{M}$. The Gelfand mapping $A \to \hat{A}$ is an isometric $*$-isomorphism of A onto $C(\mathfrak{M})$.*

Proof. Since every element is normal in a commutative B^*-algebra, we have $\|x\|^2 = \|x^2\|$ for $x \in A$, by Theorem 6.1(c). Thus A is isometrically isomorphic to $\hat{A}$, by Theorem 4.7. Given $x \in A$, let $u = (x + x^*)/2$ and $v = (x - x^*)/2i$. Then u and v are self-adjoint, $x = u + iv$ and $x^* = u - iv$. For $\phi \in \mathfrak{M}$, both $\phi(u)$ and $\phi(v)$ are real, by Theorem 6.1(d) and Theorem 4.4. Hence $\overline{\phi(x)} = \overline{\phi(u) + i\phi(v)} = \phi(u) - i\phi(v) = \phi(u - iv) = \phi(x^*)$. It follows that the complex conjugate of $\hat{x}$ is the transform of x^*. Thus the Gelfand mapping preserves involutions and is a $*$-isomorphism. In particular, the algebra $\hat{A}$ of functions is closed under complex conjugation. Recall from Theorem 4.6 that $\hat{A}$ contains the constant functions and separates the points of $\mathfrak{M}$. Also, $\hat{A}$ is closed in $C(\mathfrak{M})$ because $\hat{A}$ is isometrically isomorphic to A and hence is complete. Thus $\hat{A} = C(\mathfrak{M})$, by the Stone–Weierstrass theorem (Theorem III.11.9). □

An immediate consequence of the above theorem is that the self-adjoint elements in a commutative B^*-algebra are precisely those whose Gelfand transforms are real valued.

Theorem 6.2 was proved in Gelfand and Naimark [1]. In the same paper the authors also proved that an arbitrary (and possibly noncommutative) B^*-algebra A is isometrically $*$-isomorphic to a C^*-algebra of operators on some Hilbert space. (The B^*-algebras they considered satisfied an additional condition that later was shown to hold in every B^*-algebra.) The appropriate Hilbert space is constructed out of the intrinsic structure of A. The details are somewhat involved and, since we do not need this result in our further work, we refer the reader to Rudin [2, pages 319–323] and Arveson [1, pages 31–34].

The following theorem illustrates a typical use of the Gelfand–Naimark theorem. In this theorem we write $x \geqslant 0$ to mean that $x^* = x$ and $\sigma(x)$ consists of nonnegative real numbers. Such an x is said to be *positive*.

Theorem 6.3. *Let A be a B^*-algebra. If x is a positive element of A, then there exists a unique $y \geqslant 0$ in A such that $x = y^2 = y^*y$. Moreover, y commutes with each z in A that commutes with x. We call y the positive square root of x.*

Proof. Let B be any commutative B^*-subalgebra of A that contains x. (The closure in A of the set of polynomials in x is one choice for B.) Let $\mathfrak{M}_B$ be the carrier space of B. Since $\sigma_A(x)$ is real (by Theorem 6.1), we have from Corollary 2.7 and Theorem 4.4 that

$$\sigma_A(x) = \sigma_B(x) = \{\hat{x}(\phi) : \phi \in \mathfrak{M}_B\}.$$

Hence $\hat{x}$ is a nonnegative continuous function on $\mathfrak{M}_B$, and so $\hat{x}$ has a unique nonnegative continuous square root. By the Gelfand–Naimark theorem, there exists a unique self-adjoint y in B such that $y^2 = x$, and $\hat{y}(\phi) \geq 0$ for $\phi \in \mathfrak{M}_B$. Since $\sigma_A(y) \subset \sigma_B(y) = \{\hat{y}(\phi) : \phi \in \mathfrak{M}_B\}$, we have $y \geq 0$. Now let $y = y_1$ be the square root of x obtained by the above argument when B is the closure B_1 of the set of polynomials in x. Then if $z \in A$ commutes with x, by continuity of multiplication, z must also commute with y_1. If y_2 is another positive square root of x, let B_2 be any B^*-subalgebra containing y_2. Then $x = y_2^2 \in B_2$, and so $B_2 \supset B_1$. Applying the uniqueness result in B_2, we have $y_2 = y_1$. □

The next two theorems will be needed in § 7 when we use the Gelfand–Naimark theorem to obtain the spectral theorem for normal operators on a Hilbert space. Theorem 6.4 below should be compared with Theorem 4.9 and the discussion following it concerning finitely generated algebras.

Theorem 6.4. *Suppose a commutative B^*-algebra A is generated by an element a and its adjoint a^*. Then $\hat{a}$ is a homeomorphism of the carrier space $\mathfrak{M}$ of A onto $\sigma(a)$.*

Proof. The proof is like that of Theorem 4.9. We know that $\hat{a}$ is a continuous mapping of $\mathfrak{M}$ onto $\sigma(a)$ (cf. Theorem 4.4). Suppose that $\phi_1, \phi_2 \in \mathfrak{M}$ and $\hat{a}(\phi_1) = \hat{a}(\phi_2)$. Then $\phi_1(a^*) = \overline{\hat{a}(\phi_1)} = \overline{\hat{a}(\phi_2)} = \phi_2(a^*)$, because the Gelfand representation preserves involutions. It follows that ϕ_1 and ϕ_2 coincide on all polynomials in a and a^*. Since these polynomials are dense in A, we have $\phi_1 = \phi_2$. Thus $\hat{a}$ is one-to-one. Since $\mathfrak{M}$ is compact, $\hat{a}$ is a homeomorphism. □

Theorem 6.5. *Let A be a B^*-algebra, and let B be a B^*-subalgebra. Then an element in B has an inverse in B if and only if it has an inverse in A. Thus $\sigma_B(x) = \sigma_A(x)$ for $x \in B$.*

Proof. Any element of B invertible in B is certainly invertible in A. Conversely, if $x \in B$ and x^{-1} exists, then x^* is invertible in A (Theorem 6.1) and hence so is x^*x. Now $\sigma_A(x^*x)$ is real by Theorem 6.1, and so $\sigma_B(x^*x) = \sigma_A(x^*x)$ by Corollary 2.7. Thus $0 \notin \sigma_B(x^*x)$ and $(x^*x)^{-1}$ lies in B. From $e = (x^*x)^{-1}x^*x$ we obtain $x^{-1} = (x^*x)^{-1}x^* \in B$. The final statement of the theorem follows immediately by considering $\lambda e - x$ in place of x. □

PROBLEMS

1. Let A be a Banach algebra with a continuous involution and a unit. Then there is an equivalent norm $\|\cdot\|_1$ on A such that $\|x^*\|_1 = \|x\|_1$ for $x \in A$.
2. Every involution $x \mapsto x^*$ on a semisimple commutative Banach algebra A is continuous. [*Hint.* Show that $x \mapsto \|x^*\|$ defines a norm on A.]

3. Let A be a Banach algebra with a continuous involution and a unit e. Let $x \mapsto T_x$ be a homomorphism from A into $L(H)$, where H is a Hilbert space, such that $T_e = I$ and $T_{x^*} = (T_x)^*$ for $x \in A$. Such a mapping is called a **-representation of A on H.*
   a. Show that $\sigma(T_x) \subset \sigma(x)$, and hence $r_\sigma(T_x) \leq r_\sigma(x)$, for $x \in A$.
   b. If $x^* = x$, show that $\|T_x\| \leq \|x\|$.
   c. For arbitrary $x \in A$, consider x^*x and show that $\|T_x\|^2 \leq \|x^*\|\,\|x\|$. Deduce that the representation $x \mapsto T_x$ is necessarily continuous.
4. If A is a B^*-algebra, then $\|xx^*\| = \|x\|^2$ for $x \in A$.
5. Show that the disc algebra cannot be a B^*-algebra under any involution.
6. Let x be a normal element in a B^*-algebra A. Prove the following.
   a. $x^* = x$ if and only if $\sigma(x)$ is real.
   b. x^{-1} exists and $x^{-1} = x^*$ if and only if $\sigma(x)$ is a subset of the unit circle. Such an x is said to be *unitary.*
   c. If $x^* = x$, then $y = \exp(ix)$ is unitary.
7. Let A and B be B^*-algebras, and let Φ be a *-homomorphism of A into B.
   a. Φ is continuous and $\|\Phi\| \leq 1$. [*Hint.* First show that if $\Phi(A)$ is dense in B, then Φ maps invertible elements of A into invertible elements of B. Hence the spectral radius of x^*x is greater than or equal to the spectral radius of $\Phi(x^*x)$.]
   b. If Φ is one-to-one, then Φ is an isometry. [*Hint.* Suppose $\Phi(A)$ is dense in B and $x^* = x \in A$. If $\sigma(\Phi(x))$ is a proper subset of $\sigma(x)$, then there exists a continuous real-valued function f that is not identically zero on $\sigma(x)$ but vanishes on $\sigma(\Phi(x))$. Approximate f by polynomials and deduce a contradiction.]
   c. There is only one norm on A that makes A into a B^*-algebra.
8. Let A be a commutative Banach algebra with a unit, and let * be an involution on A. Then
   a. $\|\hat{x}\|_\infty = \|\widehat{x^*}\|_\infty$ for all $x \in A$.
   b. If $x^* = x$, then $\|x\| = \|\hat{x}\|_\infty$.
   c. If $\|x^*x\| = \|x^*\|\,\|x\|$ for all x, then $\|x^*\| = \|x\|$ for all x, which implies that A is a B^*-algebra. [*Hint.* For arbitrary $y \in A$, consider $x = y^*y$.]

VII.7 THE SPECTRAL THEOREM FOR A NORMAL OPERATOR

A commutative C^*-algebra of operators on a Hilbert space consists entirely of normal operators, because it contains the adjoint of each element and these elements all commute. Thus the theory of § 6 applies quite naturally to the study of normal operators. Our main concern in this section is to develop a spectral decomposition theorem analogous to that given in § VI.6 for self-adjoint operators. Many proofs of the spectral theorem for normal operators are known, including a fairly simple argument based on results we have proved in Chapter VI. However, we have chosen an approach that avoids reliance on the self-adjoint case and illustrates the usefulness of the Gelfand–

Naimark theorem for commutative B^*-algebras. Actually, there are two distinct forms of the spectral theorem. We shall discuss both, beginning with the classical form that pervades the early literature on the theory of operators in Hilbert space.

Throughout this section, H will denote a (complex) Hilbert space. Given a normal operator $T \in L(H)$, we let A_T be the subalgebra of $L(H)$ generated by T and T^*. That is, A_T is the closure in $L(H)$ of the set of polynomials $p(T, T^*)$ in T and T^*. The set of such polynomials is obviously a commutative self-adjoint algebra. Hence A_T is a commutative C^*-algebra. (For terminology, see Example 5, § 1.) We call A_T *the C^*-algebra generated by T*, since it is the smallest C^*-subalgebra of $L(H)$ that contains T and I.

Our first theorem shows how results of § 6 apply to the algebra A_T.

Theorem 7.1. *Let T be a normal operator in $L(H)$. The Gelfand representation of A_T is an isometric *-isomorphism of A_T onto $C(\sigma(T))$ that maps T onto the function $\hat{T}(\lambda) \equiv \lambda$, $\lambda \in \sigma(T)$. Let $\Phi: C(\sigma(T)) \to A_T$ be the inverse of this isomorphism. Then Φ is an extension of the operational calculus $f \mapsto f(T)$ defined in § V.8 for locally analytic functions f in $\mathfrak{A}(T)$.*

Proof. By Theorem 6.5, the spectrum of T equals the spectrum of T with respect to the B^*-algebra A_T. Applying Theorem 6.4 to A_T, we find that the Gelfand transform $\hat{T}$ of T is a homeomorphism of the carrier space of A_T onto $\sigma(T)$. Thus we may identify the carrier space of A_T with $\sigma(T)$ and consider the algebra $\hat{A}_T$ of transforms as an algebra of functions on $\sigma(T)$. (See the remarks following Theorem 4.9.) With this identification, we have $\hat{T}(\lambda) \equiv \lambda$. By the Gelfand–Naimark theorem, $\hat{A}_T = C(\sigma(T))$ and the Gelfand representation is an isometric *-isomorphism of A_T onto $C(\sigma(T))$.

Next, observe that $(\lambda I - T)^{-1} \in A_T$ for $\lambda \in \rho(T)$, because of Theorem 6.5. Hence, if $f \in \mathfrak{A}(T)$ and if

$$f(T) = \frac{1}{2\pi i} \int_C f(\lambda)(\lambda I - T)^{-1}\, d\lambda,$$

as in the operational calculus of § V.8, then $f(T) \in A_T$. The second assertion of the theorem now follows from the proof in § 5 that analytic functions operate on $\hat{A}_T$, for the proof of Theorem 5.4 shows that the Gelfand transform of $f(T)$ is $f \circ \hat{T}$. But since $\hat{T}(\lambda) \equiv \lambda$, the transform of $f(T)$ is just f. Thus, if Φ is the inverse of the Gelfand representation, then $\Phi(f) = f(T)$. $\square$

Theorem 7.1 easily yields the spectral theorem for a normal operator when H is finite-dimensional. Indeed, in this case $\sigma(T)$ is a finite set $\{\lambda_1, \ldots, \lambda_k\}$ and $C(\sigma(T))$ is the algebra of *all* complex-valued functions on $\sigma(T)$. In particular, for $i = 1, \ldots, k$, the characteristic function f_i of $\{\lambda_i\}$ is the transform of some operator E_i in A. Clearly $f_i(\lambda) = f_i(\lambda)^2 = \overline{f_i(\lambda)}$ for $\lambda \in \sigma(T)$.

Hence, from properties of the Gelfand representation,

$$E_i = E_i^2 = E_i^*;$$

that is, each E_i is a self-adjoint projection. Also,

$$E_i E_j = 0 \qquad \text{if } i \neq j,$$

$$E_1 + \cdots + E_k = I.$$

Now for $\lambda \in \sigma(T)$, we have $\hat{T}(\lambda) = \lambda = \Sigma_1^k \lambda_i f_i(\lambda)$. Hence

$$T = \sum_{i=1}^{k} \lambda_i E_i. \tag{7-1}$$

This representation of T is the spectral theorem for a normal operator on a finite-dimensional space.

When generalizing (7-1) to the infinite-dimensional case, we would like to show that T may be approximated in the norm of $L(H)$ by linear combinations of self-adjoint projections. However, in general these projections will not belong to A_T. Our method of proof will involve finding a larger subalgebra of $L(H)$ that contains the desired projections. To do this we shall consider the inverse of the Gelfand mapping, $\Phi: C(\sigma(T)) \to A_T$, and then extend this mapping to a *-homomorphism from a larger algebra of complex-valued functions into $L(H)$. The appropriate algebra for our purposes will be the B^*-algebra $B(\sigma(T))$ of all bounded Borel measurable functions on $\sigma(T)$ with the supremum norm. Once we have this homomorphism, we will be able to represent T in a manner similar to (7-1). The infinite-dimensional case, however, will require an integral in place of a finite sum, and the integral will be defined in terms of an operator-valued measure.

Spectral Measures

Given a locally compact Hausdorff space Δ and a Hilbert space H, a *spectral measure on* Δ is an operator-valued function E from the Borel subsets of Δ into $L(H)$ with the following properties:

1. $E(\delta)$ is a self-adjoint projection in H for each Borel set δ.
2. $E(\varnothing) = 0$, $E(\Delta) = I$.
3. For $x, y \in H$, the function $E_{x,y}(\cdot)$ defined by $E_{x,y}(\delta) = (E(\delta)x, y)$ is a regular complex Borel measure on Δ.
4. $E(\delta_1 \cup \delta_2) = E(\delta_1) + E(\delta_2)$, if $\delta_1 \cap \delta_2 = \varnothing$.
5. $E(\delta_1 \cap \delta_2) = E(\delta_1)E(\delta_2)$.

Note from (1) that $E(\delta) = E(\delta)^2 = E(\delta)^* E(\delta)$, and so

$$E_{x,x}(\delta) = (E(\delta)x, x) = \|E(\delta)x\|^2, \qquad x \in H. \tag{7-2}$$

Thus $E_{x,x}$ is a positive measure, whose total variation is given by

$$E_{x,x}(\Delta) = \|x\|^2. \tag{7-3}$$

In what follows it will be necessary to consider integrals of scalar functions with respect to a spectral measure E on Δ. First, consider a simple function f on X; that is, suppose f has the form

$$f = \sum_{i=1}^{n} c_i \chi_{\delta_i},$$

where the c_i are complex and the χ_{δ_i} are characteristic functions of mutually disjoint Borel sets δ_i in Δ. We define the integral of f with respect to E to be the operator

$$\int_\Delta f\,dE = \int_\Delta \left[\sum_{i=1}^{n} c_i \chi_{\delta_i}\right] dE \equiv \sum_{i=1}^{n} c_i E(\delta_i).$$

Since the δ_i are disjoint, property (4) of the spectral measure insures that this integral is independent of the particular representation of f. Now, for x, $y \in H$, we have

$$\left(\left[\int_\Delta f\,dE\right]x, y\right) = \sum c_i (E(\delta_i)x, y) = \sum c_i E_{x,y}(\delta_i) = \int_\Delta f\,dE_{x,y}. \tag{7-4}$$

Hence, using (7-3), we have

$$\left|\left(\left[\int_\Delta f\,dE\right]x, x\right)\right| \leq \int_\Delta |f|\,dE_{x,x} \leq \sup_{t\in\Delta} |f(t)| \cdot \|x\|^2.$$

From the formula (1-3) in Chapter VI relating quadratic and bilinear forms, and from Theorem VI.1.1, we conclude that

$$\left\|\int_\Delta f\,dE\right\| \leq 2 \cdot \sup_{t\in\Delta} |f(t)|. \tag{7-5}$$

Now the simple functions are dense in the algebra $B(\Delta)$ of bounded Borel measurable functions on Δ under the supremum norm. Hence it follows from (7-5) and the completeness of $L(H)$ that the integral of $f \in B(\Delta)$ may be defined unambiguously by

$$\int_\Delta f\,dE = \lim_{n\to\infty} \int_\Delta f_n\,dE,$$

where $\{f_n\}$ is any sequence of simple functions whose limit in $B(\Delta)$ is f. This definition of the integral is analogous to the definition in the case of a scalar-valued measure. Thus we may apply (7-4) to an approximating sequence of simple functions to conclude that

$$\left(\left[\int_\Delta f\,dE\right]x,\, y\right) = \int_\Delta f\,dE_{x,y}, \qquad x,\, y \in H, \tag{7-6}$$

for each $f \in B(\Delta)$.

The Spectral Theorem

The spectral theorem for a normal operator will be a special case of the following result.

Theorem 7.2. *Let A be a commutative C^*-subalgebra of $L(H)$, and let $\mathfrak{M}$ be the carrier space of A. Then there exists a unique spectral measure E on $\mathfrak{M}$ such that*

$$T = \int_{\mathfrak{M}} \hat{T}\,dE, \qquad T \in A, \tag{7-7}$$

where $\hat{T}$ is the Gelfand transform of T. Furthermore,

(*a*) *The isometric *-isomorphism $\hat{T} \mapsto \int_{\mathfrak{M}} \hat{T}\,dE$ of $C(\mathfrak{M})$ onto A extends to a norm-decreasing *-homomorphism $f \mapsto \int_{\mathfrak{M}} f\,dE$ of $B(\mathfrak{M})$ into $L(H)$.*

(*b*) *$TE(\delta) = E(\delta)T$ for each Borel set $\delta \subset \mathfrak{M}$ and each $T \in A$.*

(*c*) *An operator $S \in L(H)$ commutes with every $T \in A$ if and only if $SE(\delta) = E(\delta)S$ for each Borel set $\delta \subset \mathfrak{M}$.*

Proof. In this proof, all integrals will be taken over the set $\mathfrak{M}$. It follows from the Gelfand–Naimark theorem that the mapping $\hat{T} \mapsto T$ is an isometric *-isomorphism of $C(\mathfrak{M})$ onto A. Hence, for x, $y \in H$, the inequality

$$|(Tx, y)| \le \|T\|\,\|x\|\,\|y\| = \|\hat{T}\|_\infty \|x\|\,\|y\|$$

shows that the linear mapping $\hat{T} \mapsto (Tx, y)$ is a continuous linear functional on $C(\mathfrak{M})$, whose norm does not exceed $\|x\|\,\|y\|$. By the Riesz–Kakutani theorem (Theorem III.5.7), there exists a unique regular complex Borel measure $\mu_{x,y}$ on $\mathfrak{M}$, for each x, $y \in H$, such that

$$(Tx, y) = \int \hat{T}\,d\mu_{x,y}, \qquad x,\, y \in H,\ T \in A, \tag{7-8}$$

and

$$\|\mu_{x,y}\| \le \|x\|\,\|y\|. \tag{7-9}$$

For any scalar α, we have

$$\int \hat{T}\,d\mu_{\alpha x,y} = (T(\alpha x), y) = \alpha(Tx, y) = \alpha \int \hat{T}\,d\mu_{x,y}. \tag{7-10}$$

Since $\hat{T}$ ranges over all of $C(\mathfrak{M})$ as T varies in A, and since the measures $\mu_{\alpha x,y}$ and $\mu_{x,y}$ are uniquely determined by the linear functionals they represent, we conclude from (7-10) that $\mu_{\alpha x,y} = \alpha\mu_{x,y}$. Continuing with similar arguments, one easily verifies that for each Borel set δ, $\mu_{x,y}(\delta)$ is a bilinear form on H (i.e., linear in x and conjugate linear in y). It follows that if $f \in B(\mathfrak{M})$, then $\int f\, d\mu_{x,y}$ is also a bilinear form on H and, by (7-9), this form is continuous. Hence, by Theorem VI.1.2, there exists an operator T_f in $L(H)$ such that

$$(T_f x, y) = \int f\, d\mu_{x,y}, \qquad x, y \in H. \tag{7-11}$$

If $f \in C(\mathfrak{M})$, then $f = \hat{T}$ for some $T \in A$, and then (7-8) and (7-11) together imply that $T_f = T$. Thus the mapping $f \mapsto T_f$ is an extension of the inverse Gelfand mapping $\hat{T} \mapsto T$ from $C(\mathfrak{M})$ onto A. In due course we shall show that T_f is the integral of f with respect to an appropriate spectral measure.

Our immediate goal is to show that the correspondence $f \mapsto T_f$ is a norm-decreasing *-homomorphism from $B(\mathfrak{M})$ into $L(H)$. Clearly it is linear, by (7-11). The proof that it is multiplicative requires the following observation from measure theory.

Given $g \in B(\mathfrak{M})$ and given regular complex Borel measures μ, ν on $\mathfrak{M}$, if

$$\int fg\, d\mu = \int f\, d\nu \tag{7-12}$$

for all $f \in C(\mathfrak{M})$, then (7-12) also holds for all $f \in B(\mathfrak{M})$.

(First observe that by linearity and continuity of the integrals, it suffices to prove (7-12) when f is the characteristic function of a Borel set. Then approximate such an f in both the $L^1(|\mu|)$ and $L^1(|\nu|)$ norms by continuous functions.) Now take R and S in A and note that since $\hat{R}\hat{S} = \widehat{RS}$, we have from (7-8) that

$$\int \hat{R}\hat{S}\, d\mu_{x,y} = (RSx, y) = \int \hat{R}\, d\mu_{Sx,y}. \tag{7-13}$$

Since $\hat{R}$ varies over all of $C(\mathfrak{M})$, we conclude that the integrals in (7-13) are equal when $\hat{R}$ is replaced by any $f \in B(\mathfrak{M})$. Then, if $z = T_f^* y$,

$$\begin{aligned} \int f\hat{S}\, d\mu_{x,y} &= \int f\, d\mu_{Sx,y} = (T_f Sx, y) \\ &= (Sx, z) = \int \hat{S}\, d\mu_{x,z}. \end{aligned} \tag{7-14}$$

Once again, the truth of (7-14) for every $\hat{S} \in C(\mathfrak{M})$ implies the equality of the first and last terms when $\hat{S}$ is replaced by any $g \in B(\mathfrak{M})$. Hence

$$\int fg\, d\mu_{x,y} = \int g\, d\mu_{x,z} = (T_g x, z) = (T_f T_g x, y).$$

It follows easily that

$$T_{fg} = T_f T_g, \qquad f, g \in B(\mathfrak{M}). \tag{7-15}$$

Thus $f \mapsto T_f$ is an algebra homomorphism. To prove it preserves involutions, consider $T \in A$ such that $\hat{T}$ is real valued. Then T is self-adjoint (because the Gelfand mapping preserves involutions) and

$$\int \hat{T}\, d\mu_{x,y} = (Tx, y) = \overline{(Ty, x)} = \overline{\int \hat{T}\, d\mu_{y,x}} = \int \hat{T}\, d\overline{\mu_{y,x}}. \tag{7-16}$$

By considering real and imaginary parts of functions, one sees that the first and last integrals in (7-16) agree for all $\hat{T} \in C(\mathfrak{M})$. Hence $\mu_{x,y} = \overline{\mu_{y,x}}$. Now for $f \in B(\mathfrak{M})$,

$$\begin{aligned}(T_f x, y) &= \int f\, d\mu_{x,y} = \overline{\int \bar{f}\, d\overline{\mu_{x,y}}} = \overline{\int \bar{f}\, d\mu_{y,x}} \\ &= \overline{(T_{\bar{f}} y, x)} = (x, T_{\bar{f}} y).\end{aligned}$$

This holds for all $x, y \in H$, and hence

$$(T_f)^* = T_{\bar{f}}. \tag{7-17}$$

Finally, using (7-17), (7-15), and (7-9), we have

$$\begin{aligned}\|T_f x\|^2 &= (T_f^* T_f x, x) = (T_{\bar{f}f} x, x) = \int \bar{f}f\, d\mu_{x,x} \\ &\leq \|\mu_{x,x}\| \cdot \sup_{t \in \mathfrak{M}} |f(t)|^2 \leq \|x\|^2 \cdot \sup_{t \in \mathfrak{M}} |f(t)|^2,\end{aligned}$$

whence

$$\|T_f\| \leq \sup_{t \in \mathfrak{M}} |f(t)|, \qquad f \in B(\mathfrak{M}).$$

It is now easy to define and study the spectral measure E. Given a Borel set $\delta \subset \mathfrak{M}$, let f be the characteristic function of δ, and let $E(\delta) = T_f$. Since f is real valued, $E(\delta)$ is self-adjoint, by (7-17). Furthermore, since $f^2 = f$, (7-15) shows that $E(\delta)$ is a projection. Obviously if $\delta = \varnothing$, then $f \equiv 0$ and $E(\varnothing)$ is the zero operator. If $\delta = \mathfrak{M}$, then f is the Gelfand transform of the identity operator; hence $E(\mathfrak{M}) = I$, by the remark following (7-11). For $x, y \in H$ (and f the characteristic function of δ), we find that

$$E_{x,y}(\delta) = (E(\delta)x, y) = \int f\, d\mu_{x,y} = \mu_{x,y}(\delta). \tag{7-18}$$

Hence $E_{x,y}(\cdot)$ is a regular complex Borel measure. If δ_1 and δ_2 are Borel subsets of $\mathfrak{M}$, with characteristic functions f_1 and f_2, respectively, then $f_1 f_2$ is the characteristic function of $\delta_1 \cap \delta_2$, and (7-15) implies that

$$E(\delta_1 \cap \delta_2) = E(\delta_1)E(\delta_2).$$

If, in addition, δ_1 and δ_2 are disjoint, then $f_1 + f_2$ is the characteristic function of $\delta_1 \cup \delta_2$. Hence the linearity of the mapping $f \mapsto T_f$ implies that $E(\delta_1 \cup \delta_2) = E(\delta_1) + E(\delta_2)$. Consequently, we have proved that E is a spectral measure on $\mathfrak{M}$.

Now, for $f \in B(\mathfrak{M})$ and $x, y \in H$, we claim that

$$\left(\left[\int f\,dE\right]x, y\right) = \int f\,dE_{x,y} = \int f\,d\mu_{x,y} = (T_f x, y).$$

The first equality is a property of an integral with respect to a spectral measure (see (7-6)). The second equality comes from (7-18) and the third from (7-11). Since x and y were arbitrary,

$$T_f = \int f\,dE.$$

Hence both formula (7-7) and statement (a) of the theorem follow from what we have already proved about the mapping $f \mapsto T_f$. Statement (b) is obviously a special case of (c) because the algebra A is commutative.

To prove (c), suppose S commutes with each T in A. For arbitrary $x, y \in H$, let $z = S^*y$ and note that

$$\int \hat{T}\,dE_{Sx,y} = (TSx, y) = (STx, y) = (Tx, z) = \int \hat{T}\,dE_{x,z}. \tag{7-19}$$

This can hold for all $\hat{T} \in C(\mathfrak{M})$ only if the measures $E_{Sx,y}$ and $E_{x,z}$ are equal, that is, only if

$$(E(\delta)Sx, y) = (E(\delta)x, z) = (SE(\delta)x, y) \tag{7-20}$$

for all δ. Since (7-20) holds for all $x, y \in H$, the operators $E(\delta)$ and S commute. Conversely, if (7-20) holds for all δ, then $E_{Sx,y} = E_{x,z}$ and the integrals in (7-19) must be equal for all $\hat{T}$. It is easy to see that this forces S to commute with each $T \in A$. This proves (c).

A similar argument shows that the spectral measure E is unique. Suppose E' is another spectral measure satisfying condition (7-7). Then, for $x, y \in H$ and any $\hat{T} \in C(\mathfrak{M})$,

$$\int \hat{T}\,dE_{x,y} = \int \hat{T}\,dE'_{x,y}$$

by (7-6). Hence $E_{x,y} = E'_{x,y}$ for $x, y \in H$. Of course, this implies that $E(\delta) = E'(\delta)$ for each Borel set δ. $\square$

Theorem 7.3. *Let T be a normal operator in $L(H)$. Then there exists a unique spectral measure E on $\sigma(T)$ such that*

$$T = \int_{\sigma(T)} \lambda \, dE. \tag{7-21}$$

(*a*) *For each $f \in B(\sigma(T))$, let*

$$f(T) = \int_{\sigma(T)} f \, dE.$$

Then $f \mapsto f(T)$ is a norm-decreasing $*$-homomorphism of $B(\sigma(T))$ into $L(H)$ which is an extension of the inverse of the Gelfand representation of Theorem 7.1 *and hence is an extension of the operational calculus of* § *V*.8.

(*b*) *For $f \in B(\sigma(T))$ and $x \in H$,*

$$\|f(T)x\|^2 = \int_{\sigma(T)} |f|^2 \, dE_{x,x}.$$

(*c*) *If $S \in L(H)$ commutes with T and T^*, then S commutes with each $f(T)$, for $f \in B(\sigma(T))$.*

Proof. The carrier space of the C^*-algebra A_T generated by T is $\sigma(T)$, and $\hat{T}(\lambda) \equiv \lambda$. Thus the existence of the spectral measure E, the formula (7-21) for T, and statement (*a*) above all follow immediately from Theorem 7.2. The uniqueness of E is considered in problem 7. For part (*b*), we have from (*a*) and (7-6) that

$$\begin{aligned} \|f(T)x\|^2 &= (f(T)x, f(T)x) = (f(T)^* f(T)x, x) \\ &= ((\bar{f}f)(T)x, x) = \left(\left[\int |f|^2 \, dE\right]x, x\right) \\ &= \int_{\sigma(T)} |f|^2 \, dE_{x,x}. \end{aligned}$$

For (*c*), note that if S commutes with T and T^*, then S commutes with every polynomial in T and T^* and hence with each operator in A_T. By Theorem 7.2(*c*), S commutes with $E(\delta)$ for each Borel set $\delta \subset \sigma(T)$. Hence S commutes with $f(T)$ for each simple function f in $B(\sigma(T))$. A routine argument using (7-5) then shows that S commutes with $f(T)$ for all $f \in B(\sigma(T))$. □

The unique spectral measure E in Theorem 7.3 is often called *the resolution of the identity for T* because

$$I = E(\sigma(T)) = \int_{\sigma(T)} 1 \, dE.$$

The statement of Theorem 7.3(c) may be improved by the use of Fuglede's theorem, which asserts that if $S \in L(H)$ commutes with a normal operator T, then S must also commute with T^*. See problem 9. One common use of Theorem 7.3(c) occurs when $f(T)$ is a self-adjoint projection $E(\delta)$ corresponding to a Borel set δ in $\sigma(T)$. Recall from Theorem V.5.1 that an operator S commutes with $E(\delta)$ if and only if S is completely reduced by the (orthogonal) direct sum $H = \mathscr{R}(E(\delta)) \oplus \mathscr{N}(E(\delta))$. Consequently, by Theorem 7.3(c), this direct sum reduces not only T, but also every $S \in L(H)$ that commutes with T.

Note that the *-homomorphism $f \mapsto f(T)$ of Theorem 7.3 must map characteristic functions of Borel sets into self-adjoint projections. Since the function $f(\lambda) \equiv \lambda$ may be approximated in $B(\sigma(T))$ by linear combinations of characteristic functions, we conclude from the representation (7-21) and the definition of the integral that T is the limit in the norm of $L(H)$ of linear combinations of self-adjoint projections. Thus (7-21) is a generalization of the formula $T = \Sigma\, \lambda_i E_i$ discussed at the beginning of the section.

The next theorem takes another point of view. The representation $T = \Sigma\, \lambda_i E_i$ shows that T acts as multiplication by λ_i on the subspace $\mathscr{R}(E_i)$. Theorem 7.4 below considers a normal operator that may be represented, via a unitary equivalence, as "multiplication by λ" on a suitable L^2-space of functions. The general case is treated in Theorem 7.5, where it will be shown that each normal operator is unitarily equivalent to a multiplication operator of the type introduced in Example 6, § 1. For simplicity, we shall consider the C^*-algebra A_T generated by a single normal operator T. It should be fairly obvious how to rephrase the theorems and their proofs for a general commutative C^*-algebra.

Theorem 7.4. *Let T be a normal operator in $L(H)$ and suppose H contains a "cyclic vector" z such that the set $\{Sz : S \in A_T\}$ is dense in H. Then there exist a finite positive regular Borel measure μ on $\sigma(T)$ and a unitary transformation U from H onto $L^2(\sigma(T), \mu)$ such that*

$$(UTU^{-1}g)(\lambda) = \lambda g(\lambda), \qquad \lambda \in \sigma(T), \tag{7-22}$$

for $g \in L^2(\sigma(T), \mu)$. Moreover, if $f \in B(\sigma(T))$ and if $f(T)$ is the operator defined in Theorem 7.3, then

$$f(T) = U^{-1} M_f U, \tag{7-23}$$

where M_f is the operator on $L^2(\sigma(T), \mu)$ of multiplication by f.

Proof. Let E be the spectral measure from Theorem 7.3 and let $\mu = E_{z,z}$. Then μ is a finite positive regular Borel measure on $\sigma(T)$. For $f \in B(\sigma(T))$, we have (using (7-6))

$$(f(T)z, z) = \int_{\sigma(T)} f\, dE_{z,z} = \int_{\sigma(T)} f\, d\mu. \tag{7-24}$$

Let $H_0 = \{f(T)z : f \in B(\sigma(T))\}$, and note that $H_0 \supset \{f(T)z : f \in C(\sigma(T))\} = \{Sz : S \in A_T\}$, by Theorem 7.1. Thus H_0 is dense in H. We define a mapping U_0 from H_0 into $L^2(\sigma(T), \mu)$ by

$$U_0(f(T)z) = f, \qquad f \in B(\sigma(T)). \tag{7-25}$$

If $f(T)z = g(T)z$ for $f, g \in B(\sigma(T))$, then Theorem 7.3(b) shows that

$$0 = \|(f-g)(T)z\|^2 = \int_{\sigma(T)} |f-g|^2\, dE_{z,z}$$

$$= \int_{\sigma(T)} |f-g|^2\, d\mu.$$

It follows that $f = g$ μ-almost everywhere. Thus U_0 is well defined. Clearly U_0 is linear. Also, from Theorem 7.3(b),

$$\|U_0 f(T)z\|_{L^2}^2 = \int_{\sigma(T)} |f|^2\, d\mu = \|f(T)z\|^2,$$

so that U_0 is an isometry. Since H_0 is dense in H, the operator U_0 extends uniquely to an isometric linear mapping U of H onto the closure in $L^2(\sigma(T), \mu)$ of $B(\sigma(T))$, that is, onto $L^2(\sigma(T), \mu)$. By Theorem VI.7.1, U is a unitary operator. Given $f \in B(\sigma(T))$, we shall show that $Uf(T)U^{-1}$ coincides with the bounded operator M_f on a dense subspace of $L^2(\sigma(T), \mu)$. For $g \in B(\sigma(T)) \subset L^2(\sigma(T), \mu)$, we use (7-25) to find that

$$(Uf(T)U^{-1})g = U_0 f(T)(U_0^{-1} g) = U_0 f(T) g(T) z$$

$$= U_0(fg)(T)z = fg = M_f(g).$$

Hence $Uf(T)U^{-1} = M_f$ and $f(T) = U^{-1} M_f U$. This proves (7-23). Of course, (7-22) is the special case when $f(\lambda) = \lambda$ and $f(T) = T$. □

In the setting of Theorem 7.4, let $\Psi(S) = USU^{-1}$ for $S \in L(H)$. It is easily seen that Ψ is an isometric *-isomorphism of $\{f(T) : f \in B(\sigma(T))\}$ onto the algebra $\{M_f : f \in B(\sigma(T))\}$. Because it is implemented by a unitary transformation between the underlying Hilbert spaces, Ψ is said to be a *spatial isomorphism*. Rather than being only an algebraic isomorphism that preserves the norm and involution, Ψ involves a fixed unitary equivalence between the algebras so that corresponding elements behave the same *as operators*, not just as members of their respective algebras.

The generalization of Theorem 7.4 to the case of an arbitrary normal operator T will be handled by decomposing the Hilbert space H into a direct sum of orthogonal subspaces, on each of which we can apply Theorem 7.4. As a preliminary, we note that if a closed subspace H_1 of H is invariant under T and T^*, then so is $H_1^\perp$. (For if $x \in H_1^\perp$, then for every $y \in H_1$, we have $T^*y \in H_1$ and $0 = (x, T^*y) = (Tx, y)$, which implies that $Tx \in H_1^\perp$. Similarly, $T^*x \in H_1^\perp$.) Thus T and T^* are completely reduced by the orthogonal direct sum $H = H_1 \oplus H_1^\perp$. It follows easily that every $S \in A_T$ is also completely reduced by this direct sum.

Let us say that a closed subspace H_1 is a *cyclic subspace* if H_1 is invariant under T and T^* and if there exists $z \in H_1$ such that $\{Sz : S \in A_T\}$ is dense in H_1. Cyclic subspaces clearly exist—simply take the closure of the set $\{Sz : S \in A_T\}$ for any nonzero z. If H_1 is a cyclic subspace, then we may restrict the algebra A_T to $H_1^\perp$ and find a cyclic subspace H_2 in $H_1^\perp$. The orthogonal complement of H_2 in $H_1^\perp$ will also contain a cyclic subspace, and so on. An easy application of Zorn's lemma shows that there is a maximal collection $\{H_\nu\}$ of mutually orthogonal cyclic subspaces, where ν runs over some index set. The subspace K spanned by the vectors in $\bigcup_\nu H_\nu$ is clearly invariant under T and T^*. If $K^\perp$ were nonzero, there would exist a nonzero cyclic subspace in $K^\perp$ that could be added to the collection $\{H_\nu\}$, contradicting the maximality of $\{H_\nu\}$. Thus $K^\perp = (0)$, and the closed linear manifold generated by $\{H_\nu\}$ is H. Let P_ν be the orthogonal projection of H onto H_ν. Since the H_ν are mutually orthogonal, an elementary argument shows that if $x \in H$, then $x = \Sigma_\nu P_\nu x$, where $P_\nu x = 0$ for all but a countable number of indices ν, and $\|x\|^2 = \Sigma_\nu \|P_\nu x\|^2$. For this reason we call H the orthogonal direct sum of $\{H_\nu\}$.

Theorem 7.5. *Let T be a normal operator on H. Then there exist a measure space $(X, \mathcal{S}, \mu)$ with X a locally compact metric space and μ a positive regular measure, a bounded continuous complex-valued function h on X, and a unitary transformation U from H onto $L^2(X, \mu)$ such that*

$$T = U^{-1} M_h U,$$

where M_h is the operator on $L^2(X, \mu)$ of multiplication by h.

Proof. From the discussion above, we may represent H as an orthogonal direct sum of cyclic subspaces $\{H_\nu\}$. Let A_ν denote the C^*-subalgebra of $L(H_\nu)$ generated by the restriction T_ν of T to H_ν. Then Theorem 7.4 applies to A_ν, and so there exist a finite regular Borel measure μ_ν on $\sigma(T_\nu)$ and a unitary operator U_ν from H_ν onto $L^2(\sigma(T_\nu), \mu_\nu)$ such that

$$(U_\nu T_\nu x_\nu)(\lambda) = \lambda (U_\nu x_\nu)(\lambda), \qquad \lambda \in \sigma(T_\nu), \tag{7-26}$$

for $x_\nu \in H_\nu$.

Our next step is to construct a suitable measure space $(X, \mathcal{S}, \mu)$. Henceforth, let us consider each set $\sigma(T_\nu)$ as belonging to a distinct copy of the complex plane, and let X be the disjoint union $\bigcup_\nu \sigma(T_\nu)$. Define a metric on X

by using the usual metric of the complex plane for points belonging to the same set $\sigma(T_\nu)$ and by setting the distance equal to $\|T\|$ for two points belonging to distinct sets. Then X is a locally compact metric space in which the sets $\sigma(T_\nu)$ are both open and compact. Let $\mathscr{S}$ be the collection of all (disjoint) unions $\delta = \bigcup_\nu \delta_\nu$, where δ_ν is a Borel subset of $\sigma(T_\nu)$ for each ν. For $\delta \in \mathscr{S}$, define $\mu(\delta) = \Sigma_\nu \mu_\nu(\delta_\nu)$ when the series has at most a countable number of nonzero terms and converges, and $\mu(\delta) = +\infty$ otherwise. One can verify that $(X, \mathscr{S}, \mu)$ is a measure space, with μ a regular measure.

Now for each ν and each $x_\nu \in H_\nu$, $U_\nu x_\nu$ is an L^2-function on $\sigma(T_\nu)$. Let us extend the domain of this function to all of X and make the function zero on $X \backslash \sigma(T_\nu)$. Clearly then $U_\nu x_\nu \in L^2(X, \mu)$. Let $\tilde{H}$ be the subset of H of all finite sums of the form $x = \Sigma_{\nu \in F} x_\nu$, where F represents a finite set of indices, and define $\tilde{U}$ from $\tilde{H}$ into $L^2(X, \mu)$ by

$$\tilde{U}\Big(\sum_{\nu \in F} x_\nu\Big) = \sum_{\nu \in F} U_\nu x_\nu. \tag{7-27}$$

The domain of $\tilde{U}$ is obviously dense in H. The range of $\tilde{U}$ is dense in $L^2(X, \mu)$ because it includes, for example, all bounded μ-measurable functions that are zero on all but a finite number of the sets $\sigma(T_\nu)$. Since $\tilde{U}$ is obviously linear and isometric, it extends uniquely to a unitary transformation U from H onto $L^2(X, \mu)$. Finally, let h be the function on X whose values on the subset $\sigma(T_\nu)$ are given by the function $h_\nu(\lambda) = \lambda$. Since each h_ν is continuous, so is h. Also, each set $\sigma(T_\nu)$ lies inside the disc $|\lambda| \leqslant \|T\|$, because the norm of the restriction T_ν does not exceed $\|T\|$. Hence h is a bounded function. For $x = \Sigma_{\nu \in F} x_\nu$ in $\tilde{H}$, we have from (7-26) and (7-27),

$$\begin{aligned}(UTx)(\lambda) &= \sum_{\nu \in F} (U_\nu T_\nu x_\nu)(\lambda) = \sum_{\nu \in F} h(\lambda)(U_\nu x_\nu)(\lambda) \\ &= h(\lambda) \sum_{\nu \in F} (U_\nu x_\nu)(\lambda) = h(\lambda)(Ux)(\lambda) \\ &= (M_h U x)(\lambda).\end{aligned}$$

Since $\tilde{H}$ is dense in H, we conclude that $UT = M_h U$ and $T = U^{-1} M_h U$. □

The final two theorems below give simple applications of the spectral theorem for normal operators to the spectral theory developed in Chapters V and VI. An extensive survey of more significant applications of the spectral theorem may be found in Dunford and Schwartz [2, pages 937–1184]. Among the topics discussed there are compact topological groups, almost periodic functions, group algebras and Fourier transforms, classical Tauberian theorems, and Hilbert–Schmidt operators.

Theorem 7.6. *Let T be a normal operator in $L(H)$ and let σ be a relatively open and closed subset of $\sigma(T)$. Then the spectral projection P associated with the spectral set σ is self-adjoint and commutes with T and T^*.*

Proof. Recall from the operational calculus of § V.9 that $P=f(T)$, where f is a locally analytic function on $\sigma(T)$ such that $f(\lambda)=1$ on σ and $f(\lambda)=0$ on $\sigma(T)\backslash\sigma$. The restriction of f to $\sigma(T)$ belongs to $B(\sigma(T))$ and equals the characteristic function of σ. Hence, by Theorem 7.3, $f(T)$ is the self-adjoint projection $E(\sigma)$, where E is the spectral measure associated with T, and $f(T)$ commutes with T and T^*. □

Theorem 7.7. *Let T be a normal operator in $L(H)$. Then the following statements are equivalent for $\lambda_0 \in \sigma(T)$:*

(*a*) *λ_0 is an isolated point of $\sigma(T)$;*
(*b*) *$H=\mathcal{R}(\lambda_0-T)\oplus\mathcal{N}(\lambda_0-T)$;*
(*c*) *$\mathcal{R}(\lambda_0-T)$ is closed;*
(*d*) *λ_0 is a pole of the resolvent of T.*

Proof. Given (a), let P be the spectral projection associated with the set $\{\lambda_0\}$. By Theorem 7.6, T and T^* are completely reduced by the orthogonal direct sum $H=\mathcal{N}(P)\oplus\mathcal{R}(P)$. Hence the restriction T_2 of T to $\mathcal{R}(P)$ is normal. Now $\sigma(T_2)=\{\lambda_0\}$, by Theorem V.9.2, and so λ_0-T_2 has spectral radius zero. Since T_2 is normal, λ_0-T_2 must be the zero operator on $\mathcal{R}(P)$, by Theorem VI.3.5 (or Theorem 6.1(c)). From formulas (10-8) and (10-9) in § V.10, we conclude that $\mathcal{R}(\lambda_0-T)=\mathcal{N}(P)$ and $\mathcal{N}(\lambda_0-T)=\mathcal{R}(P)$. This proves that ($a$) implies ($b$). We see that ($b$) and ($c$) are equivalent by applying Theorem VI.3.6 to the normal operator λ_0-T. Next, it is not difficult to show that (b) implies that the ascent and descent of λ_0-T are both 1. (See problem 4, § V.6, or use Theorem VI.3.7 and problem 6, § VI.3.) Hence λ_0 is a pole of the resolvent of T, by Theorem V.10.2. Thus (b) implies (d). Finally, (d) obviously implies (a). □

PROBLEMS

1. Let A be a C^*-algebra of operators such that each element of A is normal. Then A is commutative.
2. Properties (4) and (5) of a spectral measure may be deduced from properties (1) and (3). Let E be a spectral measure on Δ, and let δ_1, δ_2 be Borel subsets of Δ. Show the following:
   a. If $\delta_1\cap\delta_2=\varnothing$, then $E(\delta_1\cup\delta_2)=E(\delta_1)+E(\delta_2)$.
   b. If $\delta_1\subset\delta_2$, then $E(\delta_1)E(\delta_2)=E(\delta_2)E(\delta_1)=E(\delta_1)$. [*Hint.* Use Theorem VI.2.4.]
   c. If $\delta_1\cap\delta_2=\varnothing$, then $E(\delta_1)E(\delta_2)=0$.
   d. $E(\delta_1\cap\delta_2)=E(\delta_1)E(\delta_2)$.

3. Let E be a spectral measure on δ with values in $L(H)$. Let δ be the union of a sequence $\{\delta_n\}$ of disjoint Borel sets in Δ. Then, for $x \in H$, $\Sigma_n E(\delta_n)x$ converges in H to $E(\delta)x$.
4. Let E be a spectral measure on Δ with values in $L(H)$ and, for $f \in B(\Delta)$, let $T_f = \int_\Delta f\, dE$. Show the following:
   a. The mapping $f \mapsto T_f$ is a *-homomorphism.
   b. For $x \in H$, $\|T_f x\|^2 = \int_\Delta |f|^2\, dE_{x,x}$.
   c. $\|T_f\| \leq \sup\{|f(t)| : t \in \Delta\}$.
5. Let A be a commutative C^*-subalgebra of $L(H)$, with carrier space $\mathfrak{M}$. Let E be the spectral measure given by Theorem 7.2. For $f \in B(\mathfrak{M})$, show that $\int_{\mathfrak{M}} f\, dE$ belongs to the closure of A in the strong operator topology on $L(H)$.
6. Let A and $\mathfrak{M}$ be as in problem 5. Show that $E(\delta) \neq 0$ for each nonempty open set $\delta \subset \mathfrak{M}$.
7. Show that the spectral measure in Theorem 7.3 is uniquely determined by the condition (7-21). [*Hint.* Use problem 4.]
8. Let S and T be normal operators in $L(H)$. Then $\sigma(S) = \sigma(T)$ if and only if there is an isometric *-isomorphism Φ from A_S onto A_T such that $\Phi(S) = T$. [*Hint.* Use Theorem 7.1.]
9. (Fuglede's theorem). Suppose that $S, T \in L(H)$, where T is normal and $ST = TS$. Then $ST^* = T^*S$. [*Hint.* (Rosenblum) Define $F(\lambda) = \exp(\lambda T^*) \cdot S \cdot \exp(-\lambda T^*)$, and show that $F(\lambda) = \exp(\lambda T^* - \bar{\lambda} T) \cdot S \cdot \exp(\bar{\lambda} T - \lambda T^*)$. Observe that the operator $U = \exp(\lambda T^* - \bar{\lambda} T)$ is unitary, and conclude that $\|F(\lambda)\| \leq \|S\|$ for all λ.]
10. Let T be a normal operator, and let $\{f_n\}$ be a uniformly bounded sequence of functions in $B(\sigma(T))$ that converges pointwise to $f \in B(\sigma(T))$. Then $\{f_n(T)\}$ converges to $f(T)$ in the strong operator topology.
11. Let T be a normal operator in $L(H)$, and suppose that A_T contains a projection E with $E \neq 0, I$. Show that $\sigma(T)$ is not connected.
12. a. If U in $L(H)$ is unitary, then there exists a self-adjoint A in $L(H)$ such that $U = \exp(iA)$. [*Hint.* Find a real-valued $f \in B(\sigma(U))$ such that $\exp(if(z)) = z$ on $\sigma(U)$.]
    b. The set (group) of all unitary operators in $L(H)$ is connected. [*Hint.* Given U, find an arc joining U to the identity.]
    c. The set (group) of all invertible operators in $L(H)$ is connected. [*Hint.* First show that the set of all invertible positive operators is connected, and then use the polar decomposition of an invertible operator. See problem 5 in § VI.7.]
13. A normal operator $T \in L(H)$ is compact if and only if (1) $\sigma(T)$ contains at most a countable number of points, including 0, with 0 as the only possible accumulation point, and (2) $\dim \mathcal{N}(\lambda - T) < \infty$ for all $\lambda \neq 0$. For the necessity of the conditions, let $\{\lambda_n\}$ be an enumeration of the spectrum such that $|\lambda_1| \geq |\lambda_2| \geq \cdots$, and for each n define $f_n(\lambda) = \lambda$ if $\lambda \in \{\lambda_1, \ldots, \lambda_n\}$ and $f_n(\lambda) = 0$ otherwise. Show that $f_n(T)$ is compact and $\|f_n(T) - T\| \to 0$.
14. Given $T \in L(H)$, if T^*T is a compact operator, then T is compact. [*Hint.* Consider the factorization $T = VP$ described in problem 6, § VI.7.]

BIBLIOGRAPHY

Akhiezer, N. I., and I. M. Glazman

1. *Theory of Linear Operators in Hilbert Space*, vol. I, Ungar, New York, 1961.
2. *Theory of Linear Operators in Hilbert Space*, vol. II, Ungar, New York, 1963.

Alexits, G.

1. *Convergence Problems of Orthogonal Series*, Pergamon Press, New York, 1961.

Aronszajn, N., and K. Smith

1. Invariant subspaces of completely continuous operators, *Ann. Math.* 60(1954), 345–350.

Arveson, W.

1. *An Invitation to C^*-Algebras*, Springer-Verlag, New York, 1976.

Bade, W. G.

1. An operational calculus for operators with spectrum in a strip, *Pacific J. Math.* 3(1953), 257–290.

Banach, S.

1. *Théorie des Opérations Linéaires*, Monografje Matematyczne, Warsaw, 1932.

Ben-Israel, A., and T. N. E. Greville

1. *Generalized Inverses: Theory and Applications*, Wiley-Interscience, New York, 1974.

Bernstein, A., and A. Robinson

1. Solution of an invariant subspace problem of K. T. Smith and P. R. Halmos, *Pacific J. Math.* 16(1966), 421–431.

Boas, R. P.

1. Entire functions of exponential type, *Bull. Amer. Math. Soc.* 48(1942), 839–849.

Bonsall, F. F.

1. A survey of Banach algebra theory, *Bull. London Math. Soc.* 2(1970), 257–274.

Bourbaki, N.

Each of the following is in the series, *Éléments de Mathématique*, Actualités Sci. Ind., Hermann, Paris.

1. *Topologie Générale*, No. 858(1940).

2. *Algèbre Linéaire*, No. 1032(1947).
3. *Espaces Vectoriels Topologiques*, No. 1189(1953) and No. 1229(1955).

de Branges, L.

1. The Stone–Weierstrass theorem, *Proc. Amer. Math. Soc.* 10(1959), 822–824.

Clarkson, J. A.

1. Uniformly convex spaces, *Trans. Amer. Math. Soc.* 40(1936), 396–414.

Courant, R., and D. Hilbert

1. *Methods of Mathematical Physics*, vol. I, Interscience, New York, 1953.

Davie, A. M.

1. The approximation problem for Banach spaces, *Bull. London Math. Soc.* 5(1973), 261–266.

Davis, C.

1. The Toeplitz–Hausdorff theorem explained, *Canadian Math. Bull.* 14(1971), 245–246.

Day, M. M.

1. The spaces L^p with $0<p<1$, *Bull. Amer. Math. Soc.* 46(1940), 816–823.
2. *Normed Linear Spaces*, 3rd ed., Springer-Verlag, New York, 1973.

Dixmier, J.

1. Sur les bases orthonormales dans les espaces préhilbertiens, *Acta Sci. Math. Szeged* 15(1953), 29–30.

Douglas, R. G.

1. *Banach Algebra Techniques in Operator Theory*, Academic Press, New York, 1972.

Dunford, N.

1. Spectral theory I. Convergence to projections, *Trans. Amer. Math. Soc.* 54(1943), 185–217.
2. Spectral theory, *Bull. Amer. Math. Soc.* 49(1943), 637–651.
3. Spectral theory in abstract spaces and Banach algebras, *Proceedings of the Symposium on Spectral Theory and Differential Problems*, Oklahoma Agricultural and Mechanical College, Stillwater, 1951.
4. Spectral theory II. Resolutions of the identity, *Pacific J. Math.* 2(1952), 559–614.

Dunford, N., and B. J. Pettis

1. Linear operators on summable functions, *Trans. Amer. Math. Soc.* 47(1940), 323–392.

Dunford, N., and J. T. Schwartz

1. *Linear Operators, Part I: General Theory*, Interscience, New York, 1958.

2. *Linear Operators, Part II: Spectral Theory*, Interscience, New York, 1963.

Duren, P. L.

1. *Theory of H^p Spaces*, Academic Press, New York, 1970.

Eberlein, W. F.

1. Weak compactness in Banach spaces I, *Proc. Nat. Acad. Sci., U.S.A.* 33(1947), 51–53.

Edwards, R. E.

1. *Functional Analysis Theory and Applications*, Holt, Rinehart and Winston, New York, 1965.

Enflo, P.

1. A counterexample to the approximation problem in Banach spaces, *Acta Math.* 130(1973), 309–317.

Friedman, A.

1. *Generalized Functions and Partial Differential Equations*, Prentice-Hall, Englewood Cliffs, 1963.

Gamelin, T. W.

1. *Uniform Algebras*, Prentice-Hall, Englewood Cliffs, 1969.

Gelfand, I. M.

1. On normed rings, *Doklady Akad. Nauk. SSSR* 23(1939), 430–432.
2. To the theory of normed rings, II. On absolutely convergent trigonometrical series and integrals, *Doklady Akad. Nauk SSSR* 25(1939), 571–574.
3. Normierte Ringe, *Mat. Sbornik* 9(51)(1941), 3–24.
4. Über absolut konvergente trigonometrische Reihen und Integralen, *Mat. Sbornik* 9(51)(1941), 51–66.

Gelfand, I. M., and A. N. Kolmogorov

1. On rings of continuous functions on a topological space, *Doklady Akad. Nauk SSSR* 22(1939), 11–15.

Gelfand, I. M., and M. A. Naimark

1. On the embedding of normed rings into the ring of operators in Hilbert space, *Mat. Sbornik* 12(54)(1943), 197–213.

Gelfand, I. M., and D. A. Raikov

1. On the theory of characters of commutative topological groups, *Doklady Akad. Nauk SSSR* 28(1940), 195–198.

Gelfand, I. M., D. Raikov, and G. Shilov

1. Commutative Normed Rings, Chelsea, New York, 1964.

Gelfand, I. M., and G. E. Shilov

1. Über verschiedene Methoden der Einführung der Topologie in die Menge der maximalen Ideale eines normierten Ringes, *Mat. Sbornik* 9(51)(1941), 25–38.
2. *Generalized functions*, vol. 1, 2, 3, Academic Press, New York, 1964.

Gohberg, I. C., and M. G. Krein

1. The basic propositions on defect numbers, root numbers, and indices of linear operators (Russian), *Usp. Mat. Nauk* 12(1957), 43–118. Translated in *Amer. Math. Soc. Transl.*, Ser. 2, 13(1960), 185–264.

Goldberg, S.

1. Linear operators and their conjugates, *Pacific J. Math.* 9(1959), 69–79.
2. *Unbounded Linear Operators*, McGraw-Hill, New York, 1966.

Gramsch, B., and D. C. Lay

1. Spectral mapping theorems for essential spectra, *Math. Ann.* 192(1971), 17–32.

Halberg, C. J. A., Jr.

1. *Spectral theory of linked operators in the spaces ℓ^p*, Doctoral dissertation, University of California, Los Angeles, 1955.
2. The generalized Fibonacci operator, *Fibonacci Quart.* 6(1968), 15–34.

Halberg, C. J. A., Jr., and A. Samuelsson

1. On the fine structure of spectra, *Math. Scand.* 29(1971), 37–49.

Halmos, P. R.

1. *Finite-dimensional Vector Spaces*, 2nd ed., Van Nostrand, Princeton, 1958.
2. *Introduction to Hilbert space and the Theory of Spectral Multiplicity*, 2nd ed., Chelsea, New York, 1957.
3. *A Hilbert Space Problem Book*, Van Nostrand, Princeton, 1967.
4. Some unsolved problems of unknown depth about operators on Hilbert space, *Proc. Royal Soc. Edinburgh* 76(1976), 67–76.

Hamburger, H. L., and M. E. Grimshaw

1. *Linear Transformations in n-dimensional Vector Space*, Cambridge University Press, Cambridge, 1951.

Hamel, G.

1. Eine Basis aller Zahlen und die unstetige Lösungen der Funktionalgleichung $f(x+y)=f(x)+f(y)$, *Math. Ann.* 60(1905), 459–462.

Hardy, G. H.

1. The mean value of the modulus of an analytic function, *Proc. London Math. Soc.* 14(1914), 269–277.

Hardy, G. H., J. E. Littlewood, and G. Pólya

1. *Inequalities*, 2nd ed., Cambridge University Press, Cambridge, 1952.

Harmuth, H. F.

1. *Transmission of Information by Orthogonal Functions*, 2nd ed., Springer-Verlag, New York, 1972.

Hellinger, E. D.

1. Spectra of quadratic forms in infinitely many variables, *Mathematical Monographs* I, Northwestern University Studies, 1941.

Hellinger, E., and O. Toeplitz

1. *Integralgleichungen und Gleichungen mit Unendlichvielen Unbekannten*, Chelsea, New York, 1953. (A reprint of Part II C13 of the Encyklopädie der mathematischen Wissenschaften.)

Hille, E., and R. S. Phillips

1. *Functional Analysis and Semigroups*, American Mathematical Society Colloquium Publications, vol. 31, New York, 1957, reprinted 1974.

Hirschfeld, R. A., and W. Zelazko

1. On spectral norm Banach algebras, *Bull. Acad. Polon. Sci.* 16(1968), 195–199.

Hörmander, L.

1. *An Introduction to Complex Analysis in Several Variables*, Van Nostrand, Princeton, 1966.

Hoffman, K.

1. *Banach Spaces of Analytic Functions*, Prentice-Hall, Englewood Cliffs, 1962.

Horváth, J.

1. *Topological Vector Spaces and Distributions*, I., Addison-Wesley, Reading, Mass., 1966.
2. An introduction to distributions, *Amer. Math. Monthly* 77(1970), 227–240.

Jacobson, N.

1. *Lectures in Abstract Algebra, vol.* II, *Linear Algebra*, Springer-Verlag, New York, 1975.

James, R. C.

1. A non-reflexive Banach space isometric with its second conjugate space, *Proc. Nat. Acad. Sci. U.S.A.* 37(1951), 174–177.
2. Characterizations of reflexivity, *Studia Math.* 23(1964), 205–216.

Johnson, B. E.

1. The uniqueness of the (complete) norm topology, *Bull. Amer. Math. Soc.* 73(1967), 537–539.

Kaashoek, M. A.

1. Ascent, descent, nullity and defect, a note on a paper by A. E. Taylor, *Math. Ann.* 172(1967), 105–115.

Kadets, M. I., and B. S. Mityagin

1. Complemented subspaces in Banach spaces, *Russian Math. Surveys* 28(6)(1973), 77–95.

Kato, T.

1. Perturbation theory for nullity, deficiency and other quantities of linear operators, *J. d'Analyse Math.* 6(1958), 273–322.

Katznelson, Y.

1. *An Introduction to Harmonic Analysis*, Wiley, New York, 1968.

Kelly, J. L.

1. *General Topology*, Springer-Verlag, 1975.

Kelley, J. L., and I. Namioka

1. *Linear Topological Spaces*, Van Nostrand, Princeton, 1963.

Kellog, O. D.

1. *Foundations of Potential Theory*, Springer-Verlag, New York, 1967.

Köthe, G.

1. *Topological Vector Spaces*, I, Springer-Verlag, New York, 1969.

Krein, M., and D. Milman

1. On extreme points of regular convex sets, *Studia Math.* 9(1940), 133–138.

Krishnamurthy, V.

1. On the state of a linear operator and its adjoint, *Math. Ann.* 141(1960), 153–160.

Lax, P. D.

1. On the existence of Green's function, *Proc. Amer. Math. Soc.* 3(1952), 526–531.

Lay, D. C.

1. Spectral analysis using ascent, descent, nullity and defect, *Math. Ann.* 184(1970), 197–214.
2. Spectral properties of generalized inverses of linear operators, *SIAM J. Appl. Math.* 29(1975), 103–109.

LePage, C.

1. Sur quelques conditions entraînant la commutativité dans les algebres de Banach, *C. R. Acad. Sci. Paris*, Ser. A-B 265(1967), 235–237.

Lindenstrauss, J., and L. Tzafriri

1. On the complemented subspaces problem, *Israel J. Math.* 9(1971), 263–269.
2. *Classical Banach Spaces I, Sequence Spaces*, Springer-Verlag, New York, 1977.

Liusternik, L. A., and V. J. Sobolev

1. *Elements of Functional Analysis*, 3rd English translation, Hindustan Publ., Delhi, (John Wiley, New York), 1974.

Löwig, H.

1. Über die Dimension Linearer Räume, *Studia Math.* 5(1934), 18–23.

Lomonosov, V. I.

1. On invariant subspaces of families of operators commuting with a completely continuous operator. *Funkcional. Anal. i Priložen* 7(3)(1973), 55–56.

Lorch, E. R.

1. The spectrum of a linear transformation, *Trans. Amer. Math. Soc.* 52(1942), 238–248.
2. *Spectral Theory*, Oxford University Press, New York, 1962.

Miranda, C.

1. Sul principio di Dirichlet per le funzioni armoniche, *Atti Accad. Naz. Lincei, Rend. Cl. Sci. Fis. Mat. Nat.* (8), 3(1947), 55–59.

Naimark, M. A.

1. *Normed Algebras*, Wolters-Noordhoff, Groningen, The Netherlands, 1972.

Nashed, M. Z.

1. *Generalized Inverses and Applications*, Academic Press, New York, 1976.

Pearcy, C., and A. L. Shields

1. A survey of the Lomonosov technique in the theory of invariant subspaces, *Topics in Operator Theory*, C. Pearcy (ed.), Mathematical Surveys, No. 13, American Mathematical Society, Providence, 1974.

Pelczynski, A.

1. On some problems of Banach, *Russian Math. Surveys* 28(6)(1973), 77–95.

Phillips, R. S.

1. On linear transformations, *Trans. Amer. Math. Soc.* 48(1940), 516–541.

Pietsch, A.

1. Zur Theorie der σ-Transformationen in lokalkonvexen Vectorräumen, *Math. Nachr.* 21(1960), 347–369.

Pollard, H.

1. Integral transforms, *Duke Math. J.* 13(1946), 307–330.

Rellich, F.

1. Spektraltheorie in nichtseparablen Räumen, *Math. Ann.* 110(1934), 342–356.

Rickart, C. E.

1. *General Theory of Banach Algebras*, Van Nostrand, Princeton, 1960.

Riesz, F.

1. *Les Systèmes d'Équations Lineairés à une Infinité d'Inconnues*, Gauthier-Villars, Paris, 1913.
2. Über lineare Funktionalgleichungen, *Acta Math.* 41(1917), 71–98.
3. Sur les valeurs moyennes du module des fonctions harmoniques et des fonctions analytiques, *Acta Sci. Math. Szeged* 1(1922), 27–32.
4. Über die Randwerte einer analytischen Funktion, *Math. Zeit.* 18(1923), 87–95.

Riesz, F., and B. Sz.-Nagy

1. *Functional Analysis*, Ungar, New York, 1955.

Riesz, M.

1. Sur les maxima des formes bilinéaires et sur les fonctionelles linéaires, *Acta Math.* 49(1927), 465–497.

Rosenblum, M.

1. On a theorem of Fuglede and Putnam, *J. London Math. Soc.* 33(1958), 376–377.

Rudin, W.

1. *Real and Complex Analysis*, 2nd ed., McGraw-Hill, New York, 1974.
2. *Functional Analysis*, McGraw-Hill, New York, 1973.
3. *Fourier Analysis on Groups*, Interscience, New York, 1962.

Ruston, A.

1. A note on the Caradus class $\mathscr{F}$ of bounded linear operators on a complex Banach space, *Canadian J. Math.* 21(1969), 592–594.
2. On operators with a rational resolvent, *Proc. Amer. Math. Soc.* 32(1972), 329–330.

Schaefer, H. H.

1. *Topological Vector Spaces*, Springer-Verlag, New York, 1971.

Schechter, M.

1. *Principles of Functional Analysis*, Academic Press, New York, 1971.

Schwartz, L.

1. *Théorie des Distributions*, Hermann & Cie, Paris, 1966.

Sierpinski, W.

1. *Leçons sur les Nombres Transfinis*, Gauthier-Villars, Paris, 1950.

Simmons, G. F.

1. *Introduction to Topology and Modern Analysis*, McGraw-Hill, New York, 1963.

Singer, I.

1. *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, Springer-Verlag, New York, 1970.

Stone, M. H.

1. *Linear Transformations in Hilbert Space*, American Mathematical Society Colloquium Publications, vol. 15, Providence, 1932.
2. Applications of the theory of Boolean rings to general topology, *Trans. Amer. Math. Soc.* 41(1937), 375–481.

Szegö, G.

1. *Orthogonal Polynomials*, American Mathematical Society Colloquium Publications, vol. 23, Providence, 1939.

Sz.-Nagy, B.

1. Spectral mapping theorems for semigroups of operators, *Acta Sci. Math.* 38(1976), 343–351.

Taylor, A. E.

1. Analysis in complex Banach spaces, *Bull. Amer. Math. Soc.* 49(1943), 652–669.

2. Spectral theory of closed distributive operators, *Acta Math.* 84(1950), 189–224.
3. New proofs of some theorems of Hardy by Banach space methods, *Math. Mag.* 23(1950), 115–124.
4. Banach spaces of functions analytic in the unit circle, *Studia Math.* 11(1950), 145–170, and 12(1951), 25–50.
5. *General Theory of Functions and Integration*, Blaisdell, Waltham, Mass., 1965.
6. Theorems on ascent, descent, nullity and defect of linear operators, *Math. Ann.* 163(1966), 18–49.
7. Historical notes on analyticity as a concept in functional analysis, *Problems in Analysis—A Symposium in Honor of Salomon Bochner*, R. C. Gunning (ed.), Princeton University Press, Princeton, 1970, 325–348.

Taylor, A. E., and C. J. A. Halberg, Jr.
1. General theorems about a bounded linear operator and its conjugate, *J. Reine Angew. Math.* 198(1957), 93–111.

Thorin, G. O.
1. *Convexity theorems generalizing those of M. Riesz and Hadamard, with some applications*, University of Lund thesis, Uppsala, 1948.

Thorp, E., and R. Whitley
1. The strong maximum modulus theorem for analytic functions into a Banach space, *Proc. Amer. Math. Soc.* 18(1967), 640–646.

Titchmarsh, E. C.
1. *Theory of Functions*, 2nd ed., Oxford University Press, Oxford, 1939.

Trèves, F.
1. Applications of distributions to PDE theory, *Amer. Math. Monthly* 77(1970), 241–248.

Webb, J. H.
1. Perturbation theory for a linear operator, *Proc. Camb. Phil. Soc.* 63(1967), 11–20.

Weil, A.
1. *Sur les Espaces à Structure Uniforme*, Actualités Sci. Ind., No. 551, Hermann, Paris, 1937.

Weyl, H.
1. The method of orthogonal projection in potential theory, *Duke Math. J.* 7(1940), 411–444.

Whitley, R.
1. An elementary proof of the Eberlein–Šmulian theorem, *Math. Ann.* 172(1967), 116–118.
2. Projecting m onto c_0, *Amer. Math. Monthly* 73(1966), 285–286.

Widder, D. V.

1. *The Laplace Transform*, Princeton University Press, Princeton, 1946.

Wintner, A.

1. *Spektraltheorie der Unendlichen Matrizen*, Hirzel Verlag, Leipzig, 1929.

Yosida, K.

1. *Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1968.

Zaanen, A. C.

1. *Linear Analysis*, North Holland, Amsterdam, 1953.

Zelazko, W.

1. A characterization of multiplicative linear functionals on complex Banach algebras, *Studia Math.* 30(1968), 83–85.

Zygmund, A.

1. *Trigonometric Series*, 2nd ed., vol. I, II, Cambridge University Press, Cambridge, 1968.

LIST OF SPECIAL SYMBOLS

ADDITIONAL SYMBOLS ARE INTRODUCED ON PAGES 1–2.

DISCUSSIONS RELATING TO PARTICULAR SPACES

INDEX