Graduate Texts in Mathematics

Steven Roman

Advanced Linear Algebra

Second Edition

Springer

Graduate Texts in Mathematics 135

Graduate Texts in Mathematics

1. Takeuti/Zaring. Introduction to Axiomatic Set Theory. 2nd ed.
2. Oxtoby. Measure and Category. 2nd ed.
3. Schaefer. Topological Vector Spaces. 2nd ed.
4. Hilton/Stammbach. A Course in Homological Algebra. 2nd ed.
5. Mac Lane. Categories for the Working Mathematician. 2nd ed.
6. Hughes/Piper. Projective Planes.
7. J.-P. Serre. A Course in Arithmetic.
8. Takeuti/Zaring. Axiomatic Set Theory.
9. Humphreys. Introduction to Lie Algebras and Representation Theory.
10. Cohen. A Course in Simple Homotopy Theory.
11. Conway. Functions of One Complex Variable I. 2nd ed.
12. Beals. Advanced Mathematical Analysis.
13. Anderson/Fuller. Rings and Categories of Modules. 2nd ed.
14. Golubitsky/Guillemin. Stable Mappings and Their Singularities.
15. Berberian. Lectures in Functional Analysis and Operator Theory.
16. Winter. The Structure of Fields.
17. Rosenblatt. Random Processes. 2nd ed.
18. Halmos. Measure Theory.
19. Halmos. A Hilbert Space Problem Book. 2nd ed.
20. Husemoller. Fibre Bundles. 3rd ed.
21. Humphreys. Linear Algebraic Groups.
22. Barnes/Mack. An Algebraic Introduction to Mathematical Logic.
23. Greub. Linear Algebra. 4th ed.
24. Holmes. Geometric Functional Analysis and Its Applications.
25. Hewitt/Stromberg. Real and Abstract Analysis.
26. Manes. Algebraic Theories.
27. Kelley. General Topology.
28. Zariski/Samuel. Commutative Algebra. Vol.I.
29. Zariski/Samuel. Commutative Algebra. Vol.II.
30. Jacobson. Lectures in Abstract Algebra I. Basic Concepts.
31. Jacobson. Lectures in Abstract Algebra II. Linear Algebra.
32. Jacobson. Lectures in Abstract Algebra III. Theory of Fields and Galois Theory.
33. Hirsch. Differential Topology.
34. Spitzer. Principles of Random Walk. 2nd ed.
35. Alexander/Wermer. Several Complex Variables and Banach Algebras. 3rd ed.
36. Kelley/Namioka et al. Linear Topological Spaces.
37. Monk. Mathematical Logic.
38. Grauert/Fritzsche. Several Complex Variables.
39. Arveson. An Invitation to C^*-Algebras.
40. Kemeny/Snell/Knapp. Denumerable Markov Chains. 2nd ed.
41. Apostol. Modular Functions and Dirichlet Series in Number Theory. 2nd ed.
42. J.-P. Serre. Linear Representations of Finite Groups.
43. Gillman/Jerison. Rings of Continuous Functions.
44. Kendig. Elementary Algebraic Geometry.
45. Loève. Probability Theory I. 4th ed.
46. Loève. Probability Theory II. 4th ed.
47. Moise. Geometric Topology in Dimensions 2 and 3.
48. Sachs/Wu. General Relativity for Mathematicians.
49. Gruenberg/Weir. Linear Geometry. 2nd ed.
50. Edwards. Fermat's Last Theorem.
51. Klingenberg. A Course in Differential Geometry.
52. Hartshorne. Algebraic Geometry.
53. Manin. A Course in Mathematical Logic.
54. Graver/Watkins. Combinatorics with Emphasis on the Theory of Graphs.
55. Brown/Pearcy. Introduction to Operator Theory I: Elements of Functional Analysis.
56. Massey. Algebraic Topology: An Introduction.
57. Crowell/Fox. Introduction to Knot Theory.
58. Koblitz. p-adic Numbers, p-adic Analysis, and Zeta-Functions. 2nd ed.
59. Lang. Cyclotomic Fields.
60. Arnold. Mathematical Methods in Classical Mechanics. 2nd ed.
61. Whitehead. Elements of Homotopy Theory.
62. Kargapolov/Merlzjakov. Fundamentals of the Theory of Groups.
63. Bollobas. Graph Theory.

(continued after index)

Steven Roman

Advanced Linear Algebra

Second Edition

Springer

Steven Roman
University of California, Irvine
Irvine, California 92697-3875
USA
sroman@romanpress.com

Mathematics Subject Classification (2000): 15-xx

Library of Congress Cataloging-in-Publication Data
Roman, Steven.
Advanced linear algebra / Steven Roman.—2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-387-24766-1 (acid-free paper)
1. Algebras, Linear. I. Title.
QA184.2.R66 2005
512′.5—dc22 2005040244

ISBN 0-387-24766-1 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2

springer.com

To Donna
and to my poker buddies
Rachelle, Carol and Dan

Preface to the Second Edition

Let me begin by thanking the readers of the first edition for their many helpful comments and suggestions. The second edition represents a major change from the first edition. Indeed, one might say that it is a totally new book, with the exception of the general range of topics covered.

The text has been completely rewritten. I hope that an additional 12 years and roughly 20 books worth of experience has enabled me to improve the quality of my exposition. Also, the exercise sets have been completely rewritten.

The second edition contains two new chapters: a chapter on convexity, separation and positive solutions to linear systems (Chapter 15) and a chapter on the QR decomposition, singular values and pseudoinverses (Chapter 17). The treatments of tensor products and the umbral calculus have been greatly expanded and I have included discussions of determinants (in the chapter on tensor products), the complexification of a real vector space, Schur's lemma and Geršgorin disks.

Steven Roman *Irvine, California February 2005*

Preface to the First Edition

This book is a thorough introduction to linear algebra, for the graduate or advanced undergraduate student. Prerequisites are limited to a knowledge of the basic properties of matrices and determinants. However, since we cover the basics of vector spaces and linear transformations rather rapidly, a prior course in linear algebra (even at the sophomore level), along with a certain measure of "mathematical maturity," is highly desirable.

Chapter 0 contains a summary of certain topics in modern algebra that are required for the sequel. *This chapter should be skimmed quickly and then used primarily as a reference*. Chapters 1–3 contain a discussion of the basic properties of vector spaces and linear transformations.

Chapter 4 is devoted to a discussion of modules, emphasizing a comparison between the properties of modules and those of vector spaces. Chapter 5 provides more on modules. The main goals of this chapter are to prove that any two bases of a free module have the same cardinality and to introduce noetherian modules. However, the instructor may simply skim over this chapter, omitting all proofs. Chapter 6 is devoted to the theory of modules over a principal ideal domain, establishing the cyclic decomposition theorem for finitely generated modules. This theorem is the key to the structure theorems for finite-dimensional linear operators, discussed in Chapters 7 and 8.

Chapter 9 is devoted to real and complex inner product spaces. The emphasis here is on the finite-dimensional case, in order to arrive as quickly as possible at the finite-dimensional spectral theorem for normal operators, in Chapter 10. However, we have endeavored to state as many results as is convenient for vector spaces of arbitrary dimension.

The second part of the book consists of a collection of independent topics, with the one exception that Chapter 13 requires Chapter 12. Chapter 11 is on metric vector spaces, where we describe the structure of symplectic and orthogonal geometries over various base fields. Chapter 12 contains enough material on metric spaces to allow a unified treatment of topological issues for the basic

Hilbert space theory of Chapter 13. The rather lengthy proof that every metric space can be embedded in its completion may be omitted.

Chapter 14 contains a brief introduction to tensor products. In order to motivate the universal property of tensor products, without getting too involved in categorical terminology, we first treat both free vector spaces and the familiar direct sum, in a universal way. Chapter 15 [Chapter 16 in the second edition] is on affine geometry, emphasizing algebraic, rather than geometric, concepts.

The final chapter provides an introduction to a relatively new subject, called the umbral calculus. This is an algebraic theory used to study certain types of polynomial functions that play an important role in applied mathematics. We give only a brief introduction to the subject – emphasizing the algebraic aspects, rather than the applications. This is the first time that this subject has appeared in a true textbook.

One final comment. Unless otherwise mentioned, omission of a proof in the text is a tacit suggestion that the reader attempt to supply one.

Steven Roman *Irvine, California*

Contents

Chapter 0
Preliminaries

In this chapter, we briefly discuss some topics that are needed for the sequel. This chapter should be skimmed quickly and used primarily as a reference.

Part 1 Preliminaries

Multisets

The following simple concept is much more useful than its infrequent appearance would indicate.

Definition *Let S be a nonempty set. A* **multiset** *M with* **underlying set** *S is a set of ordered pairs*

$$M = \{(s_i, n_i) \mid s_i \in S, n_i \in \mathbb{Z}^+, s_i \neq s_j \textit{ for } i \neq j\}$$

where $\mathbb{Z}^+ = \{1, 2, \dots\}$. The number n_i is referred to as the **multiplicity** *of the elements s_i in M. If the underlying set of a multiset is finite, we say that the multiset is* **finite**. *The* **size** *of a finite multiset M is the sum of the multiplicities of all of its elements.* □

For example, $M = \{(a, 2), (b, 3), (c, 1)\}$ is a multiset with underlying set $S = \{a, b, c\}$. The elements a has multiplicity 2. One often writes out the elements of a multiset according to multiplicities, as in $M = \{a, a, b, b, b, c\}$.

Of course, two mutlisets are equal if their underlying sets are equal and if the multiplicity of each element in the comon underlying set is the same in both multisets.

Matrices

The set of $m \times n$ matrices with entries in a field F is denoted by $\mathcal{M}_{m,n}(F)$ or by $\mathcal{M}_{m,n}$ when the field does not require mention. The set $\mathcal{M}_{n,n}(\mathcal{F})$ is denoted by $\mathcal{M}_n(F)$ or $\mathcal{M}_n$. If $A \in \mathcal{M}$, the (i, j)-th entry of A will be denoted by $A_{i,j}$. The identity matrix of size $n \times n$ is denoted by I_n. The elements of the base field F are called **scalars**. We expect that the reader is familiar with the basic properties of matrices, including matrix addition and multiplication.

The **main diagonal** of an $m \times n$ matrix A is the sequence of entries

$$A_{1,1}, A_{2,2}, \ldots, A_{k,k}$$

where $k = \min\{m, n\}$.

Definition *The* **transpose** *of $A \in \mathcal{M}_{m,n}$ is the matrix A^t defined by*

$$(A^t)_{i,j} = A_{j,i}$$

A matrix A is **symmetric** *if $A = A^t$ and* **skew-symmetric** *if $A^t = -A$.* □

Theorem 0.1 *(Properties of the transpose) Let A, $B \in \mathcal{M}_{m,n}$. Then*
1) $(A^t)^t = A$
2) $(A + B)^t = A^t + B^t$
3) $(rA)^t = rA^t$ *for all $r \in F$*
4) $(AB)^t = B^t A^t$ *provided that the product AB is defined*
5) $\det(A^t) = \det(A)$. □

Partitioning and Matrix Multiplication

Let M be a matrix of size $m \times n$. If $B \subseteq \{1, \ldots, m\}$ and $C \subseteq \{1, \ldots, n\}$ then the **submatrix** $M[B, C]$ is the matrix obtained from M by keeping only the rows with index in B and the columns with index in C. Thus, all other rows and columns are discarded and $M[B, C]$ has size $|B| \times |C|$.

Suppose that $M \in \mathcal{M}_{m,n}$ and $N \in \mathcal{M}_{n,k}$. Let

1) $\mathcal{P} = \{B_1, \ldots, B_p\}$ be a partition of $\{1, \ldots, m\}$
2) $\mathcal{Q} = \{C_1, \ldots, C_q\}$ be a partition of $\{1, \ldots, n\}$
3) $\mathcal{R} = \{D_1, \ldots, D_r\}$ be a partition of $\{1, \ldots, k\}$

(Partitions are defined formally later in this chapter.) Then it is a very useful fact that matrix multiplication can be performed at the block level as well as at the entry level. In particular, we have

$$[MN][B_i, D_j] = \sum_{C_h \in \mathcal{Q}} M[B_i, C_h] N[C_h, D_j]$$

When the partitions in question contain only single-element blocks, this is precisely the usual formula for matrix multiplication

$$[MN]_{i,j} = \sum_{h=1}^{m} M_{i,h} N_{h,j}$$

Block Matrices

It will be convenient to introduce the notational device of a block matrix. If $B_{i,j}$ are matrices of the appropriate sizes then by the **block matrix**

$$M = \begin{bmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,n} \\ \vdots & \vdots & & \vdots \\ B_{m,1} & B_{m,2} & \cdots & B_{m,n} \end{bmatrix}_{\text{block}}$$

we mean the matrix whose upper left *submatrix* is $B_{1,1}$, and so on. Thus, the $B_{i,j}$'s are *submatrices* of M and not entries. A square matrix of the form

$$M = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_n \end{bmatrix}_{\text{block}}$$

where each B_i is square and 0 is a zero submatrix, is said to be a **block diagonal matrix**.

Elementary Row Operations

Recall that there are three types of elementary row operations. Type 1 operations consist of multiplying a row of A by a nonzero scalar. Type 2 operations consist of interchanging two rows of A. Type 3 operations consist of adding a scalar multiple of one row of A to another row of A.

If we perform an elementary operation of type k to an identity matrix I_n, the result is called an **elementary matrix** of type k. It is easy to see that all elementary matrices are invertible.

In order to perform an elementary row operation on $A \in \mathcal{M}_{m,n}$ we can perform that operation on the identity I_m, to obtain an elementary matrix E and then take the product EA. Note that multiplying on the right by E has the effect of performing column operations.

Definition *A matrix R is said to be in* **reduced row echelon form** *if*

1) *All rows consisting only of 0's appear at the bottom of the matrix.*
2) *In any nonzero row, the first nonzero entry is a 1. This entry is called a* **leading entry**.
3) *For any two consecutive rows, the leading entry of the lower row is to the right of the leading entry of the upper row.*
4) *Any column that contains a leading entry has 0's in all other positions.* □

Here are the basic facts concerning reduced row echelon form.

Theorem 0.2 *Matrices* $A, B \in \mathcal{M}_{m,n}$ *are* **row equivalent**, *denoted by* $A \sim B$, *if either one can be obtained from the other by a series of elementary row operations.*

1) *Row equivalence is an equivalence relation. That is,*
   a) $A \sim A$
   b) $A \sim B \Rightarrow B \sim A$
   c) $A \sim B$, $B \sim C \Rightarrow A \sim C$.

2) *A matrix* A *is row equivalent to one and only one matrix* R *that is in reduced row echelon form. The matrix* R *is called the* **reduced row echelon form** *of* A. *Furthermore,*

$$A = E_1 \cdots E_k R$$

where E_i *are the elementary matrices required to reduce* A *to reduced row echelon form.*

3) A *is invertible if and only if its reduced row echelon form is an identity matrix. Hence, a matrix is invertible if and only if it is the product of elementary matrices.* □

The following definition is probably well known to the reader.

Definition *A square matrix is* **upper triangular** *if all of its entries below the main diagonal are* 0. *Similarly, a square matrix is* **lower triangular** *if all of its entries above the main diagonal are* 0. *A square matrix is* **diagonal** *if all of its entries off the main diagonal are* 0. □

Determinants

We assume that the reader is familiar with the following basic properties of determinants.

Theorem 0.3 *Let* $A \in \mathcal{M}_{n,n}(F)$. *Then* $\det(A)$ *is an element of* F. *Furthermore,*

1) *For any* $B \in \mathcal{M}_n(F)$,

$$\det(AB) = \det(A)\det(B)$$

2) A *is nonsingular (invertible) if and only if* $\det(A) \neq 0$.

3) *The determinant of an upper triangular or lower triangular matrix is the product of the entries on its main diagonal.*

4) *If a square matrix* M *has the block diagonal form*

$$M = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & B_n \end{bmatrix}_{\text{block}}$$

then $\det(M) = \prod \det(B_i)$. □

Polynomials

The set of all polynomials in the variable x with coefficients from a field F is denoted by $F[x]$. If $p(x) \in F[x]$, we say that $p(x)$ is a polynomial **over** F. If

$$p(x) = a_0 + a_1x + \cdots + a_nx^n$$

is a polynomial with $a_n \neq 0$ then a_n is called the **leading coefficient** of $p(x)$ and the **degree** of $p(x)$ is n, written $\deg p(x) = n$. For convenience, the degree of the zero polynomial is $-\infty$. A polynomial is **monic** if its leading coefficient is 1.

Theorem 0.4 (**Division algorithm**) *Let* $f(x), g(x) \in F[x]$ *where* $\deg g(x) > 0$. *Then there exist unique polynomials* $q(x), r(x) \in F[x]$ *for which*

$$f(x) = q(x)g(x) + r(x)$$

where $r(x) = 0$ *or* $0 \leq \deg r(x) < \deg g(x)$. □

If $p(x)$ **divides** $q(x)$, that is, if there exists a polynomial $f(x)$ for which

$$q(x) = f(x)p(x)$$

then we write $p(x) \mid q(x)$.

Theorem 0.5 *Let* $f(x), g(x) \in F[x]$. *The* **greatest common divisor** *of* $f(x)$ *and* $g(x)$, *denoted by* $\gcd(f(x), g(x))$, *is the unique monic polynomial* $p(x)$ *over* F *for which*

1) $p(x) \mid f(x)$ *and* $p(x) \mid g(x)$

2) *if* $r(x) \mid f(x)$ *and* $r(x) \mid g(x)$ *then* $r(x) \mid p(x)$.

Furthermore, there exist polynomials $a(x)$ *and* $b(x)$ *over* F *for which*

$$\gcd(f(x), g(x)) = a(x)f(x) + b(x)g(x) \qquad \square$$

Definition *The polynomials* $f(x), g(x) \in F[x]$ *are* **relatively prime** *if* $\gcd(f(x), g(x)) = 1$. *In particular,* $f(x)$ *and* $g(x)$ *are relatively prime if and only if there exist polynomials* $a(x)$ *and* $b(x)$ *over* F *for which*

$$a(x)f(x) + b(x)g(x) = 1 \qquad \square$$

Definition *A nonconstant polynomial* $f(x) \in F[x]$ *is* **irreducible** *if whenever* $f(x) = p(x)q(x)$ *then one of* $p(x)$ *and* $q(x)$ *must be constant.* □

The following two theorems support the view that irreducible polynomials behave like prime numbers.

Theorem 0.6 *A nonconstant polynomial* $f(x)$ *is irreducible if and only if it has the property that whenever* $f(x) \mid p(x)q(x)$ *then either* $f(x) \mid p(x)$ *or* $f(x) \mid q(x)$. □

Theorem 0.7 *Every nonconstant polynomial in $F[x]$ can be written as a product of irreducible polynomials. Moreover, this expression is unique up to order of the factors and multiplication by a scalar.* □

Functions

To set our notation, we should make a few comments about functions.

Definition *Let $f{:}\, S \to T$ be a function from a set S to a set T.*

1) *The* **domain** *of f is the set S.*
2) *The* **image** *or* **range** *of f is the set* $\operatorname{im}(f) = \{f(s) \mid s \in S\}$.
3) *f is* **injective** *(***one-to-one***), or an* **injection***, if $x \neq y \Rightarrow f(x) \neq f(y)$.*
4) *f is* **surjective** *(***onto** *T), or a* **surjection***, if* $\operatorname{im}(f) = T$.
5) *f is* **bijective***, or a* **bijection***, if it is both injective and surjective.*
6) *Assuming that $0 \in T$, the* **support** *of f is*

$$\operatorname{supp}(f) = \{s \in S \mid f(s) \neq 0\} \qquad \square$$

If $f{:}\, S \to T$ is injective then its inverse $f^{-1}{:}\, \operatorname{im}(f) \to S$ exists and is well-defined as a function on $\operatorname{im}(f)$.

It will be convenient to apply f to subsets of S and T. In particular, if $X \subseteq S$ and if $Y \subseteq T$, we set

$$f(X) = \{f(x) \mid x \in X\}$$

and

$$f^{-1}(Y) = \{s \in S \mid f(s) \in Y\}$$

Note that the latter is defined even if f is not injective.

Let $f{:}\, S \to T$. If $A \subseteq S$, the **restriction** of f to A is the function $f|_A{:}\, A \to T$ defined by

$$f|_A(a) = f(a)$$

for all $a \in A$. Clearly, the restriction of an injective map is injective.

Equivalence Relations

The concept of an equivalence relation plays a major role in the study of matrices and linear transformations.

Definition *Let S be a nonempty set. A binary relation $\sim$ on S is called an* **equivalence relation** *on S if it satisfies the following conditions:*

1) (**Reflexivity**)

$$a \sim a$$

for all $a \in S$.

2) (**Symmetry**)

$$a \sim b \Rightarrow b \sim a$$

for all $a, b \in S$.

3) (**Transitivity**)

$$a \sim b, b \sim c \Rightarrow a \sim c$$

for all $a, b, c \in S$. □

Definition *Let* $\sim$ *be an equivalence relation on* S. *For* $a \in S$, *the set of all elements equivalent to* a *is denoted by*

$$[a] = \{b \in S \mid b \sim a\}$$

and called the **equivalence class** *of* a. □

Theorem 0.8 *Let* $\sim$ *be an equivalence relation on* S. *Then*

1) $b \in [a] \Leftrightarrow a \in [b] \Leftrightarrow [a] = [b]$
2) *For any* $a, b \in S$, *we have either* $[a] = [b]$ *or* $[a] \cap [b] = \emptyset$. □

Definition *A* **partition** *of a nonempty set* S *is a collection* $\{A_1, \ldots, A_n\}$ *of* nonempty *subsets of* S, *called the* **blocks** *of the partition, for which*

1) $A_i \cap A_j = \emptyset$ *for all* $i \neq j$
2) $S = A_1 \cup \cdots \cup A_n$. □

The following theorem sheds considerable light on the concept of an equivalence relation.

Theorem 0.9

1) *Let* $\sim$ *be an equivalence relation on* S. *Then the set of* distinct *equivalence classes with respect to* $\sim$ *are the blocks of a partition of* S.
2) *Conversely, if* $\mathcal{P}$ *is a partition of* S, *the binary relation* $\sim$ *defined by*

   $$a \sim b \text{ if } a \text{ and } b \text{ lie in the same block of } \mathcal{P}$$

   is an equivalence relation on S, *whose equivalence classes are the blocks of* $\mathcal{P}$.

This establishes a one-to-one correspondence between equivalence relations on S *and partitions of* S. □

The most important problem related to equivalence relations is that of finding an efficient way to determine when two elements are equivalent. Unfortunately, in most cases, the definition does not provide an efficient test for equivalence and so we are led to the following concepts.

Definition *Let $\sim$ be an equivalence relation on S. A function $f\colon S \to T$, where T is any set, is called an* **invariant** *of $\sim$ if it is constant on the equivalence classes of $\sim$, that is,*

$$a \sim b \Rightarrow f(a) = f(b)$$

and a **complete invariant** *if it is constant and distinct on the equivalence classes of $\sim$, that is,*

$$a \sim b \Leftrightarrow f(a) = f(b)$$

A collection $\{f_1, \dots, f_n\}$ of invariants is called a **complete system of invariants** *if*

$$a \sim b \Leftrightarrow f_i(a) = f_i(b) \textit{ for all } i = 1, \dots, n \qquad \square$$

Definition *Let $\sim$ be an equivalence relation on S. A subset $C \subseteq S$ is said to be a set of* **canonical forms** *(or just a* **canonical form***) for $\sim$ if for every $s \in S$, there is* exactly one *$c \in C$ such that $c \sim s$. Put another way, each equivalence class under $\sim$ contains* exactly one *member of C.* □

Example 0.1 Define a binary relation $\sim$ on $F[x]$ by letting $p(x) \sim q(x)$ if and only if $p(x) = aq(x)$ for some nonzero constant $a \in F$. This is easily seen to be an equivalence relation. The function that assigns to each polynomial its degree is an invariant, since

$$p(x) \sim q(x) \Rightarrow \deg(p(x)) = \deg(q(x))$$

However, it is not a complete invariant, since there are inequivalent polynomials with the same degree. The set of all monic polynomials is a set of canonical forms for this equivalence relation. □

Example 0.2 We have remarked that row equivalence is an equivalence relation on $\mathcal{M}_{m,n}(F)$. Moreover, the subset of reduced row echelon form matrices is a set of canonical forms for row equivalence, since every matrix is row equivalent to a unique matrix in reduced row echelon form. □

Example 0.3 Two matrices A, $B \in \mathcal{M}_n(F)$ are row equivalent if and only if there is an invertible matrix P such that $A = PB$. Similarly, A and B are **column equivalent**, that is, A can be reduced to B using elementary column operations if and only if there exists an invertible matrix Q such that $A = BQ$.

Two matrices A and B are said to be **equivalent** if there exist invertible matrices P and Q for which

$$A = PBQ$$

Put another way, A and B are equivalent if A can be reduced to B by performing a series of elementary row and/or column operations. (The use of the term equivalent is unfortunate, since it applies to all equivalence relations, not just this one. However, the terminology is standard, so we use it here.)

It is not hard to see that an $m \times n$ matrix R that is in both reduced row echelon form and reduced column echelon form must have the block form

$$J_k = \begin{bmatrix} I_k & 0_{k,n-k} \\ 0_{m-k,k} & 0_{m-k,n-k} \end{bmatrix}_{\text{block}}$$

We leave it to the reader to show that every matrix A in $\mathcal{M}_n$ is equivalent to exactly one matrix of the form J_k and so the set of these matrices is a set of canonical forms for equivalence. Moreover, the function f defined by $f(A) = k$, where $A \sim J_k$, is a complete invariant for equivalence.

Since the rank of J_k is k and since neither row nor column operations affect the rank, we deduce that the rank of A is k. Hence, rank is a complete invariant for equivalence. In other words, two matrices are equivalent if and only if they have the same rank. □

Example 0.4 Two matrices $A, B \in \mathcal{M}_n(F)$ are said to be **similar** if there exists an invertible matrix P such that

$$A = PBP^{-1}$$

Similarity is easily seen to be an equivalence relation on $\mathcal{M}_n$. As we will learn, two matrices are similar if and only if they represent the same linear operators on a given n-dimensional vector space V. Hence, similarity is extremely important for studying the structure of linear operators. One of the main goals of this book is to develop canonical forms for similarity.

We leave it to the reader to show that the determinant function and the trace function are invariants for similarity. However, these two invariants do not, in general, form a complete system of invariants. □

Example 0.5 Two matrices $A, B \in \mathcal{M}_n(F)$ are said to be **congruent** if there exists an invertible matrix P for which

$$A = PBP^t$$

where P^t is the transpose of P. This relation is easily seen to be an equivalence relation and we will devote some effort to finding canonical forms for congruence. For some base fields F (such as $\mathbb{R}$, $\mathbb{C}$ or a finite field), this is relatively easy to do, but for other base fields (such as $\mathbb{Q}$), it is extremely difficult. □

Zorn's Lemma

In order to show that any vector space has a basis, we require a result known as *Zorn's lemma*. To state this lemma, we need some preliminary definitions.

Definition *A* **partially ordered set** *is a pair* $(P,\ \leq\ )$ *where* P *is a nonempty set and* $\leq$ *is a binary relation called a* **partial order***, read "less than or equal to," with the following properties:*

1) (**Reflexivity**) *For all* $a\in P$*,*

$$a\leq a$$

2) (**Antisymmetry**) *For all* $a,b\in P$*,*

$$a\leq b \textit{ and } b\leq a \textit{ implies } a=b$$

3) (**Transitivity**) *For all* $a,b,c\in P$*,*

$$a\leq b \textit{ and } b\leq c \textit{ implies } a\leq c$$

Partially ordered sets are also called **posets**. □

It is customary to use a phrase such as "Let P be a partially ordered set" when the partial order is understood. Here are some key terms related to partially ordered sets.

Definition Let P be a partially ordered set.

1) *A* **maximal element** *is an element* $m\in P$ *with the property that there is no larger element in* P*, that is*

$$p\in P, m\leq p\Rightarrow m=p$$

2) *A* **minimal element** *is an element* $n\in P$ *with the property that there is no smaller element in* P*, that is*

$$p\in P, p\leq n\Rightarrow p=n$$

3) *Let* $a,b\in P$*. Then* $u\in P$ *is an* **upper bound** *for* a *and* b *if*

$$a\leq u \textit{ and } b\leq u$$

The unique smallest upper bound for a *and* b*, if it exists, is called the* **least upper bound** *of* a *and* b *and is denoted by* $\operatorname{lub}\{a,b\}$.

4) *Let* $a,b\in P$*. Then* $\ell\in P$ *is a* **lower bound** *for* a *and* b *if*

$$\ell\leq a \textit{ and } \ell\leq b$$

The unique largest lower bound for a *and* b*, if it exists, is called the* **greatest lower bound** *of* a *and* b *and is denoted by* $\operatorname{glb}\{a,b\}$. □

Let S be a subset of a partially ordered set P. We say that an element $u \in P$ is an **upper bound** for S if $s \leq u$ for all $s \in S$. Lower bounds are defined similarly.

Note that in a partially ordered set, it is possible that not all elements are comparable. In other words, it is possible to have $x, y \in P$ with the property that $x \nleq y$ and $y \nleq x$.

Definition *A partially ordered set in which every pair of elements is comparable is called a* **totally ordered set***, or a* **linearly ordered set***. Any totally ordered subset of a partially ordered set P is called a* **chain** *in P.* □

Example 0.6

1) The set $\mathbb{R}$ of real numbers, with the usual binary relation $\leq$, is a partially ordered set. It is also a totally ordered set. It has no maximal elements.
2) The set $\mathbb{N} = \{0, 1, \ldots\}$ of natural numbers, together with the binary relation of divides, is a partially ordered set. It is customary to write $n \mid m$ to indicate that n divides m. The subset S of $\mathbb{N}$ consisting of all powers of 2 is a totally ordered subset of $\mathbb{N}$, that is, it is a chain in $\mathbb{N}$. The set $P = \{2, 4, 8, 3, 9, 27\}$ is a partially ordered set under $\mid$. It has two maximal elements, namely 8 and 27. The subset $Q = \{2, 3, 5, 7, 11\}$ is a partially ordered set in which every element is both maximal and minimal!
3) Let S be any set and let $\mathcal{P}(S)$ be the power set of S, that is, the set of all subsets of S. Then $\mathcal{P}(S)$, together with the subset relation $\subseteq$, is a partially ordered set. □

Now we can state Zorn's lemma, which gives a condition under which a partially ordered set has a maximal element.

Theorem 0.10 *(***Zorn's lemma***) If P is a partially ordered set in which every chain has an upper bound then P has a maximal element.* □

We will not prove Zorn's lemma. Indeed, Zorn's lemma is a result that is so fundamental that it cannot be proved or disproved in the context of ordinary set theory. (It is equivalent to the famous *Axiom of Choice*.) Therefore, Zorn's lemma (along with the Axiom of Choice) must either be accepted or rejected as an axiom of set theory. Since almost all mathematicians accept it, we will do so as well. Indeed, we will use Zorn's lemma to prove that every vector space has a basis.

Cardinality

Two sets S and T have the same **cardinality**, written

$$|S| = |T|$$

if there is a bijective function (a one-to-one correspondence) between the sets.

The reader is probably aware of the fact that

$$|\mathbb{Z}| = |\mathbb{N}| \text{ and } |\mathbb{Q}| = |\mathbb{N}|$$

where $\mathbb{N}$ denotes the natural numbers, $\mathbb{Z}$ the integers and $\mathbb{Q}$ the rational numbers.

If S is in one-to-one correspondence with a *subset* of T, we write $|S| \leq |T|$. If S is in one-to-one correspondence with a *proper* subset of T but not all of T then we write $|S| < |T|$. The second condition is necessary, since, for instance, $\mathbb{N}$ is in one-to-one correspondence with a proper subset of $\mathbb{Z}$ and yet $\mathbb{N}$ is also in one-to-one correspondence with $\mathbb{Z}$ itself. Hence, $|\mathbb{N}| = |\mathbb{Z}|$.

This is not the place to enter into a detailed discussion of cardinal numbers. The intention here is that the cardinality of a set, whatever that is, represents the "size" of the set. It is actually easier to talk about two sets having the same, or different, size (cardinality) than it is to explicitly define the size (cardinality) of a given set.

Be that as it may, we associate to each set S a cardinal number, denoted by $|S|$ or card(S), that is intended to measure the size of the set. Actually, cardinal numbers are just very special types of sets. However, we can simply think of them as vague amorphous objects that measure the size of sets.

Definition

1) *A set is* **finite** *if it can be put in one-to-one correspondence with a set of the form* $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$*, for some nonnegative integer* n*. A set that is not finite is* **infinite**. *The* **cardinal number** *(or* **cardinality***) of a finite set is just the number of elements in the set.*
2) *The* **cardinal number** *of the set* $\mathbb{N}$ *of natural numbers is* $\aleph_0$ *(read "aleph nought"), where* $\aleph$ *is the first letter of the Hebrew alphabet. Hence,*
$$|\mathbb{N}| = |\mathbb{Z}| = |\mathbb{Q}| = \aleph_0$$
3) *Any set with cardinality* $\aleph_0$ *is called a* **countably infinite** *set and any finite or countably infinite set is called a* **countable** *set. An infinite set that is not countable is said to be* **uncountable**. □

Since it can be shown that $|\mathbb{R}| > |\mathbb{N}|$, the real numbers are uncountable.

If S and T are *finite* sets then it is well known that

$$|S| \leq |T| \text{ and } |T| \leq |S| \Rightarrow |S| = |T|$$

The first part of the next theorem tells us that this is also true for infinite sets.

The reader will no doubt recall that the **power set** $\mathcal{P}(S)$ of a set S is the set of all subsets of S. For finite sets, the power set of S is always bigger than the set itself. In fact,

$$|S| = n \Rightarrow |\mathcal{P}(S)| = 2^n$$

The second part of the next theorem says that the power set of any set S is bigger (has larger cardinality) than S itself. On the other hand, the third part of this theorem says that, for infinite sets S, the set of all *finite* subsets of S is the same size as S.

Theorem 0.11

1) (**Schröder–Bernstein theorem**) *For any sets S and T,*

$$|S| \le |T| \textit{ and } |T| \le |S| \Rightarrow |S| = |T|$$

2) (**Cantor's theorem**) *If $\mathcal{P}(S)$ denotes the power set of S then*

$$|S| < |\mathcal{P}(S)|$$

3) *If $\mathcal{P}_0(S)$ denotes the set of all finite subsets of S and if S is an infinite set then*

$$|S| = |\mathcal{P}_0(S)|$$

Proof. We prove only parts 1) and 2). Let $f: S \to T$ be an injective function from S into T and let $g: T \to S$ be an injective function from T into S. We want to use these functions to create a bijective function from S to T. For this purpose, we make the following definitions. The **descendants** of an element $s \in S$ are the elements obtained by repeated alternate applications of the functions f and g, namely

$$f(s), g(f(s)), f(g(f(s))), \ldots$$

If t is a descendant of s then s is an **ancestor** of t. Descendants and ancestors of elements of T are defined similarly.

Now, by tracing an element's ancestry to its beginning, we find that there are three possibilities: the element may originate in S, or in T, or it may have no point of origin. Accordingly, we can write S as the union of three disjoint sets

$$\begin{aligned} \mathcal{S}_S &= \{s \in S \mid s \text{ originates in } S\} \\ \mathcal{S}_T &= \{s \in S \mid s \text{ originates in } T\} \\ \mathcal{S}_\infty &= \{s \in S \mid s \text{ has no originator}\} \end{aligned}$$

Similarly, T is the disjoint union of $\mathcal{T}_S$, $\mathcal{T}_T$ and $\mathcal{T}_\infty$.

Now, the restriction

$$f|_{\mathcal{S}_S}: \mathcal{S}_S \to \mathcal{T}_S$$

is a bijection. To see this, note that if $t \in \mathcal{T}_S$ then t originated in S and therefore must have the form $f(s)$ for some $s \in S$. But t and its ancestor $f(s)$ have the same point of origin and so $t \in \mathcal{T}_S$ implies $s \in \mathcal{S}_S$. Thus, $f|_{\mathcal{S}_S}$ is surjective and hence bijective. We leave it to the reader to show that the functions

$$(g|_{\mathcal{T}_T})^{-1}\colon \mathcal{S}_T \to \mathcal{T}_T \text{ and } f|_{\mathcal{S}_\infty}\colon \mathcal{S}_\infty \to \mathcal{T}_\infty$$

are also bijections. Putting these three bijections together gives a bijection between S and T. Hence, $|S| = |T|$, as desired.

We now prove Cantor's Theorem. The map $\iota\colon S \to \mathcal{P}(S)$ defined by $\iota(s) = \{s\}$ is an injection from S to $\mathcal{P}(S)$ and so $|S| \le |\mathcal{P}(S)|$. To complete the proof we must show that if no injective map $f\colon S \to \mathcal{P}(S)$ can be surjective. To this end, let

$$X = \{s \in S \mid s \notin f(s)\} \in \mathcal{P}(S)$$

We claim that X is not in $\operatorname{im}(f)$. For suppose that $X = f(x)$ for some $x \in S$. Then if $x \in X$, we have by the definition of X that $x \notin X$. On the other hand, if $x \notin X$, we have again by the definition of X that $x \in X$. This contradiction implies that $X \notin \operatorname{im}(f)$ and so f is not surjective. □

Cardinal Arithmetic

Now let us define addition, multiplication and exponentiation of cardinal numbers. If S and T are sets, the **cartesian product** $S \times T$ is the set of all ordered pairs

$$S \times T = \{(s,t) \mid s \in S, t \in T\}$$

The set of all functions from T to S is denoted by S^T.

Definition *Let κ and λ denote cardinal numbers. Let S and T be any sets for which $|S| = \kappa$ and $|T| = \lambda$.*

1) *The* **sum** *$\kappa + \lambda$ is the cardinal number of $S \cup T$.*
2) *The* **product** *$\kappa\lambda$ is the cardinal number of $S \times T$.*
3) *The* **power** *κ^λ is the cardinal number of S^T.* □

We will not go into the details of why these definitions make sense. (For instance, they seem to depend on the sets S and T, but in fact they do not.) It can be shown, using these definitions, that cardinal addition and multiplication are associative and commutative and that multiplication distributes over addition.

Theorem 0.12 *Let κ, λ and μ be cardinal numbers. Then the following properties hold:*

1) **(Associativity)**

$$\kappa + (\lambda + \mu) = (\kappa + \lambda) + \mu \text{ and } \kappa(\lambda\mu) = (\kappa\lambda)\mu$$

2) (**Commutativity**)

$$\kappa + \lambda = \lambda + \kappa \textit{ and } \kappa\lambda = \lambda\kappa$$

3) (**Distributivity**)

$$\kappa(\lambda + \mu) = \kappa\lambda + \kappa\mu$$

4) (*Properties of Exponents*)
   a) $\kappa^{\lambda+\mu} = \kappa^{\lambda}\kappa^{\mu}$
   b) $(\kappa^{\lambda})^{\mu} = \kappa^{\lambda\mu}$
   c) $(\kappa\lambda)^{\mu} = \kappa^{\mu}\lambda^{\mu}$ □

On the other hand, the arithmetic of cardinal numbers can seem a bit strange, as the next theorem shows.

Theorem 0.13 *Let κ and λ be cardinal numbers, at least one of which is infinite. Then*

$$\kappa + \lambda = \kappa\lambda = \max\{\kappa, \lambda\}$$ □

It is not hard to see that there is a one-to-one correspondence between the power set $\mathcal{P}(S)$ of a set S and the set of all functions from S to $\{0, 1\}$. This leads to the following theorem.

Theorem 0.14 *For any cardinal κ*
1) *If $|S| = \kappa$ then $|\mathcal{P}(S)| = 2^{\kappa}$*
2) $\kappa < 2^{\kappa}$ □

We have already observed that $|\mathbb{N}| = \aleph_0$. It can be shown that $\aleph_0$ is the smallest infinite cardinal, that is,

$$\kappa < \aleph_0 \Rightarrow \kappa \text{ is a natural number}$$

It can also be shown that the set $\mathbb{R}$ of real numbers is in one-to-one correspondence with the power set $\mathcal{P}(\mathbb{N})$ of the natural numbers. Therefore,

$$|\mathbb{R}| = 2^{\aleph_0}$$

The set of all points on the real line is sometimes called the **continuum** and so $2^{\aleph_0}$ is sometimes called the **power of the continuum** and denoted by c.

Theorem 0.13 shows that cardinal addition and multiplication have a kind of "absorption" quality, which makes it hard to produce larger cardinals from smaller ones. The next theorem demonstrates this more dramatically.

Theorem 0.15
1) *Addition applied a countable number of times or multiplication applied a finite number of times to the cardinal number $\aleph_0$, does not yield anything*

more than $\aleph_0$. Specifically, for any nonzero $n \in \mathbb{N}$, we have

$$\aleph_0 \cdot \aleph_0 = \aleph_0 \text{ and } \aleph_0^n = \aleph_0$$

2) *Addition and multiplication, applied a countable number of times to the cardinal number $2^{\aleph_0}$ does not yield more than $2^{\aleph_0}$. Specifically, we have*

$$\aleph_0 \cdot 2^{\aleph_0} = 2^{\aleph_0} \text{ and } (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0}$$ □

Using this theorem, we can establish other relationships, such as

$$2^{\aleph_0} \le (\aleph_0)^{\aleph_0} \le (2^{\aleph_0})^{\aleph_0} = 2^{\aleph_0}$$

which, by the Schröder–Bernstein theorem, implies that

$$(\aleph_0)^{\aleph_0} = 2^{\aleph_0}$$

We mention that the problem of evaluating κ^λ in general is a very difficult one and would take us far beyond the scope of this book.

We will have use for the following reasonable–sounding result, whose proof is omitted.

Theorem 0.16 *Let $\{A_k \mid k \in K\}$ be a collection of sets, indexed by the set K, with $|K| = \kappa$. If $|A_k| \le \lambda$ for all $k \in K$ then*

$$\left| \bigcup_{k \in K} A_k \right| \le \lambda\kappa$$ □

Let us conclude by describing the cardinality of some famous sets.

Theorem 0.17

1) *The following sets have cardinality $\aleph_0$.*
   a) *The rational numbers $\mathbb{Q}$.*
   b) *The set of all finite subsets of $\mathbb{N}$.*
   c) *The union of a countable number of countable sets.*
   d) *The set $\mathbb{Z}^n$ of all ordered n-tuples of integers.*
2) *The following sets have cardinality $2^{\aleph_0}$.*
   a) *The set of all points in $\mathbb{R}^n$.*
   b) *The set of all infinite sequences of natural numbers.*
   c) *The set of all infinite sequences of real numbers.*
   d) *The set of all finite subsets of $\mathbb{R}$.*
   e) *The set of all irrational numbers.* □

Part 2 Algebraic Structures

We now turn to a discussion of some of the many algebraic structures that play a role in the study of linear algebra.

Groups

Definition *A* **group** *is a nonempty set G, together with a binary operation denoted by *, that satisfies the following properties:*

1) (**Associativity**) *For all $a, b, c \in G$*

$$(a*b)*c = a*(b*c)$$

2) (**Identity**) *There exists an element $e \in G$ for which*

$$e*a = a*e = a$$

for all $a \in G$.

3) (**Inverses**) *For each $a \in G$, there is an element $a^{-1} \in G$ for which*

$$a*a^{-1} = a^{-1}*a = e \qquad \square$$

Definition *A group G is* **abelian**, *or* **commutative**, *if*

$$a*b = b*a$$

for all $a, b \in G$. When a group is abelian, it is customary to denote the operation $*$ by $+$, thus writing $a*b$ as $a+b$. It is also customary to refer to the identity as the **zero element** *and to denote the inverse a^{-1} by $-a$, referred to as the* **negative** *of a.* □

Example 0.7 The set $\mathcal{F}$ of all bijective functions from a set S to S is a group under composition of functions. However, in general, it is not abelian. □

Example 0.8 The set $\mathcal{M}_{m,n}(F)$ is an abelian group under addition of matrices. The identity is the zero matrix $0_{m,n}$ of size $m \times n$. The set $\mathcal{M}_n(F)$ is not a group under multiplication of matrices, since not all matrices have multiplicative inverses. However, the set of invertible matrices of size $n \times n$ is a (nonabelian) group under multiplication. □

A group G is **finite** if it contains only a finite number of elements. The cardinality of a finite group G is called its **order** and is denoted by $o(G)$ or simply $|G|$. Thus, for example, $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ is a finite group under addition modulo n, but $\mathcal{M}_{m,n}(\mathbb{R})$ is not finite.

Definition *A* **subgroup** *of a group G is a nonempty subset S of G that is a group in its own right, using the same operations as defined on G.* □

Rings

Definition *A* **ring** *is a nonempty set R, together with two binary operations, called* **addition** *(denoted by $+$) and* **multiplication** *(denoted by juxtaposition), for which the following hold:*

1) *R is an abelian group under addition*

2) (**Associativity**) *For all* $a, b, c \in R$,

$$(ab)c = a(bc)$$

3) (**Distributivity**) *For all* $a, b, c \in R$,

$$(a+b)c = ac + bc \textit{ and } c(a+b) = ca + cb$$

A ring R *is said to be* **commutative** *if* $ab = ba$ *for all* $a, b \in R$. *If a ring* R *contains an element* e *with the property that*

$$ae = ea = a$$

for all $a \in R$, *we say that* R *is a* **ring with identity**. *The identity* e *is usually denoted by* 1. □

Example 0.9 The set $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ is a commutative ring under addition and multiplication modulo n

$$a \oplus b = (a+b) \operatorname{mod} n, \quad a \odot b = ab \operatorname{mod} n$$

The element $1 \in \mathbb{Z}_n$ is the identity. □

Example 0.10 The set E of even integers is a commutative ring under the usual operations on $\mathbb{Z}$, but it has no identity. □

Example 0.11 The set $\mathcal{M}_n(F)$ is a noncommutative ring under matrix addition and multiplication. The identity matrix I_n is the identity for $\mathcal{M}_n(F)$. □

Example 0.12 Let F be a field. The set $F[x]$ of all polynomials in a single variable x, with coefficients in F, is a commutative ring, under the usual operations of polynomial addition and multiplication. What is the identity for $F[x]$? Similarly, the set $F[x_1, \ldots, x_n]$ of polynomials in n variables is a commutative ring under the usual addition and multiplication of polynomials. □

Definition *A* **subring** *of a ring* R *is a subset* S *of* R *that is a ring in its own right, using the same operations as defined on* R *and having the same multiplicative identity as* R. □

The condition that a subring S have the same multiplicative identity as R is required. For example, the set S of all 2×2 matrices of the form

$$A_a = \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}$$

for $a \in F$ is a ring under addition and multiplication of matrices (isomorphic to F). The multiplicative identity in S is the matrix A_1, which is not the identity I_2 of $\mathcal{M}_{2,2}(F)$. Hence, S is a ring under the same operations as $\mathcal{M}_{2,2}(F)$ but it is not a subring of $\mathcal{M}_{2,2}(F)$.

Applying the definition is not generally the easiest way to show that a subset of a ring is a subring. The following characterization is usually easier to apply.

Theorem 0.18 *A nonempty subset S of a ring R is a subring if and only if*
1) *The multiplicative identity 1_R of R is in S*
2) *S is closed under subtraction, that is*

$$a, b \in S \Rightarrow a - b \in S$$

3) *S is closed under multiplication, that is,*

$$a, b \in S \Rightarrow ab \in S$$ □

Ideals

Rings have another important substructure besides subrings.

Definition *Let R be a ring. A nonempty subset $\mathcal{I}$ of R is called an* **ideal** *if*
1) *$\mathcal{I}$ is a subgroup of the abelian group R, that is, $\mathcal{I}$ is closed under subtraction*

$$a, b \in \mathcal{I} \Rightarrow a - b \in \mathcal{I}$$

2) *$\mathcal{I}$ is closed under multiplication by* any *ring element, that is,*

$$a \in \mathcal{I}, r \in R \Rightarrow ar \in \mathcal{I} \textit{ and } ra \in \mathcal{I}$$ □

Note that if an ideal $\mathcal{I}$ contains the unit element 1 then $\mathcal{I} = R$.

Example 0.13 Let $p(x)$ be a polynomial in $F[x]$. The set of all multiples of $p(x)$

$$\langle p(x) \rangle = \{q(x)p(x) \mid q(x) \in F[x]\}$$

is an ideal in $F[x]$, called the *ideal generated by* $p(x)$. □

Definition *Let S be a subset of a ring R with identity. The set*

$$\langle S \rangle = \{r_1 s_1 + \cdots + r_n s_n \mid r_i \in R, s_i \in S, n \geq 1\}$$

of all finite linear combinations of elements of S, with coefficients in R, is an ideal in R, called the **ideal generated by** *S. It is the smallest (in the sense of set inclusion) ideal of R containing S. If $S = \{s_1, \ldots, s_n\}$ is a finite set, we write*

$$\langle s_1, \ldots, s_n \rangle = \{r_1 s_1 + \cdots + r_n s_n \mid r_i \in R, s_i \in S\}$$ □

Note that in the previous definition, we require that R have an identity. This is to ensure that $S \subseteq \langle S \rangle$.

Theorem 0.19 *Let R be a ring.*

1) *The intersection of any collection* $\{\mathcal{I}_k \mid k \in K\}$ *of ideals is an ideal.*
2) *If* $\mathcal{I}_1 \subseteq \mathcal{I}_2 \subseteq \cdots$ *is an ascending sequence of ideals, each one contained in the next, then the union* $\bigcup \mathcal{I}_k$ *is also an ideal.*
3) *More generally, if*

$$\mathcal{C} = \{\mathcal{I}_i \mid i \in I\}$$

is a chain of ideals in R *then the union* $\mathcal{J} = \bigcup_{i\in I}\mathcal{I}_i$ *is also an ideal in* R.

Proof. To prove 1), let $\mathcal{J} = \bigcap \mathcal{I}_k$. Then if $a, b \in \mathcal{J}$, we have $a, b \in \mathcal{I}_k$ for all $k \in K$. Hence, $a - b \in \mathcal{I}_k$ for all $k \in K$ and so $a - b \in \mathcal{J}$. Hence, $\mathcal{J}$ is closed under subtraction. Also, if $r \in R$ then $ra \in \mathcal{I}_k$ for all $k \in K$ and so $ra \in \mathcal{J}$. Of course, part 2) is a special case of part 3). To prove 3), if $a, b \in \mathcal{J}$ then $a \in \mathcal{I}_i$ and $b \in \mathcal{I}_j$ for some $i, j \in I$. Since one of $\mathcal{I}_i$ and $\mathcal{I}_j$ is contained in the other, we may assume that $\mathcal{I}_i \subseteq \mathcal{I}_j$. It follows that $a, b \in \mathcal{I}_j$ and so $a - b \in \mathcal{I}_j \subseteq \mathcal{J}$ and if $r \in R$ then $ra \in \mathcal{I}_j \subseteq \mathcal{J}$. Thus $\mathcal{J}$ is an ideal. □

Note that in general, the union of ideals is not an ideal. However, as we have just proved, the union of any *chain* of ideals is an ideal.

Quotient Rings and Maximal Ideals

Let S be a subset of a commutative ring R with identity. Let $\equiv$ be the binary relation on R defined by

$$a \equiv b \Leftrightarrow a - b \in S$$

It is easy to see that $\equiv$ is an equivalence relation. When $a \equiv b$, we say that a and b are **congruent modulo** S. The term “mod” is used as a colloquialism for modulo and $a \equiv b$ is often written

$$a \equiv b \operatorname{mod} S$$

As shorthand, we write $a \equiv b$.

To see what the equivalence classes look like, observe that

$$\begin{aligned}[a] &= \{r \in R \mid r \equiv a\}\\ &= \{r \in R \mid r - a \in S\}\\ &= \{r \in R \mid r = a + s \text{ for some } s \in S\}\\ &= \{a + s \mid s \in S\}\\ &= a + S\end{aligned}$$

The set

$$a + S = \{a + s \mid s \in S\}$$

is called a **coset** of S in R. The element a is called a **coset representative** for $a + S$.

Thus, the equivalence classes for congruence mod S are the cosets $a+S$ of S in R. The set of all cosets is denoted by

$$\frac{R}{S}=\{a+S \mid a\in R\}$$

This is read "R mod S." We would like to place a ring structure on R/S. Indeed, if S is a subgroup of the abelian group R then R/S is easily seen to be an abelian group as well under coset addition defined by

$$(a+S)+(b+S)=(a+b)+S$$

In order for the product

$$(a+S)(b+S)=ab+S$$

to be well defined we must have

$$b+S=b'+S \Rightarrow ab+S=ab'+S$$

or, equivalently,

$$b-b'\in S \Rightarrow a(b-b')\in S$$

But $b-b'$ may be any element of S and a may be any element of R and so this condition implies that S must be an ideal. Conversely, if S is an ideal then coset multiplication is well defined.

Theorem 0.20 Let R be a commutative ring with identity. Then the quotient $R/\mathcal{I}$ is a ring under coset addition and multiplication if and only if $\mathcal{I}$ is an ideal of R. In this case, $R/\mathcal{I}$ is called the **quotient ring** of R **modulo** $\mathcal{I}$, where addition and multiplication are defined by

$$\begin{gathered}(a+S)+(b+S)=(a+b)+S\\(a+S)(b+S)=ab+S\end{gathered}\qquad\square$$

Definition *An ideal $\mathcal{I}$ in a ring R is a* **maximal ideal** *if $\mathcal{I}\neq R$ and if whenever $\mathcal{J}$ is an ideal satisfying $\mathcal{I}\subseteq\mathcal{J}\subseteq R$ then either $\mathcal{J}=\mathcal{I}$ or $\mathcal{J}=R$.* □

Here is one reason why maximal ideals are important.

Theorem 0.21 Let R be a commutative ring with identity. Then the quotient ring $R/\mathcal{I}$ is a field if and only if $\mathcal{I}$ is a maximal ideal.

Proof. First, note that for any ideal $\mathcal{I}$ of R, the ideals of $R/\mathcal{I}$ are precisely the quotients $\mathcal{J}/\mathcal{I}$ where $\mathcal{J}$ is an ideal for which $\mathcal{I}\subseteq\mathcal{J}\subseteq R$. It is clear that $\mathcal{J}/\mathcal{I}$ is an ideal of $R/\mathcal{I}$. Conversely, if $\mathcal{K}'$ is an ideal of $R/\mathcal{I}$ then let

$$\mathcal{K}=\{r\in R \mid r+\mathcal{I}\in\mathcal{K}'\}$$

It is easy to see that $\mathcal{K}$ is an ideal of R for which $\mathcal{I}\subseteq\mathcal{K}\subseteq R$.

Next, observe that a commutative ring S with identity is a field if and only if S has no nonzero proper ideals. For if S is a field and $\mathcal{I}$ is an ideal of S containing a nonzero element r then $1 = r^{-1}r \in \mathcal{I}$ and so $\mathcal{I} = S$. Conversely, if S has no nonzero proper ideals and $0 \neq s \in S$ then the ideal $\langle s \rangle$ must be S and so there is an $r \in S$ for which $rs = 1$. Hence, S is a field.

Putting these two facts together proves the theorem. □

The following result says that maximal ideals always exist.

Theorem 0.22 *Any commutative ring R with identity contains a maximal ideal.*
Proof. Since R is not the zero ring, the ideal $\{0\}$ is a proper ideal of R. Hence, the set $\mathcal{S}$ of all proper ideals of R is nonempty. If

$$\mathcal{C} = \{\mathcal{I}_i \mid i \in I\}$$

is a chain of proper ideals in R then the union $\mathcal{J} = \bigcup_{i \in I} \mathcal{I}_i$ is also an ideal. Furthermore, if $\mathcal{J} = R$ is not proper, then $1 \in \mathcal{J}$ and so $1 \in \mathcal{I}_i$, for some $i \in I$, which implies that $\mathcal{I}_i = R$ is not proper. Hence, $\mathcal{J} \in \mathcal{S}$. Thus, any chain in $\mathcal{S}$ has an upper bound in $\mathcal{S}$ and so Zorn's lemma implies that $\mathcal{S}$ has a maximal element. This shows that R has a maximal ideal. □

Integral Domains

Definition *Let R be a ring. A nonzero element $r \in R$ is called a* **zero divisor** *if there exists a nonzero $s \in R$ for which $rs = 0$. A commutative ring R with identity is called an* **integral domain** *if it contains no zero divisors.* □

Example 0.14 If n is not a prime number then the ring $\mathbb{Z}_n$ has zero divisors and so is not an integral domain. To see this, observe that if n is not prime then $n = ab$ in $\mathbb{Z}$, where $a, b \geq 2$. But in $\mathbb{Z}_n$, we have

$$a \odot b = ab \bmod n = 0$$

and so a and b are both zero divisors. As we will see later, if n is a prime then $\mathbb{Z}_n$ is a field (which is an integral domain, of course). □

Example 0.15 The ring $F[x]$ is an integral domain, since $p(x)q(x) = 0$ implies that $p(x) = 0$ or $q(x) = 0$. □

If R is a ring and $rx = ry$ where $r, x, y \in R$ then we cannot in general cancel the r's and conclude that $x = y$. For instance, in $\mathbb{Z}_4$, we have $2 \cdot 3 = 2 \cdot 1$, but canceling the 2's gives $3 = 1$. However, it is precisely the integral domains in which we can cancel. The simple proof is left to the reader.

Theorem 0.23 *Let R be a commutative ring with identity. Then R is an integral domain if and only if the cancellation law*

$$rx = ry, r \neq 0 \Rightarrow x = y$$

holds. □

The Field of Quotients of an Integral Domain

Any integral domain R can be embedded in a field. The **quotient field** (or **field of quotients**) of R is a field that is constructed from R just as the field of rational numbers is constructed from the ring of integers. In particular, we set

$$R^+ = \{(p,q) \mid p,q \in R, q \neq 0\}$$

Thinking of (p,q) as the "fraction" p/q we define addition and multiplication of fractions in the same way as for rational numbers

$$(p,q) + (r,s) = (ps + qr, qs)$$

and

$$(p,q) \cdot (r,s) = (pr, qs)$$

It is customary to write (p,q) in the form p/q. Note that if R has zero divisors, then these definitions do not make sense, because qs may be 0 even if q and s are not. This is why we require that R be an integral domain.

Principal Ideal Domains

Definition *Let R be a ring with identity and let $a \in R$. The* **principal ideal** *generated by a is the ideal*

$$\langle a \rangle = \{ra \mid r \in R\}$$

An integral domain *R in which every ideal is a principal ideal is called a* **principal ideal domain**. □

Theorem 0.24 *The integers form a principal ideal domain. In fact, any ideal $\mathcal{I}$ in $\mathbb{Z}$ is generated by the smallest positive integer a that is contained in $\mathcal{I}$.* □

Theorem 0.25 *The ring $F[x]$ is a principal ideal domain. In fact, any ideal $\mathcal{I}$ is generated by the unique monic polynomial of smallest degree contained in $\mathcal{I}$. Moreover, for polynomials $p_1(x), \ldots, p_n(x)$,*

$$\langle p_1(x), \ldots, p_n(x) \rangle = \langle \text{gcd}\{p_1(x), \ldots, p_n(x)\} \rangle$$

Proof. Let $\mathcal{I}$ be an ideal in $F[x]$ and let $m(x)$ be a monic polynomial of smallest degree in $\mathcal{I}$. First, we observe that there is only one such polynomial in $\mathcal{I}$. For if $n(x) \in \mathcal{I}$ is monic and $\deg(n(x)) = \deg(m(x))$ then

$$b(x) = m(x) - n(x) \in \mathcal{I}$$

and since $\deg(b(x)) < \deg(m(x))$, we must have $b(x) = 0$ and so $n(x) = m(x)$.

We show that $\mathcal{I} = \langle m(x) \rangle$. Since $m(x) \in \mathcal{I}$, we have $\langle m(x) \rangle \subseteq \mathcal{I}$. To establish the reverse inclusion, if $p(x) \in \mathcal{I}$ then dividing $p(x)$ by $m(x)$ gives

$$p(x) = q(x)m(x) + r(x)$$

where $r(x) = 0$ or $0 \leq \deg r(x) < \deg m(x)$. But since $\mathcal{I}$ is an ideal,

$$r(x) = p(x) - q(x)m(x) \in \mathcal{I}$$

and so $0 \leq \deg r(x) < \deg m(x)$ is impossible. Hence, $r(x) = 0$ and

$$p(x) = q(x)m(x) \in \langle m(x) \rangle$$

This shows that $\mathcal{I} \subseteq \langle m(x) \rangle$ and so $\mathcal{I} = \langle m(x) \rangle$.

To prove the second statement, let $\mathcal{I} = \langle p_1(x), \dots, p_n(x) \rangle$. Then, by what we have just shown,

$$\mathcal{I} = \langle p_1(x), \dots, p_n(x) \rangle = \langle m(x) \rangle$$

where $m(x)$ is the unique monic polynomial $m(x)$ in $\mathcal{I}$ of smallest degree. In particular, since $p_i(x) \in \langle m(x) \rangle$, we have $m(x) \mid p_i(x)$ for each $i = 1, \dots, n$. In other words, $m(x)$ is a common divisor of the $p_i(x)$'s.

Moreover, if $q(x) \mid p_i(x)$ for all i, then $p_i(x) \in \langle q(x) \rangle$ for all i, which implies that

$$m(x) \in \langle m(x) \rangle = \langle p_1(x), \dots, p_n(x) \rangle \subseteq \langle q(x) \rangle$$

and so $q(x) \mid m(x)$. This shows that $m(x)$ is the *greatest* common divisor of the $p_i(x)$'s and completes the proof. □

Example 0.16 The ring $R = F[x, y]$ of polynomials in two variables x and y is not a principal ideal domain. To see this, observe that the set $\mathcal{I}$ of all polynomials with zero constant term is an ideal in R. Now, suppose that $\mathcal{I}$ is the principal ideal $\mathcal{I} = \langle p(x, y) \rangle$. Since $x, y \in \mathcal{I}$, there exist polynomials $a(x, y)$ and $b(x, y)$ for which

$$x = a(x, y)p(x, y) \text{ and } y = b(x, y)p(x, y) \tag{0.1}$$

But $p(x, y)$ cannot be a constant for then we would have $\mathcal{I} = R$. Hence, $\deg(p(x, y)) \geq 1$ and so $a(x, y)$ and $b(x, y)$ must both be constants, which implies that (0.1) cannot hold. □

Theorem 0.26 *Any principal ideal domain R satisfies the* **ascending chain condition**, *that is, R cannot have a strictly increasing sequence of ideals*

$$\mathcal{I}_1 \subset \mathcal{I}_2 \subset \cdots$$

where each ideal is properly contained in the next one.

Proof. Suppose to the contrary that there is such an increasing sequence of ideals. Consider the ideal

$$U = \bigcup \mathcal{I}_i$$

which must have the form $U = \langle a \rangle$ for some $a \in U$. Since $a \in \mathcal{I}_k$ for some k, we have $\mathcal{I}_k = \mathcal{I}_j$ for all $j \geq k$, contradicting the fact that the inclusions are proper. □

Prime and Irreducible Elements

We can define the notion of a prime element in any integral domain. For $r, s \in R$, we say that r **divides** s (written $r \mid s$) if there exists an $x \in R$ for which $s = xr$.

Definition *Let R be an integral domain.*

1) *An invertible element of R is called a* **unit**. *Thus, $u \in R$ is a unit if $uv = 1$ for some $v \in R$.*
2) *Two elements $a, b \in R$ are said to be* **associates** *if there exists a unit u for which $a = ub$.*
3) *A nonzero nonunit $p \in R$ is said to be* **prime** *if*

$$p \mid ab \Rightarrow p \mid a \text{ or } p \mid b$$

4) *A nonzero nonunit $r \in R$ is said to be* **irreducible** *if*

$$r = ab \Rightarrow a \text{ or } b \text{ is a unit}$$ □

Note that if p is prime or irreducible then so is up for any unit u.

Theorem 0.27 *Let R be a ring.*

1) *An element $u \in R$ is a unit if and only if $\langle u \rangle = R$.*
2) *r and s are associates if and only if $\langle r \rangle = \langle s \rangle$.*
3) *r divides s if and only if $\langle s \rangle \subseteq \langle r \rangle$.*
4) *r* **properly divides** *s, that is, $s = xr$ where x is not a unit, if and only if $\langle s \rangle \subset \langle r \rangle$.* □

In the case of the integers, an integer is prime if and only if it is irreducible. In any integral domain, prime elements are irreducible, but the converse need not hold. (In the ring $\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$ the prime element 2 divides the product $(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6$ but does not divide either factor.)

However, in principal ideal domains, the two concepts are equivalent.

Theorem 0.28 *Let R be a principal ideal domain.*

1) *An $r \in R$ is irreducible if and only if the ideal $\langle r \rangle$ is maximal.*
2) *An element in R is prime if and only if it is irreducible.*

3) *The elements $a, b \in R$ are* **relatively prime***, that is, have no common nonunit factors if and only if there exist $r, s \in R$ for which*

$$ra + sb = 1$$

Proof. To prove 1), suppose that r is irreducible and that $\langle r \rangle \subseteq \langle a \rangle \subseteq R$. Then $r \in \langle a \rangle$ and so $r = xa$ for some $x \in R$. The irreducibility of r implies that a or x is a unit. If a is a unit then $\langle a \rangle = R$ and if x is a unit then $\langle a \rangle = \langle xa \rangle = \langle r \rangle$. This shows that $\langle r \rangle$ is maximal. (We have $\langle r \rangle \neq R$, since r is not a unit.) Conversely, suppose that r is not irreducible, that is, $r = ab$ where neither a nor b is a unit. Then $\langle r \rangle \subseteq \langle a \rangle \subseteq R$. But if $\langle a \rangle = \langle r \rangle$ then r and a are associates, which implies that b is a unit. Hence $\langle r \rangle \neq \langle a \rangle$. Also, if $\langle a \rangle = R$ then a must be a unit. So we conclude that $\langle r \rangle$ is not maximal, as desired.

To prove 2), assume first that p is prime and $p = ab$. Then $p \mid a$ or $p \mid b$. We may assume that $p \mid a$. Therefore, $a = xp = xab$. Canceling a's gives $1 = xb$ and so b is a unit. Hence, p is irreducible. (Note that this argument applies in any integral domain.)

Conversely, suppose that r is irreducible and let $r \mid ab$. We wish to prove that $r \mid a$ or $r \mid b$. The ideal $\langle r \rangle$ is maximal and so $\langle r, a \rangle = \langle r \rangle$ or $\langle r, a \rangle = R$. In the former case, $r \mid a$ and we are done. In the latter case, we have

$$1 = xa + yr$$

for some $x, y \in R$. Thus,

$$b = xab + yrb$$

and since r divides both terms on the right, we have $r \mid b$.

To prove 3), it is clear that if $ra + sb = 1$ then a and b are relatively prime. For the converse, consider the ideal $\langle a, b \rangle$ which must be principal, say $\langle a, b \rangle = \langle x \rangle$. Then $x \mid a$ and $x \mid b$ and so x must be a unit, which implies that $\langle a, b \rangle = R$. Hence, there exists $r, s \in R$ for which $ra + sb = 1$. □

Unique Factorization Domains

Definition *An integral domain R is said to be a* **unique factorization domain** *if it has the following factorization properties:*

1) *Every nonzero nonunit element $r \in R$ can be written as a product of a finite number of irreducible elements $r = p_1 \cdots p_n$.*
2) *The factorization into irreducible elements is unique in the sense that if $r = p_1 \cdots p_n$ and $r = q_1 \cdots q_m$ are two such factorizations then $m = n$ and after a suitable reindexing of the factors, p_i and q_i are associates.* □

Unique factorization is clearly a desirable property. Fortunately, principal ideal domains have this property.

Theorem 0.29 *Every principal ideal domain R is a unique factorization domain.*

Proof. Let $r \in R$ be a nonzero nonunit. If r is irreducible then we are done. If not then $r = r_1 r_2$, where neither factor is a unit. If r_1 and r_2 are irreducible, we are done. If not, suppose that r_2 is not irreducible. Then $r_2 = r_3 r_4$, where neither r_3 nor r_4 is a unit. Continuing in this way, we obtain a factorization of the form (after renumbering if necessary)

$$r = r_1 r_2 = r_1(r_3 r_4) = (r_1 r_3)(r_5 r_6) = (r_1 r_3 r_5)(r_7 r_8) = \cdots$$

Each step is a factorization of r into a product of nonunits. However, this process must stop after a finite number of steps, for otherwise it will produce an infinite sequence $s_1, s_2, \ldots$ of nonunits of R for which s_{i+1} properly divides s_i. But this gives the ascending chain of ideals

$$\langle s_1 \rangle \subset \langle s_2 \rangle \subset \langle s_3 \rangle \subset \langle s_4 \rangle \subset \cdots$$

where the inclusions are proper. But this contradicts the fact that a principal ideal domain satisfies the ascending chain condition. Thus, we conclude that every nonzero nonunit has a factorization into irreducible elements.

As to uniqueness, if $r = p_1 \cdots p_n$ and $r = q_1 \cdots q_m$ are two such factorizations then because R is an integral domain, we may equate them and cancel like factors, so let us assume this has been done. Thus, $p_i \neq q_j$ for all i, j. If there are no factors on either side, we are done. If exactly one side has no factors left then we have expressed 1 as a product of irreducible elements, which is not possible since irreducible elements are nonunits.

Suppose that both sides have factors left, that is,

$$p_1 \cdots p_n = q_1 \cdots q_m$$

where $p_i \neq q_j$. Then $q_m \mid p_1 \cdots p_n$, which implies that $q_m \mid p_i$ for some i. We can assume by reindexing if necessary that $p_n = a_n q_m$. Since p_n is irreducible a_n must be a unit. Replacing p_n by $a_n q_m$ and canceling q_m gives

$$a_n p_1 \cdots p_{n-1} = q_1 \cdots q_{m-1}$$

This process can be repeated until we run out of q's or p's. If we run out of q's first then we have an equation of the form $u p_1 \cdots p_k = 1$ where u is a unit, which is not possible since the p_i's are not units. By the same reasoning, we cannot run out of q's first and so $n = m$ and the p's and q's can be paired off as associates. □

Fields

For the record, let us give the definition of a field (a concept that we have been using).

Definition *A* **field** *is a set F, containing at least two elements, together with two binary operations, called* **addition** *(denoted by $+$) and* **multiplication** *(denoted by juxtaposition), for which the following hold:*

1) *F is an abelian group under addition.*
2) *The set F^* of all* nonzero *elements in F is an abelian group under multiplication.*
3) (**Distributivity**) *For all $a, b, c \in F$,*

$$(a+b)c = ac + bc \text{ and } c(a+b) = ca + cb \qquad \square$$

We require that F have at least two elements to avoid the pathological case, in which $0 = 1$.

Example 0.17 The sets $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$, of all rational, real and complex numbers, respectively, are fields, under the usual operations of addition and multiplication of numbers. □

Example 0.18 The ring $\mathbb{Z}_n$ is a field if and only if n is a prime number. We have already seen that $\mathbb{Z}_n$ is not a field if n is not prime, since a field is also an integral domain. Now suppose that $n = p$ is a prime.

We have seen that $\mathbb{Z}_p$ is an integral domain and so it remains to show that every nonzero element in $\mathbb{Z}_p$ has a multiplicative inverse. Let $0 \neq a \in \mathbb{Z}_p$. Since $a < p$, we know that a and p are relatively prime. It follows that there exist integers u and v for which

$$ua + vp = 1$$

Hence,

$$ua \equiv (1 - vp) \equiv 1 \bmod p$$

and so $u \odot a = 1$ in $\mathbb{Z}_p$, that is, u is the multiplicative inverse of a. □

The previous example shows that not all fields are infinite sets. In fact, finite fields play an extremely important role in many areas of abstract and applied mathematics.

A field F is said to be **algebraically closed** if every nonconstant polynomial over F has a root in F. This is equivalent to saying that every nonconstant polynomial *splits* into linear factors over F. For example, the complex field $\mathbb{C}$ is algebraically closed but the real field $\mathbb{R}$ is not. We mention without proof that every field F is contained in an algebraically closed field $\overline{F}$, called the **algebraic closure** of F.

The Characteristic of a Ring

Let R be a ring with identity. If n is a positive integer then by $n \cdot r$, we simply mean

$$n \cdot r = \underbrace{r + \cdots + r}_{n \text{ terms}}$$

Now, it may happen that there is a positive integer n for which

$$n \cdot 1 = 0$$

For instance, in $\mathbb{Z}_n$, we have $n \cdot 1 = n = 0$. On the other hand, in $\mathbb{Z}$, the equation $n \cdot 1 = 0$ implies $n = 0$ and so no such positive integer exists.

Notice that, in any *finite* ring, there must exist such a positive integer n, since the infinite sequence of numbers

$$1 \cdot 1, 2 \cdot 1, 3 \cdot 1, \ldots$$

cannot be distinct and so $i \cdot 1 = j \cdot 1$ for some $i < j$, whence $(j - i) \cdot 1 = 0$.

Definition *Let R be a ring with identity. The smallest positive integer c for which $c \cdot 1 = 0$ is called the* **characteristic** *of R. If no such number c exists, we say that R has characteristic 0. The characteristic of R is denoted by* char(R). □

If char$(R) = c$ then for any $r \in R$, we have

$$c \cdot r = \underbrace{r + \cdots + r}_{c \text{ terms}} = (\underbrace{1 + \cdots + 1}_{c \text{ terms}})r = 0 \cdot r = 0$$

Theorem 0.30 Any finite ring has nonzero characteristic. Any finite field has prime characteristic.

Proof. We have already seen that a finite ring has nonzero characteristic. Let F be a finite field and suppose that char$(F) = c > 0$. If $c = pq$, where $p, q < c$ then $pq \cdot 1 = 0$. Hence, $(p \cdot 1)(q \cdot 1) = 0$, implying that $p \cdot 1 = 0$ or $q \cdot 1 = 0$. In either case, we have a contradiction to the fact that c is the smallest positive integer such that $c \cdot 1 = 0$. Hence, c must be prime. □

Notice that in any field F of characteristic 2, we have $2a = 0$ for all $a \in F$. Thus, in F

$$a = -a \text{ for all } a \in F$$

This property takes a bit of getting used to and makes fields of characteristic 2 quite exceptional. (As it happens, there are many important uses for fields of characteristic 2.)

Algebras

The final algebraic structure of which we will have use is a combination of a vector space and a ring. (We have not yet officially defined vector spaces, but we will do so before needing the following definition, which is placed here for easy reference.)

Definition *An* **algebra** $\mathcal{A}$ *over a field* F *is a nonempty set* $\mathcal{A}$*, together with three operations, called* **addition** *(denoted by* $+$*),* **multiplication** *(denoted by juxtaposition) and* **scalar multiplication** *(also denoted by juxtaposition), for which the following properties hold:*

1) $\mathcal{A}$ *is a vector space over* F *under addition and scalar multiplication.*
2) $\mathcal{A}$ *is a ring under addition and multiplication.*
3) *If* $r \in F$ *and* $a, b \in \mathcal{A}$ *then*

$$r(ab) = (ra)b = a(rb) \qquad \square$$

Thus, an algebra is a vector space in which we can take the product of vectors, or a ring in which we can multiply each element by a scalar (subject, of course, to additional requirements as given in the definition).

Part I—Basic Linear Algebra

Chapter 1
Vector Spaces

Vector Spaces

Let us begin with the definition of one of our principal objects of study.

Definition *Let F be a field, whose elements are referred to as* **scalars***. A* **vector space** *over F is a nonempty set V, whose elements are referred to as* **vectors***, together with two operations. The first operation, called* **addition** *and denoted by $+$, assigns to each pair (u,v) of vectors in V a vector $u+v$ in V. The second operation, called* **scalar multiplication** *and denoted by juxtaposition, assigns to each pair $(r,u)\in F\times V$ a vector ru in V. Furthermore, the following properties must be satisfied:*

1) (**Associativity of addition**) *For all vectors $u,v,w\in V$*

$$u+(v+w)=(u+v)+w$$

2) (**Commutativity of addition**) *For all vectors $u,v\in V$*

$$u+v=v+u$$

3) (**Existence of a zero**) *There is a vector $0\in V$ with the property that*

$$0+u=u+0=u$$

for all vectors $u\in V$.

4) (**Existence of additive inverses**) *For each vector $u\in V$, there is a vector in V, denoted by $-u$, with the property that*

$$u+(-u)=(-u)+u=0$$

5) (**Properties of scalar multiplication**) *For all scalars $a,b \in F$ and for all vectors $u,v \in V$*

$$\begin{aligned} a(u+v) &= au+av \\ (a+b)u &= au+bu \\ (ab)u &= a(bu) \\ 1u &= u \end{aligned}$$ □

Note that the first four properties in the definition of vector space can be summarized by saying that V is an abelian group under addition.

Any expression of the form

$$a_1v_1 + \cdots + a_nv_n$$

where $a_i \in F$ and $v_i \in V$ for all i, is called a **linear combination** of the vectors $v_1,\ldots,v_n$. If at least one of the scalars a_i is nonzero, then the linear combination is **nontrivial**.

Example 1.1

1) Let F be a field. The set F^F of all functions from F to F is a vector space over F, under the operations of ordinary addition and scalar multiplication of functions

   $$(f+g)(x) = f(x) + g(x)$$

   and

   $$(af)(x) = a(f(x))$$

2) The set $\mathcal{M}_{m,n}(F)$ of all $m \times n$ matrices with entries in a field F is a vector space over F, under the operations of matrix addition and scalar multiplication.
3) The set F^n of all ordered n-tuples, whose components lie in a field F, is a vector space over F, with addition and scalar multiplication defined componentwise

   $$(a_1,\ldots,a_n) + (b_1,\ldots,b_n) = (a_1+b_1,\ldots,a_n+b_n)$$

   and

   $$c(a_1,\ldots,a_n) = (ca_1,\ldots,ca_n)$$

   When convenient, we will also write the elements of F^n in column form. When F is a finite field F_q with q elements, we write $V(n,q)$ for F_q^n.
4) Many sequence spaces are vector spaces. The set $\mathrm{Seq}(F)$ of all infinite sequences with members from a field F is a vector space under componentwise operations

   $$(s_n) + (t_n) = (s_n + t_n)$$

and

$$a(s_n) = (as_n)$$

In a similar way, the set c_0 of all sequences of complex numbers that converge to 0 is a vector space, as is the set ℓ^∞ of all bounded complex sequences. Also, if p is a positive integer then the set ℓ^p of all complex sequences (s_n) for which

$$\sum_{n=1}^{\infty} |s_n|^p < \infty$$

is a vector space under componentwise operations. To see that addition is a binary operation on ℓ^p, one verifies **Minkowski's inequality**

$$\left(\sum_{n=1}^{\infty} |s_n + t_n|^p\right)^{1/p} \le \left(\sum_{n=1}^{\infty} |s_n|^p\right)^{1/p} + \left(\sum_{n=1}^{\infty} |t_n|^p\right)^{1/p}$$

which we will not do here. □

Subspaces

Most algebraic structures contain substructures, and vector spaces are no exception.

Definition *A* **subspace** *of a vector space V is a subset S of V that is a vector space in its own right under the operations obtained by restricting the operations of V to S.* □

Since many of the properties of addition and scalar multiplication hold a fortiori in a nonempty subset S, we can establish that S is a subspace merely by checking that S is closed under the operations of V.

Theorem 1.1 *A nonempty subset S of a vector space V is a subspace of V if and only if S is closed under addition and scalar multiplication or, equivalently, S is closed under linear combinations, that is*

$$a, b \in F, u, v \in S \Rightarrow au + bv \in S \qquad \square$$

Example 1.2 Consider the vector space $V(n,2)$ of all binary n-tuples, that is, n-tuples of 0's and 1's. The **weight** $\mathcal{W}(v)$ of a vector $v \in V(n,2)$ is the number of nonzero coordinates in v. For instance, $\mathcal{W}(101010) = 3$. Let E_n be the set of all vectors in V of even weight. Then E_n is a subspace of $V(n,2)$.

To see this, note that

$$\mathcal{W}(u+v) = \mathcal{W}(u) + \mathcal{W}(v) - 2\mathcal{W}(u \cap v)$$

where $u \cap v$ is the vector in $V(n,2)$ whose ith component is the product of the

ith components of u and v, that is,

$$(u \cap v)_i = u_i \cdot v_i$$

Hence, if $\mathcal{W}(u)$ and $\mathcal{W}(v)$ are both even, so is $\mathcal{W}(u+v)$. Finally, scalar multiplication over F_2 is trivial and so E_n is a subspace of $V(n,2)$, known as the **even weight subspace** of $V(n,2)$. □

Example 1.3 Any subspace of the vector space $V(n,q)$ is called a **linear code**. Linear codes are among the most important and most studied types of codes, because their structure allows for efficient encoding and decoding of information. □

The Lattice of Subspaces

The set $\mathcal{S}(V)$ of all subspaces of a vector space V is partially ordered by set inclusion. The **zero subspace** $\{0\}$ is the smallest element in $\mathcal{S}(V)$ and the entire space V is the largest element.

If $S,T \in \mathcal{S}(V)$ then $S \cap T$ is the largest subspace of V that is contained in both S and T. In terms of set inclusion, $S \cap T$ is the *greatest lower bound* of S and T

$$S \cap T = \text{glb}\{S,T\}$$

Similarly, if $\{S_i \mid i \in K\}$ is any collection of subspaces of V then their intersection is the greatest lower bound of the subspaces

$$\bigcap_{i \in K} S_i = \text{glb}\{S_i \mid i \in K\}$$

On the other hand, if $S,T \in \mathcal{S}(V)$ (and F is infinite) then $S \cup T \in \mathcal{S}(V)$ if and only if $S \subseteq T$ or $T \subseteq S$. Thus, the union of two subspaces is never a subspace in any "interesting" case. We also have the following.

Theorem 1.2 *A nontrivial vector space V over an infinite field F is not the union of a finite number of proper subspaces.*
Proof. Suppose that $V = S_1 \cup \cdots \cup S_n$, where we may assume that

$$S_1 \not\subseteq S_2 \cup \cdots \cup S_n$$

Let $w \in S_1 \setminus (S_2 \cup \cdots \cup S_n)$ and let $v \notin S_1$. Consider the infinite set

$$A = \{rw + v \mid r \in F\}$$

which is the "line" through v, parallel to w. We want to show that each S_i contains at most one vector from the infinite set A, which is contrary to the fact that $V = S_1 \cup \cdots \cup S_n$. This will prove the theorem.

If $rw + v \in S_1$ for $r \neq 0$ then $w \in S_1$ implies $v \in S_1$, contrary to assumption. Next, suppose that $r_1 w + v \in S_i$ and $r_2 w + v \in S_i$, for $i \geq 2$, where $r_1 \neq r_2$. Then

$$S_i \ni (r_1 w + v) - (r_2 w + v) = (r_2 - r_1)w$$

and so $w \in S_i$, which is also contrary to assumption. □

To determine the smallest subspace of V containing the subspaces S and T, we make the following definition.

Definition *Let S and T be subspaces of V. The* **sum** *$S + T$ is defined by*

$$S + T = \{u + v \mid u \in S, v \in T\}$$

More generally, the **sum** *of any collection $\{S_i \mid i \in K\}$ of subspaces is the set of all finite sums of vectors from the union $\bigcup S_i$*

$$\sum_{i \in K} S_i = \left\{ s_1 + \cdots + s_n \mid s_j \in \bigcup_{i \in K} S_i \right\} \qquad □$$

It is not hard to show that the sum of any collection of subspaces of V is a subspace of V and that in terms of set inclusion, the sum is the least upper bound

$$S + T = \text{lub}\{S, T\}$$

More generally,

$$\sum_{i \in K} S_i = \text{lub}\{S_i \mid i \in K\}$$

If a partially ordered set P has the property that every pair of elements has a least upper bound and greatest lower bound, then P is called a **lattice**. If P has a smallest element and a largest element and has the property that every collection of elements has a least upper bound and greatest lower bound, then P is called a **complete lattice**.

Theorem 1.3 *The set $\mathcal{S}(V)$ of all subspaces of a vector space V is a complete lattice under set inclusion, with smallest element $\{0\}$, largest element V,*

$$\text{glb}\{S_i \mid i \in K\} = \bigcap_{i \in K} S_i$$

and

$$\text{lub}\{S_i \mid i \in K\} = \sum_{i \in K} S_i \qquad □$$

Direct Sums

As we will see, there are many ways to construct new vector spaces from old ones.

External Direct Sums

Definition *Let $V_1, \ldots, V_n$ be vector spaces over a field F. The* **external direct sum** *of $V_1, \ldots, V_n$, denoted by*

$$V = V_1 \boxplus \cdots \boxplus V_n$$

is the vector space V whose elements are ordered n-tuples

$$V = \{(v_1, \ldots, v_n) \mid v_i \in V_i, i = 1, \ldots, n\}$$

with componentwise operations

$$(u_1, \ldots, u_n) + (v_1, \ldots, v_n) = (u_1 + v_1, \ldots, u_n + v_n)$$

and

$$r(v_1, \ldots, v_n) = (rv_1, \ldots, rv_n) \qquad \square$$

Example 1.4 The vector space F^n is the external direct sum of n copies of F, that is,

$$F^n = F \boxplus \cdots \boxplus F$$

where there are n summands on the right-hand side. □

This construction can be generalized to any collection of vector spaces by generalizing the idea that an ordered n-tuple $(v_1, \ldots, v_n)$ is just a function $f\colon \{1, \ldots, n\} \to \bigcup V_i$ from the *index set* $\{1, \ldots, n\}$ to the union of the spaces with the property that $f(i) \in V_i$.

Definition *Let $\mathcal{F} = \{V_i \mid i \in K\}$ be any family of vector spaces over F. The* **direct product** *of $\mathcal{F}$ is the vector space*

$$\prod_{i \in K} V_i = \left\{ f\colon K \to \bigcup_{i \in K} V_i \,\middle|\, f(i) \in V_i \right\}$$

thought of as a subspace of the vector space of all functions from K to $\bigcup V_i$. □

It will prove more useful to restrict the set of functions to those with finite support.

Definition *Let $\mathcal{F} = \{V_i \mid i \in K\}$ be a family of vector spaces over F. The* **support** *of a function $f\colon K \to \bigcup V_i$ is the set*

$$\operatorname{supp}(f) = \{i \in K \mid f(i) \neq 0\}$$

Thus, a function f has **finite support** *if $f(i) = 0$ for all but a finite number of $i \in K$. The* **external direct sum** *of the family $\mathcal{F}$ is the vector space*

$$\bigoplus_{i \in K}^{\text{ext}} V_i = \left\{ f\colon K \to \bigcup_{i \in K} V_i \;\middle|\; f(i) \in V_i,\ f \text{ has finite support} \right\}$$

thought of as a subspace of the vector space of all functions from K to $\bigcup V_i$. □

An important special case occurs when $V_i = V$ for all $i \in K$. If we let V^K denote the set of all functions from K to V and $(V^K)_0$ denote the set of all functions in V^K that have finite support then

$$\prod_{i \in K} V = V^K \quad \text{and} \quad \bigoplus_{i \in K}^{\text{ext}} V = (V^K)_0$$

Note that the direct product and the external direct sum are the same for a *finite* family of vector spaces.

Internal Direct Sums

An internal version of the direct sum construction is often more relevant.

Definition *Let V be a vector space. We say that V is the (***internal***)* **direct sum** *of a family $\mathcal{F} = \{S_i \mid i \in K\}$ of subspaces of V if every vector $v \in V$ can be written, in a unique way (except for order), as a finite sum of vectors from the subspaces in $\mathcal{F}$, that is, if for all $v \in V$,*

$$v = u_1 + \cdots + u_n$$

for $u_i \in S_i$ and furthermore, if

$$v = w_1 + \cdots + w_m$$

where $w_i \in S_i$ then $m = n$ and (after reindexing if necessary) $w_i = u_i$ for all $i = 1, \ldots, n$.

If V is the direct sum of $\mathcal{F}$, we write

$$V = \bigoplus_{i \in K} S_i$$

and refer to each S_i as a **direct summand** *of V. If $\mathcal{F} = \{S_1, \ldots, S_n\}$ is a finite family, we write*

$$V = S_1 \oplus \cdots \oplus S_n$$

If $V = S \oplus T$ then T is called a **complement** *of S in V.* □

Note that a sum is direct if and only if whenever $u_{i_1}+\cdots+u_{i_n}=0$ where $u_{i_j}\in S_{i_j}$ and $i_j\neq i_k$ then $u_{i_j}=0$ for all j, that is, if and only if 0 has a unique representation as a sum of vectors from distinct subspaces.

The reader will be asked in a later chapter to show that the concepts of internal and external direct sum are essentially equivalent (isomorphic). For this reason, we often use the term "direct sum" without qualification. Once we have discussed the concept of a basis, the following theorem can be easily proved.

Theorem 1.4 *Any subspace of a vector space has a complement, that is, if S is a subspace of V then there exists a subspace T for which $V=S\oplus T$.* □

It should be emphasized that a subspace generally has many complements (although they are isomorphic). The reader can easily find examples of this in $\mathbb{R}^2$. We will have more to say about the existence and uniqueness of complements later in the book.

The following characterization of direct sums is quite useful.

Theorem 1.5 *A vector space V is the direct sum of a family $\mathcal{F}=\{S_i\mid i\in K\}$ of subspaces if and only if*

1) V is the sum of the S_i

$$V=\sum_{i\in K}S_i$$

2) For each $i\in K$,

$$S_i\cap\Big(\sum_{j\neq i}S_j\Big)=\{0\}$$

Proof. Suppose first that V is the direct sum of $\mathcal{F}$. Then 1) certainly holds and if

$$v\in S_i\cap\Big(\sum_{j\neq i}S_j\Big)$$

then $v=s_i$ for some $s_i\in S_i$ and

$$v=s_{j_1}+\cdots+s_{j_n}$$

where $s_{j_k}\in S_{j_k}$ and $j_k\neq i$ for all $k=1,\ldots,n$. Hence, by the uniqueness of direct sum representations, $s_i=0$ and so $v=0$. Thus, 2) holds.

For the converse, suppose that 1) and 2) hold. We need only verify the uniqueness condition. If

$$v=s_{j_1}+\cdots+s_{j_n}$$

and

$$v = t_{k_1} + \cdots + t_{k_m}$$

where $s_{j_i} \in S_{j_i}$ and $t_{k_i} \in S_{k_i}$ then by including additional terms equal to 0 we may assume that the index sets $\{j_1, \ldots, j_n\}$ and $\{k_1, \ldots, k_m\}$ are the same set $\{i_1, \ldots, i_p\}$, that is

$$v = s_{i_1} + \cdots + s_{i_p}$$

and

$$v = t_{i_1} + \cdots + t_{i_p}$$

Thus,

$$(s_{i_1} - t_{i_1}) + \cdots + (s_{i_p} - t_{i_p}) = 0$$

Hence, each term $s_{i_u} - t_{i_u} \in S_{i_u}$ is a sum of vectors from subspaces other than S_{i_u}, which can happen only if $s_{i_u} - t_{i_u} = 0$. Thus, $s_{i_u} = t_{i_u}$ for all i_u and V is the direct sum of $\mathcal{F}$. □

Example 1.5 Any matrix $A \in \mathcal{M}_n$ can be written in the form

$$A = \frac{1}{2}(A + A^t) + \frac{1}{2}(A - A^t) = B + C \tag{1.1}$$

where A^t is the transpose of A. It is easy to verify that B is symmetric and C is skew-symmetric and so (1.1) is a decomposition of A as the sum of a symmetric matrix and a skew-symmetric matrix.

Since the sets Sym and SkewSym of all symmetric and skew-symmetric matrices in $\mathcal{M}_n$ are subspaces of $\mathcal{M}_n$, we have

$$\mathcal{M}_n = \text{Sym} + \text{SkewSym}$$

Furthermore, if $S + T = S' + T'$, where S and S' are symmetric and T and T' are skew-symmetric, then the matrix

$$U = S - S' = T' - T$$

is both symmetric and skew-symmetric. Hence, provided that $\text{char}(F) \neq 2$, we must have $U = 0$ and so $S = S'$ and $T = T'$. Thus,

$$\mathcal{M}_n = \text{Sym} \oplus \text{SkewSym} \qquad □$$

Spanning Sets and Linear Independence

A set of vectors *spans* a vector space if every vector can be written as a linear combination of some of the vectors in that set. Here is the formal definition.

Definition *The* **subspace spanned** *(or* **subspace generated***) by a set S of vectors in V is the set of all linear combinations of vectors from S*

$$\langle S\rangle = \operatorname{span}(S) = \{r_1v_1 + \cdots + r_nv_n \mid r_i \in F, v_i \in V\}$$

When $S = \{v_1,\ldots,v_n\}$ is a finite set, we use the notation $\langle v_1,\ldots,v_n\rangle$, or $\operatorname{span}(v_1,\ldots,v_n)$. *A set S of vectors in V is said to* **span** *V, or* **generate** *V, if $V = \operatorname{span}(S)$, that is, if every vector $v \in V$ can be written in the form*

$$v = r_1v_1 + \cdots + r_nv_n$$

for some scalars $r_1,\ldots,r_n$ and vectors $v_1,\ldots,v_n$. □

It is clear that any superset of a spanning set is also a spanning set. Note also that all vector spaces have spanning sets, since the entire space is a spanning set.

Definition *A nonempty set S of vectors in V is* **linearly independent** *if for any $v_1,\ldots,v_n$ in S, we have*

$$r_1v_1 + \cdots + r_nv_n = 0 \Rightarrow r_1 = \cdots = r_n = 0$$

If a set of vectors is not linearly independent, it is said to be **linearly dependent**. □

It follows from the definition that any nonempty subset of a linearly independent set is linearly independent.

Theorem 1.6 *Let S be a set of vectors in V. The following are equivalent:*

1) *S is linearly independent.*
2) *Every vector in* $\operatorname{span}(S)$ *has a* unique *expression as a linear combination of the vectors in S.*
3) *No vector in S is a linear combination of the other vectors in S.* □

The following key theorem relates the notions of spanning set and linear independence.

Theorem 1.7 *Let S be a set of vectors in V. The following are equivalent:*

1) *S is linearly independent and spans V.*
2) *For every vector $v \in V$, there is a* unique *set of vectors $v_1,\ldots,v_n$ in S, along with a* unique *set of scalars $r_1,\ldots,r_n$ in F, for which*

$$v = r_1v_1 + \cdots + r_nv_n$$

3) *S is a* **minimal spanning set**, *that is, S spans V but any proper subset of S does not span V.*
4) *S is a* **maximal linearly independent set**, *that is, S is linearly independent, but any proper superset of S is not linearly independent.*

Proof. We leave it to the reader to show that 1) and 2) are equivalent. Now suppose 1) holds. Then S is a spanning set. If some proper subset S' of S also spanned V then any vector in $S - S'$ would be a linear combination of the vectors in S', contradicting the fact that the vectors in S are linearly independent. Hence 1) implies 3).

Conversely, if S is a minimal spanning set then it must be linearly independent. For if not, some vector $s \in S$ would be a linear combination of the other vectors in S and so $S - \{s\}$ would be a proper spanning subset of S, which is not possible. Hence 3) implies 1).

Suppose again that 1) holds. If S were not maximal, there would be a vector $v \in V - S$ for which the set $S \cup \{v\}$ is linearly independent. But then v is not in the span of S, contradicting the fact that S is a spanning set. Hence, S is a maximal linearly independent set and so 1) implies 4).

Conversely, if S is a maximal linearly independent set then S must span V, for if not, we could find a vector $v \in V - S$ that is not a linear combination of the vectors in S. Hence, $S \cup \{v\}$ would be a linearly independent proper superset of S, which is a contradiction. Thus, 4) implies 1). □

Definition *A set of vectors in V that satisfies any (and hence all) of the equivalent conditions in Theorem 1.7 is called a* **basis** *for V.* □

Corollary 1.8 *A finite set $S = \{v_1, \dots, v_n\}$ of vectors in V is a basis for V if and only if*

$$V = \langle v_1 \rangle \oplus \cdots \oplus \langle v_n \rangle \qquad \square$$

Example 1.6 The ith **standard vector** in F^n is the vector e_i that has 0s in all coordinate positions except the ith, where it has a 1. Thus,

$$e_1 = (1, 0, \dots, 0), \quad e_2 = (0, 1, \dots, 0) \quad , \dots , \quad e_n = (0, \dots, 0, 1)$$

The set $\{e_1, \dots, e_n\}$ is called the **standard basis** for F^n. □

The proof that every nontrivial vector space has a basis is a classic example of the use of Zorn's lemma.

Theorem 1.9 *Let V be a nonzero vector space. Let I be a linearly independent set in V and let S be a spanning set in V containing I. Then there is a basis $\mathcal{B}$ for V for which $I \subseteq \mathcal{B} \subseteq S$. In particular,*

1) *Any vector space, except the zero space $\{0\}$, has a basis.*
2) *Any linearly independent set in V is contained in a basis.*
3) *Any spanning set in V contains a basis.*

Proof. Consider the collection $\mathcal{A}$ of all linearly independent subsets of V containing I and contained in S. This collection is not empty, since $I \in \mathcal{A}$. Now, if

$$\mathcal{C} = \{I_k \mid k \in K\}$$

is a chain in $\mathcal{A}$ then the union

$$U = \bigcup_{k \in K} I_i$$

is linearly independent and satisfies $I \subseteq U \subseteq S$, that is, $U \in \mathcal{A}$. Hence, every chain in $\mathcal{A}$ has an upper bound in $\mathcal{A}$ and according to Zorn's lemma, $\mathcal{A}$ must contain a maximal element $\mathcal{B}$, which is linearly independent.

Now, $\mathcal{B}$ is a basis for the vector space $\langle S \rangle = V$, for if any $s \in S$ is not a linear combination of the elements of $\mathcal{B}$ then $\mathcal{B} \cup \{s\} \subseteq S$ is linearly independent, contradicting the maximality of $\mathcal{B}$. Hence $S \subseteq \langle \mathcal{B} \rangle$ and so $V = \langle S \rangle \subseteq \langle \mathcal{B} \rangle$. □

The reader can now show, using Theorem 1.9, that any subspace of a vector space has a complement.

The Dimension of a Vector Space

The next result, with its classical elegant proof, says that if a vector space V has a *finite* spanning set S then the size of any linearly independent set cannot exceed the size of S.

Theorem 1.10 *Let V be a vector space and assume that the vectors $v_1, \ldots, v_n$ are linearly independent and the vectors $s_1, \ldots, s_m$ span V. Then $n \leq m$.*
Proof. First, we list the two sets of vectors: the spanning set followed by the linearly independent set

$$s_1, \ldots, s_m; v_1, \ldots, v_n$$

Then we move the first vector v_1 to the front of the first list

$$v_1, s_1, \ldots, s_m; v_2, \ldots, v_n$$

Since $s_1, \ldots, s_m$ span V, v_1 is a linear combination of the s_i's. This implies that we may remove one of the s_i's, which by reindexing if necessary can be s_1, from the first list and still have a spanning set

$$v_1, s_2, \ldots, s_m; v_2, \ldots, v_n$$

Note that the first set of vectors still spans V and the second set is still linearly independent.

Now we repeat the process, moving v_2 from the second list to the first list

$$v_1, v_2, s_2, \ldots, s_m; v_3, \ldots, v_n$$

As before, the vectors in the first list are linearly dependent, since they spanned V before the inclusion of v_2. However, since the v_i's are linearly independent, any nontrivial linear combination of the vectors in the first list that equals 0

must involve at least one of the s_i's. Hence, we may remove that vector, which again by reindexing if necessary may be taken to be s_2 and still have a spanning set

$$v_1, v_2, s_3, \ldots, s_m; v_3, \ldots, v_n$$

Once again, the first set of vectors spans V and the second set is still linearly independent.

Now, if $m < n$, then this process will eventually exhaust the s_i's and lead to the list

$$v_1, v_2, \ldots, v_m; v_{m+1}, \ldots, v_n$$

where $v_1, v_2, \ldots, v_m$ span V, which is clearly not possible since v_n is not in the span of $v_1, v_2, \ldots, v_m$. Hence, $n \le m$. □

Corollary 1.11 *If V has a* finite *spanning set then any two bases of V have the same size.* □

Now let us prove Corollary 1.11 for arbitrary vector spaces.

Theorem 1.12 *If V is a vector space then any two bases for V have the same cardinality.*

Proof. We may assume that all bases for V are infinite sets, for if any basis is finite then V has a finite spanning set and so Corollary 1.11 applies.

Let $\mathcal{B} = \{b_i \mid i \in I\}$ be a basis for V and let $\mathcal{C}$ be another basis for V. Then any vector $c \in \mathcal{C}$ can be written as a finite linear combination of the vectors in $\mathcal{B}$, where all of the coefficients are nonzero, say

$$c = \sum_{i \in U_c} r_i b_i$$

But because $\mathcal{C}$ is a basis, we must have

$$\bigcup_{c \in \mathcal{C}} U_c = I$$

for if the vectors in $\mathcal{C}$ can be expressed as finite linear combinations of the vectors in a *proper* subset $\mathcal{B}'$ of $\mathcal{B}$ then $\mathcal{B}'$ spans V, which is not the case.

Since $|U_c| < \aleph_0$ for all $c \in \mathcal{C}$, Theorem 0.16 implies that

$$|\mathcal{B}| = |I| \le \aleph_0 |\mathcal{C}| = |\mathcal{C}|$$

But we may also reverse the roles of $\mathcal{B}$ and $\mathcal{C}$, to conclude that $|\mathcal{B}| \le |\mathcal{C}|$ and so $|\mathcal{B}| = |\mathcal{C}|$ by the Schröder–Bernstein theorem. □

Theorem 1.12 allows us to make the following definition.

Definition *A vector space V is* **finite-dimensional** *if it is the zero space $\{0\}$, or if it has a finite basis. All other vector spaces are* **infinite-dimensional**. The **dimension** of the zero space is 0 and the **dimension** of any nonzero vector space V is the cardinality of any basis for V. If a vector space V has a basis of cardinality κ, we say that V is **κ-dimensional** and write $\dim(V) = \kappa$. □

It is easy to see that if S is a subspace of V then $\dim(S) \leq \dim(V)$. If in addition, $\dim(S) = \dim(V) < \infty$ then $S = V$.

Theorem 1.13 *Let V be a vector space.*

1) *If $\mathcal{B}$ is a basis for V and if $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ and $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$ then*
$$V = \langle \mathcal{B}_1 \rangle \oplus \langle \mathcal{B}_2 \rangle$$
2) *Let $V = S \oplus T$. If $\mathcal{B}_1$ is a basis for S and $\mathcal{B}_2$ is a basis for T then $\mathcal{B}_1 \cap \mathcal{B}_2 = \emptyset$ and $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ is a basis for V.* □

Theorem 1.14 *Let S and T be subspaces of a vector space V. Then*
$$\dim(S) + \dim(T) = \dim(S + T) + \dim(S \cap T)$$
In particular, if T is any complement of S in V then
$$\dim(S) + \dim(T) = \dim(V)$$
that is,
$$\dim(S \oplus T) = \dim(S) + \dim(T)$$

Proof. Suppose that $\mathcal{B} = \{b_i \mid i \in I\}$ is a basis for $S \cap T$. Extend this to a basis $\mathcal{A} \cup \mathcal{B}$ for S where $\mathcal{A} = \{a_j \mid j \in J\}$ is disjoint from $\mathcal{B}$. Also, extend $\mathcal{B}$ to a basis $\mathcal{B} \cup \mathcal{C}$ for T where $\mathcal{C} = \{c_k \mid k \in K\}$ is disjoint from $\mathcal{B}$. We claim that $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ is a basis for $S + T$. It is clear that $\langle \mathcal{A} \cup \mathcal{B} \cup \mathcal{C} \rangle = S + T$.

To see that $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ is linearly independent, suppose to the contrary that
$$\alpha_1 v_1 + \cdots + \alpha_n v_n = 0$$
where $v_i \in \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ and $\alpha_i \neq 0$ for all i. There must be vectors v_i in this expression from both $\mathcal{A}$ and $\mathcal{C}$, since $\mathcal{A} \cup \mathcal{B}$ and $\mathcal{B} \cup \mathcal{C}$ are linearly independent. Isolating the terms involving the vectors from $\mathcal{A}$ on one side of the equality shows that there is a nonzero vector in $x \in \langle \mathcal{A} \rangle \cap \langle \mathcal{B} \cup \mathcal{C} \rangle$. But then $x \in S \cap T$ and so $x \in \langle \mathcal{A} \rangle \cap \langle \mathcal{B} \rangle$, which implies that $x = 0$, a contradiction. Hence, $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$ is linearly independent and a basis for $S + T$.

Now,

$$\begin{aligned}\dim(S)+\dim(T) &= |\mathcal{A}\cup\mathcal{B}|+|\mathcal{B}\cup\mathcal{C}| \\ &= |\mathcal{A}|+|\mathcal{B}|+|\mathcal{B}|+|\mathcal{C}| \\ &= |\mathcal{A}|+|\mathcal{B}|+|\mathcal{C}|+\dim(S\cap T) \\ &= \dim(S+T)+\dim(S\cap T)\end{aligned}$$

as desired. □

It is worth emphasizing that while the equation

$$\dim(S)+\dim(T)=\dim(S+T)+\dim(S\cap T)$$

holds for all vector spaces, we cannot write

$$\dim(S+T)=\dim(S)+\dim(T)-\dim(S\cap T)$$

unless $S+T$ is finite-dimensional.

Ordered Bases and Coordinate Matrices

It will be convenient to consider bases that have an order imposed upon their members.

Definition *Let V be a vector space of dimension n. An* **ordered basis** *for V is an ordered n-tuple $(v_1,\dots,v_n)$ of vectors for which the set $\{v_1,\dots,v_n\}$ is a basis for V.* □

If $\mathcal{B}=(v_1,\dots,v_n)$ is an ordered basis for V then for each $v\in V$ there is a unique ordered n-tuple $(r_1,\dots,r_n)$ of scalars for which

$$v=r_1v_1+\cdots+r_nv_n$$

Accordingly, we can define the **coordinate map** $\phi_{\mathcal{B}}\colon V\to F^n$ by

$$\phi_{\mathcal{B}}(v)=[v]_{\mathcal{B}}=\begin{bmatrix} r_1\\ \vdots\\ r_n\end{bmatrix} \tag{1.3}$$

where the column matrix $[v]_{\mathcal{B}}$ is known as the **coordinate matrix** of v with respect to the ordered basis $\mathcal{B}$. Clearly, knowing $[v]_{\mathcal{B}}$ is equivalent to knowing v (assuming knowledge of $\mathcal{B}$).

Furthermore, it is easy to see that the coordinate map $\phi_{\mathcal{B}}$ is bijective and preserves the vector space operations, that is,

$$\phi_{\mathcal{B}}(r_1v_1+\cdots+r_nv_n)=r_1\phi_{\mathcal{B}}(v_1)+\cdots+r_n\phi_{\mathcal{B}}(v_n)$$

or equivalently

$$[r_1v_1+\cdots+r_nv_n]_{\mathcal{B}}=r_1[v_1]_{\mathcal{B}}+\cdots+r_n[v_n]_{\mathcal{B}}$$

Functions from one vector space to another that preserve the vector space operations are called *linear transformations* and form the objects of study in the next chapter.

The Row and Column Spaces of a Matrix

Let A be an $m\times n$ matrix over F. The rows of A span a subspace of F^n known as the **row space** of A and the columns of A span a subspace of F^m known as the **column space** of A. The dimensions of these spaces are called the **row rank** and **column rank**, respectively. We denote the row space and row rank by rs(A) and rrk(A) and the column space and column rank by cs(A) and crk(A).

It is a remarkable and useful fact that the row rank of a matrix is always equal to its column rank, despite the fact that if $m\neq n$, the row space and column space are not even in the same vector space!

Our proof of this fact hinges upon the following simple observation about matrices.

Lemma 1.15 *Let A be an $m\times n$ matrix. Then elementary column operations do not affect the row rank of A. Similarly, elementary row operations do not affect the column rank of A.*
Proof. The second statement follows from the first by taking transposes. As to the first, the row space of A is

$$\text{rs}(A)=\langle e_1A,\ldots,e_nA\rangle$$

where e_i are the standard basis vectors in F^m. Performing an elementary column operation on A is equivalent to multiplying A on the right by an elementary matrix E. Hence the row space of AE is

$$\text{rs}(AE)=\langle e_1AE,\ldots,e_nAE\rangle$$

and since E is invertible,

$$\text{rr}(A)=\dim(\text{rs}(A))=\dim(\text{rs}(AE))=\text{rr}(AE)$$

as desired. □

Theorem 1.16 *If $A\in\mathcal{M}_{m,n}$ then* rrk$(A)=$ crk(A). *This number is called the* **rank** *of A and is denoted by* rk(A).
Proof. According to the previous lemma, we may reduce A to reduced column echelon form without affecting the row rank. But this reduction does not affect the column rank either. Then we may further reduce A to reduced row echelon form without affecting either rank. The resulting matrix M has the same row and column ranks as A. But M is a matrix with 1's followed by 0's on the main

diagonal (entries $M_{1,1}, M_{2,2}, \dots$) and 0's elsewhere. Hence,

$$\text{rrk}(A) = \text{rrk}(M) = \text{crk}(M) = \text{crk}(A)$$

as desired. □

The Complexification of a Real Vector Space

If W is a complex vector space (that is, a vector space over $\mathbb{C}$), then we can think of W as a real vector space simply by restricting all scalars to the field $\mathbb{R}$. Let us denote this real vector space by $W_{\mathbb{R}}$ and call it the **real version** of W.

On the other hand, to each real vector space V, we can associate a complex vector space $V^{\mathbb{C}}$. This "complexification" process will play a useful role when we discuss the structure of linear operators on a real vector space. (Throughout our discussion V will denote a real vector space.)

Definition *If V is a real vector space then the set $V^{\mathbb{C}} = V \times V$ of ordered pairs, with componentwise addition*

$$(u, v) + (x, y) = (u + x, v + y)$$

and scalar multiplication over $\mathbb{C}$ defined by

$$(a + bi)(u, v) = (au - bv, av + bu)$$

for $a, b \in \mathbb{R}$ is a complex vector space, called the ***complexification*** *of V.* □

It is convenient to introduce a notation for vectors in $V^{\mathbb{C}}$ that resembles complex numbers. In particular, we denote $(u, v) \in V^{\mathbb{C}}$ by $u + vi$ and so

$$V^{\mathbb{C}} = \{u + vi \mid u, v \in V\}$$

Addition now looks like ordinary addition of complex numbers

$$(u + vi) + (x + yi) = (u + x) + (v + y)i$$

and scalar multiplication looks like ordinary multiplication of complex numbers

$$(a + bi)(u + vi) = (au - bv) + (av + bu)i$$

Thus, for example, we immediately have for $a, b \in \mathbb{R}$

$$\begin{aligned} a(u + vi) &= au + avi \\ bi(u + vi) &= -bv + bui \\ (a + bi)u &= au + bui \\ (a + bi)vi &= -bv + avi \end{aligned}$$

The **real part** of $z = u + vi$ is $u \in V$ and the **imaginary part** of z is $v \in V$. The essence of the fact that $z = u + vi \in V^{\mathbb{C}}$ is really an ordered pair is that z is 0 if and only if its real and imaginary parts are both 0.

We can define the **complexification map** cpx: $V \to V^{\mathbb{C}}$ by

$$\text{cpx}(v) = v + 0i$$

Let us refer to $v + 0i$ as the **complexification**, or **complex version** of $v \in V$. Note that this map is a group homomorphism, that is,

$$\text{cpx}(0) = 0 + 0i \quad \text{and} \quad \text{cpx}(u \pm v) = \text{cpx}(u) \pm \text{cpx}(v)$$

and it is injective

$$\text{cpx}(u) = \text{cpx}(v) \Leftrightarrow u = v$$

Also, it preserves multiplication by *real* scalars

$$\text{cpx}(au) = au + 0i = a(u + 0i) = a\text{cpx}(u)$$

for $a \in \mathbb{R}$. However, the complexification map is not surjective, since it gives only "real" vectors in $V^{\mathbb{C}}$.

The complexification map is an injective linear transformation from the real vector space V to the real version $(V^{\mathbb{C}})_{\mathbb{R}}$ of the complexification $V^{\mathbb{C}}$, that is, to the complex vector space $V^{\mathbb{C}}$ provided that scalars are restricted to real numbers. In this way, we see that $V^{\mathbb{C}}$ contains an embedded copy of V.

The Dimension of $V^{\mathbb{C}}$

The vector-space dimensions of V and $V^{\mathbb{C}}$ are the same. This should not necessarily come as a surprise because although $V^{\mathbb{C}}$ may seem "bigger" than V, the field of scalars is also "bigger."

Theorem 1.17 *If $\mathcal{B} = \{v_j \mid j \in I\}$ is a basis for V over $\mathbb{R}$ then the* **complexification** *of $\mathcal{B}$*

$$\text{cpx}(\mathcal{B}) = \{v_j + 0i \mid v_j \in \mathcal{B}\}$$

is a basis for the vector space $V^{\mathbb{C}}$ over $\mathbb{C}$. Hence,

$$\dim(V^{\mathbb{C}}) = \dim(V)$$

Proof. To see that $\text{cpx}(\mathcal{B})$ spans $V^{\mathbb{C}}$ over $\mathbb{C}$, let $x + iy \in V^{\mathbb{C}}$. Then $x, y \in V$ and so there exist real numbers a_i and b_i (some of which may be 0) for which

$$\begin{aligned} x+yi &= \sum_{j=1}^{J} a_j v_j + \left[\sum_{j=1}^{J} b_j v_j\right] i \\ &= \sum_{j=1}^{J} (a_j v_j + b_j v_j i) \\ &= \sum_{j=1}^{J} (a_j + b_j i)(v_j + 0i) \end{aligned}$$

To see that $\operatorname{cpx}(\mathcal{B})$ is linearly independent, if

$$\sum_{j=1}^{J} (a_j + b_j i)(v_j + 0i) = 0 + 0i$$

then the previous computations show that

$$\sum_{j=1}^{J} a_j v_j = 0 \text{ and } \sum_{j=1}^{J} b_j v_j = 0$$

The independence of $\mathcal{B}$ then implies that $a_i = 0$ and $b_i = 0$ for all i. □

If $v \in V$ and $\mathcal{B}$ is a basis for V then we may write

$$v = \sum_{i=1}^{n} a_i v_i$$

for $a_i \in \mathbb{R}$. Since the coefficients are real, we have

$$v + 0i = \sum_{i=1}^{n} a_i (v_i + 0i)$$

and so the coordinate matrices are equal

$$[v + 0i]_{\operatorname{cpx}(\mathcal{B})} = [v]_{\mathcal{B}}$$

Exercises

1. Let V be a vector space over F. Prove that $0v = 0$ and $r0 = 0$ for all $v \in V$ and $r \in F$. Describe the different 0's in these equations. Prove that if $rv = 0$ then $r = 0$ or $v = 0$. Prove that $rv = v$ implies that $v = 0$ or $r = 1$.
2. Prove Theorem 1.3.
3. a) Find an abelian group V and a field F for which V is a vector space over F in at least two different ways, that is, there are two different definitions of scalar multiplication making V a vector space over F.

b) Find a vector space V over F and a subset S of V that is (1) a subspace of V and (2) a vector space using operations that differ from those of V.

4. Suppose that V is a vector space with basis $\mathcal{B} = \{b_i \mid i \in I\}$ and S is a subspace of V. Let $\{B_1, \ldots, B_k\}$ be a partition of $\mathcal{B}$. Then is it true that

$$S = \bigoplus_{i=1}^{k} (S \cap \langle B_i \rangle)$$

What if $S \cap \langle B_i \rangle \neq \{0\}$ for all i?

5. Prove Corollary 1.8.

6. Let $S, T, U \in \mathcal{S}(V)$. Show that if $U \subseteq S$ then

$$S \cap (T + U) = (S \cap T) + U$$

This is called the **modular law** for the lattice $\mathcal{S}(V)$.

7. For what vector spaces does the distributive law of subspaces

$$S \cap (T + U) = (S \cap T) + (S \cap U)$$

hold?

8. A vector $v = (a_1, \ldots, a_n) \in \mathbb{R}^n$ is called **strongly positive** if $a_i > 0$ for all $i = 1, \ldots, n$.

a) Suppose that v is strongly positive. Show that any vector that is "close enough" to v is also strongly positive. (Formulate carefully what "close enough" should mean.)

b) Prove that if a subspace S of $\mathbb{R}^n$ contains a strongly positive vector, then S has a basis of strongly positive vectors.

9. Let M be an $m \times n$ matrix whose rows are linearly independent. Suppose that the k columns $c_{i_1}, \ldots, c_{i_k}$ of M span the column space of M. Let C be the matrix obtained from M by deleting all columns except $c_{i_1}, \ldots, c_{i_k}$. Show that the rows of C are also linearly independent.

10. Prove that the first two statements in Theorem 1.7 are equivalent.

11. Show that if S is a subspace of a vector space V then $\dim(S) \leq \dim(V)$. Furthermore, if $\dim(S) = \dim(V) < \infty$ then $S = V$. Give an example to show that the finiteness is required in the second statement.

12. Let $\dim(V) < \infty$ and suppose that $V = U \oplus S_1 = U \oplus S_2$. What can you say about the relationship between S_1 and S_2? What can you say if $S_1 \subseteq S_2$?

13. What is the relationship between $S \oplus T$ and $T \oplus S$? Is the direct sum operation commutative? Formulate and prove a similar statement concerning associativity. Is there an "identity" for direct sum? What about "negatives"?

14. Let V be a finite-dimensional vector space over an infinite field F. Prove that if $S_1, \ldots, S_k$ are subspaces of V of equal dimension then there is a subspace T of V for which $V = S_i \oplus T$ for all $i = 1, \ldots, k$. In other words, T is a common complement of the subspaces S_i.

15. Prove that the vector space $\mathcal{C}$ of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ is infinite-dimensional.
16. Show that Theorem 1.2 need not hold if the base field F is finite.
17. Let S be a subspace of V. The set $v+S=\{v+s \mid s\in S\}$ is called an **affine subspace** of V.
    a) Under what conditions is an affine subspace of V a subspace of V?
    b) Show that any two affine subspaces of the form $v+S$ and $w+S$ are either equal or disjoint.
18. If V and W are vector spaces over F for which $|V|=|W|$ then does it follow that $\dim(V)=\dim(W)$?
19. Let V be an n-dimensional real vector space and suppose that S is a subspace of V with $\dim(S)=n-1$. Define an equivalence relation $\equiv$ on the set $V\setminus S$ by $v\equiv w$ if the "line segment"

$$L(v,w)=\{rv+(1-r)w \mid 0\le r\le 1\}$$

    has the property that $L(v,w)\cap S=\emptyset$. Prove that $\equiv$ is an equivalence relation and that it has exactly two equivalence classes.
20. Let F be a field. A **subfield** of F is a subset K of F that is a field in its own right using the same operations as defined on F.
    a) Show that F is a vector space over any subfield K of F.
    b) Suppose that F is an m-dimensional vector space over a subfield K of F. If V is an n-dimensional vector space over F, show that V is also a vector space over K. What is the dimension of V as a vector space over K?
21. Let F be a finite field of size q and let V be an n-dimensional vector space over F. The purpose of this exercise is to show that the number of subspaces of V of dimension k is

$$\binom{n}{k}_q=\frac{(q^n-1)\cdots(q-1)}{(q^k-1)\cdots(q-1)(q^{n-k}-1)\cdots(q-1)}$$

    The expressions $\binom{n}{k}_q$ are called **Gaussian coefficients** and have properties similar to those of the binomial coefficients. Let $S(n,k)$ be the number of k-dimensional subspaces of V.
    a) Let $N(n,k)$ be the number of k-tuples of linearly independent vectors $(v_1,\ldots,v_k)$ in V. Show that

$$N(n,k)=(q^n-1)(q^n-q)\cdots(q^n-q^{k-1})$$

    b) Now, each of the k-tuples in a) can be obtained by first choosing a subspace of V of dimension k and then selecting the vectors from this subspace. Show that for any k-dimensional subspace of V, the number of k-tuples of independent vectors in this subspace is

$$(q^k-1)(q^k-q)\cdots(q^k-q^{k-1})$$

c) Show that

$$N(n,k) = S(n,k)(q^k - 1)(q^k - q)\cdots(q^k - q^{k-1})$$

How does this complete the proof?

22. Prove that any subspace S of $\mathbb{R}^n$ is a closed set or, equivalently, that its set complement $S^c = \mathbb{R}^n \setminus S$ is open, that is, for any $x \in S^c$ there is an open ball $B(s,\epsilon)$ centered at x with radius $\epsilon > 0$ for which $B(x,\epsilon) \subseteq S^c$.
23. Let $\mathcal{B} = \{b_1, \ldots, b_n\}$ and $\mathcal{C} = \{c_1, \ldots, c_n\}$ be bases for a vector space V. Let $1 \le m \le n-1$. Show that there is a permutation σ of $\{1, \ldots, n\}$ such that

$$b_1, \ldots, b_m, c_{\sigma(m+1)}, \ldots, c_{\sigma(n)}$$

and

$$c_{\sigma(1)}, \ldots, c_{\sigma(m)}, b_{m+1}, \ldots, b_n$$

are both bases for V.
24. Let $\dim(V) = n$ and suppose that $S_1, \ldots, S_k$ are subspaces of V with $\dim(S_i) \le m < n$. Prove that there is a subspace T of V of dimension $n-m$ for which $T \cap S_i = \{0\}$ for all i.
25. What is the dimension of the complexification $V^{\mathbb{C}}$ thought of as a real vector space?
26. (When is a subspace of a complex vector space a complexification?) Let V be a real vector space with complexification $V^{\mathbb{C}}$ and let U be a subspace of $V^{\mathbb{C}}$. Prove that there is a subspace S of V for which

$$U = S^{\mathbb{C}} = \{s + ti \mid s, t \in S\}$$

if and only if U is closed under complex conjugation $\chi\colon V^{\mathbb{C}} \to V^{\mathbb{C}}$ defined by $\chi(u + iv) = u - iv$.

Chapter 2
Linear Transformations

Linear Transformations

Loosely speaking, a linear transformation is a function from one vector space to another that *preserves* the vector space operations. Let us be more precise.

Definition *Let V and W be vector spaces over a field F. A function $\tau: V \to W$ is a* **linear transformation** *if*

$$\tau(ru+sv)=r\tau(u)+s\tau(v)$$

for all scalars $r,s\in F$ and vectors $u,v\in V$. A linear transformation $\tau: V \to V$ is called a **linear operator** *on V. The set of all linear transformations from V to W is denoted by $\mathcal{L}(V,W)$ and the set of all linear operators on V is denoted by $\mathcal{L}(V)$.* □

We should mention that some authors use the term linear operator for any linear transformation from V to W.

Definition *The following terms are also employed:*

1) **homomorphism** *for linear transformation*
2) **endomorphism** *for linear operator*
3) **monomorphism** *(or* **embedding***) for injective linear transformation*
4) **epimorphism** *for surjective linear transformation*
5) **isomorphism** *for bijective linear transformation.*
6) **automorphism** *for bijective linear operator.* □

Example 2.1

1) The derivative $D: V \to V$ is a linear operator on the vector space V of all infinitely differentiable functions on $\mathbb{R}$.

2) The integral operator $\tau: F[x] \to F[x]$ defined by

$$\tau(f) = \int_0^x f(t)dt$$

is a linear operator on $F[x]$.

3) Let A be an $m \times n$ matrix over F. The function $\tau_A: F^n \to F^m$ defined by $\tau_A(v) = Av$, where all vectors are written as column vectors, is a linear transformation from F^n to F^m. This function is just multiplication by A.
4) The coordinate map $\phi: V \to F^n$ of an n-dimensional vector space is a linear transformation from V to F^n. □

The set $\mathcal{L}(V,W)$ is a vector space in its own right and $\mathcal{L}(V)$ has the structure of an algebra, as defined in Chapter 0.

Theorem 2.1

1) *The set $\mathcal{L}(V,W)$ is a vector space under ordinary addition of functions and scalar multiplication of functions by elements of F.*
2) *If $\sigma \in \mathcal{L}(U,V)$ and $\tau \in \mathcal{L}(V,W)$ then the composition $\tau\sigma$ is in $\mathcal{L}(U,W)$.*
3) *If $\tau \in \mathcal{L}(V,W)$ is bijective then $\tau^{-1} \in \mathcal{L}(W,V)$.*
4) *The vector space $\mathcal{L}(V)$ is an algebra, where multiplication is composition of functions. The identity map $\iota \in \mathcal{L}(V)$ is the multiplicative identity and the zero map $0 \in \mathcal{L}(V)$ is the additive identity.*

Proof. We prove only part 3). Let $\tau: V \to W$ be a bijective linear transformation. Then $\tau^{-1}: W \to V$ is a well-defined function and since any two vectors w_1 and w_2 in W have the form $w_1 = \tau(v_1)$ and $w_2 = \tau(v_2)$, we have

$$\begin{aligned}\tau^{-1}(aw_1 + bw_2) &= \tau^{-1}(a\tau(v_1) + b\tau(v_2)) \\ &= \tau^{-1}(\tau(av_1 + bv_2)) \\ &= av_1 + bv_2 \\ &= a\tau^{-1}(w_1) + b\tau^{-1}(w_2)\end{aligned}$$

which shows that τ^{-1} is linear. □

One of the easiest ways to define a linear transformation is to give its values on a basis. The following theorem says that we may assign these values arbitrarily and obtain a unique linear transformation by linear extension to the entire domain.

Theorem 2.2 *Let V and W be vector spaces and let $\mathcal{B} = \{v_i \mid i \in I\}$ be a basis for V. Then we can define a linear transformation $\tau \in \mathcal{L}(V,W)$ by specifying the values of $\tau(v_i) \in W$* arbitrarily *for all $v_i \in \mathcal{B}$ and extending the domain of τ to V using linearity, that is,*

$$\tau(a_1v_1 + \cdots + a_nv_n) = a_1\tau(v_1) + \cdots + a_n\tau(v_n)$$

This process uniquely *defines a linear transformation, that is, if $\tau, \sigma \in \mathcal{L}(V, W)$ satisfy $\tau(v_i) = \sigma(v_i)$ for all $v_i \in \mathcal{B}$ then $\tau = \sigma$.*
Proof. The crucial point is that the extension by linearity is well-defined, since each vector in V has a unique representation as a linear combination of a finite number of vectors in $\mathcal{B}$. We leave the details to the reader. □

Note that if $\tau \in \mathcal{L}(V, W)$ and if S is a subspace of V, then the restriction $\tau|_S$ of τ to S is a linear transformation from S to W.

The Kernel and Image of a Linear Transformation

There are two very important vector spaces associated with a linear transformation τ from V to W.

Definition *Let $\tau \in \mathcal{L}(V, W)$. The subspace*

$$\ker(\tau) = \{v \in V \mid \tau(v) = 0\}$$

is called the **kernel** *of τ and the subspace*

$$\operatorname{im}(\tau) = \{\tau(v) \mid v \in V\}$$

is called the **image** *of τ. The dimension of* $\ker(\tau)$ *is called the* **nullity** *of τ and is denoted by* $\operatorname{null}(\tau)$*. The dimension of* $\operatorname{im}(\tau)$ *is called the* **rank** *of τ and is denoted by* $\operatorname{rk}(\tau)$. □

It is routine to show that $\ker(\tau)$ is a subspace of V and $\operatorname{im}(\tau)$ is a subspace of W. Moreover, we have the following.

Theorem 2.3 *Let $\tau \in \mathcal{L}(V, W)$. Then*
1) τ is surjective if and only if $\operatorname{im}(\tau) = W$
2) τ is injective if and only if $\ker(\tau) = \{0\}$
Proof. The first statement is merely a restatement of the definition of surjectivity. To see the validity of the second statement, observe that

$$\tau(u) = \tau(v) \Leftrightarrow \tau(u - v) = 0 \Leftrightarrow u - v \in \ker(\tau)$$

Hence, if $\ker(\tau) = \{0\}$ then $\tau(u) = \tau(v) \Leftrightarrow u = v$, which shows that τ is injective. Conversely, if τ is injective and $u \in \ker(\tau)$ then $\tau(u) = \tau(0)$ and so $u = 0$. This shows that $\ker(\tau) = \{0\}$. □

Isomorphisms

Definition *A bijective linear transformation $\tau: V \to W$ is called an* **isomorphism** *from V to W. When an isomorphism from V to W exists, we say that V and W are* **isomorphic** *and write $V \approx W$.* □

Example 2.2 Let $\dim(V) = n$. For any ordered basis $\mathcal{B}$ of V, the coordinate map $\phi_{\mathcal{B}}: V \to F^n$ that sends each vector $v \in V$ to its coordinate matrix

$[v]_{\mathcal{B}} \in F^n$ is an isomorphism. Hence, any n-dimensional vector space over F is isomorphic to F^n. □

Isomorphic vector spaces share many properties, as the next theorem shows. If $\tau \in \mathcal{L}(V,W)$ and $S \subseteq V$ we write

$$\tau(S) = \{\tau(s) \mid s \in S\}$$

Theorem 2.4 *Let $\tau \in \mathcal{L}(V,W)$ be an isomorphism. Let $S \subseteq V$. Then*

1) S spans V if and only if $\tau(S)$ spans W.

2) S is linearly independent in V if and only if $\tau(S)$ is linearly independent in W.

3) S is a basis for V if and only if $\tau(S)$ is a basis for W. □

An isomorphism can be characterized as a linear transformation $\tau: V \to W$ that maps a basis for V to a basis for W.

Theorem 2.5 *A linear transformation $\tau \in \mathcal{L}(V,W)$ is an isomorphism if and only if there is a basis $\mathcal{B}$ of V for which $\tau(\mathcal{B})$ is a basis of W. In this case, τ maps any basis of V to a basis of W.* □

The following theorem says that, up to isomorphism, there is only one vector space of any given dimension.

Theorem 2.6 *Let V and W be vector spaces over F. Then $V \approx W$ if and only if* $\dim(V) = \dim(W)$. □

In Example 2.2, we saw that any n-dimensional vector space is isomorphic to F^n. Now suppose that B is a set of cardinality κ and let $(F^B)_0$ be the vector space of all functions from B to F with finite support. We leave it to the reader to show that the functions $\delta_b \in (F^B)_0$ defined for all $b \in B$, by

$$\delta_b(x) = \begin{cases} 1 & \text{if } x = b \\ 0 & \text{if } x \neq b \end{cases}$$

form a basis for $(F^B)_0$, called the **standard basis**. Hence, $\dim((F^B)_0) = |B|$.

It follows that for any cardinal number κ, there is a vector space of dimension κ. Also, any vector space of dimension κ is isomorphic to $(F^B)_0$.

Theorem 2.7 *If n is a natural number then any n-dimensional vector space over F is isomorphic to F^n. If κ is any cardinal number and if B is a set of cardinality κ then any κ-dimensional vector space over F is isomorphic to the vector space $(F^B)_0$ of all functions from B to F with finite support.* □

The Rank Plus Nullity Theorem

Let $\tau \in \mathcal{L}(V,W)$. Since any subspace of V has a complement, we can write

$$V = \ker(\tau) \oplus \ker(\tau)^c$$

where $\ker(\tau)^c$ is a complement of $\ker(\tau)$ in V. It follows that

$$\dim(V) = \dim(\ker(\tau)) + \dim(\ker(\tau)^c)$$

Now, the restriction of τ to $\ker(\tau)^c$

$$\tau^c\colon \ker(\tau)^c \to W$$

is injective, since

$$\ker(\tau^c) = \ker(\tau) \cap \ker(\tau)^c = \{0\}$$

Also, $\operatorname{im}(\tau^c) \subseteq \operatorname{im}(\tau)$. For the reverse inclusion, if $\tau(v) \in \operatorname{im}(\tau)$ then since $v = u + w$ for $u \in \ker(\tau)$ and $w \in \ker(\tau)^c$, we have

$$\tau(v) = \tau(u) + \tau(w) = \tau(w) = \tau^c(w) \in \operatorname{im}(\tau^c)$$

Thus $\operatorname{im}(\tau^c) = \operatorname{im}(\tau)$. It follows that

$$\ker(\tau)^c \approx \operatorname{im}(\tau)$$

From this, we deduce the following theorem.

Theorem 2.8 *Let* $\tau \in \mathcal{L}(V,W)$.

1) *Any complement of* $\ker(\tau)$ *is isomorphic to* $\operatorname{im}(\tau)$
2) **(The rank plus nullity theorem)**

$$\dim(\ker(\tau)) + \dim(\operatorname{im}(\tau)) = \dim(V)$$

or, in other notation,

$$\operatorname{rk}(\tau) + \operatorname{null}(\tau) = \dim(V) \qquad \square$$

Theorem 2.8 has an important corollary.

Corollary 2.9 *Let* $\tau \in \mathcal{L}(V,W)$, *where* $\dim(V) = \dim(W) < \infty$. *Then* τ *is injective if and only if it is surjective.* $\square$

Note that this result fails if the vector spaces are not finite-dimensional.

Linear Transformations from F^n to F^m

Recall that for any $m \times n$ matrix A over F the multiplication map

$$\tau_A(v) = Av$$

is a linear transformation. In fact, any linear transformation $\tau \in \mathcal{L}(F^n, F^m)$ has this form, that is, τ is just multiplication by a matrix, for we have

$$\big(\tau(e_1) \mid \cdots \mid \tau(e_n)\big)e_i = \big(\tau(e_1) \mid \cdots \mid \tau(e_n)\big)^{(i)} = \tau(e_i)$$

and so $\tau = \tau_A$ where

$$A = \big(\tau(e_1) \mid \cdots \mid \tau(e_n)\big)$$

Theorem 2.10
1) If A is an $m \times n$ matrix over F then $\tau_A \in \mathcal{L}(F^n, F^m)$.
2) If $\tau \in \mathcal{L}(F^n, F^m)$ then $\tau = \tau_A$ where

$$A = (\tau(e_1) \mid \cdots \mid \tau(e_n))$$

The matrix A is called the **matrix** *of τ.* □

Example 2.3 Consider the linear transformation $\tau: F^3 \to F^3$ defined by

$$\tau(x, y, z) = (x - 2y, z, x + y + z)$$

Then we have, in column form

$$\tau\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x-2y \\ z \\ x+y+z \end{bmatrix} = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

and so the standard matrix of τ is

$$A = \begin{bmatrix} 1 & -2 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$ □

If $A \in \mathcal{M}_{m,n}$ then since the image of τ_A is the column space of A, we have

$$\dim(\ker(\tau_A)) + \mathrm{rk}(A) = \dim(F^n)$$

This gives the following useful result.

Theorem 2.11 *Let A be an $m \times n$ matrix over F.*
1) $\tau_A: F^n \to F^m$ is injective if and only if $\mathrm{rk}(A) = n$.
2) $\tau_A: F^n \to F^m$ is surjective if and only if $\mathrm{rk}(A) = m$. □

Change of Basis Matrices

Suppose that $\mathcal{B} = (b_1, \ldots, b_n)$ and $\mathcal{C} = (c_1, \ldots, c_n)$ are ordered bases for a vector space V. It is natural to ask how the coordinate matrices $[v]_{\mathcal{B}}$ and $[v]_{\mathcal{C}}$ are related. The map that takes $[v]_{\mathcal{B}}$ to $[v]_{\mathcal{C}}$ is $\phi_{\mathcal{B},\mathcal{C}} = \phi_{\mathcal{C}}\phi_{\mathcal{B}}^{-1}$ and is called the **change of basis operator** (or **change of coordinates operator**). Since $\phi_{\mathcal{B},\mathcal{C}}$ is an operator on F^n, it has the form τ_A where

$$\begin{aligned}A &= (\phi_{\mathcal{B},\mathcal{C}}(e_1),\ldots,\phi_{\mathcal{B},\mathcal{C}}(e_n))\\ &= (\phi_{\mathcal{C}}\phi_{\mathcal{B}}^{-1}([b_1]_{\mathcal{B}}),\ldots,\phi_{\mathcal{C}}\phi_{\mathcal{B}}^{-1}([b_n]_{\mathcal{B}}))\\ &= ([b_1]_{\mathcal{C}},\ldots,[b_n]_{\mathcal{C}}))\end{aligned}$$

We denote A by $M_{\mathcal{B},\mathcal{C}}$ and call it the **change of basis matrix** from $\mathcal{B}$ to $\mathcal{C}$.

Theorem 2.12 *Let $\mathcal{B}=(b_1,\ldots,b_n)$ and $\mathcal{C}$ be ordered bases for a vector space V. Then the change of basis operator $\phi_{\mathcal{B},\mathcal{C}}=\phi_{\mathcal{C}}\phi_{\mathcal{B}}^{-1}$ is an automorphism of F^n, whose standard matrix is*

$$M_{\mathcal{B},\mathcal{C}}=([b_1]_{\mathcal{C}},\ldots,[b_n]_{\mathcal{C}}))$$

Hence

$$[v]_{\mathcal{C}}=M_{\mathcal{B},\mathcal{C}}[v]_{\mathcal{B}}$$

and $M_{\mathcal{C},\mathcal{B}}=M_{\mathcal{B},\mathcal{C}}^{-1}$. □

Consider the equation

$$A=M_{\mathcal{B},\mathcal{C}}$$

or equivalently,

$$A=([b_1]_{\mathcal{C}},\ldots,[b_n]_{\mathcal{C}}))$$

Then given any two of A (an invertible $n\times n$ matrix), $\mathcal{B}$ (an ordered basis for F^n) and $\mathcal{C}$ (an order basis for F^n), the third component is uniquely determined by this equation. This is clear if $\mathcal{B}$ and $\mathcal{C}$ are given or if A and $\mathcal{C}$ are given. If A and $\mathcal{B}$ are given then there is a unique $\mathcal{C}$ for which $A^{-1}=M_{\mathcal{C},\mathcal{B}}$ and so there is a unique $\mathcal{C}$ for which $A=M_{\mathcal{B},\mathcal{C}}$.

Theorem 2.13 *If we are given any two of the following:*
1) An invertible $n\times n$ matrix A.
2) An ordered basis $\mathcal{B}$ for F^n.
3) An ordered basis $\mathcal{C}$ for F^n.
then the third is uniquely determined by the equation

$$A=M_{\mathcal{B},\mathcal{C}}$$

□

The Matrix of a Linear Transformation

Let $\tau\colon V\to W$ be a linear transformation, where $\dim(V)=n$ and $\dim(W)=m$ and let $\mathcal{B}=(b_1,\ldots,b_n)$ be an ordered basis for V and $\mathcal{C}$ an ordered basis for W. Then the map

$$\theta\colon [v]_{\mathcal{B}}\to[\tau(v)]_{\mathcal{C}}$$

is a *representation* of τ as a linear transformation from F^n to F^m, in the sense that knowing θ (along with $\mathcal{B}$ and $\mathcal{C}$, of course) is equivalent to knowing τ. Of course, this representation depends on the choice of ordered bases $\mathcal{B}$ and $\mathcal{C}$.

Since θ is a linear transformation from F^n to F^m, it is just multiplication by an $m \times n$ matrix A, that is

$$[\tau(v)]_{\mathcal{C}} = A[v]_{\mathcal{B}}$$

Indeed, since $[b_i]_{\mathcal{B}} = e_i$, we get the columns of A as follows:

$$A^{(i)} = Ae_i = A[v]_{\mathcal{B}} = [\tau(b_i)]_{\mathcal{C}}$$

Theorem 2.14 *Let $\tau \in \mathcal{L}(V,W)$ and let $\mathcal{B} = (b_1, \ldots, b_n)$ and $\mathcal{C}$ be ordered bases for V and W, respectively. Then τ can be represented with respect to $\mathcal{B}$ and $\mathcal{C}$ as matrix multiplication, that is*

$$[\tau(v)]_{\mathcal{C}} = [\tau]_{\mathcal{B},\mathcal{C}}[v]_{\mathcal{B}}$$

where

$$[\tau]_{\mathcal{B},\mathcal{C}} = ([\tau(b_1)]_{\mathcal{C}} \mid \cdots \mid [\tau(b_n)]_{\mathcal{C}})$$

is called the **matrix of τ with respect to the bases** *$\mathcal{B}$ and $\mathcal{C}$. When $V = W$ and $\mathcal{B} = \mathcal{C}$, we denote $[\tau]_{\mathcal{B},\mathcal{B}}$ by $[\tau]_{\mathcal{B}}$ and so*

$$[\tau(v)]_{\mathcal{B}} = [\tau]_{\mathcal{B}}[v]_{\mathcal{B}}$$ □

Example 2.4 Let $D\colon \mathcal{P}_2 \to \mathcal{P}_2$ be the derivative operator, defined on the vector space of all polynomials of degree at most 2. Let $\mathcal{B} = \mathcal{C} = (1, x, x^2)$. Then

$$[D(1)]_{\mathcal{C}} = [0]_{\mathcal{C}} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, [D(x)]_{\mathcal{C}} = [1]_{\mathcal{C}} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, [D(x^2)]_{\mathcal{C}} = [2x]_{\mathcal{C}} = \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

and so

$$[D]_{\mathcal{B}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

Hence, for example, if $p(x) = 5 + x + 2x^2$ then

$$[Dp(x)]_{\mathcal{C}} = [D]_{\mathcal{B}}\,[p(x)]_{\mathcal{B}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \\ 0 \end{bmatrix}$$

and so $Dp(x) = 1 + 4x$. □

The following result shows that we may work equally well with linear transformations or with the matrices that represent them (with respect to fixed ordered bases $\mathcal{B}$ and $\mathcal{C}$). This applies not only to addition and scalar multiplication, but also to matrix multiplication.

Theorem 2.15 *Let V and W be vector spaces over F, with ordered bases $\mathcal{B} = (b_1, \ldots, b_n)$ and $\mathcal{C} = (c_1, \ldots, c_m)$, respectively.*

1) *The map $\mu\colon \mathcal{L}(V,W) \to \mathcal{M}_{m,n}(F)$ defined by*

$$\mu(\tau) = [\tau]_{\mathcal{B},\mathcal{C}}$$

is an isomorphism and so $\mathcal{L}(V,W) \approx \mathcal{M}_{m,n}(F)$.

2) *If $\sigma \in \mathcal{L}(U,V)$ and $\tau \in \mathcal{L}(V,W)$ and if $\mathcal{B}$, $\mathcal{C}$ and $\mathcal{D}$ are ordered bases for U, V and W, respectively then*

$$[\tau\sigma]_{\mathcal{B},\mathcal{D}} = [\tau]_{\mathcal{C},\mathcal{D}}[\sigma]_{\mathcal{B},\mathcal{C}}$$

Thus, the matrix of the product (composition) $\tau\sigma$ is the product of the matrices of τ and σ. In fact, this is the primary motivation for the definition of matrix multiplication.

Proof. To see that μ is linear, observe that for all i

$$\begin{aligned}[s\sigma + t\tau]_{\mathcal{B},\mathcal{C}}[b_i]_{\mathcal{B}} &= [(s\sigma + t\tau)(b_i)]_{\mathcal{C}} \\ &= [s\sigma(b_i) + t\tau(b_i)]_{\mathcal{C}} \\ &= s[\sigma(b_i)]_{\mathcal{C}} + t[\tau(b_i)]_{\mathcal{C}} \\ &= s[\sigma]_{\mathcal{B},\mathcal{C}}[b_i]_{\mathcal{B}} + t[\tau]_{\mathcal{B},\mathcal{C}}[b_i]_{\mathcal{B}} \\ &= (s[\sigma]_{\mathcal{B},\mathcal{C}} + t[\tau]_{\mathcal{B},\mathcal{C}})[b_i]_{\mathcal{B}}\end{aligned}$$

and since $[b_i]_{\mathcal{B}} = e_i$ is a standard basis vector, we conclude that

$$[s\sigma + t\tau]_{\mathcal{B},\mathcal{C}} = s[\sigma]_{\mathcal{B},\mathcal{C}} + t[\tau]_{\mathcal{B},\mathcal{C}}$$

and so μ is linear. If $A \in \mathcal{M}_{m,n}$, we define τ by the condition $[\tau(b_i)]_{\mathcal{C}} = A^{(i)}$, whence $\mu(\tau) = A$ and μ is surjective. Since $\dim(\mathcal{L}(V,W)) = \dim(\mathcal{M}_{m,n}(F))$, the map μ is an isomorphism. To prove part 2), we have

$$[\tau\sigma]_{\mathcal{B},\mathcal{D}}[v]_{\mathcal{B}} = [\tau(\sigma(v))]_{\mathcal{D}} = [\tau]_{\mathcal{C},\mathcal{D}}[\sigma(v)]_{\mathcal{C}} = [\tau]_{\mathcal{C},\mathcal{D}}[\sigma]_{\mathcal{B},\mathcal{C}}[v]_{\mathcal{B}} \qquad \square$$

Change of Bases for Linear Transformations

Since the matrix $[\tau]_{\mathcal{B},\mathcal{C}}$ that represents τ depends on the ordered bases $\mathcal{B}$ and $\mathcal{C}$, it is natural to wonder how to choose these bases in order to make this matrix as simple as possible. For instance, can we always choose the bases so that τ is represented by a diagonal matrix?

As we will see in Chapter 7, the answer to this question is no. In that chapter, we will take up the general question of how best to represent a linear operator by a matrix. For now, let us take the first step and describe the relationship between the matrices $[\tau]_{\mathcal{B},\mathcal{C}}$ and $[\tau]_{\mathcal{B}',\mathcal{C}'}$ of τ with respect to two different pairs $(\mathcal{B},\mathcal{C})$ and $(\mathcal{B}',\mathcal{C}')$ of ordered bases. Multiplication by $[\tau]_{\mathcal{B}',\mathcal{C}'}$ sends $[v]_{\mathcal{B}'}$ to

$[\tau(v)]_{\mathcal{C}'}$. This can be reproduced by first switching from $\mathcal{B}'$ to $\mathcal{B}$, then applying $[\tau]_{\mathcal{B},\mathcal{C}}$ and finally switching from $\mathcal{C}$ to $\mathcal{C}'$, that is,

$$[\tau]_{\mathcal{B}',\mathcal{C}'} = M_{\mathcal{C},\mathcal{C}'}[\tau]_{\mathcal{B},\mathcal{C}}M_{\mathcal{B}',\mathcal{B}} = M_{\mathcal{C},\mathcal{C}'}[\tau]_{\mathcal{B},\mathcal{C}}M_{\mathcal{B},\mathcal{B}'}^{-1}$$

Theorem 2.16 *Let $\tau \in \mathcal{L}(V,W)$ and let $(\mathcal{B},\mathcal{C})$ and $(\mathcal{B}',\mathcal{C}')$ be pairs of ordered bases of V and W, respectively. Then*

$$[\tau]_{\mathcal{B}',\mathcal{C}'} = M_{\mathcal{C},\mathcal{C}'}[\tau]_{\mathcal{B},\mathcal{C}}M_{\mathcal{B}',\mathcal{B}} \tag{2.1}$$ □

When $\tau \in \mathcal{L}(V)$ is a linear operator on V, it is generally more convenient to represent τ by matrices of the form $[\tau]_{\mathcal{B}}$, where the ordered bases used to represent vectors in the domain and image are the same. When $\mathcal{B} = \mathcal{C}$, Theorem 2.16 takes the following important form.

Corollary 2.17 *Let $\tau \in \mathcal{L}(V)$ and let $\mathcal{B}$ and $\mathcal{C}$ be ordered bases for V. Then the matrix of τ with respect to $\mathcal{C}$ can be expressed in terms of the matrix of τ with respect to $\mathcal{B}$ as follows*

$$[\tau]_{\mathcal{C}} = M_{\mathcal{B},\mathcal{C}}[\tau]_{\mathcal{B}}M_{\mathcal{B},\mathcal{C}}^{-1} \tag{2.2}$$ □

Equivalence of Matrices

Since the change of basis matrices are precisely the invertible matrices, (2.1) has the form

$$[\tau]_{\mathcal{B}',\mathcal{C}'} = P[\tau]_{\mathcal{B},\mathcal{C}}Q^{-1}$$

where P and Q are invertible matrices. This motivates the following definition.

Definition *Two matrices A and B are* **equivalent** *if there exist invertible matrices P and Q for which*

$$B = PAQ^{-1}$$ □

We remarked in Chapter 0 that B is equivalent to A if and only if B can be obtained from A by a series of elementary row and column operations. Performing the row operations is equivalent to multiplying the matrix A on the left by P and performing the column operations is equivalent to multiplying A on the right by Q^{-1}.

In terms of (2.1), we see that performing row operations (premultiplying by P) is equivalent to changing the basis used to represent vectors in the image and performing column operations (postmultiplying by Q^{-1}) is equivalent to changing the basis used to represent vectors in the domain.

According to Theorem 2.16, if A and B are matrices that represent τ with respect to possibly different ordered bases then A and B are equivalent. The converse of this also holds.

Theorem 2.18 *Let V and W be vector spaces with $\dim(V) = n$ and $\dim(W) = m$. Then two $m \times n$ matrices A and B are equivalent if and only if they represent the same linear transformation $\tau \in \mathcal{L}(V, W)$, but possibly with respect to different ordered bases. In this case, A and B represent exactly the same set of linear transformations in $\mathcal{L}(V, W)$.*

Proof. If A and B represent τ, that is, if

$$A = [\tau]_{\mathcal{B},\mathcal{C}} \text{ and } B = [\tau]_{\mathcal{B}',\mathcal{C}'}$$

for ordered bases $\mathcal{B}, \mathcal{C}, \mathcal{B}'$ and $\mathcal{C}'$ then Theorem 2.16 shows that A and B are equivalent. Now suppose that A and B are equivalent, say

$$B = PAQ^{-1}$$

where P and Q are invertible. Suppose also that A represents a linear transformation $\tau \in \mathcal{L}(V, W)$ for some ordered bases $\mathcal{B}$ and $\mathcal{C}$, that is,

$$A = [\tau]_{\mathcal{B},\mathcal{C}}$$

Theorem 2.13 implies that there is a unique ordered basis $\mathcal{B}'$ for V for which $Q = M_{\mathcal{B},\mathcal{B}'}$ and a unique ordered basis $\mathcal{C}'$ for W for which $P = M_{\mathcal{C},\mathcal{C}'}$. Hence

$$B = M_{\mathcal{C},\mathcal{C}'}[\tau]_{\mathcal{B},\mathcal{C}}M_{\mathcal{B}',\mathcal{B}} = [\tau]_{\mathcal{B}',\mathcal{C}'}$$

Hence, B also represents τ. By symmetry, we see that A and B represent the same set of linear transformations. This completes the proof. □

We remarked in Example 0.3 that every matrix is equivalent to exactly one matrix of the block form

$$J_k = \begin{bmatrix} I_k & 0_{k,n-k} \\ 0_{m-k,k} & 0_{m-k,n-k} \end{bmatrix}_{\text{block}}$$

Hence, the set of these matrices is a set of canonical forms for equivalence. Moreover, the rank is a complete invariant for equivalence. In other words, two matrices are equivalent if and only if they have the same rank.

Similarity of Matrices

When a linear operator $\tau \in \mathcal{L}(V)$ is represented by a matrix of the form $[\tau]_{\mathcal{B}}$, equation (2.2) has the form

$$[\tau]_{\mathcal{B}'} = P[\tau]_{\mathcal{B}}P^{-1}$$

where P is an invertible matrix. This motivates the following definition.

Definition *Two matrices A and B are* **similar** *if there exists an invertible matrix P for which*

$$B = PAP^{-1}$$

The equivalence classes associated with similarity are called **similarity classes**. □

The analog of Theorem 2.18 for square matrices is the following.

Theorem 2.19 *Let V be a vector space of dimension n. Then two $n \times n$ matrices A and B are similar if and only if they represent the same linear operator $\tau \in \mathcal{L}(V)$, but possibly with respect to different ordered bases. In this case, A and B represent exactly the same set of linear operators in $\mathcal{L}(V)$.*

Proof. If A and B represent $\tau \in \mathcal{L}(V)$, that is, if

$$A = [\tau]_{\mathcal{B}} \text{ and } B = [\tau]_{\mathcal{C}}$$

for ordered bases $\mathcal{B}$ and $\mathcal{C}$ then Corollary 2.17 shows that A and B are similar. Now suppose that A and B are similar, say

$$B = PAP^{-1}$$

Suppose also that A represents a linear operator $\tau \in \mathcal{L}(V)$ for some ordered basis $\mathcal{B}$, that is,

$$A = [\tau]_{\mathcal{B}}$$

Theorem 2.13 implies that there is a unique ordered basis $\mathcal{C}$ for V for which $P = M_{\mathcal{B},\mathcal{C}}$. Hence

$$B = M_{\mathcal{B},\mathcal{C}}[\tau]_{\mathcal{B}}M_{\mathcal{B},\mathcal{C}}^{-1} = [\tau]_{\mathcal{C}}$$

Hence, B also represents τ. By symmetry, we see that A and B represent the same set of linear operators. This completes the proof. □

We will devote much effort in Chapter 7 to finding a canonical form for similarity.

Similarity of Operators

We can also define similarity of operators.

Definition *Two linear operators $\tau, \sigma \in \mathcal{L}(V)$ are* **similar** *if there exists an automorphism $\phi \in \mathcal{L}(V)$ for which*

$$\sigma = \phi\tau\phi^{-1}$$

The equivalence classes associated with similarity are called **similarity classes**. □

The analog of Theorem 2.19 in this case is the following.

Theorem 2.20 *Let V be a vector space of dimension n. Then two linear operators τ and σ on V are similar if and only if there is a matrix $A \in \mathcal{M}_n$ that represents both operators (but with respect to possibly different ordered bases). In this case, τ and σ are represented by exactly the same set of matrices in $\mathcal{M}_n$.*
Proof. If τ and σ are represented by $A \in \mathcal{M}_n$, that is, if

$$[\tau]_{\mathcal{B}} = A = [\sigma]_{\mathcal{C}}$$

for ordered bases $\mathcal{B}$ and $\mathcal{C}$ then

$$[\sigma]_{\mathcal{C}} = [\tau]_{\mathcal{B}} = M_{\mathcal{C},\mathcal{B}}[\tau]_{\mathcal{C}}M_{\mathcal{B},\mathcal{C}}$$

Let $\phi \in \mathcal{L}(V)$ be the automorphism of V defined by $\phi(c_i) = b_i$, where $\mathcal{B} = \{b_1, \ldots, b_n\}$ and $\mathcal{C} = \{c_1, \ldots, c_n\}$. Then

$$[\phi]_{\mathcal{C}} = ([\phi(c_1)]_{\mathcal{C}} \mid \cdots \mid [\phi(c_n)]_{\mathcal{C}}) = ([b_1]_{\mathcal{C}} \mid \cdots \mid [b_n]_{\mathcal{C}}) = M_{\mathcal{B},\mathcal{C}}$$

and so

$$[\sigma]_{\mathcal{C}} = [\phi]_{\mathcal{C}}^{-1}[\tau]_{\mathcal{C}}[\phi]_{\mathcal{C}} = [\phi^{-1}\tau\phi]_{\mathcal{C}}$$

from which it follows that σ and τ are similar. Conversely, suppose that τ and σ are similar, say

$$\sigma = \phi\tau\phi^{-1}$$

Suppose also that τ is represented by the matrix $A \in \mathcal{M}_n$, that is,

$$A = [\tau]_{\mathcal{B}}$$

for some ordered basis $\mathcal{B}$. Then

$$[\sigma]_{\mathcal{B}} = [\phi\tau\phi^{-1}]_{\mathcal{B}} = [\phi]_{\mathcal{B}}[\tau]_{\mathcal{B}}[\phi]_{\mathcal{B}}^{-1}$$

If we set $c_i = \phi(b_i)$ then $\mathcal{C} = (c_1, \ldots, c_n)$ is an ordered basis for V and

$$[\phi]_{\mathcal{B}} = ([\phi(b_1)]_{\mathcal{B}} \mid \cdots \mid [\phi(b_n)]_{\mathcal{B}}) = ([c_1]_{\mathcal{B}} \mid \cdots \mid [c_n]_{\mathcal{B}}) = M_{\mathcal{C},\mathcal{B}}$$

Hence

$$[\sigma]_{\mathcal{B}} = M_{\mathcal{C},\mathcal{B}}[\tau]_{\mathcal{B}}M_{\mathcal{C},\mathcal{B}}^{-1}$$

It follows that

$$A = [\tau]_{\mathcal{B}} = M_{\mathcal{B},\mathcal{C}}[\sigma]_{\mathcal{B}}M_{\mathcal{B},\mathcal{C}}^{-1} = [\sigma]_{\mathcal{C}}$$

and so A also represents σ. By symmetry, we see that τ and σ are represented by the same set of matrices. This completes the proof. □

Invariant Subspaces and Reducing Pairs

The restriction of a linear operator $\tau \in \mathcal{L}(V)$ to a subspace S of V is not necessarily a linear operator on S. This prompts the following definition.

Definition *Let $\tau \in \mathcal{L}(V)$. A subspace S of V is said to be* **invariant under** *τ or τ-***invariant** *if $\tau(S) \subseteq S$, that is, if $\tau(s) \in S$ for all $s \in S$. Put another way, S is invariant under τ if the restriction $\tau|_S$ is a linear operator on S.* □

If

$$V = S \oplus T$$

then the fact that S is τ-invariant does not imply that the complement T is also τ-invariant. (The reader may wish to supply a simple example with $V = \mathbb{R}^2$.)

Definition *Let $\tau \in \mathcal{L}(V)$. If $V = S \oplus T$ and if both S and T are τ-invariant, we say that the pair (S,T)* **reduces** *τ.* □

A reducing pair can be used to decompose a linear operator into a direct sum as follows.

Definition *Let $\tau \in \mathcal{L}(V)$. If (S,T) reduces τ we write*

$$\tau = \tau|_S \oplus \tau|_T$$

and call τ the **direct sum** *of $\tau|_S$ and $\tau|_T$. Thus, the expression*

$$\rho = \sigma \oplus \tau$$

means that there exist subspaces S and T of V for which (S,T) reduces ρ and

$$\sigma = \rho|_S \text{ and } \tau = \rho|_T \qquad \square$$

The concept of the direct sum of linear operators will play a key role in the study of the structure of a linear operator.

Topological Vector Spaces

This section is for readers with some familiarity with point-set topology. The **standard topology** on $\mathbb{R}^n$ is the topology for which the set of **open rectangles**

$$B = \{I_1 \times \cdots \times I_n \mid I_i\text{'s are open intervals in } \mathbb{R}\}$$

is a basis (in the sense of topology), that is, a subset of $\mathbb{R}^n$ is open if and only if it is a union of sets in B. The standard topology is the topology induced by the Euclidean metric on $\mathbb{R}^n$.

The standard topology on $\mathbb{R}^n$ has the property that the addition function

$$\mathcal{A}: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n: (v, w) \to v + w$$

and the scalar multiplication function

$$\mathcal{M}: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n: (r, v) \to rv$$

are continuous. As such, $\mathbb{R}^n$ is a *topological vector space*. Also, any linear functional $f: \mathbb{R}^n \to \mathbb{R}$ is a continuous map.

More generally, any real vector space V endowed with a topology $\mathcal{T}$ is called a **topological vector space** if the operations of addition $\mathcal{A}: V \times V \to V$ and scalar multiplication $\mathcal{M}: \mathbb{R} \times V \to V$ are continuous under $\mathcal{T}$.

Let V be a real vector space of dimension n and fix an ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$ for V. Consider the coordinate map

$$\phi = \phi_{\mathcal{B}}: V \to \mathbb{R}^n: v \to [v]_{\mathcal{B}}$$

and its inverse

$$\psi_{\mathcal{B}} = \phi_{\mathcal{B}}^{-1}: \mathbb{R}^n \to V: (a_1, \ldots, a_n) \to \sum a_i v_i$$

We claim that there is precisely one topology $\mathcal{T} = \mathcal{T}_V$ on V for which V becomes a topological vector space and for which all linear functionals are continuous. This is called the **natural topology** on V. In fact, the natural topology is the topology for which $\phi_{\mathcal{B}}$ (and therefore also $\psi_{\mathcal{B}}$) is a homeomorphism, for any basis $\mathcal{B}$. (Recall that a **homeomorphism** is a bijective map that is continuous and has a continuous inverse.)

Once this has been established, it will follow that the open sets in $\mathcal{T}$ are precisely the images of the open sets in $\mathbb{R}^n$ under the map $\psi_{\mathcal{B}}$. A basis for the natural topology is given by

$$\begin{aligned}&\{\psi_{\mathcal{B}}(I_1 \times \cdots \times I_n) \mid I_i\text{'s are open intervals in } \mathbb{R}\}\\ &\qquad = \Big\{\sum_{r_i \in I_i} r_i v_i \mid I_i\text{'s are open intervals in } \mathbb{R}\Big\}\end{aligned}$$

First, we show that if V is a topological vector space under a topology $\mathcal{T}$ then ψ is continuous. Since $\psi = \sum \psi_i$ where $\psi_i: \mathbb{R}^n \to V$ is defined by

$$\psi_i(a_1, \ldots, a_n) = a_i v_i$$

it is sufficient to show that these maps are continuous. (The sum of continuous maps is continuous.) Let O be an open set in $\mathcal{T}$. Then

$$\mathcal{M}^{-1}(O) = \{(r, x) \in \mathbb{R} \times V \mid rx \in O\}$$

is open in $\mathbb{R} \times V$. We need to show that the set

$$\psi_i^{-1}(O) = \{(a_1, \ldots, a_n) \in \mathbb{R}^n \mid a_i v_i \in O\}$$

is open in $\mathbb{R}^n$, so let $(a_1, \ldots, a_n) \in \psi_i^{-1}(O)$. Thus, $a_i v_i \in O$. It follows that $(a_i, v_i) \in \mathcal{M}^{-1}(O)$, which is open, and so there is an open interval $I \subseteq \mathbb{R}$ and an open set $B \in \mathcal{T}$ of V for which

$$(a_i, v_i) \in I \times B \subseteq \mathcal{M}^{-1}(O)$$

Then the open set $U = \mathbb{R} \times \cdots \times \mathbb{R} \times I \times \mathbb{R} \times \cdots \times \mathbb{R}$, where the factor I is in the ith position, has the property that $\psi_i(U) \subseteq O$. Thus

$$(a_1, \ldots, a_n) \in U \subseteq \psi_i^{-1}(O)$$

and so $\psi_i^{-1}(O)$ is open. Hence, ψ_i, and therefore also ψ, is continuous.

Next we show that if every linear functional on V is continuous under a topology $\mathcal{T}$ on V then the coordinate map ϕ is continuous. If $v \in V$ denote by $[v]_{\mathcal{B},i}$ the ith coordinate of $[v]_{\mathcal{B}}$. The map $\mu\colon V \to \mathbb{R}$ defined by $\mu(v) = [v]_{\mathcal{B},i}$ is a linear functional and so is continuous by assumption. Hence, for any open interval $I_i \in \mathbb{R}$ the set

$$A_i = \{v \in V \mid [v]_{\mathcal{B},i} \in I_i\}$$

is open. Now, if I_i are open intervals in $\mathbb{R}$ then

$$\phi^{-1}(I_1 \times \cdots \times I_n) = \{v \in V \mid [v]_{\mathcal{B}} \in I_1 \times \cdots \times I_n\} = \bigcap A_i$$

is open. Thus, ϕ is continuous.

Thus, if a topology $\mathcal{T}$ has the property that V is a topological vector space and every linear functional is continuous, then ϕ and $\psi = \phi^{-1}$ are homeomorphisms. This means that $\mathcal{T}$, if it exists, must be unique.

It remains to prove that the topology $\mathcal{T}$ on V that makes ϕ a homeomorphism has the property that V is a topological vector space under $\mathcal{T}$ and that any linear functional f on V continuous.

As to addition, the maps $\phi\colon V \to \mathbb{R}^n$ and $(\phi \times \phi)\colon V \times V \to \mathbb{R}^n \times \mathbb{R}^n$ are homeomorphisms and the map $\mathcal{A}'\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and so the map $\mathcal{A}\colon V \times V \to V$, being equal to $\phi^{-1} \circ \mathcal{A}' \circ (\phi \times \phi)$, is also continuous.

As to scalar multiplication, the maps $\phi\colon V \to \mathbb{R}^n$ and $(\iota \times \phi)\colon \mathbb{R} \times V \to \mathbb{R} \times \mathbb{R}^n$ are homeomorphisms and the map $\mathcal{M}'\colon \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ is continuous and so the map $\mathcal{M}\colon V \times V \to V$, being equal to $\phi^{-1} \circ \mathcal{M}' \circ (\iota \times \phi)$, is also continuous.

Now let f be a linear functional. Since ϕ is continuous if and only if $f \circ \phi^{-1}$ is continuous, we can confine attention to $V = \mathbb{R}^n$. In this case, if $e_1, \ldots, e_n$ is the standard basis for $\mathbb{R}^n$ and $|f(e_i)| \leq M$, then for any $x = (a_1, \ldots, a_n) \in \mathbb{R}^n$ we have

$$|f(x)| = \left|\sum a_i f(e_i)\right| \leq \sum |a_i||f(e_i)| \leq M \sum |a_i|$$

Now, if $|x| < \epsilon/Mn$ then $|a_i| < \epsilon/Mn$ and so $|f(x)| < \epsilon$, which implies that f is continuous.

According to the Riesz representation theorem and the Cauchy–Schwarz inequality, we have

$$\|f(x)\| \leq \|\mathcal{R}_f\|\|x\|$$

Hence, $x_n \to 0$ implies $f(x_n) \to 0$ and so by linearity, $x_n \to x$ implies $f(x_n) \to x$ and so f is continuous.

Theorem 2.21 *Let V be a real vector space of dimension n. There is a unique topology on V, called the* **natural topology** *for which V is a topological vector space and for which all linear functionals on V are continuous. This topology is determined by the fact that the coordinate map $\phi: V \to \mathbb{R}^n$ is a homeomorphism.* □

Linear Operators on $V^{\mathbb{C}}$

A linear operator τ on a real vector space V can be extended to a linear operator $\tau^{\mathbb{C}}$ on the complexification $V^{\mathbb{C}}$ by defining

$$\tau^{\mathbb{C}}(u + vi) = \tau(u) + \tau(v)i$$

Here are the basic properties of this **complexification** of τ.

Theorem 2.22 *If $\tau, \sigma \in \mathcal{L}(V)$ then*
1) $(a\tau)^{\mathbb{C}} = a\tau^{\mathbb{C}}$, $a \in \mathbb{R}$
2) $(\tau + \sigma)^{\mathbb{C}} = \tau^{\mathbb{C}} + \sigma^{\mathbb{C}}$
3) $(\tau\sigma)^{\mathbb{C}} = \tau^{\mathbb{C}}\sigma^{\mathbb{C}}$
4) $[\tau(v)]^{\mathbb{C}} = \tau^{\mathbb{C}}(v^{\mathbb{C}})$. □

Let us recall that for any ordered basis $\mathcal{B}$ for V and any vector $v \in V$ we have

$$[v + 0i]_{\text{cpx}(\mathcal{B})} = [v]_{\mathcal{B}}$$

Now, if $\mathcal{B}$ is a basis for V, then the ith column of $[\tau]_{\mathcal{B}}$ is

$$[\tau(b_i)]_{\mathcal{B}} = [\tau(b_i) + 0i]_{\text{cpx}(\mathcal{B})} = [\tau^{\mathbb{C}}(b_i + 0i)]_{\text{cpx}(\mathcal{B})}$$

which is the ith column of the coordinate matrix of $\tau^{\mathbb{C}}$ with respect to the basis cpx$(\mathcal{B})$. Thus we have the following theorem.

Theorem 2.23 *Let $\tau \in \mathcal{L}(V)$ where V is a real vector space. The matrix of $\tau^{\mathbb{C}}$ with respect to the basis* $\operatorname{cpx}(\mathcal{B})$ *is equal to the matrix of τ with respect to the basis $\mathcal{B}$*

$$[\tau^{\mathbb{C}}]_{\operatorname{cpx}(\mathcal{B})} = [\tau]_{\mathcal{B}}$$

Hence, if a real matrix A represents a linear operator τ on V then A also represents the complexification $\tau^{\mathbb{C}}$ of τ on $V^{\mathbb{C}}$. □

Exercises

1. Let $A \in \mathcal{M}_{m,n}$ have rank k. Prove that there are matrices $X \in \mathcal{M}_{m,k}$ and $Y \in \mathcal{M}_{k,n}$, both of rank k, for which $A = XY$. Prove that A has rank 1 if and only if it has the form $A = x^t y$ where x and y are row matrices.
2. Prove Corollary 2.9 and find an example to show that the corollary does not hold without the finiteness condition.
3. Let $\tau \in \mathcal{L}(V,W)$. Prove that τ is an isomorphism if and only if it carries a basis for V to a basis for W.
4. If $\tau \in \mathcal{L}(V_1,W_1)$ and $\sigma \in \mathcal{L}(V_2,W_2)$ we define the external direct sum $\tau \boxplus \sigma \in \mathcal{L}(V_1 \boxplus V_2, W_1 \boxplus W_2)$ by
$$(\tau \boxplus \sigma)((v_1,v_2)) = (\tau(v_1), \sigma(v_2))$$
Show that $\tau \boxplus \sigma$ is a linear transformation.
5. Let $V = S \oplus T$. Prove that $S \oplus T \approx S \boxplus T$. Thus, internal and external direct sums are equivalent up to isomorphism.
6. Let $V = A + B$ and consider the external direct sum $E = A \boxplus B$. Define a map $\tau: A \boxplus B \to V$ by $\tau(v,w) = v + w$. Show that τ is linear. What is the kernel of τ? When is τ an isomorphism?
7. Let $\mathcal{T}$ be a subset of $\mathcal{L}(V)$. A subspace S of V is **$\mathcal{T}$-invariant** if S is τ-invariant for every $\tau \in \mathcal{T}$. Also, V is **$\mathcal{T}$-irreducible** if the only $\mathcal{T}$-invariant subspaces of V are $\{0\}$ and V. Prove the following form of *Schur's lemma.* Suppose that $\mathcal{T}_V \subseteq \mathcal{L}(V)$ and $\mathcal{T}_W \subseteq \mathcal{L}(W)$ and V is $\mathcal{T}_V$-irreducible and W is $\mathcal{T}_W$-irreducible. Let $\alpha \in \mathcal{L}(V,W)$ satisfy $\alpha\mathcal{T}_V = \mathcal{T}_W\alpha$, that is, for any $\mu \in \mathcal{T}_V$ there is a $\lambda \in \mathcal{T}_W$ such that $\alpha\mu = \lambda\alpha$. Prove that $\alpha = 0$ or α is an isomorphism.
8. Let $\tau \in \mathcal{L}(V)$ where $\dim(V) < \infty$. If $\operatorname{rk}(\tau^2) = \operatorname{rk}(\tau)$ show that $\operatorname{im}(\tau) \cap \ker(\tau) = \{0\}$.
9. Let $\tau \in \mathcal{L}(U,V)$ and $\sigma \in \mathcal{L}(V,W)$. Show that
$$\operatorname{rk}(\tau\sigma) \le \min\{\operatorname{rk}(\tau), \operatorname{rk}(\sigma)\}$$
10. Let $\tau \in \mathcal{L}(U,V)$ and $\sigma \in \mathcal{L}(V,W)$. Show that
$$\operatorname{null}(\tau\sigma) \le \operatorname{null}(\tau) + \operatorname{null}(\sigma)$$
11. Let $\tau, \sigma \in \mathcal{L}(V)$ where τ is invertible. Show that
$$\operatorname{rk}(\tau\sigma) = \operatorname{rk}(\sigma\tau) = \operatorname{rk}(\sigma)$$

12. Let $\tau, \sigma \in \mathcal{L}(V, W)$. Show that
$$\text{rk}(\tau + \sigma) \leq \text{rk}(\tau) + \text{rk}(\sigma)$$
13. Let S be a subspace of V. Show that there is a $\tau \in \mathcal{L}(V)$ for which $\ker(\tau) = S$. Show also that there exists a $\sigma \in \mathcal{L}(V)$ for which $\text{im}(\sigma) = S$.
14. Suppose that $\tau, \sigma \in \mathcal{L}(V)$.
    a) Show that $\sigma = \tau\mu$ for some $\mu \in \mathcal{L}(V)$ if and only if $\text{im}(\sigma) \subseteq \text{im}(\tau)$.
    b) Show that $\sigma = \mu\tau$ for some $\mu \in \mathcal{L}(V)$ if and only if $\ker(\tau) \subseteq \ker(\sigma)$.
15. Let $V = S_1 \oplus S_2$. Define linear operators ρ_i on V by $\rho_i(s_1 + s_2) = s_i$ for $i = 1, 2$. These are referred to as **projection operators**. Show that
    1) $\rho_i^2 = \rho_i$
    2) $\rho_1 + \rho_2 = I$, where I is the identity map on V.
    3) $\rho_i \rho_j = 0$ for $i \neq j$ where 0 is the zero map.
    4) $V = \text{im}(\rho_1) \oplus \text{im}(\rho_2)$
16. Let $\dim(V) < \infty$ and suppose that $\tau \in \mathcal{L}(V)$ satisfies $\tau^2 = 0$. Show that $2\text{rk}(\tau) \leq \dim(V)$.
17. Let A be an $m \times n$ matrix over F. What is the relationship between the linear transformation $\tau_A \colon F^n \to F^m$ and the system of equations $AX = B$? Use your knowledge of linear transformations to state and prove various results concerning the system $AX = B$, especially when $B = 0$.
18. Let V have basis $\mathcal{B} = \{v_1, \ldots, v_n\}$. Suppose that for each $1 \leq i, j \leq n$ we define $\tau_{i,j} \in \mathcal{L}(V)$ by
$$\tau_{i,j}(v_k) = \begin{cases} v_k & \text{if } k \neq i \\ v_i + v_j & \text{if } k = i \end{cases}$$
    Prove that the $\tau_{i,j}$ are invertible and form a basis for $\mathcal{L}(V)$.
19. Let $\tau \in \mathcal{L}(V)$. If S is a τ-invariant subspace of V must there be a subspace T of V for which (S, T) reduces τ?
20. Find an example of a vector space V and a proper subspace S of V for which $V \approx S$.
21. Let $\dim(V) < \infty$. If τ, $\sigma \in \mathcal{L}(V)$ prove that $\sigma\tau = \iota$ implies that τ and σ are invertible and that $\sigma = p(\tau)$ for some polynomial $p(x) \in F[x]$.
22. Let $\tau \in \mathcal{L}(V)$ where $\dim(V) < \infty$. If $\tau\sigma = \sigma\tau$ for all $\sigma \in \mathcal{L}(V)$ show that $\tau = a\iota$, for some $a \in F$, where ι is the identity map.
23. Let $A, B \in \mathcal{M}_n(F)$. Let K be a field containing F. Show that if A and B are similar over K, that is, if $B = PAP^{-1}$ where $P \in \mathcal{M}_n(K)$ then A and B are also similar over F, that is, there exists $Q \in \mathcal{M}_n(F)$ for which $B = QAQ^{-1}$. *Hint*: consider the equation $XA - BX = 0$ as a homogeneous system of linear equations with coefficients in F. Does it have a solution? Where?
24. Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a continuous function with the property that
$$f(x + y) = f(x) + f(y)$$
    Prove that f is a linear functional on $\mathbb{R}^n$.

25. Prove that any linear functional $f\colon\mathbb{R}^n \to \mathbb{R}$ is a continuous map.
26. Prove that any subspace S of $\mathbb{R}^n$ is a closed set or, equivalently, that $S^c = \mathbb{R}^n \setminus S$ is open, that is, for any $x \in S^c$ there is an open ball $B(s,\epsilon)$ centered at x with radius $\epsilon > 0$ for which $B(x,\epsilon) \subseteq S^c$.
27. Prove that any linear transformation $\tau\colon V \to W$ is continuous under the natural topologies of V and W.
28. Prove that any surjective linear transformation τ from V to W (both finite-dimensional topological vector spaces under the natural topology) is an open map, that is, τ maps open sets to open sets.
29. Prove that any subspace S of a finite-dimensional vector space V is a closed set or, equivalently, that S^c is open, that is, for any $x \in S^c$ there is an open ball $B(s,\epsilon)$ centered at x with radius $\epsilon > 0$ for which $B(x,\epsilon) \subseteq S^c$.
30. Let S be a subspace of V with $\dim(V) < \infty$.
    a) Show that the subspace topology on S inherited from V is the natural topology.
    b) Show that the natural topology on V/S is the topology for which the natural projection map $\pi\colon V \to V/S$ continuous and open.
31. If V is a real vector space then $V^{\mathbb{C}}$ is a complex vector space. Thinking of $V^{\mathbb{C}}$ as a vector space $(V^{\mathbb{C}})_{\mathbb{R}}$ over $\mathbb{R}$, show that $(V^{\mathbb{C}})_{\mathbb{R}}$ is isomorphic to the external direct product $V \boxplus V$.
34. (When is a complex linear map a complexification?) Let V be a real vector space with complexification $V^{\mathbb{C}}$ and let $\sigma \in \mathcal{L}(V^{\mathbb{C}})$. Prove that σ is a complexification, that is, σ has the form $\tau^{\mathbb{C}}$ for some $\tau \in \mathcal{L}(V)$ if and only if σ commutes with the conjugate map $\chi\colon V^{\mathbb{C}} \to V^{\mathbb{C}}$ defined by $\chi(u+iv) = u - iv$.
35. Let W be a complex vector space.
    a) Consider replacing the scalar multiplication on W by the operation

    $$(z,w) \to \overline{z}w$$

    where $z \in \mathbb{C}$ and $w \in W$. Show that the resulting set with the addition defined for the vector space W and with this scalar multiplication is a complex vector space, which we denote by $\overline{W}$.
    b) Show, without using dimension arguments, that $(W_{\mathbb{R}})^{\mathbb{C}} \approx W \boxplus \overline{W}$.
36. a) Let τ be a linear operator on the real vector space U with the property that $\tau^2 = -\iota$. Define a scalar multiplication on U by complex numbers as follows

    $$(a+bi)\cdot v = av + b\tau(v)$$

    for $a,b \in \mathbb{R}$ and $v \in U$. Prove that under the addition of U and this scalar multiplication U is a complex vector space, which we denote by U_τ.
    b) What is the relationship between U_τ and $V^{\mathbb{C}}$? Hint: consider $U = V \boxplus V$ and $\tau(u,v) = (-v,u)$.

Chapter 3
The Isomorphism Theorems

Quotient Spaces

Let S be a subspace of a vector space V. It is easy to see that the binary relation on V defined by

$$u \equiv v \Leftrightarrow u - v \in S$$

is an equivalence relation. When $u \equiv v$, we say that u and v are **congruent modulo** S. The term *mod* is used as a colloquialism for modulo and $u \equiv v$ is often written

$$u \equiv v \operatorname{mod} S$$

When the subspace in question is clear, we will simply write $u \equiv v$.

To see what the equivalence classes look like, observe that

$$\begin{aligned} [v] &= \{u \in V \mid u \equiv v\} \\ &= \{u \in V \mid u - v \in S\} \\ &= \{u \in V \mid u = v + s \text{ for some } s \in S\} \\ &= \{v + s \mid s \in S\} \\ &= v + S \end{aligned}$$

The set

$$[v] = v + S = \{v + s \mid s \in S\}$$

is called a **coset** of S in V and v is called a **coset representative** for $v + S$. (Thus, any member of a cost is a coset representative.)

The set of all cosets of S in V is denoted by

$$\frac{V}{S} = \{v + S \mid v \in V\}$$

This is read "V mod S" and is called the **quotient space of** V **modulo** S. Of course, the term space is a hint that we intend to define vector space operations on V/S.

Note that congruence modulo S is preserved under the vector space operations on V, for if $u_1 \equiv v_1$ and $u_2 \equiv v_2$ then

$$\begin{aligned} u_1 - v_1 \in S, u_2 - v_2 \in S &\Rightarrow r(u_1 - v_1) + s(u_2 - v_2) \in S \\ &\Rightarrow (ru_1 + su_2) - (rv_1 + sv_2) \in S \\ &\Rightarrow ru_1 + su_2 \equiv rv_1 + sv_2 \end{aligned}$$

A natural choice for vector space operations on V/S is

$$\begin{aligned} r(u + S) &= ru + S \\ (u + S) + (v + S) &= (u + v) + S \end{aligned}$$

However, in order to show that these operations are well-defined, it is necessary to show that they do not depend on the choice of coset representatives, that is, if

$$u_1 + S = u_2 + S \text{ and } v_1 + S = v_2 + S$$

then

$$\begin{aligned} ru_1 + S &= ru_2 + S \\ (u_1 + v_1) + S &= (u_2 + v_2) + S \end{aligned}$$

The straightforward details of this are left to the reader. Let us summarize.

Theorem 3.1 *Let S be a subspace of V. The binary relation*

$$u \equiv v \Leftrightarrow u - v \in S$$

is an equivalence relation on V, whose equivalence classes are the **cosets**

$$v + S = \{v + s \mid s \in S\}$$

of S in V. The set V/S of all cosets of S in V, called the **quotient space** of V modulo S, is a vector space under the well-defined operations

$$\begin{aligned} ru_1 + S &= ru_2 + S \\ (u_1 + v_1) + S &= (u_2 + v_2) + S \end{aligned}$$

The zero vector in V/S is the coset $0 + S = S$. □

The Natural Projection and the Correspondence Theorem

If S is a subspace of V then we can define a map $\pi_S: V \to V/S$ by sending each vector to the coset containing it

$$\pi_S(v) = v + S$$

This map is called the **canonical projection** or **natural projection** of V onto V/S, or simply **projection modulo** S. It is easily seen to be linear, for we have

(writing π for π_S)

$$\pi(ru+sv) = (ru+sv)+S = r(u+S)+s(v+S) = r\pi(u)+s\pi(v)$$

The canonical projection is clearly surjective. To determine the kernel of π, note that

$$v \in \ker(\pi) \Leftrightarrow \pi(v)=0 \Leftrightarrow v+S=S \Leftrightarrow v \in S$$

and so

$$\ker(\pi)=S$$

Theorem 3.2 *The canonical projection $\pi_S\colon V \to V/S$ defined by*

$$\pi_S(v)=v+S$$

is a surjective linear transformation with $\ker(\pi_S)=S$. □

If S is a subspace of V then the subspaces of the quotient space V/S have the form T/S for some intermediate subspace T satisfying $S \subseteq T \subseteq V$. In fact, as shown in Figure 3.1, the projection map π_S provides a one-to-one correspondence between intermediate subspaces $S \subseteq T \subseteq V$ and subspaces of the quotient space V/S. The proof of the following theorem is left as an exercise.

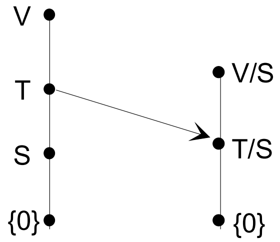

Figure 3.1: The correspondence theorem

Theorem 3.3 *(***The correspondence theorem***) Let S be a subspace of V. Then the function that assigns to each intermediate subspace $S \subseteq T \subseteq V$ the subspace T/S of V/S is an order preserving (with respect to set inclusion) one-to-one correspondence between the set of all subspaces of V containing S and the set of all subspaces of V/S.* □

The Universal Property of Quotients and the First Isomorphism Theorem

Let S be a subspace of V. The pair $(V/S, \pi_S)$ has a very special property, known as the *universal property*—a term that comes from the world of category theory.

Figure 3.2 shows a linear transformation $\tau \in \mathcal{L}(V,W)$, along with the canonical projection π_S from V to the quotient space V/S.

$$\begin{tikzcd}
V \arrow[r, "\tau"] \arrow[d, "\pi_s"'] & W \\
V/S \arrow[ur, dashed, "\tau'"'] &
\end{tikzcd}$$

Figure 3.2: The universal property

The universal property states that if $\ker(\tau) \supseteq S$ then there is a unique $\tau'\colon V/S \to W$ for which

$$\tau' \circ \pi_S = \tau$$

Another way to say this is that any such $\tau \in \mathcal{L}(V,W)$ can be *factored through* the canonical projection π_S.

Theorem 3.4 *Let S be a subspace of V and let $\tau \in \mathcal{L}(V,W)$ satisfy $S \subseteq \ker(\tau)$. Then, as pictured in Figure 3.2, there is a unique linear transformation $\tau'\colon V/S \to W$ with the property that*

$$\tau' \circ \pi_S = \tau$$

Moreover, $\ker(\tau') = \ker(\tau)/S$ *and* $\operatorname{im}(\tau') = \operatorname{im}(\tau)$.

Proof. We have no other choice but to define τ' by the condition $\tau' \circ \pi_S = \tau$, that is,

$$\tau'(v+S) = \tau(v)$$

This function is well-defined if and only if

$$v+S = u+S \Rightarrow \tau'(v+S) = \tau'(u+S)$$

which is equivalent to each of the following statements:

$$\begin{gathered}
v+S = u+S \Rightarrow \tau(v) = \tau(u) \\
v-u \in S \Rightarrow \tau(v-u) = 0 \\
x \in S \Rightarrow \tau(x) = 0 \\
S \subseteq \ker(\tau)
\end{gathered}$$

Thus, $\tau'\colon V/S \to W$ is well-defined. Also,

$$\operatorname{im}(\tau') = \{\tau'(v+S) \mid v \in V\} = \{\tau(v) \mid v \in V\} = \operatorname{im}(\tau)$$

and

$$\begin{aligned}\ker(\tau') &= \{v+S \mid \tau'(v+S)=0\}\\ &= \{v+S \mid \tau(v)=0\}\\ &= \{v+S \mid v\in\ker(\tau)\}\\ &= \ker(\tau)/S\end{aligned}$$

The uniqueness of τ' is evident. □

Theorem 3.4 has a very important corollary, which is often called the *first isomorphism theorem* and is obtained by taking $S=\ker(\tau)$.

Theorem 3.5 *(***The first isomorphism theorem***) Let $\tau: V\to W$ be a linear transformation. Then the linear transformation $\tau': V/\ker(\tau)\to W$ defined by*

$$\tau'(v+\ker(\tau))=\tau(v)$$

is injective and

$$\frac{V}{\ker(\tau)}\approx \operatorname{im}(\tau) \qquad \square$$

According to Theorem 3.5, the image of any linear transformation on V is isomorphic to a quotient space of V. Conversely, any quotient space V/S of V is the image of a linear transformation on V: the canonical projection π_S. Thus, up to isomorphism, quotient spaces are equivalent to homomorphic images.

Quotient Spaces, Complements and Codimension

The first isomorphism theorem gives some insight into the relationship between complements and quotient spaces. Let S be a subspace of V and let T be a complement of S, that is

$$V=S\oplus T$$

Since every vector $v\in V$ has the form $v=s+t$, for unique vectors $s\in S$ and $t\in T$, we can define a linear operator $\rho: V\to V$ by setting

$$\rho(s+t)=t$$

Because s and t are unique, ρ is well-defined. It is called **projection onto** T **along** S. (Note the word onto, rather than modulo, in the definition; this is not the same as projection modulo a subspace.) It is clear that

$$\operatorname{im}(\rho)=T$$

and

$$\ker(\rho)=\{s+t\in V \mid t=0\}=S$$

Hence, the first isomorphism theorem implies that

$$T \approx \frac{V}{S}$$

Theorem 3.6 Let S be a subspace of V. All complements of S in V are isomorphic to V/S and hence to each other. □

The previous theorem can be rephrased by writing

$$A \oplus B = A \oplus C \Rightarrow B \approx C$$

On the other hand, quotients and complements do not behave as nicely with respect to isomorphisms as one might casually think. We leave it to the reader to show the following:

1) It is possible that

$$A \oplus B = C \oplus D$$

with $A \approx C$ but $B \not\approx D$. Hence, $A \approx C$ does *not* imply that a complement of A is isomorphic to a complement of C.

2) It is possible that $V \approx W$ and

$$V = S \oplus B \text{ and } W = S \oplus D$$

but $B \not\approx D$. Hence, $V \approx W$ does *not* imply that $V/S \not\approx W/S$. (However, according to the previous theorem, if V *equals* W then $B \approx D$.)

Corollary 3.7 *Let S be a subspace of a vector space V. Then*

$$\dim(V) = \dim(S) + \dim(V/S) \qquad \square$$

Definition *If S is a subspace of V then* $\dim(V/S)$ *is called the* **codimension** *of S in V and is denoted by* $\operatorname{codim}(S)$ *or* $\operatorname{codim}_V(S)$. □

Thus, the codimension of S in V is the dimension of any complement of S in V and when V is *finite-dimensional*, we have

$$\operatorname{codim}_V(S) = \dim(V) - \dim(S)$$

(This makes no sense, in general, if V is not finite-dimensional, since infinite cardinal numbers cannot be subtracted.)

Additional Isomorphism Theorems

There are several other isomorphism theorems that are consequences of the first isomorphism theorem. As we have seen, if $V = S \oplus T$ then $V/T \approx S$. This can be written

$$\frac{S \oplus T}{T} \approx \frac{S}{S \cap T}$$

This applies to nondirect sums as well.

Theorem 3.8 *(***The second isomorphism theorem***) Let V be a vector space and let S and T be subspaces of V. Then*

$$\frac{S + T}{T} \approx \frac{S}{S \cap T}$$

Proof. Let $\tau\colon (S+T) \to S/(S \cap T)$ be defined by

$$\tau(s+t) = s + (S \cap T)$$

We leave it to the reader to show that τ is a well-defined surjective linear transformation, with kernel T. An application of the first isomorphism theorem then completes the proof. □

The following theorem demonstrates one way in which the expression V/S behaves like a fraction.

Theorem 3.9 *(***The third isomorphism theorem***) Let V be a vector space and suppose that $S \subseteq T \subseteq V$ are subspaces of V. Then*

$$\frac{V/S}{T/S} \approx \frac{V}{T}$$

Proof. Let $\tau\colon V/S \to V/T$ be defined by $\tau(v+S) = v+T$. We leave it to the reader to show that τ is a well-defined surjective linear transformation whose kernel is T/S. The rest follows from the first isomorphism theorem. □

The following theorem demonstrates one way in which the expression V/S does not behave like a fraction.

Theorem 3.10 *Let V be a vector space and let S be a subspace of V. Suppose that $V = V_1 \oplus V_2$ and $S = S_1 \oplus S_2$ with $S_i \subseteq V_i$. Then*

$$\frac{V}{S} = \frac{V_1 \oplus V_2}{S_1 \oplus S_2} \approx \frac{V_1}{S_1} \boxplus \frac{V_2}{S_2}$$

Proof. Let $\tau\colon V \to (V_1/S_1) \boxplus (V_2/S_2)$ be defined by

$$\tau(v_1 + v_2) = (v_1 + S_1, v_2 + S_2)$$

This map is well-defined, since the sum $V = V_1 \oplus V_2$ is direct. We leave it to the reader to show that τ is a surjective linear transformation, whose kernel is $S_1 \oplus S_2$. The rest follows from the first isomorphism theorem. □

Linear Functionals

Linear transformations from V to the base field F (thought of as a vector space over itself) are extremely important.

Definition *Let V be a vector space over F. A linear transformation $f \in \mathcal{L}(V,F)$, whose values lie in the base field F is called a* **linear functional** *(or simply* **functional***) on V. (Some authors use the term* linear function*.) The vector space of all linear functionals on V is denoted by V^* and is called the* **algebraic dual space** *of V.* □

The adjective *algebraic* is needed here, since there is another type of dual space that is defined on general normed vector spaces, where continuity of linear transformations makes sense. We will discuss the so-called *continuous dual space* briefly in Chapter 13. However, until then, the term "dual space" will refer to the algebraic dual space.

To help distinguish linear functionals from other types of linear transformations, we will usually denote linear functionals by lower case italic letters, such as f, g and h.

Example 3.1 The map $f\colon F[x] \to F$, defined by $f(p(x)) = p(0)$ is a linear functional, known as **evaluation at** 0. □

Example 3.2 Let $\mathcal{C}[a,b]$ denote the vector space of all continuous functions on $[a,b] \subseteq \mathbb{R}$. Let $f\colon \mathcal{C}[a,b] \to \mathbb{R}$ be defined by

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

Then $f \in \mathcal{C}[a,b]^*$. □

For any $f \in V^*$, the rank plus nullity theorem is

$$\dim(\ker(f)) + \dim(\operatorname{im}(f)) = \dim(V)$$

But since $\operatorname{im}(f) \subseteq F$, we have either $\operatorname{im}(f) = \{0\}$, in which case f is the zero linear functional, or $\operatorname{im}(f) = F$, in which case f is surjective. In other words, a nonzero linear functional is surjective. Moreover, if $f \neq 0$ then

$$\operatorname{codim}(\ker(f)) = \dim\left(\frac{V}{\ker(f)}\right) = 1$$

and if $\dim(V) < \infty$ then

$$\dim(\ker(f)) = \dim(V) - 1$$

Thus, in dimensional terms, the kernel of a linear functional is a very "large" subspace of the domain V.

The following theorem will prove very useful.

Theorem 3.11

1) *For any nonzero vector $v \in V$, there exists a linear functional $f \in V^*$ for which $f(v) \neq 0$.*
2) *A vector $v \in V$ is zero if and only if $f(v) = 0$ for all $f \in V^*$.*
3) *Let $f \in V^*$. If $f(x) \neq 0$ then*

$$V = \langle x \rangle \oplus \ker(f)$$

4) *Two nonzero linear functionals $f, g \in V^*$ have the same kernel if and only if there is a nonzero scalar λ such that $f = \lambda g$.*

Proof. For part 3), if $0 \neq v \in \langle x \rangle \cap \ker(f)$ then $f(v) = 0$ and $v = ax$ for $0 \neq a \in F$, whence $f(x) = 0$, which is false. Hence, $\langle x \rangle \cap \ker(f) = \{0\}$ and the direct sum $S = \langle x \rangle \oplus \ker(f)$ exists. Also, for any $v \in V$ we have

$$v = \frac{f(v)}{f(x)}x + \left(v - \frac{f(v)}{f(x)}x\right) \in \langle x \rangle + \ker(f)$$

and so $V = \langle x \rangle \oplus \ker(f)$.

For part 4), if $f = \lambda g$ for $\lambda \neq 0$ then $\ker(f) = \ker(g)$. Conversely, if $K = \ker(f) = \ker(g)$ then for $x \notin K$ we have by part 3),

$$V = \langle x \rangle \oplus K$$

Of course, $f|_K = \lambda g|_K$ for any λ. Therefore, if $\lambda = f(x)/g(x)$, it follows that $\lambda g(x) = f(x)$ and hence $f = \lambda g$. □

Dual Bases

Let V be a vector space with basis $\mathcal{B} = \{v_i \mid i \in I\}$. For each $i \in I$, we can define a linear functional $v_i^* \in V^*$, by the orthogonality condition

$$v_i^*(v_j) = \delta_{i,j}$$

where $\delta_{i,j}$ is the **Kronecker delta function**, defined by

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Then the set $\mathcal{B}^* = \{v_i^* \mid i \in I\}$ is linearly independent, since applying the equation

$$0 = a_{i_1}v_{i_1}^* + \cdots + a_{i_n}v_{i_n}^*$$

to the basis vector v_{i_k} gives

$$0 = \sum_{j=1}^{k} a_{i_j} v_{i_j}^*(v_{i_k}) = \sum_{j=1}^{k} a_{i_j} \delta_{i_j,i_k} = a_{i_k}$$

for all i_k.

Theorem 3.12 *Let V be a vector space with basis $\mathcal{B} = \{v_i \mid i \in I\}$.*
1) *The set $\mathcal{B}^* = \{v_i^* \mid i \in I\}$ is linearly independent.*
2) *If V is finite-dimensional then $\mathcal{B}^*$ is a basis for V^*, called the* **dual basis** *of $\mathcal{B}$.*

Proof. For part 2), for any $f \in V^*$, we have

$$\sum_j f(v_j) v_j^*(v_i) = \sum_j f(v_j) \delta_{i,j} = f(v_i)$$

and so $f = \sum f(v_j) v_j^*$ is in the span of $\mathcal{B}^*$. Hence, $\mathcal{B}^*$ is a basis for V^*. □

Corollary 3.13 *If* $\dim(V) < \infty$ *then* $\dim(V^*) = \dim(V)$. □

The next example shows that Corollary 3.13 does not hold without the finiteness condition.

Example 3.3 Let V be an infinite-dimensional vector space over the field $F = \mathbb{Z}_2 = \{0, 1\}$, with basis $\mathcal{B}$. Since the only coefficients in F are 0 and 1, a finite linear combination over F is just a finite sum. Hence, V is the set of all finite sums of vectors in $\mathcal{B}$ and so according to Theorem 0.11,

$$|V| \leq |\mathcal{P}_0(\mathcal{B})| = |\mathcal{B}|$$

On the other hand, each linear functional $f \in V^*$ is uniquely defined by specifying its values on the basis $\mathcal{B}$. Since these values must be either 0 or 1, specifying a linear functional is equivalent to specifying the subset of $\mathcal{B}$ on which f takes the value 1. In other words, there is a one-to-one correspondence between linear functionals on V and all subsets of $\mathcal{B}$. Hence,

$$|V^*| = |\mathcal{P}(\mathcal{B})| > |\mathcal{B}| \geq |V|$$

This shows that V^* cannot be isomorphic to V, nor to any proper subset of V. Hence, $\dim(V^*) > \dim(V)$. □

Reflexivity

If V is a vector space then so is the dual space V^*. Hence, we may form the **double (algebraic) dual space** V^{**}, which consists of all linear functionals $\sigma: V^* \to F$. In other words, an element σ of V^{**} is a linear map that assigns a scalar to each linear functional on V.

With this firmly in mind, there is one rather obvious way to obtain an element of V^{**}. Namely, if $v \in V$, consider the map $\overline{v}: V^* \to F$ defined by

$$\overline{v}(f) = f(v)$$

which sends the linear functional f to the scalar $f(v)$. The map $\overline{v}$ is called **evaluation at** v. To see that $\overline{v} \in V^{**}$, if $f, g \in V^*$ and $a, b \in F$ then

$$\overline{v}(af + bg) = (af + bg)(v) = af(v) + bg(v) = a\overline{v}(f) + b\overline{v}(g)$$

and so $\overline{v}$ is indeed linear.

We can now define a map $\tau: V \to V^{**}$ by

$$\tau(v) = \overline{v}$$

This is called the **canonical map** (or the **natural map**) from V to V^{**}. This map is injective and hence in the finite-dimensional case, it is also surjective.

Theorem 3.14 *The canonical map $\tau: V \to V^{**}$ defined by $\tau(v) = \overline{v}$, where $\overline{v}$ is evaluation at v, is a monomorphism. If V is finite-dimensional then τ is an isomorphism.*
Proof. The map τ is linear since

$$\overline{au + bv}(f) = f(au + bv) = af(u) + bf(v) = (a\overline{u} + b\overline{v})(f)$$

for all $f \in V^*$. To determine the kernel of τ, observe that

$$\begin{aligned}\tau(v) = 0 &\Rightarrow \overline{v} = 0\\ &\Rightarrow \overline{v}(f) = 0 \text{ for all } f \in V^*\\ &\Rightarrow f(v) = 0 \text{ for all } f \in V^*\\ &\Rightarrow\ v = 0\end{aligned}$$

by Theorem 3.11 and so $\ker(\tau) = \{0\}$.

In the finite-dimensional case, since $\dim(V^{**}) = \dim(V^*) = \dim(V)$, it follows that τ is also surjective, hence an isomorphism. □

Note that if $\dim(V) < \infty$ then since the dimensions of V and V^{**} are the same, we deduce immediately that $V \approx V^{**}$. This is not the point of Theorem 3.14. The point is that the *natural map* $v \to \overline{v}$ is an isomorphism. Because of this, V is said to be **algebraically reflexive**. Thus, Theorem 3.14 implies that all finite-dimensional vector spaces are algebraically reflexive.

If V is finite-dimensional, it is customary to identify the double dual space V^{**} with V and to think of the elements of V^{**} simply as vectors in V. Let us consider an example of a vector space that is not algebraically reflexive.

Example 3.4 Let V be the vector space over $\mathbb{Z}_2$ with basis

$$e_k = (0, \ldots, 0, 1, 0, \ldots)$$

where the 1 is in the kth position. Thus, V is the set of all infinite binary sequences with a finite number of 1's. Define the **order** $o(v)$ of any $v \in V$ to be the largest coordinate of v with value 1. Then $o(v) < \infty$ for all $v \in V$.

Consider the dual vectors e_k^*, defined (as usual) by

$$e_k^*(e_j) = \delta_{k,j}$$

For any $v \in V$, the evaluation functional $\overline{v}$ has the property that

$$\overline{v}(e_k^*) = e_k^*(v) = 0 \text{ if } k > o(v)$$

However, since the dual vectors e_k^* are linearly independent, there is a linear functional $f \in V^{**}$ for which

$$f(e_k^*) = 1$$

for all $k \geq 1$. Hence, f does not have the form $\overline{v}$ for any $v \in V$. This shows that the canonical map is not surjective and so V is not algebraically reflexive. □

Annihilators

The functions $f \in V^*$ are defined on vectors in V, but we may also define f on subsets M of V by letting

$$f(M) = \{f(v) \mid v \in M\}$$

Definition *Let M be a nonempty subset of a vector space V. The* **annihilator** *M^0 of M is*

$$M^0 = \{f \in V^* \mid f(M) = \{0\}\}$$ □

The term annihilator is quite descriptive, since M^0 consists of all linear functionals that *annihilate* (send to 0) every vector in M. It is not hard to see that M^0 is a subspace of V^*, even when M is not a subspace of V.

The basic properties of annihilators are contained in the following theorem, whose proof is left to the reader.

Theorem 3.15

1) (**Order-reversing**) *For any subsets M and N of V,*

$$M \subseteq N \Rightarrow N^0 \subseteq M^0$$

2) *If* $\dim(V) < \infty$ *then we have*

$$M^{00} \approx \text{span}(M)$$

under the natural map. In particular, if S *is a subspace of* V *then* $S^{00} \approx S$.

3) *If* $\dim(V) < \infty$ *and* S *and* T *are subspaces of* V *then*

$$(S \cap T)^0 = S^0 + T^0 \textit{ and } (S + T)^0 = S^0 \cap T^0 \qquad \square$$

Consider a direct sum decomposition

$$V = S \oplus T$$

Then any linear functional $f \in T^*$ can be extended to a linear functional $\overline{f}$ on V by setting $f(S) = 0$. Let us call this **extension by** 0. Clearly, $\overline{f} \in S^0$ and it is easy to see that the extension by 0 map $f \to \overline{f}$ is an isomorphism from T^* to S^0, whose inverse is restriction to T.

Theorem 3.16 *Let* $V = S \oplus T$.

a) *The extension by* 0 *map is an isomorphism from* T^* *to* S^0 *and so*

$$T^* \approx S^0$$

b) *If* V *is finite-dimensional then*

$$\dim(S^0) = \text{codim}_V(S) = \dim(V) - \dim(S) \qquad \square$$

Example 3.5 Part b) of Theorem 3.16 may fail in the infinite-dimensional case, since it may easily happen that $S^0 \approx V^*$. As an example, let V be the vector space over $\mathbb{Z}_2$ with a countably infinite ordered basis $\mathcal{B} = (e_1, e_2, \ldots)$. Let $S = \langle e_1 \rangle$ and $T = \langle e_2, e_3, \ldots \rangle$. It is easy to see that $S^0 \approx T^* \approx V^*$ and that $\dim(V^*) > \dim(V)$. $\square$

The annihilator provides a way to describe the dual space of a direct sum.

Theorem 3.17 *A linear functional on the direct sum* $V = S \oplus T$ *can be written as a direct sum of a linear functional that annihilates* S *and a linear functional that annihilates* T*, that is,*

$$(S \oplus T)^* = S^0 \oplus T^0$$

Proof. Clearly $S^0 \cap T^0 = \{0\}$, since any functional that annihilates both S and T must annihilate $S \oplus T = V$. Hence, the sum $S^0 + T^0$ is direct. If $f \in V^*$ then we can write

$$f = (f \circ \rho_T) + (f \circ \rho_S) \in S^0 \oplus T^0$$

and so $V = S^0 \oplus T^0$. $\square$

Operator Adjoints

If $\tau \in \mathcal{L}(V, W)$ then we may define a map $\tau^\times: W^* \to V^*$ by

$$\tau^\times(f) = f \circ \tau = f\tau$$

for $f \in W^*$. (We will write composition as juxtaposition.) Thus, for any $v \in V$

$$[\tau^\times(f)](v) = f(\tau(v))$$

The map $\tau^\times$ is called the **operator adjoint** of τ and can be described by the phrase "apply τ first."

Theorem 3.18 (Properties of the Operator Adjoint)

1) *For $\tau, \sigma \in \mathcal{L}(V, W)$ and $a, b \in F$*

$$(a\tau + b\sigma)^\times = a\tau^\times + b\sigma^\times$$

2) *For $\sigma \in \mathcal{L}(V, W)$ and $\tau \in \mathcal{L}(W, U)$*

$$(\tau\sigma)^\times = \sigma^\times\tau^\times$$

3) *For any invertible $\tau \in \mathcal{L}(V)$*

$$(\tau^{-1})^\times = (\tau^\times)^{-1}$$

Proof. Proof of part 1) is left for the reader. For part 2), we have for all $f \in U^*$

$$(\tau\sigma)^\times(f) = f(\tau\sigma) = \sigma^\times(f\tau) = \tau^\times(\sigma^\times(f)) = (\tau^\times\sigma^\times)(f)$$

Part 3) follows from part 2) and

$$\tau^\times(\tau^{-1})^\times = (\tau^{-1}\tau)^\times = \iota^\times = \iota$$

and in the same way, $(\tau^{-1})^\times\tau^\times = \iota$. Hence $(\tau^{-1})^\times = (\tau^\times)^{-1}$. □

If $\tau \in \mathcal{L}(V, W)$ then $\tau^\times \in \mathcal{L}(W^*, V^*)$ and so $\tau^{\times\times} \in \mathcal{L}(V^{**}, W^{**})$. Of course, $\tau^{\times\times}$ is not equal to τ. However, in the finite-dimensional case, if we use the natural maps to identify V^{**} with V and W^{**} with W then we can think of $\tau^{\times\times}$ as being in $\mathcal{L}(V, W)$. Using these identifications, we do have equality in the finite-dimensional case.

Theorem 3.19 *Let V and W be finite-dimensional and let $\tau \in \mathcal{L}(V, W)$. If we identify V^{**} with V and W^{**} with W using the natural maps then $\tau^{\times\times}$* is identified with τ.

Proof. For any $x \in V$ let the corresponding element of V^{**} be denoted by $\overline{x}$ and similarly for W. Then before making any identifications, we have for $v \in V$

$$\tau^{\times\times}(\overline{v})(f) = \overline{v}[\tau^\times(f)] = \overline{v}(f\tau) = f(\tau(v)) = \overline{\tau(v)}(f)$$

for all $f \in W^*$ and so

$$\tau^{\times\times}(\overline{v}) = \overline{\tau(v)} \in W^{**}$$

Therefore, using the canonical identifications for both V^{**} and W^{**} we have

$$\tau^{\times\times}(v) = \tau(v)$$

for all $v \in V$. □

The next result describes the kernel and image of the operator adjoint.

Theorem 3.20 *Let* $\tau \in \mathcal{L}(V,W)$. *Then*

1) $\ker(\tau^{\times}) = \operatorname{im}(\tau)^0$

2) $\operatorname{im}(\tau^{\times}) = \ker(\tau)^0$

Proof. For part 1),

$$\begin{aligned}\ker(\tau^{\times}) &= \{f \in W^* \mid \tau^{\times}(f) = 0\}\\ &= \{f \in W^* \mid f(\tau(V)) = \{0\}\}\\ &= \{f \in W^* \mid f(\operatorname{im}(\tau)) = \{0\}\}\\ &= \operatorname{im}(\tau)^0\end{aligned}$$

For part 2), if $f = g\tau = \tau^{\times}g \in \operatorname{im}(\tau^{\times})$ then $\ker(\tau) \subseteq \ker(f)$ and so $f \in \ker(\tau)^0$.

For the reverse inclusion, let $f \in \ker(\tau)^0 \subseteq V^*$. On $K = \ker(\tau)$, there is no problem since f and $\tau^{\times}g = g\tau$ agree on K for any $g \in W^*$. Let S be a complement of $\ker(\tau)$. Then τ maps a basis $\mathcal{B} = \{b_i \mid i \in I\}$ for S to a linearly independent set

$$\tau(\mathcal{B}) = \{\tau(b_i) \mid i \in I\}$$

in W and so we can define $g \in W^*$ any way we want on $\tau(\mathcal{B})$. In particular, let $g \in W^*$ be defined by setting

$$g(\tau(b_i)) = f(b_i)$$

and extending in any manner to all of W^*. Then $f = g\tau = \tau^{\times}g$ on $\mathcal{B}$ and therefore on S. Thus, $f = \tau^{\times}g \in \operatorname{im}(\tau^{\times})$. □

Corollary 3.21 *Let* $\tau \in \mathcal{L}(V,W)$*, where* V *and* W *are finite-dimensional. Then* $\operatorname{rk}(\tau) = \operatorname{rk}(\tau^{\times})$. □

In the finite-dimensional case, τ and $\tau^{\times}$ can both be represented by matrices. Let

$$\mathcal{B} = (b_1, \ldots, b_n) \text{ and } \mathcal{C} = (c_1, \ldots, c_m)$$

be ordered bases for V and W, respectively and let

$$\mathcal{B}^* = (b_1^*, \ldots, b_n^*) \text{ and } \mathcal{C}^* = (c_1^*, \ldots, c_m^*)$$

be the corresponding dual bases. Then

$$([\tau]_{\mathcal{B},\mathcal{C}})_{i,j} = ([\tau(b_j)]_{\mathcal{C}})_i = c_i^*[\tau(b_j)]$$

and

$$([\tau^\times]_{\mathcal{C}^*,\mathcal{B}^*})_{i,j} = ([\tau^\times(c_j^*)]_{\mathcal{B}^*})_i = b_i^{**}[\tau^\times(c_j^*)] = \tau^\times(c_j^*)(b_i) = c_j^*(\tau(b_i))$$

Comparing the last two expressions we see that they are the same except that the roles of i and j are reversed. Hence, the matrices in question are transposes.

Theorem 3.22 *Let $\tau \in \mathcal{L}(V,W)$, where V and W are finite-dimensional. If $\mathcal{B}$ and $\mathcal{C}$ are ordered bases for V and W, respectively and $\mathcal{B}^*$ and $\mathcal{C}^*$ are the corresponding dual bases then*

$$[\tau^\times]_{\mathcal{C}^*,\mathcal{B}^*} = ([\tau]_{\mathcal{B},\mathcal{C}})^t$$

In words, the matrices of τ and its operator adjoint $\tau^\times$ are transposes of one another. □

Exercises

1. If V is infinite-dimensional and S is an infinite-dimensional subspace, must the dimension of V/S be finite? Explain.
2. Prove the correspondence theorem.
3. Prove the first isomorphism theorem.
4. Complete the proof of Theorem 3.10.
5. Let S be a subspace of V. Starting with a basis $\{s_1, \ldots, s_k\}$ for S, how would you find a basis for V/S?
6. Use the first isomorphism theorem to prove the rank-plus-nullity theorem
$$\text{rk}(\tau) + \text{null}(\tau) = \dim(V)$$
for $\tau \in \mathcal{L}(V,W)$.
7. Let $\tau \in \mathcal{L}(V)$ and suppose that S is a subspace of V. Define a map $\tau' \colon V/S \to V/S$ by
$$\tau'(v+S) = \tau(v) + S$$
When is τ' well-defined? If τ' is well-defined, is it a linear transformation? What are $\text{im}(\tau')$ and $\ker(\tau')$?
8. Show that for any nonzero vector $v \in V$, there exists a linear functional $f \in V^*$ for which $f(v) \neq 0$.
9. Show that a vector $v \in V$ is zero if and only if $f(v) = 0$ for all $f \in V^*$.
10. Let S be a proper subspace of a finite-dimensional vector space V and let $v \in V \setminus S$. Show that there is a linear functional $f \in V^*$ for which $f(v) = 1$ and $f(s) = 0$ for all $s \in S$.

11. Find a vector space V and decompositions
$$V = A \oplus B = C \oplus D$$
with $A \approx C$ but $B \not\approx D$. Hence, $A \approx C$ does *not* imply that $A^c \approx C^c$.
12. Find isomorphic vectors spaces V and W with
$$V = S \oplus B \text{ and } W = S \oplus D$$
but $B \not\approx D$. Hence, $V \approx W$ does *not* imply that $V/S \not\approx W/S$.
13. Let V be a vector space with
$$V = S_1 \oplus T_1 = S_2 \oplus T_2$$
Prove that if S_1 and S_2 have finite codimension in V then so does $S_1 \cap S_2$ and
$$\text{codim}(S_1 \cap S_2) \leq \dim(T_1) + \dim(T_2)$$
14. Let V be a vector space with
$$V = S_1 \oplus T_1 = S_2 \oplus T_2$$
Suppose that S_1 and S_2 have finite codimension. Hence, by the previous exercise, so does $S_1 \cap S_2$. Find a direct sum decomposition $V = W \oplus X$ for which (1) W has finite codimension, (2) $W \subseteq S_1 \cap S_2$ and (3) $X \supseteq T_1 + T_2$.
15. Let $\mathcal{B}$ be a basis for an infinite-dimensional vector space V and define, for all $b \in \mathcal{B}$, the map $b' \in V^*$ by $b'(c) = 1$ if $c = b$ and 0 otherwise. Does $\{b' \mid b \in \mathcal{B}\}$ form a basis for V^*? What do you conclude about the concept of a dual basis?
16. Prove that $(S \oplus T)^* \approx S^* \oplus T^*$.
17. Prove that $0^\times = 0$ and $\iota^\times = \iota$ where 0 is the zero linear operator and ι is the identity.
18. Let S be a subspace of V. Prove that $(V/S)^* \approx S^0$.
19. Verify that
    a) $(\tau + \sigma)^\times = \tau^\times + \sigma^\times$ for $\tau, \sigma \in \mathcal{L}(V, W)$.
    b) $(r\tau)^\times = r\tau^\times$ for any $r \in F$ and $\tau \in \mathcal{L}(V, W)$
20. Let $\tau \in \mathcal{L}(V, W)$, where V and W are finite-dimensional. Prove that $\text{rk}(\tau) = \text{rk}(\tau^\times)$.
21. Prove that if $\sigma \in \mathcal{L}(V)$ has the property that
$$V = \text{im}(\sigma) \oplus \ker(\sigma) \text{ and } \sigma|_{\text{im}(\sigma)} = \iota$$
then σ is projection on $\text{im}(\sigma)$ along $\ker(\sigma)$.
22. a) Let $\rho: V \to S$ be projection onto a subspace S of V along a subspace T of V. Show that ρ is **idempotent**, that is $\rho^2 = \rho$.
    b) Prove that if $\rho \in \mathcal{L}(V)$ is idempotent then it is a projection.
    c) Is the adjoint of a projection also a projection?

Chapter 4
Modules I: Basic Properties

Motivation

Let V be a vector space over a field F and let $\tau \in \mathcal{L}(V)$. Then for any polynomial $p(x) \in F[x]$, the operator $p(\tau)$ is well-defined. For instance, if $p(x) = 1 + 2x + x^3$ then

$$p(\tau) = \iota + 2\tau + \tau^3$$

where ι is the identity operator and τ^3 is the threefold composition $\tau \circ \tau \circ \tau$.

Thus, using the operator τ we can define the product of a polynomial $p(x) \in F[x]$ and a vector $v \in V$ by

$$p(x)v = p(\tau)(v) \tag{4.1}$$

This product satisfies the usual properties of scalar multiplication, namely, for all $r(x),\ s(x) \in F[x]$ and $u, v \in V$,

$$\begin{aligned} r(x)(u+v) &= r(x)u + r(x)v \\ (r(x)+s(x))u &= r(x)u + s(x)u \\ [r(x)s(x)]u &= r(x)[s(x)u] \\ 1u &= u \end{aligned}$$

Thus, for a fixed $\tau \in \mathcal{L}(V)$, we can think of V as being endowed with the operations of (the usual) addition along with multiplication of an element of V by a *polynomial* in $F[x]$. However, since $F[x]$ is not a field, these two operations do not make V into a vector space. Nevertheless, the situation in which the scalars form a ring but not a field is extremely important, not only in our context but in many others.

Modules

Definition *Let R be a commutative ring with identity, whose elements are called* **scalars**. *An R-***module** *(or a* **module over** *R) is a nonempty set M,* *together with two operations. The first operation, called* **addition** *and denoted by* $+$*, assigns to each pair* $(u,v)\in M\times M$*, an element* $u+v\in M$*. The second operation, denoted by juxtaposition, assigns to each pair* $(r,u)\in R\times M$*, an element* $rv\in M$*. Furthermore, the following properties must hold:*

1) M *is an abelian group under addition.*
2) *For all* $r,s\in R$ *and* $u,v\in M$

$$\begin{aligned} r(u+v)&=ru+rv\\ (r+s)u&=ru+su\\ (rs)u&=r(su)\\ 1u&=u \end{aligned}$$

The ring R *is called the* **base ring** *of* M*.* □

Note that vector spaces are just special types of modules: a vector space is a module over a field.

When we turn in a later chapter to the study of the structure of a linear transformation $\tau\in\mathcal{L}(V)$, we will think of V as having the structure of a vector space over F as well as a module over $F[x]$. Put another way, V is an abelian group under addition, with two scalar multiplications—one whose scalars are elements of F and one whose scalars are polynomials over F. This viewpoint will be of tremendous benefit for the study of τ. For now, we concentrate only on modules.

Example 4.1

1) If R is a ring, the set R^n of all ordered n-tuples, whose components lie in R, is an R-module, with addition and scalar multiplication defined componentwise (just as in F^n),
$$(a_1,\dots,a_n)+(b_1,\dots,b_n)=(a_1+b_1,\dots,a_n+b_n)$$
and
$$r(a_1,\dots,a_n)=(ra_1,\dots,ra_n)$$
for a_i, b_i, $r\in R$. For example, $\mathbb{Z}^n$ is the $\mathbb{Z}$-module of all ordered n-tuples of integers.
2) If R is a ring, the set $\mathcal{M}_{m,n}(R)$ of all matrices of size $m\times n$, is an R-module, under the usual operations of matrix addition and scalar multiplication over R. Since R is a ring, we can also take the product of matrices in $\mathcal{M}_{m,n}(R)$. One important example is $R=F[x]$, whence $\mathcal{M}_{m,n}(F[x])$ is the $F[x]$-module of all $m\times n$ matrices whose entries are polynomials.
3) Any commutative ring R with identity is a module over itself, that is, R is an R-module. In this case, scalar multiplication is just multiplication by

elements of R, that is, scalar multiplication is the ring multiplication. The defining properties of a ring imply that the defining properties of an R-module are satisfied. We shall use this example many times in the sequel. □

Importance of the Base Ring

Our definition of a module requires that the ring R of scalars be commutative. Modules over noncommutative rings can exhibit quite a bit more unusual behavior than modules over commutative rings. Indeed, as one would expect, the general behavior of R-modules improves as we impose more structure on the base ring R. If we impose the very strict structure of a field, the result is the very well-behaved vector space structure.

To illustrate, if we allow the base ring R to be noncommutative then, as we will see, it is possible for an R-module to have bases of different sizes! Since modules over noncommutative rings will not be needed for the sequel, we require commutativity in the definition of module.

As another example, if the base ring is an integral domain then whenever $v_1, \ldots, v_n$ are linearly independent over R so are $rv_1, \ldots, rv_n$ for any nonzero $r \in R$. This fails when R is not an integral domain.

We will also consider the property on the base ring R that all of its ideals are finitely generated. In this case, any finitely generated R-module M has the desirable property that all of its submodules are also finitely generated. This property of R-modules fails if R does not have the stated property.

When R is a principal ideal domain (such as $\mathbb{Z}$ or $F[x]$), not only are all of its ideals finitely generated, but each is generated by a single element. In this case, the R-modules are "reasonably" well behaved. For instance, in general a module may have a basis but one or more of its submodules may not. However, if R is a principal ideal domain, this cannot happen.

Nevertheless, even when R is a principal ideal domain, R-modules are less well behaved than vector spaces. For example, there are modules over a principal ideal domain that do not have any linearly independent elements. Of course, such modules cannot have a basis.

Many of the basic concepts that we defined for vector spaces can also be defined for modules, although their properties are often quite different. We begin with submodules.

Submodules

The definition of submodule parallels that of subspace.

Definition *A* **submodule** *of an R-module M is a nonempty subset S of M that is an R-module in its own right, under the operations obtained by restricting the operations of M to S.* □

Theorem 4.1 *A nonempty subset S of an R-module M is a submodule if and only if it is closed under the taking of linear combinations, that is,*

$$r, s \in R, u, v \in S \Rightarrow ru + sv \in S \qquad \square$$

Theorem 4.2 *If S and T are submodules of M then $S \cap T$ and $S + T$* are also submodules of M. □

We have remarked that a commutative ring R with identity is a module over itself. As we will see, this type of module provides some good examples of non-vector space like behavior.

When we think of a ring R as an R-module rather than as a ring, multiplication is treated as *scalar* multiplication. This has some important implications. In particular, if S is a submodule of R then it is closed under scalar multiplication, which means that it is closed under multiplication by *all* elements of the ring R. In other words, S is an ideal of the ring R. Conversely, if $\mathcal{I}$ is an ideal of the ring R then $\mathcal{I}$ is also a submodule of the module R. Hence, *the submodules of the R-module R are precisely the ideals of the ring R.*

Spanning Sets

The concept of spanning set carries over to modules as well.

Definition *The* **submodule spanned** *(or* **generated***) by a subset S of a module M is the set of all* **linear combinations** *of elements of S:*

$$\langle S \rangle = \text{span}(S) = \{r_1v_1 + \cdots + r_nv_n \mid r_i \in R, v_i \in S, n \geq 1\}$$

A subset $S \subseteq M$ is said to **span** M or **generate** M if

$$M = \text{span}(S) \qquad \square$$

One very important point to note is that if a nontrivial linear combination of the elements $v_1, \ldots, v_n$ in an R-module M is 0,

$$r_1v_1 + \cdots + r_nv_n = 0$$

where not all of the coefficients are 0 then we *cannot* conclude, as we could in a vector space, that one of the elements v_i is a linear combination of the others. After all, this involves dividing by one of the coefficients, which may not be possible in a ring. For instance, for the $\mathbb{Z}$-module $\mathbb{Z} \times \mathbb{Z}$ we have

$$2(3,6) - 3(2,4) = (0,0)$$

but neither $(3,6)$ nor $(2,4)$ is an integer multiple of the other.

The following simple submodules play a special role in the theory.

Definition *Let M be an R-module. A submodule of the form*

$$\langle v\rangle = Rv = \{rv \mid r \in R\}$$

for $v \in M$ is called the **cyclic submodule** *generated by v.* □

Of course, any finite-dimensional vector space is the direct sum of cyclic submodules, that is, one-dimensional subspaces. One of our main goals is to show that a finitely generated module over a principal ideal domain has this property as well.

For reasons that will become clear soon, we need the following definition.

Definition *An R-module M is said to be* **finitely generated** *if it contains a finite set that generates M.* □

Of course, a vector space is finitely generated if and only if it has a finite basis, that is, if and only if it is finite-dimensional. For modules, life is more complicated. The following is an example of a finitely generated module that has a submodule that is not finitely generated.

Example 4.2 Let R be the ring $F[x_1, x_2, \ldots]$ of all polynomials in infinitely many variables over a field F. It will be convenient to use X to denote $x_1, x_2, \ldots$ and write a polynomial in R in the form $p(X)$. (Each polynomial in R, being a finite sum, involves only finitely many variables, however.) Then R is an R-module and as such, is finitely generated by the identity element $p(X) = 1$.

Now, consider the submodule S of all polynomials with zero constant term. This module is generated by the variables themselves,

$$S = \langle x_1, x_2, \ldots\rangle$$

However, S is not finitely generated. To see this, suppose that $G = \{p_1, \ldots, p_n\}$ is a finite generating set for S. Choose a variable x_k that does not appear in any of the polynomials in G. Then no linear combination of the polynomials in G can be equal to x_k. For if

$$x_k = \sum_{i=1}^{n} a_i(X)p_i(X)$$

then let $a_i(X) = x_k q_i(X) + r_i(X)$ where $r_i(X)$ does not involve x_k. This gives

$$\begin{aligned} x_k &= \sum_{i=1}^{n}[x_k q_i(X) + r_i(X)]p_i(X) \\ &= x_k \sum_{i=1}^{n} q_i(X)p_i(X) + \sum_{i=1}^{n} r_i(X)p_i(X) \end{aligned}$$

The last sum does not involve x_k and so it must equal 0. Hence, the first sum must equal 1, which is not possible since $p_i(X)$ has no constant term. □

Linear Independence

The concept of linear independence also carries over to modules.

Definition *A subset S of a module M is* **linearly independent** *if for any $v_1, \ldots, v_n \in S$ and $r_1, \ldots, r_n \in R$, we have*

$$r_1 v_1 + \cdots + r_n v_n = 0 \Rightarrow r_i = 0 \textit{ for all } i$$

A set S that is not linearly independent is **linearly dependent**. □

It is clear from the definition that any subset of a linearly independent set is linearly independent.

Recall that, in a vector space, a set S of vectors is linearly dependent if and only if some vector in S is a linear combination of the other vectors in S. For arbitrary modules, this is not true.

Example 4.3 Consider $\mathbb{Z}$ as a $\mathbb{Z}$-module. The elements $2, 3 \in \mathbb{Z}$ are linearly dependent, since

$$3(2) - 2(3) = 0$$

but neither one is a linear combination (i.e., integer multiple) of the other. □

The problem in the previous example (as noted earlier) is that

$$r_1 v_1 + \cdots + r_n v_n = 0$$

implies that

$$r_1 v_1 = -r_2 v_2 - \cdots - r_n v_n$$

but, in general, we cannot divide both sides by r_1, since it may not have a multiplicative inverse in the ring R.

Torsion Elements

In a vector space V over a field F, singleton sets $\{v\}$ where $v \neq 0$ are linearly independent. Put another way, $r \neq 0$ and $v \neq 0$ imply $rv \neq 0$. However, in a module, this need not be the case.

Example 4.4 The abelian group $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ is a $\mathbb{Z}$-module, with scalar multiplication defined by $za = (z \cdot a) \bmod n$, for all $n \in \mathbb{Z}$ and $a \in \mathbb{Z}_n$. However, since $na = 0$ for all $a \in \mathbb{Z}_n$, no singleton set $\{a\}$ is linearly independent. Indeed, $\mathbb{Z}_n$ has no linearly independent sets. □

This example motivates the following definition.

Definition *Let M be an R-module. A nonzero element $v \in M$ for which $rv = 0$ for some nonzero $r \in R$ is called a* **torsion element** *of M. A module that has no nonzero torsion elements is said to be* **torsion-free**. *If all elements of M are torsion elements then M is a* **torsion module**. *The set of all torsion elements of M, together with the zero element, is denoted by M_{tor}.* □

If M is a module over an *integral domain*, it is not hard to see that M_{tor} is a submodule of M and that M/M_{tor} is torsion-free. (We will define quotient modules shortly: they are defined in the same way as for vector spaces.)

Annihilators

Closely associated with the notion of a torsion element is that of an annihilator.

Definition *Let M be an R-module. The* **annihilator** *of an element $v \in M$ is*

$$\text{ann}(v) = \{r \in R \mid rv = 0\}$$

and the **annihilator** *of a submodule N of M is*

$$\text{ann}(N) = \{r \in R \mid rN = \{0\}\}$$

where $rN = \{rv \mid v \in N\}$. Annihilators are also called **order ideals**. □

It is easy to see that $\text{ann}(v)$ and $\text{ann}(N)$ are ideals of R. Clearly, $v \in M$ is a torsion element if and only if $\text{ann}(v) \neq \{0\}$.

Let $M = \langle u_1, \ldots, u_n \rangle$ be a finitely generated *torsion module* over an integral domain R. Then for each i there is a nonzero $a_i \in \text{ann}(u_i)$. Hence, the nonzero product $a = a_1 \cdots a_n$ annihilates each generator of M and therefore every element of M, that is, $a \in \text{ann}(M)$. This shows that $\text{ann}(M) \neq \{0\}$.

Free Modules

The definition of a basis for a module parallels that of a basis for a vector space.

Definition *Let M be an R-module. A subset $\mathcal{B}$ of M is a* **basis** *if $\mathcal{B}$ is linearly independent and spans M. An R-module M is said to be* **free** *if $M = \{0\}$ or if M has a basis. If $\mathcal{B}$ is a basis for M, we say that M is* **free on** $\mathcal{B}$. □

Theorem 4.3 *A subset $\mathcal{B}$ of a module M is a basis if and only if for every $v \in M$, there are* unique *elements $v_1, \dots, v_n \in \mathcal{B}$ and* unique *scalars $r_1, \dots, r_n \in R$ for which*

$$v = r_1 v_1 + \cdots + r_n v_n$$ □

In a vector space, a set of vectors is a basis if and only if it is a minimal spanning set, or equivalently, a maximal linearly independent set. For modules, the following is the best we can do in general. We leave proof to the reader.

Theorem 4.4 *Let $\mathcal{B}$ be a basis for an R-module M. Then*

1) $\mathcal{B}$ is a minimal spanning set.

2) $\mathcal{B}$ is a maximal linearly independent set. □

The $\mathbb{Z}$-module $\mathbb{Z}_n$ has no basis since it has no linearly independent sets. But since the entire module is a spanning set, we deduce that a minimal spanning set need not be a basis. In the exercises, the reader is asked to give an example of a module M that has a finite basis, but with the property that not every spanning set in M contains a basis and not every linearly independent set in M is contained in a basis. It follows in this case that a maximal linearly independent set need not be a basis.

The next example shows that even free modules are not very much like vector spaces. It is an example of a free module that has a submodule that is not free.

Example 4.5 The set $\mathbb{Z} \times \mathbb{Z}$ is a free module over itself, using componentwise scalar multiplication

$$(n, m)(a, b) = (na, mb)$$

with basis $\{(1, 1)\}$. But the submodule $\mathbb{Z} \times \{0\}$ is not free since it has no linearly independent elements and hence no basis. □

Homomorphisms

The term *linear transformation* is special to vector spaces. However, the concept applies to most algebraic structures.

Definition *Let M and N be R-modules. A function $\tau: M \to N$ is an R-***homomorphism** *if it preserves the module operations, that is,*

$$\tau(ru + sv) = r\tau(u) + s\tau(v)$$

for all $r, s \in R$ and $u, v \in M$. The set of all R-homomorphisms from M to N is denoted by $\operatorname{hom}_R(M, N)$. The following terms are also employed:

*1) An R-***endomorphism** *is an R-homomorphism from M to itself.*
*2) A R-***monomorphism** *or R-***embedding** *is an injective R-homomorphism.*
*3) An R-***epimorphism** *is a surjective R-homomorphism.*
*4) An R-***isomorphism** *is a bijective R-homomorphism.* □

It is easy to see that $\hom_R(M,N)$ is itself an R-module under addition of functions and scalar multiplication defined by

$$(r\tau)(v) = r(\tau(v)) = \tau(rv)$$

Theorem 4.5 *Let $\tau \in \hom_R(M,N)$. The kernel and image of τ, defined as for linear transformations by*

$$\ker(\tau) = \{v \in M \mid \tau(v) = 0\}$$

and

$$\operatorname{im}(\tau) = \{\tau(v) \mid v \in M\}$$

are submodules of M and N, respectively. Moreover, τ is a monomorphism if and only if $\ker(\tau) = \{0\}$. □

If N is a submodule of the R-module M then the map $j: N \to M$ defined by $j(v) = v$ is evidently an R-monomorphism, called **injection** of N into M.

Quotient Modules

The procedure for defining quotient modules is the same as that for defining quotient vector spaces. We summarize in the following theorem.

Theorem 4.6 *Let S be a submodule of an R-module M. The binary relation*

$$u \equiv v \Leftrightarrow u - v \in S$$

is an equivalence relation on M, whose equivalence classes are the **cosets**

$$v + S = \{v + s \mid s \in S\}$$

of S in M. The set M/S of all cosets of S in M, called the **quotient module** *of M* **modulo** *S, is an R-module under the well-defined operations*

$$\begin{aligned}(u+S)+(v+S) &= (u+v)+S\\ r(u+S) &= ru+S\end{aligned}$$

The zero element in M/S is the coset $0 + S = S$. □

One question that immediately comes to mind is whether a quotient space of a free module need be free. As the next example shows, the answer is no.

Example 4.6 As a module over itself, $\mathbb{Z}$ is free on the set $\{1\}$. For any $n > 0$, the set $\mathbb{Z}n = \{zn \mid z \in \mathbb{Z}\}$ is a free cyclic submodule of $\mathbb{Z}$, but the quotient $\mathbb{Z}$-

module $\mathbb{Z}/\mathbb{Z}n$ is isomorphic to $\mathbb{Z}_n$ via the map

$$\tau(u+\mathbb{Z}n) = u \bmod n$$

and since $\mathbb{Z}_n$ is not free as a $\mathbb{Z}$-module, neither is $\mathbb{Z}/\mathbb{Z}n$. $\square$

The Correspondence and Isomorphism Theorems

The correspondence and isomorphism theorems for vector spaces have analogs for modules.

Theorem 4.7 *(***The correspondence theorem***) Let S be a submodule of M. Then the function that assigns to each intermediate submodule $S \subseteq T \subseteq M$ the quotient submodule T/S of M/S is an order-preserving (with respect to set inclusion) one-to-one correspondence between submodules of M containing S and all submodules of M/S.* $\square$

Theorem 4.8 *(***The first isomorphism theorem***) Let $\tau\colon M \to N$ be an R-homomorphism. Then the map $\tau'\colon M/\mathrm{ker}(\tau) \to N$ defined by*

$$\tau'(v+\mathrm{ker}(\tau)) = \tau(v)$$

is an R-embedding and so

$$\frac{M}{\mathrm{ker}(\tau)} \approx \mathrm{im}(\tau) \qquad \square$$

Theorem 4.9 *(***The second isomorphism theorem***) Let M be an R-module and let S and T be submodules of M. Then*

$$\frac{S+T}{T} \approx \frac{S}{S\cap T} \qquad \square$$

Theorem 4.10 *(***The third isomorphism theorem***) Let M be an R-module and suppose that $S \subseteq T$ are submodules of M. Then*

$$\frac{M/S}{T/S} \approx \frac{M}{T} \qquad \square$$

Direct Sums and Direct Summands

The definition of direct sum is the same for modules as for vector spaces. We will confine our attention to the direct sum of a finite number of modules.

Definition *An R-module M is the* **direct sum** *of the submodules $S_1,\dots,S_n$, written*

$$M = S_1 \oplus \cdots \oplus S_n$$

if every $v \in M$ can be written, in a unique way *(except for order), as a sum of* *one element from each of the submodules S_i, that is, there are unique $u_i \in S_i$ for which*

$$v = u_1 + \cdots + u_n$$

In this case, each S_i is called a **direct summand** *of M. If $M = S \oplus T$ then S is said to be* **complemented** *and T is called a* **complement** *of S in M.* □

Note that a sum is direct if and only if whenever $u_1 + \cdots + u_n = 0$ where $u_i \in S_i$ then $u_i = 0$ for all i, that is, if and only if 0 has a unique representation as a sum of vectors from distinct submodules.

As with vector spaces, we have the following useful characterization of direct sums.

Theorem 4.11 *A module M is the direct sum of submodules $S_1, \ldots, S_n$ if and only if*
1) $M = S_1 + \cdots + S_n$
2) For each $i = 1, \ldots, n$

$$S_i \cap \Big(\sum_{j \neq i} S_j\Big) = \{0\}$$ □

In the case of vector spaces, every subspace is a direct summand, that is, every subspace has a complement. However, as the next example shows, this is not true for modules.

Example 4.7 The set $\mathbb{Z}$ of integers is a $\mathbb{Z}$-module. Since the submodules of $\mathbb{Z}$ are precisely the ideals of the ring $\mathbb{Z}$ and since $\mathbb{Z}$ is a principal ideal domain, the submodules of $\mathbb{Z}$ are the sets

$$\langle n \rangle = \mathbb{Z}n = \{zn \mid z \in \mathbb{Z}\}$$

Hence, any two nonzero proper submodules of $\mathbb{Z}$ have nonzero intersection, for if $n \neq m > 0$ then

$$\mathbb{Z}n \cap \mathbb{Z}m = \mathbb{Z}k$$

where $k = \text{lcm}\{n, m\}$. It follows that the only complemented submodules of $\mathbb{Z}$ are $\mathbb{Z}$ and $\{0\}$. □

In the case of vector spaces, there is an intimate connection between subspaces and quotient spaces, as we saw in Theorem 3.6. The problem we face in generalizing this to modules in general is that not all submodules have a complement. However, this is the only problem.

Theorem 4.12 *Let S be a complemented submodule of M. All complements of S are isomorphic to M/S and hence to each other.*

Proof. For any complement T of S, the first isomorphism theorem applied to the projection $\rho\colon M \to T$ onto T along S gives $T \approx M/S$. □

Direct Summands and One-Sided Invertibility

The next theorem characterizes direct summands, but first a definition.

Definition *A submodule S of an R-module M is a* **retract** *of M* **by** *an R-homomorphism $\tau\colon M \to S$ if τ fixes each element of S, that is, $\tau(s) = s$ for all $s \in S$.* □

Note that the homomorphism τ in the definition of a retract is similar to a projection map onto S, in that both types of maps are the identity when restricted to S. In fact, if S is complemented and T is a complement of S then S is a retract of M by the projection onto S along T. Indeed, more can be said.

Theorem 4.13 *A submodule S of the R-module M is complemented if and only if it is a retract of M. In this case, if S is a retract of M by τ then τ is projection onto S along* $\ker(\tau)$ *and so*

$$M = S \oplus \ker(\tau) = \operatorname{im}(\tau) \oplus \ker(\tau)$$

Proof. If $M = S \oplus T$ then S is a retract of M by the projection map $\rho_S\colon M \to S$. Conversely, if $\tau\colon M \to S$ is an R-homomorphism that fixes S then clearly τ is surjective and $S = \operatorname{im}(\tau)$. Also,

$$v \in S \cap \ker(\tau) \Rightarrow v = \tau(v) = 0$$

and for any $v \in M$ we have

$$v = [v - \tau(v)] + \tau(v) \in \ker(\tau) + S$$

Hence, $M = S \oplus \ker(\tau)$. □

Definition *Let $\tau\colon A \to B$ be a module homomorphism. Then a* **left inverse** *of τ is a module homomorphism $\tau_L\colon B \to A$ for which $\tau_L \circ \tau = \iota$. A* **right inverse** *of τ is a module homomorphism $\tau_R\colon B \to A$ for which $\tau \circ \tau_R = \iota$.* □

It is easy to see that in order for τ to have a left inverse τ_L, it must be injective since

$$\tau(a) = \tau(b) \Rightarrow \tau_L \circ \tau(a) = \tau_L \circ \tau(b) \Rightarrow a = b$$

and in order for τ to have a right inverse τ_R, it must be surjective, since if $b \in B$ then $b = \tau[\tau_R(b)] \in \operatorname{im}(\tau)$.

Now, if we were dealing with functions between sets, then the converses of these statements would hold: τ is left-invertible if and only if it is injective and

τ is right-invertible if and only if it is surjective. However, for modules things are more complicated.

Theorem 4.14

1) *An R-homomorphism $\tau: A \to B$ has a left inverse τ_L if and only if it is injective and* $\mathrm{im}(\tau)$ *is a direct summand of B, in which case*
$$B = \mathrm{im}(\tau) \oplus \ker(\tau_L) \approx \mathrm{im}(\tau_L) \oplus \ker(\tau_L)$$
2) *An R-homomorphism $\tau: A \to B$ has a right inverse τ_R if and only if it is surjective and* $\ker(\tau)$ *is a direct summand of B, in which case*
$$A = \ker(\tau) \oplus \mathrm{im}(\tau_R) \approx \ker(\tau) \oplus \mathrm{im}(\tau)$$

Proof. For part 1), suppose first that $\tau_L \tau = \iota_A$. Then τ is injective since applying τ_L to the expression $\tau(x) = \tau(y)$ gives $x = y$. Also, $x \in \mathrm{im}(\tau) \cap \ker(\tau_L)$ implies that $x = \tau(a)$ and
$$0 = \tau_L(x) = \tau_L(\tau(a)) = a$$
and so $x = 0$. Hence, the direct sum $\mathrm{im}(\tau) \oplus \ker(\tau_L)$ exists. For any $b \in B$, we can write
$$b = \tau\tau_L(b) + [b - \tau\tau_L(b)]$$
where $\tau\tau_L(b) \in \mathrm{im}(\tau)$ and $b - \tau\tau_L(b) \in \ker(\tau_L)$ and so $B = \mathrm{im}(\tau) \oplus \ker(\tau_L)$.

Conversely, if τ is injective and $B = K \oplus \mathrm{im}(\tau)$ for some submodule K then let
$$\tau_L = \tau^{-1} \circ \rho_{\mathrm{im}(\tau)}$$
where $\rho_{\mathrm{im}(\tau)}$ is projection onto $\mathrm{im}(A)$. This is well-defined since $\tau: A \to \mathrm{im}(\tau)$ is an isomorphism. Then
$$\tau_L \circ \tau(b) = \tau^{-1} \circ \rho_{\mathrm{im}(\tau)} \circ \tau(b) = \tau^{-1} \circ \tau(b) = b$$
and so τ_L is a left-inverse of τ. It is clear that $K = \ker(\tau_L)$ and since $\tau_L: \mathrm{im}(\tau) \to \mathrm{im}(\tau_L)$ is injective, $\mathrm{im}(\tau) \approx \mathrm{im}(\tau_L)$.

For part 2), if τ has a right inverse τ_R then τ_R has a left inverse τ and so τ_R must be injective and
$$A = \mathrm{im}(\tau_R) \oplus \ker(\tau) \approx \mathrm{im}(\tau) \oplus \ker(\tau)$$

Conversely, suppose that $A = X \oplus \ker(\tau)$. Since the elements of different cosets of $\ker(\tau)$ are mapped to different elements of B, it is natural to define $\tau_R: B \to A$ by taking $\tau_R(b)$ to be a particular element in the coset $b + \ker(\tau)$ that is sent to b by τ. However, in general we cannot pick just any element of the coset and expect to get a module morphism. (We get a right inverse but only as a set function.)

However, the condition $A = X \oplus \ker(\tau)$ is precisely what we need, because it says that the elements of the *submodule* X form a set of distinct coset representatives; that is, each $x \in X$ belongs to exactly one coset and each coset contains exactly one element of X.

In addition, if $\tau(x_1) = b_1$ and $\tau(x_2) = b_2$ for $x_1, x_2 \in X$ then

$$\tau(rx_1 + sx_2) = r\tau(x_1) + s\tau(x_2) = rb_1 + sb_2$$

Thus, we can define τ_R as follows. For any $b \in B$ there is a unique $x \in X$ for which $\tau(x) = b$. Let $\tau_R(b) = x$. Then we have

$$\tau(\tau_R(b)) = \tau(x) = b$$

and so $\tau \circ \tau_R = \iota_B$. Also,

$$\tau_R(rb_1 + sb_2) = rx_1 + sx_2 = r\tau_R(b_1) + s\tau_R(b_2)$$

and so τ_R is a module morphism. Thus τ is right-invertible. □

The last part of the previous theorem is worth further comment. Recall that if $\tau: V \to W$ is a linear transformation on vector spaces then

$$V \approx \ker(\tau) \oplus \operatorname{im}(\tau)$$

This does not hold in general for modules, but it does hold if $\ker(\tau)$ is a direct summand.

Modules Are Not As Nice As Vector Spaces

Here is a list of some of the properties of modules (over commutative rings with identity) that emphasize the differences between modules and vector spaces.

1) A submodule of a module need not have a complement.
2) A submodule of a finitely generated module need not be finitely generated.
3) There exist modules with no linearly independent elements and hence with no basis.
4) A minimal spanning set or maximal linearly independent set is not necessarily a basis.
5) There exist free modules with submodules that are not free.
6) There exist free modules with linearly independent sets that are not contained in a basis and spanning sets that do not contain a basis.

Recall also that a module over a *noncommutative* ring may have bases of different sizes. However, all bases for a free module over a commutative ring with identity have the same size, as we will prove in the next chapter.

Exercises

1. Give the details to show that any commutative ring with identity is a module over itself.
2. Let $S = \{v_1, \ldots, v_n\}$ be a subset of a module M. Prove that $N = \langle S \rangle$ is the *smallest* submodule of M containing S. First you will need to formulate precisely what it means to be the smallest submodule of M containing S.
3. Let M be an R-module and let I be an ideal in R. Let IM be the set of all finite sums of the form

$$r_1 v_1 + \cdots + r_n v_n$$

where $r_i \in I$ and $v_i \in M$. Is IM a submodule of M?
4. Show that if S and T are submodules of M then (with respect to set inclusion)

$$S \cap T = \text{glb}\{S, T\} \text{ and } S + T = \text{lub}\{S, T\}$$

5. Let $S_1 \subseteq S_2 \subseteq \cdots$ be an ascending sequence of submodules of an R-module M. Prove that the union $\bigcup S_i$ is a submodule of M.
6. Give an example of a module M that has a finite basis but with the property that not every spanning set in M contains a basis and not every linearly independent set in M is contained in a basis.
7. Show that, just as in the case of vector spaces, an R-homomorphism can be defined by assigning arbitrary values on the elements of a basis and extending by linearity.
8. Let $\tau \in \text{hom}_R(M, N)$ be an R-isomorphism. If $\mathcal{B}$ is a basis for M, prove that $\tau(\mathcal{B}) = \{\tau(b) \mid b \in \mathcal{B}\}$ is a basis for N.
9. Let M be an R-module and let $\tau \in \text{hom}_R(M, M)$ be an R-endomorphism. If τ is **idempotent**, that is, if $\tau^2 = \tau$ show that

$$M = \ker(\tau) \oplus \text{im}(\tau)$$

Does the converse hold?
10. Consider the ring $R = F[x, y]$ of polynomials in two variables. Show that the set M consisting of all polynomials in R that have zero constant term is an R-module. Show that M is not a free R-module.
11. Prove that R is an integral domain if and only if all R-modules M have the following property: If $v_1, \ldots, v_n$ is linearly independent over R then so is $rv_1, \ldots, rv_n$ for any nonzero $r \in R$.
12. Prove that if a commutative ring R with identity has the property that every finitely generated R-module is free then R is a field.
13. Let M and N be R-modules. If S is a submodule of M and T is a submodule of N show that

$$\frac{M \oplus N}{S \oplus T} \approx \frac{M}{S} \oplus \frac{N}{T}$$

14. If R is a commutative ring with identity and $\mathcal{I}$ is an ideal of R then $\mathcal{I}$ is an R-module. What is the maximum size of a linearly independent set in $\mathcal{I}$? Under what conditions is $\mathcal{I}$ free?
15. a) Show that for any module M over an integral domain the set M_{tor} of all torsion elements in a module M is a submodule of M.
    b) Find an example of a ring R with the property that for some R-module M the set M_{tor} is not a submodule.
    c) Show that for any module M over an integral domain, the quotient module M/M_{tor} is torsion-free.
16. Fix a prime p and let

$$M = \left\{ \frac{a}{p^k} \mid a, k \in \mathbb{Z}, k \geq 0 \right\}$$

Show that M is a $\mathbb{Z}$-module and that the set $E = \{1/p^k \mid k \geq 0\}$ is a minimal spanning set for M. Is the set E linearly independent?

17. Let N be an abelian group together with a scalar multiplication over a ring R that satisfies all of the properties of an R-module except that $1v$ does not necessarily equal v for all $v \in N$. Show that N can be written as a direct sum of an R-module N_0 and another "pseudo R-module" N_1.
18. Prove that $\hom_R(M, N)$ is an R-module under addition of functions and scalar multiplication defined by

$$(r\tau)(v) = r(\tau(v)) = \tau(rv)$$

19. Prove that any R-module M is isomorphic to the R-module $\hom_R(R, M)$.
20. Let R and S be commutative rings with identity and let $f\colon R \to S$ be a ring homomorphism. Show that any S-module is also an R-module under the scalar multiplication

$$rv = f(r)v$$

21. Prove that $\hom_{\mathbb{Z}}(\mathbb{Z}_n, \mathbb{Z}_m) \approx \mathbb{Z}_d$ where $d = \gcd(n, m)$.
22. Suppose that R is a commutative ring with identity. If $\mathcal{I}$ and $\mathcal{J}$ are ideals of R for which $R/\mathcal{I} \approx R/\mathcal{J}$ as R-modules then prove that $\mathcal{I} = \mathcal{J}$. Is the result true if $R/\mathcal{I} \approx R/\mathcal{J}$ as rings?

Chapter 5
Modules II: Free and Noetherian Modules

The Rank of a Free Module

Since all bases for a vector space V have the same cardinality, the concept of vector space dimension is well-defined. A similar statement holds for free R-modules when the base ring is commutative (but not otherwise).

Theorem 5.1 *Let M be a free module over a commutative ring R with identity.*
1) Then any two bases of M have the same cardinality.
2) The cardinality of a spanning set is greater than or equal to that of a basis.
Proof. The plan is to find a vector space V with the property that, for any basis for M, there is a basis of the same cardinality for V. Then we can appeal to the corresponding result for vector spaces.

Let $\mathcal{I}$ be a maximal ideal of R, which exists by Theorem 0.22. Then $R/\mathcal{I}$ is a field. Our first thought might be that M is a vector space over $R/\mathcal{I}$ but that is not the case. In fact, scalar multiplication using the field $R/\mathcal{I}$

$$(r+\mathcal{I})v = rv$$

is not even well-defined, since this would require that $\mathcal{I}M=\{0\}$. On the other hand, we can fix precisely this problem by factoring out the submodule

$$\mathcal{I}M=\{a_1v_1+\cdots+a_nv_n \mid a_i\in\mathcal{I}, v_i\in M\}$$

Indeed, $M/\mathcal{I}M$ is a vector space over $R/\mathcal{I}$, with scalar multiplication defined by

$$(r+\mathcal{I})(u+\mathcal{I}M)=ru+\mathcal{I}M$$

To see that this is well-defined, we must show that the conditions

$$\begin{aligned} r+\mathcal{I}&=r'+\mathcal{I}\\ u+\mathcal{I}M&=u'+\mathcal{I}M \end{aligned}$$

imply

$$ru + \mathcal{I}M = r'u' + \mathcal{I}M$$

But this follows from the fact that

$$ru - r'u' = r(u - u') + (r - r')u' \in \mathcal{I}M$$

Hence, scalar multiplication is well-defined. We leave it to the reader to show that $M/\mathcal{I}M$ is a vector space over $R/\mathcal{I}$.

Consider now a set $\mathcal{B} = \{b_i \mid i \in I\} \subseteq M$ and the corresponding set

$$\mathcal{B} + \mathcal{I}M = \{b_i + \mathcal{I}M \mid i \in I\} \subseteq \frac{M}{\mathcal{I}M}$$

If $\mathcal{B}$ spans M over R then $\mathcal{B} + \mathcal{I}M$ spans $M/\mathcal{I}M$ over $R/\mathcal{I}$. To see this, note that any $v \in M$ has the form $v = \Sigma r_i b_i$ for $r_i \in R$ and so

$$v + \mathcal{I}M = \left(\sum r_i b_i\right) + \mathcal{I}M = \sum r_i(b_i + \mathcal{I}M) = \sum (r_i + \mathcal{I})(b_i + \mathcal{I}M)$$

which shows that $\mathcal{B} + \mathcal{I}M$ spans $M/\mathcal{I}M$.

Now suppose that $\mathcal{B} = \{b_i \mid i \in I\}$ is a basis for M over R. We claim that $\mathcal{B} + \mathcal{I}M$ is a basis for $M/\mathcal{I}M$ over $R/\mathcal{I}$. We have seen that $\mathcal{B} + \mathcal{I}M$ spans $M/\mathcal{I}M$. Also, if

$$\sum (r_i + \mathcal{I})(b_i + \mathcal{I}M) = \mathcal{I}M$$

then $\sum r_j b_j \in \mathcal{I}M$ and so

$$\sum_{j=1}^{n} r_j b_j = \sum_{j=1}^{m} a_j b_j$$

where $a_j \in \mathcal{I}$. From the linear independence of $\mathcal{B}$ we deduce that $r_i \in \mathcal{I}$ for all i and so $r_i + \mathcal{I} = \mathcal{I}$. Hence $\mathcal{B} + \mathcal{I}M$ is linearly independent and therefore a basis, as desired.

To see that $|\mathcal{B}| = |\mathcal{B} + \mathcal{I}M|$, note that if $b_i + \mathcal{I}M = b_k + \mathcal{I}M$ then

$$b_i - b_k = \sum_{j=1}^{m} a_j b_j$$

where $a_j \in \mathcal{I}$. If $b_i \neq b_k$ then $1 = a_i \in \mathcal{I}$, which is not possible since $\mathcal{I}$ is a maximal ideal. Hence, $b_i = b_k$.

Thus, if $\mathcal{B}$ is a basis for M over R then

$$|\mathcal{B}| = |\mathcal{B} + \mathcal{I}M| = \dim_{R/\mathcal{I}}(M/\mathcal{I}M)$$

and so all bases for M over R have the same cardinality, which proves part 1). Moreover, if $\mathcal{B}$ spans M over R then $\mathcal{B}+\mathcal{I}M$ spans $M/\mathcal{I}M$ and so

$$\dim_{R/\mathcal{I}}(M/\mathcal{I}M) \leq |\mathcal{B}+\mathcal{I}M| \leq |\mathcal{B}|$$

Thus, $\mathcal{B}$ has cardinality at least as great as that of any basis for M over R. □

The previous theorem allows us to define the *rank* of a free module. (The term *dimension* is not used for modules in general.)

Definition *Let R be a commutative ring with identity. The* **rank** $\operatorname{rk}(M)$ *of a nonzero free R-module M is the cardinality of any basis for M. The rank of the trivial module $\{0\}$ is* 0. □

Theorem 5.1 fails if the underlying ring of scalars is not commutative. The next example describes a module over a noncommutative ring that has the remarkable property of possessing a basis of size n for any positive integer n.

Example 5.1 Let V be a vector space over F with a countably infinite basis $\mathcal{B}=\{b_1,b_2,\ldots\}$. Let $\mathcal{L}(V)$ be the ring of linear operators on V. Observe that $\mathcal{L}(V)$ is not commutative, since composition of functions is not commutative.

The ring $\mathcal{L}(V)$ is an $\mathcal{L}(V)$-module and as such, the identity map ι forms a basis for $\mathcal{L}(V)$. However, we can also construct a basis for $\mathcal{L}(V)$ of any desired finite size n. To understand the idea, consider the case $n=2$ and define the operators β_1 and β_2 by

$$\beta_1(b_{2k})=b_k, \beta_1(b_{2k+1})=0$$

and

$$\beta_2(b_{2k})=0, \beta_2(b_{2k+1})=b_k$$

These operators are linearly independent essentially because they are surjective and their supports are disjoint. In particular, if

$$f\beta_1+g\beta_2=0$$

then

$$0=(f\beta_1+g\beta_2)(b_{2k})=f(b_k)$$

and

$$0=(f\beta_1+g\beta_2)(b_{2k+1})=g(b_k)$$

which shows that $f=0$ and $g=0$. Moreover, if $h\in\mathcal{L}(V)$ then we define f and g by

$$f(b_k) = h(b_{2k})$$
$$g(b_k) = h(b_{2k+1})$$

from which it follows easily that

$$h = f\beta_1 + g\beta_2$$

which shows that $\{\beta_1, \beta_2\}$ is a basis for $\mathcal{L}(V)$.

More generally, we begin by partitioning $\mathcal{B}$ into n blocks. For each $s = 0, \ldots, n-1$, let

$$\mathcal{B}_s = \{b_i \mid i \equiv s \operatorname{mod} n\}$$

Now we define elements $\beta_s \in \mathcal{L}(V)$ by

$$\beta_s(b_{kn+t}) = \delta_{t,s} b_k$$

where $0 \le t < n$ and where $\delta_{t,s}$ is the Kronecker delta function. These functions are surjective and have disjoint support. It follows that $\mathcal{C}_n = \{\beta_0, \ldots, \beta_{n-1}\}$ is linearly independent. For if $\alpha_s \in \mathcal{L}(V)$ and

$$0 = \alpha_0\beta_0 + \cdots + \alpha_{n-1}\beta_{n-1}$$

then, applying this to b_{kn+t} gives

$$0 = \alpha_t\beta_t(b_{kn+t}) = \alpha_t(b_k)$$

for all k. Hence, $\alpha_t = 0$.

Also, $\mathcal{C}_n$ spans $\mathcal{L}(V)$, for if $\tau \in \mathcal{L}(V)$, we define $\alpha_s \in \mathcal{L}(V)$ by

$$\alpha_s(b_k) = \tau(b_{kn+s})$$

to get

$$(\alpha_0\beta_0 + \cdots + \alpha_{n-1}\beta_{n-1})(b_{kn+t}) = \alpha_t\beta_t(b_{kn+t}) = \alpha_t(b_k) = \tau(b_{kn+t})$$

and so

$$\tau = \alpha_0\beta_0 + \cdots + \alpha_{n-1}\beta_{n-1}$$

Thus, $\mathcal{C}_n = \{\beta_0, \ldots, \beta_{n-1}\}$ is a basis for $\mathcal{L}(V)$ of size n. □

We have spoken about the cardinality of minimal spanning sets. Let us now speak about the cardinality of maximal linearly independent sets.

Theorem 5.2 *Let R be an integral domain and let M be a free R-module of finite rank n. Then all linearly independent sets have size at most* $\operatorname{rk}(M)$.

Proof. Since $M \approx R^n$ if we prove the result for R^n it will hold for M. Let R^+ be the field of quotients of R. Then R^n is a subset of the vector space $(R^+)^n$ and

1) R^n is a subgroup of $(R^+)^n$ under addition
2) scalar multiplication by elements of R is defined and R^n is an R-module,
3) scalar multiplication by elements of R^+ is defined but R^n is not closed under this scalar multiplication.

Now, if $\mathcal{B} = \{v_1, \ldots, v_k\} \subseteq R^n$ is linearly dependent over R^+ then $\mathcal{B}$ is clearly linearly dependent over R. Conversely, suppose that $\mathcal{B}$ is linearly independent over R and

$$\frac{r_1}{s_1}v_1 + \cdots + \frac{r_k}{s_k}v_k = 0$$

where $s_i \neq 0$ for all i and $r_j \neq 0$ for some j. Multiplying by $s = s_1 \cdots s_k \neq 0$ produces a nontrivial linear dependency over R

$$\frac{s}{s_1}r_1v_1 + \cdots + \frac{s}{s_k}r_kv_k = 0$$

which implies that $r_i = 0$ for all i. Thus $\mathcal{B}$ is linearly dependent over R if and only if it is linearly dependent over R^+. Of course, in the vector space $(R^+)^n$ all sets of size $n+1$ or larger are linearly dependent over R^+ and hence all subsets of R^n of size $n+1$ or larger are linearly dependent over R. □

Recall that if B is a basis for a vector space V over F then V is isomorphic to the vector space $(F^B)_0$ of all functions from B to F that have finite support. A similar result holds for free R-modules. We begin with the fact that $(R^B)_0$ is a free R-module. The simple proof is left to the reader.

Theorem 5.3 *Let B be any set and let R be a ring. The set $(R^B)_0$ of all functions from B to R that have finite support is a free R-module of rank $|B|$ with basis $\mathcal{B} = \{\delta_b\}$ where*

$$\delta_b(x) = \begin{cases} 1 & \text{if } x = b \\ 0 & \text{if } x \neq b \end{cases}$$

This basis is referred to as the **standard basis** for $(R^B)_0$. □

Theorem 5.4 *Let M be an R-module. If B is a basis for M then M is isomorphic to $(R^B)_0$.*

Proof. Consider the map $\tau: M \to (R^B)_0$ defined by setting

$$\tau(b) = \delta_b$$

where δ_b is defined in Theorem 5.3 and extending this to all of M by linearity, that is,

$$\tau(r_1b_1+\cdots+r_nb_n)=r_1\delta_{b_1}+\cdots+r_n\delta_{b_n}$$

Since τ maps a basis for M to a basis $\mathcal{B}=\{\delta_b\}$ for $(R^B)_0$, it follows that τ is an isomorphism from M to $(R^B)_0$. □

Theorem 5.5 *Two free R-modules (over a commutative ring) are isomorphic if and only if they have the same rank.*

Proof. If $M\approx N$ then any isomorphism τ from M to N maps a basis for M to a basis for N. Since τ is a bijection, we have $\text{rk}(M)=\text{rk}(N)$. Conversely, suppose that $\text{rk}(M)=\text{rk}(N)$. Let $\mathcal{B}$ be a basis for M and let $\mathcal{C}$ be a basis for N. Since $|\mathcal{B}|=|\mathcal{C}|$, there is a bijective map $\tau\colon\mathcal{B}\to\mathcal{C}$. This map can be extended by linearity to an isomorphism of M onto N and so $M\approx N$. □

Free Modules and Epimorphisms

Homomorphic images that are free have some behavior reminiscent of vector spaces.

Theorem 5.6

1) *If $\sigma\colon M\to F$ is a surjective R-homomorphism and F is free then* $\ker(\sigma)$ *is complemented and*

$$M=\ker(\sigma)\oplus N\approx\ker(\sigma)\boxplus F$$

where $N\approx F$.

2) *If S is a submodule of M and if M/S is free then S is complemented and*

$$M\approx S\boxplus\frac{M}{S}$$

If in addition, M,S and M/S are free then

$$\text{rk}(M)=\text{rk}(S)+\text{rk}\left(\frac{M}{S}\right)$$

and if the ranks are all finite then

$$\text{rk}\left(\frac{M}{S}\right)=\text{rk}(M)-\text{rk}(S)$$

Proof. For part 1), we prove that σ is right-invertible. Let $\mathcal{B}=\{v_i\mid i\in I\}$ be a basis for F. Define $\sigma_R\colon F\to B$ by setting $\sigma_R(v_i)$ equal to any member of the nonempty set $\sigma^{-1}(v_i)$ and extending σ_R to an R-homomorphism. Then σ_R is a right inverse of σ and so Theorem 4.14 implies that $\ker(\sigma)$ is a direct summand of M and $M\approx\ker(\sigma)\boxplus F$. Part 2) follows from part 1), where $\sigma=\pi_S$ is projection onto S. □

Noetherian Modules

One of the most desirable properties of a finitely generated R-module M is that all of its submodules be finitely generated. Example 4.2 shows that this is not always the case and leads us to search for conditions on the ring R that will guarantee that all submodules of a finitely generated module are themselves finitely generated.

Definition *An R-module M is said to satisfy the* **ascending chain condition** *on submodules* (abbreviated a.c.c.) *if any ascending sequence of submodules*

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots$$

of M is eventually constant, that is, there exists an index k for which

$$S_k = S_{k+1} = S_{k+2} = \cdots$$

Modules with the ascending chain condition on submodules are also called **noetherian modules** *(after Emmy Noether, one of the pioneers of module theory).* □

Theorem 5.7 *An R-module M is noetherian if and only if every submodule of M is finitely generated.*

Proof. Suppose that all submodules of M are finitely generated and that M contains an infinite ascending sequence

$$S_1 \subseteq S_2 \subseteq S_3 \subseteq \cdots \tag{5.1}$$

of submodules. Then the union

$$S = \bigcup_j S_j$$

is easily seen to be a submodule of M. Hence, S is finitely generated, say $S = \langle u_1, \ldots, u_n \rangle$. Since $u_i \in S$, there exists an index k_i such that $u_i \in S_{k_i}$. Therefore, if $k = \max\{k_1, \ldots, k_n\}$, we have

$$\{u_1, \ldots, u_n\} \subseteq S_k$$

and so

$$S = \langle u_1, \ldots, u_n \rangle \subseteq S_k \subseteq S_{k+1} \subseteq S_{k+2} \subseteq \cdots \subseteq S$$

which shows that the chain (5.1) is eventually constant.

For the converse, suppose that M satisfies the a.c.c on submodules and let S be a submodule of M. Pick $u_1 \in S$ and consider the submodule $S_1 = \langle u_1 \rangle \subseteq S$ generated by u_1. If $S_1 = S$ then S is finitely generated. If $S_1 \neq S$ then there is a $u_2 \in S - S_1$. Now let $S_2 = \langle u_1, u_2 \rangle$. If $S_2 = S$ then S is finitely generated. If $S_2 \neq S$ then pick $u_3 \in S - S_2$ and consider the submodule $S_3 = \langle u_1, u_2, u_3 \rangle$.

Continuing in this way, we get an ascending chain of submodules

$$\langle u_1\rangle \subseteq \langle u_1,u_2\rangle \subseteq \langle u_1,u_2,u_3\rangle \subseteq \cdots \subseteq S$$

If none of these submodules is equal to S, we would have an infinite ascending chain of submodules, each properly contained in the next, which contradicts the fact that M satisfies the a.c.c. on submodules. Hence, $S=\langle u_1,\ldots,u_n\rangle$, for some n and so S is finitely generated. □

Since a ring R is a module over itself and since the submodules of the module R are precisely the ideals of the ring R, the preceding discussion may be formulated for rings as follows.

Definition *A ring R is said to satisfy the* **ascending chain condition** *on ideals if any ascending sequence*

$$\mathcal{I}_1 \subseteq \mathcal{I}_2 \subseteq \mathcal{I}_3 \subseteq \cdots$$

of ideals of R is eventually constant, that is, there exists an index k for which

$$\mathcal{I}_k = \mathcal{I}_{k+1} = \mathcal{I}_{k+2} = \cdots$$

A ring that satisfies the ascending chain condition on ideals is called a **noetherian ring**. □

Theorem 5.8 *A ring R is noetherian if and only if every ideal of R is finitely generated.* □

Note that a ring R is noetherian *as a ring* if and only if it is noetherian *as a module* over itself. More generally, a ring R is noetherian if and only if every finitely generated R-module is noetherian.

Theorem 5.9 *Let R be a commutative ring with identity.*

1) *R is noetherian if and only if every finitely generated R-module is noetherian.*
2) *If, in addition, R is a principal ideal domain then if M is generated by n elements any submodule of M is generated by at most n elements.*

Proof. For part 1), one direction is evident. Assume that R is noetherian and let $M=\langle u_1,\ldots,u_n\rangle$ be a finitely generated R-module. Consider the epimorphism $\tau\colon R^n \to M$ defined by

$$\tau(r_1,\ldots,r_n)=r_1u_1+\cdots+r_nu_n$$

Let S be a submodule of M. Then

$$\tau^{-1}(S)=\{u\in R^n \mid \tau(u)\in S\}$$

is a submodule of R^n and $\tau(\tau^{-1}(S))=S$. If every submodule of R^n is finitely generated, then $\tau^{-1}(S)$ is finitely generated and so $\tau^{-1}(S)=\langle v_1,\ldots,v_k\rangle$. Then

S is finitely generated by $\{\tau(v_1),\dots,\tau(v_k)\}$. Hence, it is sufficient to show that every submodule M of R^n is finitely generated. We proceed by induction on n.

If $n=1$, then M is an ideal of R and is thus finitely generated by assumption. Assume that every submodule of R^k is finitely generated for all $1\le k<n$ and let S be a submodule of R^n.

If $n>1$, we can extract from S something that is isomorphic to an ideal of R and so will be finitely generated. In particular, let S_1 be the "last coordinates" in S, specifically, let

$$S_1=\{(0,\dots,0,a_n)\mid(a_1,\dots,a_{n-1},a_n)\in S\text{ for some }a_1,\dots,a_{n-1}\in R\}$$

The set S_1 is isomorphic to an ideal of R and is therefore finitely generated, say $S_1=\langle\mathcal{G}_1\rangle$, where $\mathcal{G}_1=\{g_1,\dots,g_k\}$ is a finite subset of S_1.

Also, let

$$S_2=\{v\in S\mid v=(a_1,\dots,a_{n-1},0)\text{ for some }a_1,\dots,a_{n-1}\in R\}$$

be the set of all elements of S that have last coordinate equal to 0. Note that S_2 is a nonempty submodule of R^n and is isomorphic to a submodule of R^{n-1}. Hence, the inductive hypothesis implies that S_2 is finitely generated, say $S_2=\langle\mathcal{G}_2\rangle$, where $\mathcal{G}_2$ is a finite subset of S.

By definition of S_1, each $g_i\in\mathcal{G}_1$ has the form

$$g_i=(0,\dots,0,g_{i,n})$$

for $g_{i,n}\in R$ where there is a $\overline{g}_i\in S$ of the form

$$\overline{g}_i=(g_{i,1},\dots,g_{i,n-1},g_{i,n})$$

Let $\overline{\mathcal{G}}_1=\{\overline{g}_1,\dots,\overline{g}_k\}$. We claim that S is generated by the finite set $\overline{\mathcal{G}}_1\cup\mathcal{G}_2$.

To see this, let $v=(a_1,\dots,a_n)\in S$. Then $(0,\dots,0,a_n)\in S_1$ and so

$$(0,\dots,0,a_n)=\sum_{i=1}^{k}r_ig_i$$

for $r_i\in R$. Consider now the sum

$$w=\sum_{i=1}^{k}r_i\overline{g}_i\in\langle\overline{\mathcal{G}}_1\rangle$$

The last coordinate of this sum is

$$\sum_{i=1}^{k} r_i g_{i,n} = a_n$$

and so the difference $v - w$ has last coordinate 0 and is thus in $S_2 = \langle \mathcal{G}_2 \rangle$. Hence

$$v = (v - w) + w \in \langle \overline{\mathcal{G}}_1 \rangle + \langle \mathcal{G}_2 \rangle = \langle \overline{\mathcal{G}}_1 \cup \mathcal{G}_2 \rangle$$

as desired.

For part 2), we leave it to the reader to review the proof and make the necessary changes. The key fact is that S_1 is isomorphic to an ideal of R, which is principal. Hence, S_1 is generated by a single element of M. □

The Hilbert Basis Theorem

Theorem 5.9 naturally leads us to ask which familiar rings are noetherian. The following famous theorem describes one very important case.

Theorem 5.10 *(***Hilbert basis theorem***) If a ring R is noetherian then so is the polynomial ring $R[x]$.*
Proof. We wish to show that any ideal $\mathcal{I}$ in $R[x]$ is finitely generated. Let L denote the set of all leading coefficients of polynomials in $\mathcal{I}$, together with the 0 element of R. Then L is an ideal of R.

To see this, observe that if $\alpha \in L$ is the leading coefficient of $f(x) \in \mathcal{I}$ and if $r \in R$ then either $r\alpha = 0$ or else $r\alpha$ is the leading coefficient of $rf(x) \in \mathcal{I}$. In either case, $r\alpha \in L$. Similarly, suppose that $\beta \in L$ is the leading coefficient of $g(x) \in \mathcal{I}$. We may assume that $\deg f(x) = i$ and $\deg g(x) = j$, with $i \leq j$. Then $h(x) = x^{j-i} f(x)$ is in $\mathcal{I}$, has leading coefficient α and has the same degree as $g(x)$. Hence, $\alpha - \beta$ is either 0 or it is the leading coefficient of $h(x) - g(x) \in \mathcal{I}$. In either case $\alpha - \beta \in L$.

Since L is an ideal of the noetherian ring R, it must be finitely generated, say $L = \langle a_1, \ldots, a_m \rangle$. Since $a_i \in L$, there exist polynomials $f_i(x) \in \mathcal{I}$ with leading coefficient a_i. By multiplying each $f_i(x)$ by a suitable power of x, we may assume that

$$\deg f_i(x) = d = \max\{\deg f_i(x)\}$$

for all $i = 1, \ldots, m$.

Now for $k = 0, \ldots, d - 1$ let L_k be the set of all leading coefficients of polynomials in $\mathcal{I}$ of degree k, together with the 0 element of R. A similar argument shows that L_k is an ideal of R and so L_k is also finitely generated. Hence, we can find polynomials $P_k = \{p_{k,1}(x), \ldots, p_{k,n_k}(x)\}$ in $\mathcal{I}$ whose leading coefficients constitute a generating set for L_k.

Consider now the finite set

$$P = \left(\bigcup_{k=0}^{d-1} P_k \right) \cup \{f_1(x), \ldots, f_m(x)\}$$

If $\mathcal{J}$ is the ideal generated by P then $\mathcal{J} \subseteq \mathcal{I}$. An induction argument can be used to show that $\mathcal{J} = \mathcal{I}$. If $g(x) \in \mathcal{I}$ has degree 0 then it is a linear combination of the elements of P_0 (which are constants) and is thus in $\mathcal{J}$. Assume that any polynomial in $\mathcal{I}$ of degree less than k is in $\mathcal{J}$ and let $g(x) \in \mathcal{I}$ have degree k.

If $k < d$ then some linear combination $h(x)$ over R of the polynomials in P_k has the same leading coefficient as $g(x)$ and if $k \geq d$ then some linear combination $h(x)$ of the polynomials

$$\left\{x^{k-d} f_1(x), \ldots, x^{k-d} f_m(x)\right\} \subseteq \mathcal{J}$$

has the same leading coefficient as $g(x)$. In either case, there is a polynomial $h(x) \in \mathcal{J}$ that has the same leading coefficient as $g(x)$. Since $g(x) - h(x) \in \mathcal{I}$ has degree strictly smaller than that of $g(x)$ the induction hypothesis implies that

$$g(x) - h(x) \in \mathcal{J}$$

and so

$$g(x) = [g(x) - h(x)] + h(x) \in \mathcal{J}$$

This completes the induction and shows that $\mathcal{I} = \mathcal{J}$ is finitely generated. □

Exercises

1. If M is a free R-module and $\tau: M \to N$ is an epimorphism then must N also be free?
2. Let $\mathcal{I}$ be an ideal of R. Prove that if $R/\mathcal{I}$ is a free R-module then $\mathcal{I}$ is the zero ideal.
3. Prove that the union of an ascending chain of submodules is a submodule.
4. Let S be a submodule of an R-module M. Show that if M is finitely generated, so is the quotient module M/S.
5. Let S be a submodule of an R-module. Show that if both S and M/S are finitely generated then so is M.
6. Show that an R-module M satisfies the a.c.c. for submodules if and only if the following condition holds. Every nonempty collection $\mathcal{S}$ of submodules of M has a maximal element. That is, for every nonempty collection $\mathcal{S}$ of submodules of M there is an $S \in \mathcal{S}$ with the property that $T \in \mathcal{S} \Rightarrow T \subseteq S$.
7. Let $\tau: M \to N$ be an R-homomorphism.
   a) Show that if M is finitely generated then so is $\operatorname{im}(\tau)$.

 b) Show that if $\ker(\tau)$ and $\operatorname{im}(\tau)$ are finitely generated then $M=\ker(\tau)+S$ where S is a finitely generated submodule of M. Hence, M is finitely generated.

8. If R is noetherian and $\mathcal{I}$ is an ideal of R show that $R/\mathcal{I}$ is also noetherian.
9. Prove that if R is noetherian then so is $R[x_1,\ldots,x_n]$.
10. Find an example of a commutative ring with identity that does not satisfy the ascending chain condition.
11. a) Prove that an R-module M is cyclic if and only if it is isomorphic to $R/\mathcal{I}$ where $\mathcal{I}$ is an ideal of R.
 b) Prove that an R-module M is **simple** ($M\neq\{0\}$ and M has no proper nonzero submodules) if and only if it is isomorphic to $R/\mathcal{I}$ where $\mathcal{I}$ is a maximal ideal of R.
 c) Prove that for any nonzero commutative ring R with identity, a simple R-module exists.
12. Prove that the condition that R be a principal ideal domain in part 2) of Theorem 5.9 is required.
13. Prove Theorem 5.9 in the following way.
 a) Show that if $T\subseteq S$ are submodules of M and if T and S/T are finitely generated then so is S.
 b) The proof is again by induction. Assuming it true for any module generated by n elements, let $M=\langle v_1,\ldots,v_{n+1}\rangle$ and let $M'=\langle v_1,\ldots,v_n\rangle$. Then let $T=S\cap M'$ in part a).
14. Prove that any R-module M is isomorphic to the quotient of a free module F. If M is finitely generated then F can also be taken to be finitely generated.
15. Prove that if S and T are isomorphic submodules of a module M it does not necessarily follow that the quotient modules M/S and M/T are isomorphic. Prove also that if $S\oplus T_1\approx S\oplus T_2$ as modules it does not necessarily follow that $T_1\approx T_2$. Prove that these statements do hold if all modules are free and have finite rank.

Chapter 6
Modules over a Principal Ideal Domain

We remind the reader of a few of the basic properties of principal ideal domains.

Theorem 6.1 *Let R be a principal ideal domain.*

1) *An element $r \in R$ is irreducible if and only if the ideal $\langle r \rangle$ is maximal.*
2) *An element in R is prime if and only if it is irreducible.*
3) *R is a unique factorization domain.*
4) *R satisfies the ascending chain condition on ideals. Hence, so does any finitely generated R-module M. Moreover, if M is generated by n elements any submodule of M is generated by at most n elements.*

Annihilators and Orders

When R is a principal ideal domain all annihilators are generated by a single element. This permits the following definition.

Definition *Let R be a principal ideal domain and let M be an R-module, with submodule N. Any generator of* $\operatorname{ann}(N)$ *is called an* **order** *of N. An* **order** *of an element $v \in M$ is an order of the submodule $\langle v \rangle$.* □

Note that any two orders μ and ν of N (or of an element $v \in M$) are associates, since $\langle \mu \rangle = \langle \nu \rangle$. Hence, an order of N is uniquely determined up to multiplication by a unit of R. For this reason, we may occasionally abuse the terminology and refer to "the" order of an element or submodule.

Also, if $A \subseteq B \subseteq M$ are submodules of M then $\operatorname{ann}(B) \subseteq \operatorname{ann}(A)$ and so any order of A divides any order of B. Thus, just as with finite groups, the order of an element/submodule divides the order of the module.

Cyclic Modules

The simplest type of nonzero module is clearly a cyclic module. Despite their simplicity, cyclic modules are extremely important and so we want to explore some of their basic properties.

Theorem 6.2 *Let R be a principal ideal domain.*

1) *If $\langle v\rangle$ is a cyclic R-module with $\operatorname{ann}(v)=\langle\alpha\rangle$ then the map $\tau\colon R\to\langle v\rangle$ defined by $\tau(r)=rv$ is a surjective R-homomorphism with kernel $\langle\alpha\rangle$. Hence*

$$\langle v\rangle\approx\frac{R}{\langle\alpha\rangle}$$

In other words, cyclic R-modules are isomorphic to quotient modules of the base ring R. If α is a prime then $\langle\alpha\rangle$ is a maximal ideal in R and so $R/\langle\alpha\rangle$ is a field.

2) *Any submodule of a cyclic R-module is cyclic.*
3) *Let $\langle v\rangle$ be a cyclic submodule of M of order α. Then $\langle\beta v\rangle$ has order $\alpha/\gcd(\alpha,\beta)$. Hence, if β and α are relatively prime then $\langle\beta v\rangle$ also has order α.*
4) *If $u_1,\ldots,u_n$ are nonzero elements of M with orders $\alpha_1,\ldots,\alpha_n$ that are pairwise relatively prime, then the sum*

$$v=u_1+\cdots+u_n$$

has order $\mu=\alpha_1\cdots\alpha_n$. Consequently, if M is an R-module and

$$M=A_1+\cdots+A_n$$

where the submodules A_i have orders α_i that are pairwise relatively prime, then the sum is direct.

Proof. We leave proof of part 1) as an exercise. Part 2) follows from part 2) of Theorem 5.9. For part 3), we first consider the two extremes: when β is relatively prime to α and when $\beta\mid\alpha$. As to the first, let β and α be relatively prime. If $\gamma(\beta v)=0$ then $\alpha\mid\gamma\beta$. Hence, $(\alpha,\beta)=1$ implies that $\alpha\mid\gamma$ and so $\alpha=\alpha/\gcd(\alpha,\beta)$ is an order of βv.

Next, if $\alpha=\beta d$ then $d(\beta v)=0$ and so any order $o(\beta v)$ of βv divides d. But if $o(\beta v)$ properly divides d then $\epsilon\beta$ properly divides $d\beta=\alpha$ and yet annihilates v, which contradicts the fact that α is an order of v. Hence $o(\beta v)$ and d are associates and so $d=\alpha/\beta=\alpha/\gcd(\alpha,\beta)$ is an order of βv.

Now we can combine these two extremes to finish the proof. Write $\beta=d(\beta/d)$ where $d=\gcd(\beta,\alpha)$ divides α and β/d is relatively prime to α. Using the previous results, we find that $(\beta/d)v$ has order α and so $\beta v=d(\beta/d)v$ has order α/d.

For part 4), since μ annihilates v, the order of v divides μ. If the order of v is a proper divisor of μ then for some index k, there is a prime p dividing α_k for which μ/p annihilates v. But μ/p annihilates each u_i for $i \neq k$ and so

$$0 = \frac{\mu}{p}v = \frac{\mu}{p}u_k = \frac{\alpha_k}{p}\left(\frac{\mu}{\alpha_k}\right)u_k$$

However, α_k and μ/α_k are relatively prime and so the order of $(\mu/\alpha_k)u_k$ is equal to the order of u_k, which contradicts the equation above. Hence, the order of v is μ. Finally, to see that the sum above is direct, note that if

$$v_1 + \cdots + v_n = 0$$

where $v_i \in A_i$ then each v_i must be 0, for otherwise the order of the sum on the left would be different from 1. □

Free Modules over a Principal Ideal Domain

Example 4.5 showed that a submodule of a free module need not be free. (The submodule $\mathbb{Z} \times \{0\}$ of $\mathbb{Z} \times \mathbb{Z}$ is not free.) However, if R is a principal ideal domain this cannot happen.

Theorem 6.3 *Let M be a free module over a principal ideal domain R. Then any submodule S of M is also free and* $\text{rk}(S) \leq \text{rk}(M)$.
Proof. We will give the proof only for modules of finite rank, although the theorem is true for all free modules. Thus, since $M \approx R^n$ where $n = \text{rk}(M)$ we may in fact assume that $M = R^n$. Our plan is to proceed by induction on n.

For $n = 1$, we have $M = R$ and any submodule S of R is just an ideal of R. If $S = \{0\}$ then S is free by definition. Otherwise, $S = \langle a \rangle$ for some $a \neq 0$. But since R is an integral domain, we have $ra \neq 0$ for all $r \neq 0$ and so $\{a\}$ is a basis for S. Thus, S is free and $\text{rk}(S) = 1 = \text{rk}(M)$.

Now assume that if $k < n$ then any submodule S of R^k is free and $\text{rk}(S) \leq k$. Let S be a submodule of R^n. Let

$$S_1 = \{v \in S \mid v = (a_1, \ldots, a_{n-1}, 0) \text{ for some } a_1, \ldots, a_{n-1} \in R\}$$

and

$$S_2 = \{(0, \ldots, 0, a_n) \mid (a_1, \ldots, a_{n-1}, a_n) \in S \text{ for some } a_1, \ldots, a_{n-1} \in R\}$$

Note that S_1 and S_2 are nonempty.

Since S_1 is isomorphic to a submodule of R^{n-1}, the inductive hypothesis implies that S_1 is free. Let $\mathcal{B}$ be a basis for S_1 with $|\mathcal{B}| \leq n - 1$. If $S_1 = \{0\}$ then take $\mathcal{B} = \emptyset$.

Now, S_2 is isomorphic to a submodule (ideal) of R and is therefore also free of rank at most 1. If $S_2 = \{0\}$ then all elements of S have zero final coordinate, which means that $S = S_1$, which is free with rank at most n, as desired. So assume that S_2 is not trivial and let $\{g\}$ be a basis for S_2 where

$$g = (0, \ldots, 0, r)$$

for $0 \neq r \in R$. Let $\overline{g} \in S$ satisfy

$$\overline{g} = (r_1, \ldots, r_{n-1}, r)$$

We claim that $\mathcal{B} \cup \{\overline{g}\}$ is a basis for S. To see that $\mathcal{B} \cup \{\overline{g}\}$ generates S let $v = (a_1, \ldots, a_n) \in S$. Then $(0, \ldots, 0, a_n) \in S_2$ and so

$$(0, \ldots, 0, a_n) = sg = (0, \ldots, 0, sr)$$

for $s \in R$. Thus $a_n = sr$ and

$$s\overline{g} = (sr_1, \ldots, sr_{n-1}, sr) = (sr_1, \ldots, sr_{n-1}, a_n)$$

Hence the difference $v - s\overline{g}$ is in $S_1 = \langle \mathcal{B} \rangle$. We then have

$$v = (v - s\overline{g}) + s\overline{g} \in \langle \mathcal{B} \rangle + \langle \overline{g} \rangle = \langle \mathcal{B} \cup \{\overline{g}\} \rangle$$

and so $\mathcal{B} \cup \{\overline{g}\}$ generates S. Finally, to see that $\mathcal{B} \cup \{\overline{g}\}$ is linearly independent, note that if $\mathcal{B} = \{v_1, \ldots, v_k\}$ and if

$$a_1 v_1 + \cdots + a_{k-1} v_k + a\overline{g} = 0$$

then comparing nth coefficients gives $ar = 0$. Since R is an integral domain and $r \neq 0$ we deduce that $a = 0$. It follows that $a_i = 0$ for all i. Thus $\mathcal{B} \cup \{\overline{g}\}$ is a basis for S and the proof is complete. □

If V is a vector space of dimension n then any set of n linearly independent vectors in V is a basis for V. This fails for modules. For example, $\mathbb{Z}$ is a $\mathbb{Z}$-module of rank 1 but the independent set $\{2\}$ is not a basis. On the other hand, the fact that a spanning set of size n is a basis does hold for modules over a principal ideal domain, as we now show.

Theorem 6.4 *Let M be a free R-module of rank n, where R is a principal ideal domain. Let $S = \{s_1, \ldots, s_n\}$ be a spanning set for M. Then S is a basis for M.*

Proof. Let $\mathcal{B} = \{b_1, \ldots, b_n\}$ be a basis for M and define the map $\tau: M \to M$ by $\tau(b_i) = s_i$ and extending to a surjective R-homomorphism. Since M is free, Theorem 5.6 implies that

$$M \approx \ker(\tau) \boxplus \operatorname{im}(\tau) = \ker(\tau) \boxplus M$$

Since $\ker(\tau)$ is a submodule of the free module and since R is a principal ideal domain, we know that $\ker(\tau)$ is free of rank at most n. It follows that

$$\text{rk}(M) = \text{rk}(\ker(\tau)) + \text{rk}(M)$$

and so $\text{rk}(\ker(\tau)) = 0$, that is, $\ker(\tau) = \{0\}$, which implies that τ is an R-isomorphism and so S is a basis. □

In general, a basis for a submodule of a free module over a principal ideal domain cannot be extended to a basis for the entire module. For example, the set $\{2\}$ is a basis for the submodule $2\mathbb{Z}$ of the $\mathbb{Z}$-module $\mathbb{Z}$, but this set cannot be extended to a basis for $\mathbb{Z}$ itself. We state without proof the following result along these lines.

Theorem 6.5 *Let M be a free R-module of rank n, where R is a principal ideal domain. Let N be a submodule of M that is free of rank $k \le n$. Then there is a basis $\mathcal{B}$ for M that contains a subset $S = \{v_1, \dots, v_k\}$ for which $\{r_1 v_1, \dots, r_k v_k\}$ is a basis for N, for some nonzero elements $r_1, \dots, r_k$ of R.* □

Torsion-Free and Free Modules

Let us explore the relationship between the concepts of torsion-free and free. It is not hard to see that any free module over an integral domain is torsion-free. The converse does not hold, unless we strengthen the hypotheses by requiring that the module be finitely generated.

Theorem 6.6 *Let M be a torsion-free finitely generated module over a principal ideal domain R. Then M is free. Thus, a finitely generated module over a principal ideal domain is free if and only if it is torsion free.*
Proof. Let $G = \{v_1, \dots, v_n\}$ be a generating set for M. Consider first the case $n = 1$, whence $G = \{v\}$. Then G is a basis for M since singleton sets are linearly independent in a torson-free module. Hence, M is free.

Now suppose that $G = \{u, v\}$ is a generating set with $u, v \ne 0$. If G is linearly independent, we are done. If not, then there exist nonzero $r, s \in R$ for which $ru = sv$. It follows that $sM = s\langle u, v\rangle \subseteq \langle u\rangle$ and so sM is a submodule of a free module and is therefore free by Theorem 6.3. But the map $\tau\colon M \to sM$ defined by $\tau(v) = sv$ is an isomorphism because M is torsion-free. Thus M is also free.

Now we can do the general case. Write

$$G = \{u_1, \dots, u_k, v_1, \dots, v_{n-k}\}$$

where $S = \{u_1, \dots, u_k\}$ is a maximal linearly independent subset of G. (Note that S is nonempty because singleton sets are linearly independent.)

For each v_i, the set $\{u_1, \dots, u_k, v_i\}$ is linearly dependent and so there exist $a_i \in R$ and $r_1, \dots, r_k \in R$ for which

$$a_i v_i + r_1 u_1 + \cdots + r_k u_k = 0$$

If $a = a_1 \cdots a_{n-k}$ then

$$aM = a\langle u_1, \ldots, u_k, v_1, \ldots, v_{n-k} \rangle \subseteq \langle u_1, \ldots, u_k \rangle$$

and since the latter is a free module, so is aM, and therefore so is M. □

Prelude to Decomposition: Cyclic Modules

The following result shows how cyclic modules can be composed and decomposed.

Theorem 6.7 *Let M be an R-module.*

1) *If $u_1, \ldots, u_n$ are nonzero elements of M with orders $\alpha_1, \ldots, \alpha_n$ that are pairwise relatively prime, then*

$$\langle u_1 + \cdots + u_n \rangle = \langle u_1 \rangle \oplus \cdots \oplus \langle u_n \rangle$$

2) *If $v \in M$ has order $\mu = \alpha_1 \cdots \alpha_n$ where $\alpha_1, \ldots, \alpha_n$ are pairwise relatively prime, then v can be written in the form*

$$v = u_1 + \cdots + u_n$$

where u_i has order α_i. Moreover,

$$\langle v \rangle = \langle u_1 \rangle \oplus \cdots \oplus \langle u_n \rangle$$

Proof. According to Theorem 6.2, the order of v is μ and the sum on the right is direct. It is clear that $\langle u_1 + \cdots + u_n \rangle \subseteq \langle u_1 \rangle \oplus \cdots \oplus \langle u_n \rangle$. For the reverse inclusion, since α_1 and μ/α_1 are relatively prime, there exist $r, s \in R$ for which

$$r\alpha_1 + s\frac{\mu}{\alpha_1} = 1$$

Hence

$$u_1 = \left(r\alpha_1 + s\frac{\mu}{\alpha_1} \right) u_1 = s\frac{\mu}{\alpha_1} u_1 = s\frac{\mu}{\alpha_1}(u_1 + \cdots + u_n) \in \langle u_1 + \cdots + u_n \rangle$$

Similarly, $u_k \in \langle u_1 + \cdots + u_n \rangle$ for all k and so we get the reverse inclusion.

For part 2), the scalars $\beta_k = \mu/\alpha_k$ are relatively prime and so there exist $a_i \in R$ for which

$$a_1\beta_1 + \cdots + a_n\beta_n = 1$$

Hence,

$$v = (a_1\beta_1 + \cdots + a_n\beta_n)v = a_1\beta_1 v + \cdots + a_n\beta_n v$$

Since the order of $a_k\beta_k v$ divides α_k, these orders are pairwise relatively prime.

Hence, the order of the sum on the right is the product of the orders of the terms and so $a_k\beta_k v$ must have order α_k. The second statement follows from part 1). □

The First Decomposition

The first step in the decomposition of a finitely generated module M over a principal ideal domain R is an easy one.

Theorem 6.8 *Any finitely generated module M over a principal ideal domain R is the direct sum of a free R-module and a torsion R-module*

$$M = M_{\text{free}} \oplus M_{\text{tor}}$$

As to uniqueness, the torsion part M_{tor} is unique (it must be the set of all torsion elements of M) whereas the free part M_{free} is not unique. However, all possible free summands are isomorphic and thus have the same rank.

Proof. As to existence, the set M_{tor} of all torsion elements is easily seen to be a submodule of M. Since M is finitely generated, so is the torsion-free quotient module M/M_{tor}. Hence, according to Theorem 6.6, M/M_{tor} is free. Consider now the canonical projection $\pi: M \to M/M_{\text{tor}}$ onto M_{tor}. Since M/M_{tor} is free, Theorem 5.6 implies that

$$M = M_{\text{tor}} \oplus F$$

where $F \approx M/M_{\text{tor}}$ is free.

As to uniqueness, suppose that $M = T \oplus G$ where T is torsion and G is free. Then $T \subseteq M_{\text{tor}}$. But if $v \in M_{\text{tor}}$ and $v = t + g$ where $t \in T$ and $g \in G$ then $av = 0$ and $bt = 0$ for some nonzero $a, b \in R$ and so $(ab)g = 0$, which implies that $g = 0$, that is, $v \in T$. Thus, $T = M_{\text{tor}}$.

For the free part, since $M = M_{\text{tor}} \oplus F = M_{\text{tor}} \oplus G$, the submodules F and G are both complements of M_{tor} and hence are isomorphic. Hence, all free summands are isomorphic and therefore have the same rank. □

Note that if $\{w_1, \ldots, w_m\}$ is a basis for M_{free} we can write

$$M = \langle w_1 \rangle \oplus \cdots \oplus \langle w_m \rangle \oplus M_{\text{tor}}$$

where each cyclic submodule $\langle w_i \rangle$ has zero annihilator. This is a partial decomposition of M into a direct sum of cyclic submodules.

A Look Ahead

So now we turn our attention to the decomposition of finitely generated torsion modules M over a principal ideal domain. We will develop two decompositions. One decomposition has the form

$$M = \langle v_1 \rangle \oplus \cdots \oplus \langle v_n \rangle$$

where the annihilators of the cyclic submodules form an ascending chain

$$\text{ann}(\langle v_1 \rangle) \subseteq \cdots \subseteq \text{ann}(\langle v_n \rangle)$$

This decomposition is called an *invariant factor decomposition* of M.

Although we will not approach it in quite this manner, the second decomposition can be obtained by further decomposing each cyclic submodule in the invariant factor decomposition into cyclic submodules whose annihilators have the form $\langle p^e \rangle$ where p is a prime. Submodules with annihilators of this form are called *primary submodules* and so the second decomposition is referred to as a *primary cyclic decomposition*.

Our plan will be to derive the primary cyclic decomposition first and then obtain the invariant factor decomposition from the primary cyclic decomposition by a piecing-together process, as described in Theorem 6.7.

As we will see, while neither of these decompositions is unique, the sequences of annihilators are unique, that is, these sequences are completely determined by the module M.

The Primary Decomposition

The first step in the primary cyclic decomposition is to decompose the torsion module into a direct sum of primary submodules.

Definition *Let p be a prime in R. A p-***primary** *(or just* **primary***) module is a module whose order is a power of p.* □

Note that a p-primary module M with order p^k must have an element of order p^k.

Theorem 6.9 *(***The primary decomposition theorem***) Let M be a nonzero torsion module over a principal ideal domain R, with order*

$$\mu = p_1^{e_1} \cdots p_n^{e_n}$$

where the p_i's are distinct nonassociate primes in R.

1) Then M is the direct sum

$$M = M_{p_1} \oplus \cdots \oplus M_{p_n}$$

where

$$M_{p_i} = \{v \in M \mid p_i^{e_i} v = 0\}$$

is a primary submodule with order $p_i^{e_i}$ and annihilator

$$\operatorname{ann}(M_{p_i}) = \langle p_i^{e_i} \rangle$$

2) *This decomposition of M into primary submodules is unique up to order of the summands. That is, if*

$$M = N_{q_1} \oplus \cdots \oplus N_{q_m}$$

where N_{q_i} is primary of order $q_i^{f_i}$ and $q_1, \ldots, q_m$ are distinct nonassociate primes then $m = n$ and after a suitable reindexing of the summands we have $N_{q_i} = M_{p_i}$. Hence, q_i and p_i are associates and $f_i = e_i$ (and so $\mu = q_1^{f_1} \cdots q_n^{f_n}$ is also a prime factorization of μ).

Proof. For part 1), let us write $\mu_i = \mu / p_i^{e_i}$. We claim that

$$M_{p_i} = \mu_i M = \{\mu_i v \mid v \in M\}$$

Since $\mu_i v$ is annihilated by $p_i^{e_i}$, we have $\mu_i M \subseteq M_{p_i}$. On the other hand, since μ_i and $p_i^{e_i}$ are relatively prime, there exist $a, b \in R$ for which

$$a\mu_i + bp_i^{e_i} = 1$$

and so if $x \in M_{p_i}$ then

$$x = 1x = (a\mu_i + bp_i^{e_i})x = a\mu_i x \in \mu_i M$$

Hence $M_{p_i} \subseteq \mu_i M$.

Now, since $\gcd(\mu_1, \ldots, \mu_n) = 1$, there exist scalars a_i for which

$$a_1\mu_1 + \cdots + a_n\mu_n = 1$$

and so for any $x \in M$

$$x = 1x = (a_1\mu_1 + \cdots + a_n\mu_n)x \in \sum_{i=1}^{n} \mu_i M$$

Moreover, since the order of $\mu_i M$ divides $p_i^{e_i}$ and the $p_i^{e_i}$'s are pairwise relatively prime, it follows that the sum of the submodules $\mu_i M$ is direct, that is,

$$M = \mu_1 M \oplus \cdots \oplus \mu_n M = M_{p_1} \oplus \cdots \oplus M_{p_n}$$

As to the annihilators, it is clear that $\langle p_i^{e_i} \rangle \subseteq \operatorname{ann}(\mu_i M)$. For the reverse inclusion, if $r \in \operatorname{ann}(\mu_i M)$ then $r\mu_i \in \operatorname{ann}(M)$ and so $p_i^{e_i}\mu_i \mid r\mu_i$, that is, $p_i^{e_i} \mid r$ and so $r \in \langle p_i^{e_i} \rangle$. Thus $\operatorname{ann}(\mu_i M) = \langle p_i^{e_i} \rangle$.

As to uniqueness, we claim that $q = q_1^{f_1} \cdots q_m^{f_m}$ is an order of M. This follows from the fact that N_{q_i} contains an element u_i of order $q_i^{f_i}$ and so the sum $v = u_1 + \cdots + u_m$ has order q. Hence, q divides μ. But μ divides q and so q and μ are associates.

Unique factorization in R now implies that $m = n$ and, after a suitable reindexing, that $f_i = e_i$ and q_i and p_i are associates. Hence, N_{q_i} is primary of order $p_i^{e_i}$. For convenience, we can write N_{q_i} as N_{p_i}. Hence,

$$N_{p_i} \subseteq \{v \in M \mid p_i^{e_i} v = 0\} = M_{p_i}$$

But if

$$N_{p_1} \oplus \cdots \oplus N_{p_n} = M_{p_1} \oplus \cdots \oplus M_{p_n}$$

and $N_{p_i} \subseteq M_{p_i}$ for all i, we must have $N_{p_i} = M_{p_i}$ for all i. □

The Cyclic Decomposition of a Primary Module

The next step in the decomposition process is to show that a primary module can be decomposed into a direct sum of cyclic submodules. While this decomposition is not unique (see the exercises), the set of annihilator ideals is unique, as we will see. To establish this uniqueness, we use the following result.

Lemma 6.10 *Let M be a module over a principal ideal domain R and let $p \in R$ be a prime.*

1) *If $pM = \{0\}$ then M is a vector space over the field $R/\langle p\rangle$ with scalar multiplication defined by*

$$(r + \langle p\rangle)v = rv$$

for all $v \in M$.

2) *For any submodule S of M the set*

$$S^{(p)} = \{v \in S \mid pv = 0\}$$

is also a submodule of M and if $M = S \oplus T$ then

$$M^{(p)} = S^{(p)} \oplus T^{(p)}$$

Proof. For part 1), since p is prime, the ideal $\langle p\rangle$ is maximal and so $R/\langle p\rangle$ is a field. We leave the proof that M is a vector space over $R/\langle p\rangle$ to the reader. For part 2), it is straightforward to show that $S^{(p)}$ is a submodule of M. Since $S^{(p)} \subseteq S$ and $T^{(p)} \subseteq T$ we see that $S^{(p)} \cap T^{(p)} = \{0\}$. Also, if $v \in M^{(p)}$ then $pv = 0$. But $v = s + t$ for some $s \in S$ and $t \in T$ and so $0 = pv = ps + pt$. Since $ps \in S$ and $pt \in T$ we deduce that $ps = pt = 0$, whence $v \in S^{(p)} \oplus T^{(p)}$. Thus, $M^{(p)} \subseteq S^{(p)} \oplus T^{(p)}$ and since the reverse inequality is manifest, the result is proved. □

Theorem 6.11 *(***The cyclic decomposition theorem of a primary module***) Let M be a nonzero primary finitely generated torsion module over a principal ideal domain R, with order p^e.*

1) *Then M is the direct sum*

$$M = \langle v_1 \rangle \oplus \cdots \oplus \langle v_n \rangle \tag{6.1}$$

of cyclic submodules with annihilators $\text{ann}(\langle v_i \rangle) = \langle p^{e_i} \rangle$ *that can be arranged in ascending order*

$$\text{ann}(\langle v_1 \rangle) \subseteq \cdots \subseteq \text{ann}(\langle v_n \rangle)$$

or equivalently

$$e = e_1 \geq e_2 \geq \cdots \geq e_n$$

2) *As to uniqueness, suppose that M is also the direct sum*

$$M = \langle u_1 \rangle \oplus \cdots \oplus \langle u_m \rangle$$

of cyclic submodules with annihilators $\text{ann}(\langle u_i \rangle) = \langle q^{f_i} \rangle$ *arranged in ascending order*

$$\text{ann}(\langle u_1 \rangle) \subseteq \cdots \subseteq \text{ann}(\langle u_m \rangle)$$

or equivalently

$$f_1 \geq f_2 \geq \cdots \geq f_m$$

Then the two chains of annihilators are identical, that is

$$\text{ann}(\langle u_i \rangle) = \text{ann}(\langle v_i \rangle)$$

for all i. Thus, $m = n$, p and q are associates and $f_i = e_i$ for all i.

Proof. Note first that if (6.1) holds then $p^e v_i = 0$. Hence, the order of v_i divides p^e and so must have the form p^{e_i} for $e_i \leq e$. To prove (6.1), let $v \in M$ be an element with order equal to the order of M, that is

$$\text{ann}(v) = \text{ann}(M) = \langle p^e \rangle$$

(We remarked earlier that such an element must exist.)

If we show that $\langle v \rangle$ is complemented, that is, $M = \langle v \rangle \oplus S$ for some submodule S then since S is also a finitely generated primary torsion module over R, we can repeat the process to get

$$M = \langle v \rangle \oplus \langle v_2 \rangle \oplus \cdots \oplus \langle v_k \rangle \oplus S_k$$

where $\text{ann}(v_i) = \langle p^{e_i} \rangle$. We can continue this decomposition as long as $S_k \neq \{0\}$. But the ascending sequence of submodules

$$\langle v \rangle \subseteq \langle v \rangle \oplus \langle v_2 \rangle \subseteq \cdots$$

must terminate since M is noetherian and so eventually we must have $S_k = \{0\}$, giving (6.1).

Now, the direct sum $M_1 = \langle v \rangle \oplus \{0\}$ clearly exists. Suppose that the direct sum

$$M_k = \langle v \rangle \oplus S_k$$

exists. We claim that if $M_k \neq M$ then it is possible to find a submodule S_{k+1} that properly contains S_k for which the direct sum $M_{k+1} = \langle v \rangle \oplus S_{k+1}$ exists.

Once this claim is established, then since M is finitely generated, this process must stop after a finite number of steps, leading to $M = \langle v \rangle \oplus S$ for some submodule S, as desired.

If $M_k \neq M$ then there is a $u \in M \setminus M_k$. We claim that for some scalar $\alpha \in R$, we can take $S_{k+1} = \langle S_k, u - \alpha v \rangle$, which is a proper superset of S_k since $u \notin \langle v \rangle \oplus S_k$.

Thus, we must show that there is an $\alpha \in R$ for which

$$x \in \langle v \rangle \cap \langle S_k, u - \alpha v \rangle \Rightarrow x = 0$$

Now, there exist scalars a and b for which

$$x = av = s + b(u - \alpha v)$$

What can we say about the scalars a and b?

First, although $u \notin M_k$, we do have

$$bu = (a + \alpha b)v - s \in \langle v \rangle \oplus S_k$$

So let us consider the ideal of all such scalars

$$\mathcal{I} = \{r \in R \mid ru \in \langle v \rangle \oplus S_k\}$$

Since $p^e \in \mathcal{I}$ and $\mathcal{I}$ is principal, we have

$$\mathcal{I} = \langle p^f \rangle = \{r \in R \mid ru \in \langle v \rangle \oplus S_k\}$$

for some $f \leq e$. Moreover, f is not 0 since that would imply that $\mathcal{I} = R$ and so $u = 1u \in \langle v \rangle \oplus S_k$, contrary to assumption.

Since $p^f \in \mathcal{I}$, we have $p^f u = cv + t$ for some $t \in S_k$. Then

$$0 = p^e u = p^{e-f}(p^f u) = p^{e-f}cv + p^{e-f}t$$

and since $\langle v \rangle \cap S = \{0\}$, it follows that $p^{e-f}cv = 0$. Hence $c = \delta p^f$ for some $\delta \in R$ and

$$p^f u = cv + t = \delta p^f v + t$$

Since $b \in \mathcal{I}$, we have $b = \beta p^f$ and so

$$bu = \beta p^f u = \delta\beta p^f v + \beta t$$

Thus,

$$x = av = s + b(u - \alpha v) = s + \delta\beta p^f v + \beta t - \alpha\beta p^f v$$

Now it appears that $\alpha = \delta$ would be a good choice, since then

$$x = av = s + \beta t \in S_k$$

and since $\langle v\rangle \cap S_k = \{0\}$ we get $x = 0$. This completes the proof of (6.1).

For uniqueness, note first that M has orders p^{e_1} and q^{f_1} and so p and q are associates and $e_1 = f_1$. Next we show that $n = m$. According to part 2) of Lemma 6.10,

$$M^{(p)} = \langle v_1\rangle^{(p)} \oplus \cdots \oplus \langle v_n\rangle^{(p)}$$

and

$$M^{(p)} = \langle u_1\rangle^{(p)} \oplus \cdots \oplus \langle u_m\rangle^{(p)}$$

where all summands are nonzero. Since $pM^{(p)} = \{0\}$, it follows from part 1) of Lemma 6.10 that $M^{(p)}$ is a vector space over $R/\langle p\rangle$ and so each of the preceding decompositions expresses $M^{(p)}$ as a direct sum of one-dimensional vector subspaces. Hence, $m = \dim(M^{(p)}) = n$.

Finally, we show that the exponents e_i and f_i are equal using induction on e_1. If $e_1 = 1$ then $e_i = 1$ for all i and since $f_1 = e_1$, we also have $f_i = 1$ for all i. Suppose the result is true whenever $e_1 \le k - 1$ and let $e_1 = k$. Suppose that

$$(e_1, \ldots, e_n) = (e_1, \ldots, e_s, 1, \ldots, 1), e_s > 1$$

and

$$(f_1, \ldots, f_n) = (f_1, \ldots, f_t, 1, \ldots, 1), f_t > 1$$

Then

$$pM = p\langle v_1\rangle \oplus \cdots \oplus p\langle v_s\rangle$$

and

$$pM = p\langle u_1\rangle \oplus \cdots \oplus p\langle u_t\rangle$$

But $p\langle v_1\rangle = \langle pv_1\rangle$ is a cyclic submodule of M with annihilator $\langle p^{e_i - 1}\rangle$ and so by the induction hypothesis

$$s = t \text{ and } e_1 = f_1, \ldots, e_s = f_s$$

which concludes the proof of uniqueness. □

The Primary Cyclic Decomposition Theorem

Now we can combine the various decompositions.

Theorem 6.12 *(***The primary cyclic decomposition theorem***) Let M be a nonzero finitely generated module over a principal ideal domain R.*

1) Then

$$M = M_{\text{free}} \oplus M_{\text{tor}}$$

where M_{free} is free and M_{tor} is a torsion module. If M_{tor} has order

$$\mu = p_1^{e_1} \cdots p_n^{e_n}$$

where the p_i's are distinct nonassociate primes in R then M_{tor} can be uniquely decomposed (up to the order of the summands) into the direct sum

$$M = M_{p_1} \oplus \cdots \oplus M_{p_n}$$

where

$$M_{p_i} = \{v \in M_{\text{tor}} \mid p_i^{e_i} v = 0\}$$

is a primary submodule with annihilator $\langle p_i^{e_i} \rangle$. Finally, each primary submodule M_{p_i} can be written as a direct sum of cyclic submodules, so that

$$M = M_{\text{free}} \oplus \big[\underbrace{\langle v_{1,1} \rangle \oplus \cdots \oplus \langle v_{1,k_1} \rangle}_{M_{p_1}} \big] \oplus \cdots \oplus \big[\underbrace{\langle v_{n,1} \rangle \oplus \cdots \oplus \langle v_{n,k_n} \rangle}_{M_{p_n}} \big]$$

where $\operatorname{ann}(\langle v_{i,j} \rangle) = \langle p_i^{e_{i,j}} \rangle$ and the terms in each cyclic decomposition can be arranged so that, for each i,

$$\operatorname{ann}(\langle v_{i,1} \rangle) \subseteq \cdots \subseteq \operatorname{ann}(\langle v_{i,k_i} \rangle)$$

or, equivalently,

$$e_i = e_{i,1} \geq e_{i,2} \geq \cdots \geq e_{i,k_i}$$

2) As for uniqueness, suppose that

$$M = N_{\text{free}} \oplus \langle x_1 \rangle \oplus \cdots \oplus \langle x_\ell \rangle$$

is a decomposition of M into the direct sum of a free module N_{free} and primary cyclic submodules $\langle x_i \rangle$. Then

a) $\operatorname{rk}(N_{\text{free}}) = \operatorname{rk}(M_{\text{free}})$

b) The number of summands is the same in both decompositions, that is $\ell = k_1 + \cdots + k_n$

c) The summands in this decomposition can be reordered to get

$$M = N_{\text{free}} \oplus [\langle u_{1,1} \rangle \oplus \cdots \oplus \langle u_{1,k_1} \rangle] \oplus \cdots \oplus [\langle u_{n,1} \rangle \oplus \cdots \oplus \langle u_{n,k_n} \rangle]$$

where the primary submodules are the same

$$\langle u_{i,1}\rangle \oplus \cdots \oplus \langle u_{i,k_i}\rangle = \langle v_{i,1}\rangle \oplus \cdots \oplus \langle v_{i,k_i}\rangle$$

for $i = 1, \ldots, n$ and the annihilator chains are the same, that is,

$$\operatorname{ann}(\langle u_{i,j}\rangle) = \operatorname{ann}(\langle v_{i,j}\rangle)$$

for all i, j.

In summary, the free rank, primary submodules and annihilator chain are uniquely determined by the module M.

Proof. We need only prove the uniqueness. We have seen that N_{free} and M_{free} are isomorphic and thus have the same rank. Let us look at the torsion part.

Since the order of each primary cyclic submodule $\langle x_i\rangle$ must divide the order of M, this order is a power of one of the primes $p_1, \ldots, p_n$. Let us group the summands by like primes p_i to get

$$M = N_{\text{free}} \oplus \big[\underbrace{\langle u_{1,1}\rangle \oplus \cdots \oplus \langle u_{1,j_1}\rangle}_{N_{p_1}\text{(primary of order } p_1^{f_1})}\big] \oplus \cdots \oplus \big[\underbrace{\langle u_{n,1}\rangle \oplus \cdots \oplus \langle u_{n,j_n}\rangle}_{N_{p_1}\text{(primary of order } p_n^{f_n})}\big]$$

Then each group N_{p_i} is a primary submodule of M with order $p_i^{f_i}$. The uniqueness of primary decompositions of M_{tor} implies that $N_{p_i} = M_{p_i}$. Then the uniqueness of cyclic decompositions implies that the annihilator chains for the decompositions of N_{p_i} and M_{p_i} are the same. □

We have seen that in a primary cyclic decomposition

$$M = M_{\text{free}} \oplus [\langle v_{1,1}\rangle \oplus \cdots \oplus \langle v_{1,k_1}\rangle] \oplus \cdots \oplus [\langle v_{n,1}\rangle \oplus \cdots \oplus \langle v_{n,k_n}\rangle]$$

the chain of annihilators

$$\operatorname{ann}(\langle v_{i,j}\rangle) = \langle p_i^{e_{i,j}}\rangle$$

is unique except for order. The sequence $p_i^{e_{i,j}}$ of generators is uniquely determined up to order and multiplication by units. This sequence is called the sequence of **elementary divisors** of M. Note that the elementary divisors are not quite as unique as the annihilators: the multiset of annihilators is unique but the multiset of generators is not since if $p_i^{e_{i,j}}$ is a generator then so is $up_i^{e_{i,j}}$ for any unit u in R.

The Invariant Factor Decomposition

According to Theorem 6.7, if S and T are cyclic submodules with relatively prime orders, then $S \oplus T$ is a cyclic submodule whose order is the product of the orders of S and T. Accordingly, in the primary cyclic decomposition of M

$$M = M_{\text{free}} \oplus \big[\underbrace{\langle v_{1,1}\rangle \oplus \cdots \oplus \langle v_{1,k_1}\rangle}_{M_{p_1}}\big] \oplus \cdots \oplus \big[\underbrace{\langle v_{n,1}\rangle \oplus \cdots \oplus \langle v_{n,k_n}\rangle}_{M_{p_n}}\big]$$

with elementary divisors $p_i^{e_{i,j}}$ satisfying

$$e_i = e_{i,1} \geq e_{i,2} \geq \cdots \geq e_{i,k_i} \tag{6.4}$$

we can feel free to regroup and combine cyclic summands with relatively prime orders. One judicious way to do this is to take the leftmost (highest order) cyclic submodules from each group to get

$$D_1 = \langle v_{1,1}\rangle \oplus \cdots \oplus \langle v_{n,1}\rangle$$

and repeat the process

$$\begin{aligned} D_2 &= \langle v_{1,2}\rangle \oplus \cdots \oplus \langle v_{n,2}\rangle \\ D_3 &= \langle v_{1,3}\rangle \oplus \cdots \oplus \langle v_{n,3}\rangle \\ &\vdots \end{aligned}$$

Of course, some summands may be missing here since the primary modules M_{p_i} do not necessarily have the same number of summands. In any case, the result of this regrouping and combining is a decomposition of the form

$$M = M_{\text{free}} \oplus D_1 \oplus \cdots \oplus D_m$$

which is called an *invariant factor decomposition* of M.

For example, suppose that

$$M = M_{\text{free}} \oplus [\langle v_{1,1}\rangle \oplus \langle v_{1,2}\rangle] \oplus [\langle v_{2,1}\rangle] \oplus [\langle v_{3,1}\rangle \oplus \langle v_{3,2}\rangle \oplus \langle v_{3,3}\rangle]$$

Then the resulting regrouping and combining gives

$$M = M_{\text{free}} \oplus [\underbrace{\langle v_{1,1}\rangle \oplus \langle v_{2,1}\rangle \oplus \langle v_{3,1}\rangle}_{D_1}] \oplus [\underbrace{\langle v_{1,2}\rangle \oplus \langle v_{3,2}\rangle}_{D_2}] \oplus [\underbrace{\langle v_{3,3}\rangle}_{D_3}]$$

As to the orders of the summands, referring to (6.4), if D_i has order d_i then since the highest powers of each prime p_i are taken for d_1, the second–highest for d_2 and so on, we conclude that

$$d_m \mid d_{m-1} \mid \cdots \mid d_2 \mid d_1 \tag{6.5}$$

or equivalently,

$$\text{ann}(D_1) \subseteq \text{ann}(D_2) \subseteq \cdots$$

The numbers d_i are called *invariant factors* of the decomposition.

For instance, in the example above suppose that the elementary divisors are

$$p_1^3, p_1^2, p_2, p_3^3, p_3^3, p_3$$

Then the invariant factors are

$$\begin{aligned} d_1 &= p_1^3 p_2 p_3^3 \\ d_2 &= p_1^2 p_3^3 \\ d_3 &= p_3 \end{aligned}$$

The process described above that passes from a sequence $p_i^{e_{i,j}}$ of elementary divisors in order (6.4) to a sequence of invariant factors in order (6.5) is reversible. The inverse process takes a sequence $d_1, \ldots, d_m$ satisfying (6.5), factors each d_i into a product of distinct nonassociate prime powers with the primes in the same order and then "peels off" like prime powers from the left. (The reader may wish to try it on the example above.)

This fact, together with Theorem 6.7, implies that primary cyclic decompositions and invariant factor decompositions are essentially equivalent. In particular, given a primary cyclic decomposition of M we can produce an invariant factor decomposition of M whose invariant factors are products of the elementary divisors and for which each elementary divisor appears in exactly one invariant factor. Conversely, given an invariant factor decomposition of M we can obtain a primary cyclic decomposition of M whose elementary divisors are precisely the multiset of prime power factors of the invariant factors.

It follows that since the elementary divisors of M are unique up to multiplication by units, the invariant factors of M are also unique up to multiplication by units.

Theorem 6.13 *(***The invariant factor decomposition theorem***) Let M be a finitely generated module over a principal ideal domain R. Then*

$$M = M_{\text{free}} \oplus D_1 \oplus \cdots \oplus D_m$$

where M_{free} is a free submodule and D_i is a cyclic submodule of M, with order d_i, where

$$d_m \mid d_{m-1} \mid \cdots \mid d_2 \mid d_1$$

This decomposition is called an **invariant factor decomposition** *of M and the scalars d_i, are called the* **invariant factors** *of M. The invariant factors are uniquely determined, up to multiplication by a unit, by the module M. Also, the rank of M_{free} is uniquely determined by M.* □

The annihilators of an invariant factor decomposition are called the **invariant ideals** of M. The chain of invariant ideals is unique, as is the chain of annihilators in the primary cyclic decomposition. Note that d_1 is an order of M, that is

$$\text{ann}(M) = \langle d_1 \rangle$$

Note also that the product

$$\gamma = d_1 \cdots d_m$$

of the invariant factors of M has some nice properties. For example, γ is the product of all the elementary divisors of M. We will see in a later chapter that in the context of a linear operator τ on a vector space, γ is the characteristic polynomial of τ.

Exercises

1. Show that any free module over an integral domain is torsion-free.
2. Let R be a principal ideal domain and R^+ the field of quotients. Then R^+ is an R-module. Prove that any nonzero finitely generated submodule of R^+ is a free module of rank 1.
3. Let R be a principal ideal domain. Let M be a finitely generated torson-free R-module. Suppose that N is a submodule of M for which N is a free R-module of rank 1 and M/N is a torsion module. Prove that M is a free R-module of rank 1. *Hint*: Use the results of the previous exercise.
4. Show that the primary cyclic decomposition of a torsion module over a principal ideal domain is not unique (even though the elementary divisors are).
5. Show that if M is a finitely generated R-module where R is a principal ideal domain, then the free summand in the decomposition $M = F \oplus M_{\text{tor}}$ need not be unique.
6. If $\langle v \rangle$ is a cyclic R-module or order a show that the map $\tau\colon R \to \langle v \rangle$ defined by $\tau(r) = rv$ is a surjective R-homomorphism with kernel $\langle a \rangle$ and so
$$\langle v \rangle \approx \frac{R}{\langle a \rangle}$$
7. If R is a ring with the property that all submodules of cyclic R-modules are cyclic, show that R is a principal ideal domain.
8. Suppose that F is a finite field and let F^* be the set of all nonzero elements of F.
   a) Show that F^* is an abelian group under multiplication.
   b) Show that $p(x) \in F[x]$ is a nonconstant polynomial over F and if $r \in F$ is a root of $p(x)$ then $x - r$ is a factor of $P(x)$.
   c) Prove that a nonconstant polynomial $p(x) \in F[x]$ of degree n can have at most n distinct roots in F.
   d) Use the invariant factor or primary cyclic decomposition of a finite $\mathbb{Z}$-module to prove that F^* is cyclic.
9. Let R be a principal ideal domain. Let $M = \langle v \rangle$ be a cyclic R-module with order α. We have seen that any submodule of M is cyclic. Prove that for each $\beta \in R$ such that $\beta \mid \alpha$ there is a unique submodule of M of order β.
10. Suppose that M is a free module of finite rank over a principal ideal domain R. Let N be a submodule of M. If M/N is torsion, prove that $\text{rk}(N) = \text{rk}(M)$.

11. Let $F[x]$ be the ring of polynomials over a field F and let $F'[x]$ be the ring of all polynomials in $F[x]$ that have coefficient of x equal to 0. Then $F[x]$ is an $F'[x]$-module. Show that $F[x]$ is finitely generated and torsion-free but not free. Is $F'[x]$ a principal ideal domain?
12. Show that the rational numbers $\mathbb{Q}$ form a torsion-free $\mathbb{Z}$-module that is not free.

More on Complemented Submodules

13. Let R be a principal ideal domain and let M be a free R-module.
    a) Prove that a submodule N of M is complemented if and only if M/N is free.
    b) If M is also finitely generated, prove that N is complemented if and only if M/N is torsion-free.
14. Let M be a free module of finite rank over a principal ideal domain R.
    a) Prove that if N is a complemented submodule of M then $\text{rk}(N)=\text{rk}(M)$ if and only if $N=M$.
    b) Show that this need not hold if N is not complemented.
    c) Prove that N is complemented if and only if any basis for N can be extended to a basis for M.
15. Let M and N be free modules of finite rank over a principal ideal domain R. Let $\tau\colon M\to N$ be an R-homomorphism.
    a) Prove that $\ker(\tau)$ is complemented.
    b) What about $\text{im}(\tau)$?
    c) Prove that

    $$\text{rk}(M)=\text{rk}(\ker(\tau))+\text{rk}(\text{im}(\tau))=\text{rk}(\ker(\tau))+\text{rk}\left(\frac{M}{\ker(\tau)}\right)$$

    d) If τ is surjective then τ is an isomorphism if and only if $\text{rk}(M)=\text{rk}(N)$.
    e) If M/L is free then

    $$\text{rk}\left(\frac{M}{L}\right)=\text{rk}(M)-\text{rk}(L)$$

16. A submodule N of a module M is said to be **pure in** M if whenever $v\notin M\setminus N$ then $rv\notin N$ for all nonzero $r\in R$.
    a) Show that N is pure if and only if $v\in N$ and $v=rw$ for $r\in R$ implies $w\in N$.
    b) Show that N is pure if and only if M/N is torsion-free.
    c) If R is a principal ideal domain and M is finitely generated, prove that N is pure if and only if M/N is free.
    d) If L and N are pure submodules of M then so are $L\cap N$ and $L\cup N$. What about $L+N$?
    e) If N is pure in M then show that $L\cap N$ is pure in L for any submodule L of M.

17. Let M be a free module of finite rank over a principal ideal domain R. Let L and N be submodules of M with L complemented in M. Prove that

$$\mathrm{rk}(L+N)+\mathrm{rk}(L\cap N)=\mathrm{rk}(L)+\mathrm{rk}(N)$$

Chapter 7
The Structure of a Linear Operator

In this chapter, we study the structure of a linear operator on a finite-dimensional vector space, using the powerful module decomposition theorems of the previous chapter. *Unless otherwise noted, all vector spaces will be assumed to be finite-dimensional.*

A Brief Review

We have seen that any linear operator on a finite-dimensional vector space can be represented by matrix multiplication. Let us restate Theorem 2.14 for linear operators.

Theorem 7.1 *Let $\tau \in \mathcal{L}(V)$ and let $\mathcal{B} = (b_1, \ldots, b_n)$ be an ordered basis for V. Then τ can be represented by matrix multiplication*

$$[\tau(v)]_{\mathcal{B}} = [\tau]_{\mathcal{B}}\,[v]_{\mathcal{B}}$$

where

$$[\tau]_{\mathcal{B}} = ([\tau(b_1)]_{\mathcal{B}} \mid \cdots \mid [\tau(b_n)]_{\mathcal{B}}) \qquad \square$$

Since the matrix $[\tau]_{\mathcal{B}}$ depends on the ordered basis $\mathcal{B}$, it is natural to wonder how to choose this basis in order to make the matrix $[\tau]_{\mathcal{B}}$ as simple as possible. That is the subject of this chapter.

Let us also restate the relationship between the matrices of τ with respect to different ordered bases.

Theorem 7.2 *Let $\tau \in \mathcal{L}(V)$ and let $\mathcal{B}$ and $\mathcal{B}'$ be ordered bases for V. Then the matrix of τ with respect to $\mathcal{B}'$ can be expressed in terms of the matrix of τ with respect to $\mathcal{B}$ as follows*

$$[\tau]_{\mathcal{B}'} = M_{\mathcal{B},\mathcal{B}'}[\tau]_{\mathcal{B}}(M_{\mathcal{B},\mathcal{B}'})^{-1}$$

where

$$M_{\mathcal{B},\mathcal{B}'} = ([b_1]_{\mathcal{B}'}, \ldots, [b_n]_{\mathcal{B}'})) \qquad \square$$

Finally, we recall the definition of similarity and its relevance to the current discussion.

Definition *Two matrices A and B are* **similar** *if there exists an invertible matrix P for which*

$$B = PAP^{-1}$$

The equivalence classes associated with similarity are called **similarity classes**. □

Theorem 7.3 *Let V be a vector space of dimension n. Then two $n \times n$ matrices A and B are similar if and only if they represent the same linear operator $\tau \in \mathcal{L}(V)$, but possibly with respect to different ordered bases. In this case, A and B represent exactly the same set of linear operators in $\mathcal{L}(V)$.* □

According to Theorem 7.3, the matrices that represent a given linear operator $\tau \in \mathcal{L}(V)$ are precisely the matrices that lie in one particular similarity class. Hence, in order to uniquely represent all linear operators on V we would like to find a simple representative of each similarity class, that is, a set of simple canonical forms for similarity.

The simplest type of useful matrices is the diagonal matrices. However, not all linear operators can be represented by diagonal matrices, that is, the set of diagonal matrices does not form a set of canonical forms for similarity.

This gives rise to two different directions for further study. First, we can search for a characterization of those linear operators that can be represented by diagonal matrices. Such operators are called *diagonalizable*. Second, we can search for a different type of "simple" matrix that does provide a set of canonical forms for similarity. We will pursue both of these directions at the same time.

The Module Associated with a Linear Operator

Throughout this chapter, we fix a nonzero linear operator $\tau \in \mathcal{L}(V)$ and think of V not only as a vector space over a field F but also as a module over $F[x]$, with scalar multiplication defined by

$$p(x)v = p(\tau)(v)$$

We call V the $F[x]$-module **defined by** τ and write V_τ to indicate the dependence on τ (when necessary). Thus, V_τ and V_σ are modules over the same ring $F[x]$, although the scalar multiplication is different if $\tau \neq \sigma$.

Our plan is to interpret the concepts of the previous chapter for the module/vector space V. First, since V is a finite-dimensional vector space, so is $\mathcal{L}(V)$. It follows that V_τ is a torsion module. To see this, note that since $\dim(\mathcal{L}(V))=n^2$, the n^2+1 vectors

$$\iota,\tau,\tau^2,\ldots,\tau^{n^2}$$

are linearly dependent in $\mathcal{L}(V)$, which implies that $p(\tau)=0$ for some polynomial $p(x)\in F[x]$. Hence, $p(x)V=\{0\}$, which shows that $\operatorname{ann}(V)\neq\{0\}$.

Also, since V is finitely generated as a vector space, it is, a fortiori, finitely generated as an $F[x]$-module defined by τ. Thus, V is a finitely generated torsion module over a principal ideal domain $F[x]$ and so we may apply the decomposition theorems of the previous chapter.

Next we take a look at the connection between module isomorphisms and vector space isomorphisms. This also describes the connection with similarity.

Theorem 7.4 *Let τ and σ be linear operators on V. Then V_τ and V_σ are isomorphic as $F[x]$-modules if and only if τ and σ are similar as linear operators. In particular, a function $\phi: V_\tau\to V_\sigma$ is a module isomorphism if and only if it is a vector space automorphism of V satisfying*

$$\sigma=\phi\tau\phi^{-1}$$

Proof. Suppose that $\phi: V_\tau\to V_\sigma$ is a module isomorphism. Then for $v\in V$

$$\phi(xv)=x\phi(v)$$

which is equivalent to

$$\phi(\tau(v))=\sigma(\phi(v))$$

and since ϕ is bijective this is equivalent to

$$(\phi\tau\phi^{-1})v=\sigma(v)$$

that is, $\sigma=\phi\tau\phi^{-1}$. Since a module isomorphism from V_τ to V_σ is a vector space isomorphism as well, the result follows.

For the converse, suppose that $\sigma=\phi\tau\phi^{-1}$ for a vector space automorphism ϕ on V. This condition is equivalent to $\sigma\phi=\phi\tau$ and so

$$\phi(x^kv)=\phi(\tau^k(v))=\sigma^k(\phi(v))=x^k\phi(v)$$

and by the F-linearity of ϕ, for any polynomial $p(x)\in F[x]$ we have

$$\phi(p(\tau)v) = p(\sigma)\phi(v)$$

which shows that ϕ is a module isomorphism from V_τ to V_σ. □

Submodules and Invariant Subspaces

There is a simple connection between the submodules of the $F[x]$-module V_τ and the subspaces of the vector space V. Recall that a subspace S of V is τ-invariant if $\tau(S) \subseteq S$.

Theorem 7.5 *A subset S of V is a submodule of the $F[x]$-module V_τ if and only if it is a τ-*invariant *subspace of the vector space V.* □

Orders and the Minimal Polynomial

We have seen that since V is finite-dimensional, the annihilator

$$\operatorname{ann}(V) = \{p(x) \in F[x] \mid p(x)V = \{0\}\}$$

of V is a nonzero ideal of $F[x]$ and since $F[x]$ is a principal ideal domain, this ideal is principal, say

$$\operatorname{ann}(V) = \langle p(x) \rangle$$

Since all orders of V are associates and since the units of $F[x]$ are precisely the nonzero elements of F, there is a unique *monic* order of V.

Definition *Let V_τ be an $F[x]$-module defined by τ. The unique monic order of V_τ, that is, the unique monic polynomial that generates* $\operatorname{ann}(V_\tau)$ *is called the* **minimal polynomial** *for τ and is denoted by $m_\tau(x)$ or* $\min(\tau)$*. Thus,*

$$\operatorname{ann}(V_\tau) = \langle m_\tau(x) \rangle$$

and

$$p(x)V_\tau = \{0\} \Leftrightarrow p(\tau) = 0 \Leftrightarrow m_\tau(x) \mid p(x) \qquad \square$$

In treatments of linear algebra that do not emphasize the role of the *module V_τ* the minimal polynomial of a linear operator τ is simply defined as the unique monic polynomial $m_\tau(x)$ of *smallest degree* for which $m_\tau(\tau) = 0$. It is not hard to see that this definition is equivalent to the previous definition.

The concept of minimal polynomial is also defined for matrices. If A is a square matrix over F the **minimal polynomial** $m_{\mathrm{A}}(x)$ of A is defined as the unique monic polynomial $p(x) \in F[x]$ of smallest degree for which $p(A) = 0$. We leave it to the reader to verify that this concept is well-defined and that the following holds.

Theorem 7.6

1) *If A and B are similar matrices then $m_A(x) = m_B(x)$. Thus, the minimal polynomial is an invariant under similarity.*
2) *The minimal polynomial of $\tau \in \mathcal{L}(V)$ is the same as the minimal polynomial of any matrix that represents τ.* □

Cyclic Submodules and Cyclic Subspaces

For an $F[x]$-module V_τ, consider the cyclic submodule

$$\langle v \rangle = \{p(x)v \mid p(x) \in F[x]\} = \{p(\tau)(v) \mid p(x) \in F[x]\}$$

We would like to characterize these simple but important submodules in terms of vector space notions.

As we have seen, $\langle v \rangle$ is a τ-invariant subspace of V, but more can be said. Let $m(x)$ be the minimal polynomial of $\tau|_{\langle v \rangle}$ and suppose that deg$(m(x)) = n$. Any element of $\langle v \rangle$ has the form $p(x)v$. Dividing $p(x)$ by $m(x)$ gives

$$p(x) = q(x)m(x) + r(x)$$

where deg $r(x) <$ deg $m(x)$. Since $m(x)v = 0$, we have

$$p(x)v = q(x)m(x)v + r(x)v = r(x)v$$

Thus,

$$\langle v \rangle = \{r(x)v \mid \deg r(x) < n\}$$

Put another way, the ordered set

$$\mathcal{B} = (v, xv, \ldots, x^{n-1}v) = (v, \tau(v), \ldots, \tau^{n-1}(v))$$

spans $\langle v \rangle$ *as a vector space over* F. But it is also the case that $\mathcal{B}$ is linearly independent over F, for if

$$r_0 v + r_1 xv + \cdots + r_{n-1}x^{n-1}v = 0$$

then $r(x)v = 0$ where

$$r(x) = r_0 + r_1 x + \cdots + r_{n-1}x^{n-1}$$

has degree less than n. Hence, $r(x) = 0$, that is, $r_i = 0$ for all $i = 0, \ldots, n-1$. Thus, $\mathcal{B}$ is an ordered basis for $\langle v \rangle$.

To determine the matrix of $\tau|_{\langle v \rangle}$ with respect to $\mathcal{B}$, write $w_i = \tau^i(v)$. Then

$$\tau(w_i) = \tau(\tau^i(v)) = \tau^{i+1}(v) = w_{i+1}$$

for $i = 0, \ldots, n-2$ and so τ simply “shifts” each basis vector in $\mathcal{B}$, except the last one, to the next basis vector in $\mathcal{B}$. For the last vector w_{n-1}, if

$$m(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$$

then since $m(\tau) = 0$ we have

$$0 = m(\tau) = a_0 + a_1\tau + \cdots + a_{n-1}\tau^{n-1} + \tau^n$$

and so

$$\begin{aligned} \tau(w_{n-1}) &= \tau(\tau^{n-1}(v)) = \tau^n(v) \\ &= -(a_0 + a_1\tau + \cdots + a_{n-1}\tau^{n-1})(v) \\ &= -a_0v - a_1\tau(v) - \cdots - a_{n-1}\tau^{n-1}(v) \\ &= -a_0w_0 - a_1w_1 - \cdots - a_{n-1}w_{n-1} \end{aligned}$$

Hence, the matrix of $\tau|_{\langle v\rangle}$ with respect to $\mathcal{B}$ is

$$C[m(x)] = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & 0 & -a_{n-2} \\ 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix}$$

This is known as the **companion matrix** for the polynomial $m(x)$. Note that companion matrices are defined only for *monic* polynomials.

Definition *Let $\tau \in \mathcal{L}(V)$. A subspace S of V is τ-***cyclic** *if there exists a vector $v \in S$ for which the set*

$$\{v, \tau(v), \ldots, \tau^{m-1}(v)\}$$

is a basis for S. □

Theorem 7.7 *Let V_τ be an $F[x]$-module defined by $\tau \in \mathcal{L}(V)$.*

1) **(Characterization of cyclic submodules)** *A subset $S \subseteq V$ is a cyclic submodule of V_τ if and only if it is a τ-cyclic subspace of the vector space V.*
2) *Suppose that $\langle v\rangle$ is a cyclic submodule of V. If the monic order of $\langle v\rangle$ is*

$$m(x) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$$

then

$$\mathcal{B} = (v, xv, \ldots, x^{n-1}v) = (v, \tau(v), \ldots, \tau^{n-1}(v))$$

is an ordered basis for $\langle v\rangle$ and the matrix $[\tau|_{\langle v\rangle}]_{\mathcal{B}}$ is the companion matrix $C[m(x)]$ of $m(x)$. Hence,

$$\dim(\langle v\rangle) = \deg(m(x))$$ □

Summary

The following table summarizes the connection between the module concepts and the vector space concepts that we have discussed.

$F[x]$-**Module** V_τ	F-**Vector Space** V
Scalar multiplication: $p(x)v$	Action of $p(\tau)$: $p(\tau)(v)$
Submodule of V_τ	τ-Invariant subspace of V
Annihilator: $\operatorname{ann}(V_\tau) = \{p(x) \mid p(x)V_\tau = \{0\}\}$	Annihilator: $\operatorname{ann}(V) = \{p(x) \mid p(\tau)(V) = \{0\}\}$
Monic order $m(x)$ of V_τ: $\operatorname{ann}(V_\tau) = \langle m(x)\rangle$	Minimal polynomial of τ: $m(x)$ has smallest deg with $m(\tau) = 0$
Cyclic submodule of V_τ: $\langle v\rangle = \{p(x)v \mid \deg p(x) < \deg m(x)\}$	τ-cyclic subspace of V: $\langle v\rangle = \operatorname{span}\{v, \tau(v), \ldots, \tau^{m-1}(v)\}$

The Decomposition of V_τ

We are now ready to translate the cyclic decomposition theorem into the language of V_τ. First, we define the **elementary divisors** and **invariant factors** of an operator τ to be the elementary divisors and invariant factors, respectively, of the module V_τ. Also, the **elementary divisors** and **invariant factors** of a matrix A are defined to be the elementary divisors and invariant factors, respectively, of the operator τ_A.

We will soon see that the multiset of elementary divisors and the multiset of invariant factors are complete invariants under similarity and so the multiset of elementary divisors (or invariant factors) of an operator τ is the same as the multiset of elementary divisors (or invariant factors) of any matrix that represents τ.

Theorem 7.8 *(***The cyclic decomposition theorem for** V) *Let τ be a linear operator on a finite-dimensional vector space V. Let*

$$m_r(x) = p_1^{e_1}(x)\cdots p_n^{e_n}(x)$$

be the minimal polynomial of τ, where the monic polynomials $p_i(x)$ are distinct and irreducible.

1) (**Primary decomposition**) *The $F[x]$-module V_τ is the direct sum*

$$V_\tau = V_{p_1} \oplus \cdots \oplus V_{p_n}$$

where

$$V_{p_i} = \{v \in V \mid p_i^{e_i}(\tau)(v) = 0\}$$

is a primary submodule of V_τ of order $p_i^{e_i}(x)$. In vector space terms, V_{p_i} is a

τ-invariant subspace of V and the minimal polynomial of $\tau|_{V_{p_i}}$ is

$$\min(\tau|_{V_{p_i}}) = p_i^{e_i}(x)$$

2) (**Cyclic decomposition**) *Each primary summand V_{p_i} can be decomposed into a direct sum*

$$V_{p_i} = \langle v_{i,1}\rangle \oplus \cdots \oplus \langle v_{i,k_i}\rangle$$

of cyclic submodules $\langle v_{i,j}\rangle$ of order $p_i^{e_{i,j}}(x)$ with

$$e_i = e_{i,1} \geq e_{i,2} \geq \cdots \geq e_{i,k_i}$$

In vector space terms, $\langle v_{i,j}\rangle$ is a τ-cyclic subspace of V_{p_i} and the minimal polynomial of $\tau|_{\langle v_{i,j}\rangle}$ is

$$\min(\tau|_{\langle v_{i,j}\rangle}) = p_i^{e_{i,j}}(x)$$

3) (**The complete decomposition**) *This yields the decomposition of V into a direct sum of τ-cyclic subspaces*

$$V = (\langle v_{1,1}\rangle \oplus \cdots \oplus \langle v_{1,k_1}\rangle) \oplus \cdots \oplus (\langle v_{n,1}\rangle \oplus \cdots \oplus \langle v_{n,k_n}\rangle)$$

4) (**Elementary divisors and dimensions**) *The multiset of elementary divisors $\{p_i^{e_{i,j}}(x)\}$ of τ is uniquely determined by τ. If $\deg(p_i^{e_{i,j}}(x)) = d_{i,j}$ then the τ-cyclic subspace $\langle v_{i,j}\rangle$ has basis*

$$\mathcal{B}_{i,j} = (v_{i,j}, \tau(v_{i,j}), \ldots, \tau^{d_{i,j}-1}(v_{i,j}))$$

and so $\dim(\langle v_{i,j}\rangle) = \deg(p_i^{e_{i,j}})$. Hence,

$$\dim(V_{p_i}) = \sum_{j=1}^{k_i} \deg(p_i^{e_{i,j}}) \qquad \square$$

The Rational Canonical Form

The cyclic decomposition theorem can be used to determine a set of canonical forms for similarity. Recall that if $V = S \oplus T$ and if both S and T are invariant under τ, the pair (S,T) is said to **reduce** τ. Put another way, (S,T) reduces τ if the restrictions $\tau|_S$ and $\tau|_T$ are linear operators on S and T, respectively.

Recall also that we write $\tau = \rho \oplus \sigma$ if there exist subspaces S and T of V for which (S,T) reduces τ and

$$\rho = \tau|_S \text{ and } \sigma = \tau|_T$$

If $\tau = \sigma \oplus \rho$ then any matrix representations of σ and ρ can be used to construct a matrix representation of τ.

Theorem 7.9 Suppose that $\tau=\tau_1\oplus\tau_2\in\mathcal{L}(V)$ has a reducing pair (S,T). Let

$$\mathcal{B}=(c_1,\ldots,c_s,d_1,\ldots,d_t)$$

be an ordered basis for V, where $\mathcal{C}=(c_1,\ldots,c_s)$ is an ordered basis for S and $\mathcal{D}=(d_1,\ldots,d_t)$ is an ordered basis for T. Then the matrix $[\tau]_\mathcal{B}$ has the block diagonal form

$$[\tau]_\mathcal{B}=\begin{bmatrix}[\tau_1]_\mathcal{C} & 0\\ 0 & [\tau_2]_\mathcal{D}\end{bmatrix}_{\text{block}} \qquad \square$$

Of course, this theorem may be extended to apply to multiple direct summands and this is especially relevant to our situation, since according to Theorem 7.8

$$\tau=\tau|_{\langle v_{1,1}\rangle}\oplus\cdots\oplus\tau|_{\langle v_{n,kn}\rangle}$$

In particular, if $\mathcal{B}_{i,j}$ is an ordered basis for the cyclic submodule $\langle v_{i,j}\rangle$ and if

$$\mathcal{B}=(\mathcal{B}_{1,1},\ldots,\mathcal{B}_{n,k_n})$$

denotes the ordered basis for V obtained from these ordered bases (as we did in Theorem 7.9) then

$$[\tau]_\mathcal{B}=\begin{bmatrix}[\tau_{1,1}]_{\mathcal{B}_{1,1}} & & \\ & \ddots & \\ & & [\tau_{n,k_n}]_{\mathcal{B}_{n,kn}}\end{bmatrix}_{\text{block}}$$

where $\tau_{i,j}=\tau|_{\langle v_{i,j}\rangle}$.

According to Theorem 7.8, the cyclic submodule $\langle v_{i,j}\rangle$ has ordered basis

$$\mathcal{B}_{i,j}=(v_{i,j},\tau(v_{i,j}),\ldots,\tau^{d_{i,j}-1}(v_{i,j}))$$

where $d_{i,j}=\deg(p_i^{e_{i,j}}(x))$. Hence, we arrive at the matrix representation of τ described in the following theorem.

Theorem 7.10 *(***The rational canonical form***) Let* $\dim(V)<\infty$ *and suppose that* $\tau\in\mathcal{L}(V)$ *has minimal polynomial*

$$m_\tau(x)=p_1^{e_1}(x)\cdots p_n^{e_n}(x)$$

where the monic polynomials $p_i(x)$ *are distinct and irreducible. Then* V *has an ordered basis* $\mathcal{B}$ *under which*

$$[\tau]_{\mathcal{B}} = \begin{bmatrix} C[p_1^{e_{1,1}}(x)] & & & & & & \\ & \ddots & & & & & \\ & & C[p_1^{e_{1,k_1}}(x)] & & & & \\ & & & \ddots & & & \\ & & & & C[p_n^{e_{n,n_1}}(x)] & & \\ & & & & & \ddots & \\ & & & & & & C[p_n^{e_{n,k_n}}(x)] \end{bmatrix}_{\text{block}}$$

where the polynomials $p_k^{e_{k,i}}(x)$ are the elementary divisors of τ. We also write this in the form

$$\operatorname{diag}\Big(C[p_1^{e_{1,1}}(x)], \ldots, C[p_1^{e_{1,k_1}}(x)], \ldots, C[p_n^{e_{n,n_1}}(x)], \ldots, C[p_n^{e_{n,k_n}}(x)]\Big)$$

This block diagonal matrix is said to be in **rational canonical form** *and is called a* **rational canonical form of** *τ. Except for the order of the blocks in the matrix, the rational canonical form is a canonical form for similarity, that is, up to order of the blocks, each similarity class contains exactly one matrix in rational canonical form. Put another way, the multiset of elementary divisors is a complete invariant for similarity.*

Proof. It remains to prove that if two matrices in rational canonical form are similar, then they must be equal, up to order of the blocks. Let A be the matrix $[\tau]_{\mathcal{B}}$ above. The ordered basis $\mathcal{B}$ clearly gives a decomposition of V_τ into a direct sum of primary cyclic submodules for which the elementary divisors are the polynomials $p_i^{e_{i,j}}(x)$.

Now suppose that B is another matrix in rational canonical form

$$B = \operatorname{diag}\Big(C[q_1^{f_{1,1}}(x)], \ldots, C[q_1^{q_{1,j_1}}(x)], \ldots, C[q_m^{f_{m,m_1}}(x)], \ldots, C[q_m^{f_{m,j_m}}(x)]\Big)$$

If B is similar to A then we get another primary cyclic decomposition of τ for which the elementary divisors are the polynomials $q_i^{f_{i,j}}(x)$. It follows that the two sets of elementary divisors are the same and so A and B are the same up to the order of their blocks. □

Corollary 7.11

1) *Any square matrix A is similar to a unique (except for the order of the blocks on the diagonal) matrix that is in rational canonical form. Any such matrix is called a* **rational canonical form** *of A.*
2) *Two square matrices over the same field F are similar if and only if they have the same multiset of elementary divisors.* □

Here are some examples of rational canonical forms.

Example 7.1 Let τ be a linear operator on the vector space $\mathbb{R}^7$ and suppose that τ has minimal polynomial

$$m_\tau(x) = (x-1)(x^2+1)^2$$

Noting that $x-1$ and $(x^2+1)^2$ are elementary divisors and that the sum of the degrees of all elementary divisors must equal 7, we have two possibilities

1) $x-1,\ (x^2+1)^2,\ x^2+1$
2) $x-1,\ x-1,\ x-1,\ (x^2+1)^2$

These correspond to the following rational canonical forms

$$1)\quad \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

$$2)\quad \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$ □

Exercises

1. We have seen that any $\tau \in \mathcal{L}(V)$ can be used to make V into an $F[x]$-module. Does every module V over $F[x]$ come from some $\tau \in \mathcal{L}(V)$? Explain.
2. Show that if A and B are block diagonal matrices with the same blocks, but in possibly different order, then A and B are similar.
3. Let A be a square matrix over a field F. Let K be the smallest subfield of F containing the entries of A. Prove that any rational canonical form for A has coefficients in the field K. This means that the coefficients of any rational canonical form for A are "rational" expressions in the coefficients of A, hence the origin of the term "rational canonical form." Given an operator $\tau \in \mathcal{L}(V)$ what is the smallest field K for which any rational canonical form must have entries in K?
4. Let K be a subfield of F. Prove that two matrices $A, B \in \mathcal{M}(K)$ are similar over K if and only if they are similar over F, that is $A = PBP^{-1}$

for some $P \in \mathcal{M}(K)$ if and only if $A = QBQ^{-1}$ for some $Q \in \mathcal{M}(F)$. *Hint*: Use the results of the previous exercise.

5. Prove that the minimal polynomial of $\tau \in \mathcal{L}(V)$ is the least common multiple of its elementary divisors.
6. Let $\mathbb{Q}$ be the field of rational numbers. Consider the linear operator $\tau \in \mathcal{L}(\mathbb{Q}^2)$ defined by $\tau(e_1) = e_2,\ \tau(e_2) = -e_1$.
   a) Find the minimal polynomial for τ and show that the rational canonical form for τ is
   $$R = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$
   What are the elementary divisors of τ?
   b) Now consider the map $\sigma \in \mathcal{L}(\mathbb{C}^2)$ defined by the same rules as τ, namely, $\sigma(e_1) = e_2,\ \sigma(e_2) = -e_1$. Find the minimal polynomial for σ and the rational canonical form for σ. What are the elementary divisors of σ?
   c) The invariant factors of τ are defined using the elementary divisors of τ in the same way as we did at the end of Chapter 6, for a module M. Describe the invariant factors for the operators in parts a) and b).
7. Find all rational canonical forms (up to the order of the blocks on the diagonal) for a linear operator on $\mathbb{R}^6$ having minimal polynomial $(x-1)^2(x+1)^2$.
8. How many possible rational canonical forms (up to order of blocks) are there for linear operators on $\mathbb{R}^6$ with minimal polynomial $(x-1)(x+1)^2$?
9. Prove that if C is the companion matrix of $p(x)$ then $p(C) = 0$ and C has minimal polynomial $p(x)$.
10. Let τ be a linear operator on F^4 with minimal polynomial $m_\tau(x) = (x^2+1)(x^2-2)$. Find the rational canonical form for τ if $F = \mathbb{Q}$, $F = \mathbb{R}$ or $F = \mathbb{C}$.
11. Suppose that the minimal polynomial of $\tau \in \mathcal{L}(V)$ is irreducible. What can you say about the dimension of V?

Chapter 8
Eigenvalues and Eigenvectors

Unless otherwise noted, we will assume throughout this chapter that all vector spaces are finite-dimensional.

The Characteristic Polynomial of an Operator

It is clear from our discussion of the rational canonical form that elementary divisors and their companion matrices are important. Let $C[p_n(x)]$ be the companion matrix of a monic polynomial

$$p_n(x;a_0,\ldots,a_{n-1}) = a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n$$

By way of motivation, note that when $n = 2$, we can write the polynomial $p_2(x)$ as follows

$$p_2(x;a_0,a_1) = a_0 + a_1x + x^2 = x(x + a_1) + a_0$$

which looks suspiciously like a determinant, namely,

$$\begin{aligned} p_2(x;a_0,a_1) &= \det\begin{bmatrix} x & a_0 \\ -1 & x+a_1 \end{bmatrix} \\ &= \det\left(xI - \begin{bmatrix} 0 & -a_0 \\ 1 & -a_1 \end{bmatrix}\right) \\ &= \det(xI - C[p_2(x)]) \end{aligned}$$

So, let us define

$$\begin{aligned} A(x;a_0,\ldots,a_{n-1}) &= xI - C[p_n(x)] \\ &= \begin{bmatrix} x & 0 & \cdots & 0 & a_0 \\ -1 & x & \cdots & 0 & a_1 \\ 0 & -1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & x & a_{n-2} \\ 0 & 0 & \cdots & -1 & x+a_{n-1} \end{bmatrix} \end{aligned}$$

where x is an independent variable. The determinant of this matrix is a polynomial in x whose degree equals the number of parameters $a_0,\ldots,a_{n-1}$. We have just seen that

$$\det(A(x;a_0,a_1)) = p_2(x;a_0,a_1)$$

and this is also true for $n=1$. As a basis for induction, suppose that

$$\det(A(x;a_0,\ldots,a_{n-1})) = p_n(x;a_0,\ldots,a_{n-1})$$

Then, expanding along the first row gives

$$\begin{aligned}
&\det(A(x,a_0,\ldots,a_n))\\
&\quad= x\det(A(x,a_1,\ldots,a_n)) + (-1)^n a_0 \det\begin{bmatrix} -1 & x & \cdots & 0\\ 0 & -1 & \ddots & \\ \vdots & \vdots & \ddots & x\\ 0 & 0 & \cdots & -1\end{bmatrix}_{n\times n}\\
&\quad= x\det(A(x,a_1,\ldots,a_n)) + a_0\\
&\quad= x\,p_n(x;a_1,\ldots,a_n) + a_0\\
&\quad= a_1x + a_2x^2 + \cdots + a_nx^n + x^{n+1} + a_0\\
&\quad= p_{n+1}(x;a_0,\ldots,a_n)
\end{aligned}$$

We have proved the following.

Lemma 8.1 *If $C[p(x)]$ is the companion matrix of the polynomial $p(x)$, then*

$$\det(xI - C[p(x)]) = p(x) \qquad \square$$

Since the determinant of a block diagonal matrix is the product of the determinants of the blocks on the diagonal, if R is a matrix in rational canonical form then

$$\det(xI - R) = \prod_{i,j} p_i^{e_{i,j}}(x)$$

is the product of the elementary divisors of τ. Moreover, if M is similar to R, say $M = PRP^{-1}$ then

$$\begin{aligned}
\det(xI - M) &= \det(xI - PRP^{-1})\\
&= \det\left[P(xI - R)P^{-1}\right]\\
&= \det(P)\det(xI - R)\det(P^{-1})\\
&= \det(xI - R)
\end{aligned}$$

and so $C_M(x)$ is the product of the elementary divisors of M. The polynomial $C_M(x) = \det(xI - M)$ is known as the **characteristic polynomial** of M. Since the characteristic polynomial is an invariant under similarity, we have the following.

Theorem 8.2 *Let τ be a linear operator on a finite-dimensional vector space V. The* **characteristic polynomial** *$C_\tau(x)$ of τ is defined to be the product of the elementary divisors of τ. If M is any matrix that represents τ, then*

$$C_\tau(x) = C_M(x) = \det(xI - M) \qquad \square$$

Note that the characteristic polynomial is not a *complete* invariant under similarity. For example, the matrices

$$A = \begin{bmatrix} \delta & 0 \\ 0 & \delta \end{bmatrix} \text{ and } B = \begin{bmatrix} \delta & 0 \\ 1 & \delta \end{bmatrix}$$

have the same characteristic polynomial but are not similar. (The reader might wish to provide an example of two nonsimilar matrices with the same characteristic and minimal polynomials.)

We shall have several occasions to use the fact that the minimal polynomial

$$m_\tau(x) = p_1^{e_1}(x) \cdots p_n^{e_n}(x)$$

and characteristic polynomial

$$C_\tau(x) = \prod_{i,j} p_i^{e_{i,j}}(x)$$

of a linear operator $\tau \in \mathcal{L}(V)$ have the same set of prime factors. This implies, for example, that these two polynomials have the same *set* of roots (not counting multiplicity).

Eigenvalues and Eigenvectors

Let $\tau \in \mathcal{L}(V)$ and let M be a matrix that represents τ. A scalar $\lambda \in F$ is a root of the characteristic polynomial $C_\tau(x)$ of τ if and only if

$$\det(\lambda I - M) = 0 \tag{8.1}$$

that is, if and only if the matrix $\lambda I - M$ is singular. In particular, if $\dim(V) = n$ then (8.1) holds if and only if there exists a nonzero vector $x \in F^n$ for which

$$(\lambda I - M)x = 0$$

or, equivalently

$$\tau_M(x) = \lambda x$$

If $M = [\tau]_\mathcal{B}$ and $[v]_\mathcal{B} = x$, then this is equivalent to

$$[\tau]_\mathcal{B}[v]_\mathcal{B} = \lambda[v]_\mathcal{B}$$

or, in operator language

$$\tau(v) = \lambda v$$

This prompts the following definition.

Definition *Let V be a vector space over F.*

1) *A scalar $\lambda \in F$ is an* **eigenvalue** *(or* **characteristic value***) of an operator $\tau \in \mathcal{L}(V)$ if there exists a* nonzero *vector $v \in V$ for which*

$$\tau(v) = \lambda v$$

In this case, v is an **eigenvector** *(or* **characteristic vector***) of τ associated with λ.*

2) *A scalar $\lambda \in F$ is an* **eigenvalue** *for a matrix A if there exists a* nonzero *column vector x for which*

$$Ax = \lambda x$$

In this case, x is an **eigenvector** *(or* **characteristic vector***) for A associated with λ.*

3) *The set of all eigenvectors associated with a given eigenvalue λ, together with the zero vector, forms a subspace of V, called the* ***eigenspace*** *of λ, denoted by $\mathcal{E}_\lambda$. This applies to both linear operators and matrices.*
4) *The set of all eigenvalues of an operator or matrix is called the* **spectrum** *of the operator or matrix.* □

The following theorem summarizes some key facts.

Theorem 8.3 *Let $\tau \in \mathcal{L}(V)$ have minimal polynomial $m_\tau(x)$ and characteristic polynomial $C_\tau(x)$.*

1) *The polynomials $m_\tau(x)$ and $C_\tau(x)$ have the same prime factors and hence the same set of roots, called the spectrum of τ.*
2) (**The Cayley–Hamilton theorem**) *The minimal polynomial divides the characteristic polynomial. Another way to say this is that an operator τ satisfies its own characteristic polynomial, that is,*

$$C_\tau(\tau) = 0$$

3) *The eigenvalues of a matrix are invariants under similarity.*
4) *If λ is an eigenvalue of a matrix A then the eigenspace $\mathcal{E}_\lambda$ is the solution space to the homogeneous system of equations*

$$(\lambda I - A)(x) = 0$$ □

One way to compute the eigenvalues of a linear operator τ is to first represent τ by a matrix A and then solve the **characteristic equation**

$$C_A(x) = 0$$

Unfortunately, it is quite likely that this equation cannot be solved when

$\dim(V) \geq 3$. As a result, the art of approximating the eigenvalues of a matrix is a very important area of applied linear algebra.

The following theorem describes the relationship between eigenspaces and eigenvectors of distinct eigenvalues.

Theorem 8.4 *Suppose that $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of a linear operator $\tau \in \mathcal{L}(V)$.*

1) The eigenspaces meet only in the 0 vector, that is

$$\mathcal{E}_{\lambda_i} \cap \mathcal{E}_{\lambda_j} = \{0\}$$

2) *Eigenvectors associated with distinct eigenvalues are linearly independent. That is, if $v_i \in \mathcal{E}_{\lambda_i}$ then the vectors $\{v_1, \ldots, v_k\}$ are linearly independent.*

Proof. We leave the proof of part 1) to the reader. For part 2), assume that the eigenvectors $v_i \in \mathcal{E}_{\lambda_i}$ are linearly dependent. By renumbering if necessary, we may also assume that among all nontrivial linear combinations of these vectors that equal 0, the equation

$$r_1 v_1 + \cdots + r_j v_j = 0 \tag{8.2}$$

has the fewest number of terms. Applying τ gives

$$r_1 \lambda_1 v_1 + \cdots + r_j \lambda_j v_j = 0 \tag{8.3}$$

Now we multiply (8.2) by λ_1 and subtract from (8.3), to get

$$r_2(\lambda_2 - \lambda_1)v_2 + \cdots + r_j(\lambda_j - \lambda_1)v_j = 0$$

But this equation has fewer terms than (8.2) and so all of the coefficients must equal 0. Since the λ_i's are distinct we deduce that $r_i = 0$ for $i = 2, \ldots, j$ and so $r_1 = 0$ as well. This contradiction implies that the v_i's are linearly independent. □

Geometric and Algebraic Multiplicities

Eigenvalues have two forms of multiplicity, as described in the next definition.

Definition *Let λ be an eigenvalue of a linear operator $\tau \in \mathcal{L}(V)$.*

*1) The **algebraic multiplicity** of λ is the multiplicity of λ as a root of the characteristic polynomial $C_\tau(x)$.*

*2) The **geometric multiplicity** of λ is the dimension of the eigenspace $\mathcal{E}_\lambda$.* □

Theorem 8.5 *The geometric multiplicity of an eigenvalue λ of $\tau \in \mathcal{L}(V)$ is less than or equal to its algebraic multiplicity.*

Proof. Suppose that λ is an eigenvalue of τ with eigenspace $\mathcal{E}_\lambda$. Given any basis $\mathcal{B}_1 = \{v_1, \ldots, v_k\}$ of $\mathcal{E}_\lambda$ we can extend it to a basis $\mathcal{B}$ for V. Since $\mathcal{E}_\lambda$ is invariant under τ, the matrix of τ with respect to $\mathcal{B}$ has the block form

$$[\tau]_{\mathcal{B}} = \begin{pmatrix} \lambda I_k & A \\ 0 & B \end{pmatrix}_{\text{block}}$$

where A and B are matrices of the appropriate sizes and so

$$\begin{aligned} C_\tau(x) &= \det(xI - [\tau]_{\mathcal{B}}) \\ &= \det(xI_k - \lambda I_k)\det(xI_{n-k} - A) \\ &= (x-\lambda)^k \det(xI_{n-k} - A) \end{aligned}$$

(Here n is the dimension of V.) Hence, the algebraic multiplicity of λ is at least k, which is the geometric multiplicity of τ. □

The Jordan Canonical Form

One of the virtues of the rational canonical form is that every linear operator τ on a finite-dimensional vector space has a rational canonical form. However, the rational canonical form may be far from the ideal of simplicity that we had in mind for a set of simple canonical forms.

We can do better when the minimal polynomial of τ **splits** over F, that is, factors into a product of linear factors

$$m_\tau(x) = (x-\lambda_1)^{e_1} \cdots (x-\lambda_n)^{e_n} \tag{8.4}$$

In some sense, the difficulty in the rational canonical form is the basis for the cyclic submodules $\langle v_{i,j} \rangle$. Recall that since $\langle v_{i,j} \rangle$ is a τ-cyclic subspace of V we have chosen the ordered basis

$$\mathcal{B}_{i,j} = \left(v_{i,j}, \tau(v_{i,j}), \ldots, \tau^{d_{i,j}-1}(v_{i,j})\right)$$

where $d_{i,j} = \deg(p_i^{e_{i,j}})$. With this basis, all of the complexity comes at the end, when we attempt to express

$$\tau(\tau^{d_{i,j}-1}(v_{i,j})) = \tau^{d_{i,j}}(v_{i,j})$$

as a linear combination of the basis vectors.

When the minimal polynomial $m_\tau(x)$ has the form (8.4), the elementary divisors are

$$p_i^{e_{i,j}}(x) = (x-\lambda_i)^{e_{i,j}}$$

In this case, we can choose the ordered basis

$$\mathcal{C}_{i,j} = \left(v_{i,j}, (\tau-\lambda_i)(v_{i,j}), \ldots, (\tau-\lambda_i)^{e_{i,j}-1}(v_{i,j})\right)$$

for $\langle v_{i,j} \rangle$. Denoting the kth basis vector in $\mathcal{C}_{i,j}$ by b_k, we have for $k = 0, \ldots, e_{i,j} - 2$,

$$\begin{aligned}\tau(b_k) &= \tau[(\tau-\lambda_i)^k(v_{i,j})]\\ &= (\tau-\lambda_i+\lambda_i)[(\tau-\lambda_i)^k(v_{i,j})]\\ &= (\tau-\lambda_i)^{k+1}(v_{i,j})+\lambda_i(\tau-\lambda_i)^k(v_{i,j})\\ &= b_{k+1}+\lambda_i b_k\end{aligned}$$

For $k=e_{i,j}-1$, a similar computation, using the fact that

$$(\tau-\lambda_i)^{k+1}(v_{i,j})=(\tau-\lambda_i)^{e_{i,j}}(v_{i,j})=0$$

gives

$$\tau(b_{e_{i,j}-1})=\lambda_i b_{e_{i,j}-1}$$

In this case, the complexity is more or less spread out evenly, and the matrix of $\tau|_{\langle v_{i,j}\rangle}$ with respect to $\mathcal{C}_{i,j}$ is the $e_{i,j}\times e_{i,j}$ matrix

$$\mathcal{J}(\lambda_i,e_{i,j})=\begin{bmatrix}\lambda_i & 0 & \cdots & \cdots & 0\\ 1 & \lambda_i & \ddots & & \vdots\\ 0 & 1 & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & 1 & \lambda_i\end{bmatrix}$$

which is called a **Jordan block** associated with the scalar λ_i. Note that a Jordan block has λ_i's on the main diagonal, 1's on the subdiagonal and 0's elsewhere. This matrix is, in general, simpler (or at least more aesthetically pleasing) than a companion matrix.

Now we can state the analog of Theorem 7.10 for this choice of ordered basis.

Theorem 8.6 *(***The Jordan canonical form***) Let* $\dim(V)<\infty$ *and suppose that the minimal polynomial of* $\tau\in\mathcal{L}(V)$ *splits over the base field F, that is,*

$$m_\tau(x)=(x-\lambda_1)^{e_1}\cdots(x-\lambda_n)^{e_n}$$

Then V *has an ordered basis* $\mathcal{C}$ *under which*

$$[\tau]_{\mathcal{C}}=\begin{bmatrix}\mathcal{J}(\lambda_1,e_{1,1}) & & & & \\ & \ddots & & & \\ & & \mathcal{J}(\lambda_1,e_{1,k_1}) & & \\ & & & \ddots & \\ & & & & \mathcal{J}(\lambda_n,e_{n,k_n}) \\ & & & & & \ddots \\ & & & & & & \mathcal{J}(\lambda_n,e_{n,k_n})\end{bmatrix}_{\text{block}}$$

where the polynomials $(x-\lambda_i)^{e_{i,j}}$ *are the elementary divisors of* τ*. This block diagonal matrix is said to be in* **Jordan canonical form** *and is called the* **Jordan canonical form of** τ.

If the base field F is algebraically closed, then except for the order of the blocks in the matrix, Jordan canonical form is a canonical form for similarity, that is, up to order of the blocks, each similarity class contains exactly *one matrix in Jordan canonical form.*

Proof. As to the uniqueness, suppose that $\mathcal{J}$ is a matrix in Jordan canonical form that represents the operator τ with respect to some ordered basis $\mathcal{B}$, and that $\mathcal{J}$ has Jordan blocks $\mathcal{J}_1(\lambda_1, f_1), \ldots, \mathcal{J}_m(\lambda_m, f_m)$, where the λ_i's may not be distinct. Then V is the direct sum of τ-invariant subspaces, that is, submodules of V_τ, say

$$V = V_1 \oplus \cdots \oplus V_m$$

Consider a particular submodule V_k. it is easy to see from the matrix representation that $\tau|_{V_k}$ satisfies the polynomial $(x-\lambda_k)^{f_k}$ on V_k, but no polynomial of the form $(x-\lambda_k)^d$ for $d < f_k$, and so the order of V_k is $(x-\lambda_k)^{f_k}$. In particular, each V_k is a primary submodule of V_τ.

We claim that V_k is also a cyclic submodule of V_τ. To see this, let $(v_1, \ldots, v_{f_k})$ be the ordered basis that gives the Jordan block $\mathcal{J}(\lambda_k, f_k)$. Then it is easy to see by induction that $\tau^j v_1$ is a linear combination of $v_1, \ldots, v_{j+1}$, with coefficient of v_{j+1} equal to 1 or -1. Hence, the set

$$\{v_1, \tau v_1, \tau^2 v_1, \ldots, \tau^{f_k-1} v_1\}$$

is also a basis for V_k, from which it follows that V_k is a τ-cyclic subspace of V, that is, a cyclic submodule of V_τ.

Thus, the Jordan matrix $\mathcal{J}$ corresponds to a primary cyclic decomposition of V_τ with elementary divisors $(x-\lambda_k)^{f_k}$. Since the multiset of elementary divisors is unique, so is the Jordan matrix representation of τ, up to order of the blocks. □

Note that if τ has Jordan canonical form $\mathcal{J}$ then the diagonal elements of $\mathcal{J}$ are precisely the eigenvalues of τ, each appearing a number of times equal to its algebraic multiplicity.

Triangularizability and Schur's Lemma

We have now discussed two different canonical forms for similarity: the rational canonical form, which applies in all cases and the Jordan canonical form, which applies only when the base field is algebraically closed. Let us now drop the rather strict requirements of canonical forms and look at two classes of matrices that are too large to be canonical forms (the upper triangular matrices and the almost upper triangular matrices) and a class of matrices that is too small to be a canonical form (the diagonal matrices).

The upper triangular matrices (or lower triangular matrices) have some nice properties and it is of interest to know when an arbitrary matrix is similar to a triangular matrix. We confine our attention to upper triangular matrices, since there are direct analogs for lower triangular matrices as well.

It will be convenient to make the following, somewhat nonstandard, definition.

Definition *A linear operator τ on V is* **upper triangular** *with respect to an ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$ if the matrix $[\tau]_{\mathcal{B}}$ is upper triangular, that is, if for all $i = 1, \ldots, n$*

$$\tau(v_i) \in \langle v_1, \ldots, v_i \rangle$$

The operator τ is **upper triangularizable** *if there is an ordered basis with respect to which τ is upper triangular.* □

As we will see next, when the base field is algebraically closed, all operators are upper triangularizable. However, since two distinct upper triangular matrices can be similar, the class of upper triangular matrices is not a canonical form for similarity. Simply put, there are just too many upper triangular matrices.

Theorem 8.7 (Schur's Lemma) *Let V be a finite-dimensional vector space over a field F.*

1) *If $\tau \in \mathcal{L}(V)$ has the property that its characteristic polynomial $C_\tau(x)$ splits over F then τ is upper triangularizable.*
2) *If F is algebraically closed then all operators are upper triangularizable.*

Proof. Part 2) follows from part 1). The proof of part 1) is most easily accomplished by matrix means, namely, we prove that every square matrix $A \in M_n(F)$ whose characteristic polynomial splits over F is similar to an upper triangular matrix. If $n = 1$ there is nothing to prove, since all 1×1 matrices are upper triangular. Assume the result is true for $n - 1$ and let $A \in M_n(F)$.

Let v_1 be an eigenvector associated with the eigenvalue $\lambda_1 \in F$ of A and extend $\{v_1\}$ to an ordered basis $(v_1, ..., v_n)$ for $\mathbb{R}^n$. The matrix of A with respect to $\mathcal{B}$ has the form

$$[A]_{\mathcal{B}} = \begin{bmatrix} \lambda_1 & * \\ 0 & A_1 \end{bmatrix}_{\text{block}}$$

for some $A_1 \in M_{n-1}(F)$. Since $[A]_{\mathcal{B}}$ and A are similar, we have

$$\det(xI - A) = \det(xI - [A]_{\mathcal{B}}) = (x - \lambda_1)\det(xI - A_1)$$

Hence, the characteristic polynomial of A_1 also splits over F and, by the induction hypothesis, there exists an invertible matrix $P \in M_{n-1}(F)$ for which

$$U = PA_1P^{-1}$$

is upper triangular. Hence, if

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & P \end{bmatrix}_{\text{block}}$$

then Q is invertible and

$$Q[A]_{\mathcal{B}}Q^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & P \end{bmatrix}\begin{bmatrix} \lambda_1 & * \\ 0 & A_1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & P^{-1} \end{bmatrix} = \begin{bmatrix} \lambda_1 & * \\ 0 & U \end{bmatrix}$$

is upper triangular. This completes the proof. □

When the base field is $F = \mathbb{R}$, not all operators are triangularizable. We can, however, achieve a form that is close to triangular. For the sake of the exposition, we make the following *nonstandard* definition (that is, the reader should not expect to find this definition in other books).

Definition A matrix $A \in M_n(F)$ is **almost upper triangular** if it has the form

$$A = \begin{bmatrix} A_1 & & * & \\ & A_2 & & \\ & & \ddots & \\ & 0 & & A_k \end{bmatrix}_{block}$$

where each matrix A_i either has size 1×1 or else has size 2×2 with an irreducible characteristic polynomial. A linear operator $\tau \in \mathcal{L}(V)$ is **almost upper triangularizable** if there is an ordered basis $\mathcal{B}$ for which $[\tau]_{\mathcal{B}}$ is almost upper triangular. □

We will prove that every real linear operator is almost upper triangularizable. In the case of a complex vector space V, any complex linear operator $\tau \in \mathcal{L}(V)$ has an eigenvalue and hence V contains a one-dimensional τ-invariant subspace. The analog for the real case is that for any real linear operator $\tau \in \mathcal{L}(V)$, the vector space V contains either a one-dimensional or a "nonreducible" two-dimensional τ-invariant subspace.

Theorem 8.8 *Let $\tau \in \mathcal{L}(V)$ be a real linear operator. Then V contains at least one of the following:*

1) *A one-dimensional τ-invariant subspace,*
2) *A two-dimensional τ-invariant subspace W for which $\sigma = \tau|_W$ has the property that $m_\sigma(x) = C_\sigma(x)$ is an irreducible quadratic. Hence, W is not the direct sum of two one-dimensional τ-invariant subspaces.*

Proof. The minimal polynomial $m_\tau(x)$ of τ factors into a product of linear and quadratic factors over $\mathbb{R}$. If there is a linear factor $x - \lambda$, then λ is an eigenvalue for τ and if $\tau v = \lambda v$ then $\langle v \rangle$ is the desired one-dimensional τ-invariant subspace.

Otherwise, let $p(x) = x^2 + ax + b$ be an irreducible quadratic factor of $m_\tau(x)$ and write

$$m_\tau(x) = p(x)q(x)$$

Since $q(\tau) \neq 0$, we may choose a nonzero vector $v \in V$ such that $q(\tau)v \neq 0$. Let

$$W = \langle q(\tau)v, \tau q(\tau)v \rangle$$

This subspace is τ-invariant, for we have $\tau[q(\tau)v] \in W$ and

$$\tau[\tau q(\tau)v] = \tau^2 q(\tau)v = -(a\tau + b)q(\tau)v \in W$$

Hence, $\sigma = \tau|_W$ is a linear operator on W. Also,

$$p(\tau)W = \{0\}$$

and so σ has minimal polynomial dividing $p(x)$. But since $p(x)$ is irreducible and monic, $m_\sigma(x) = p(x)$ is quadratic. It follows that W is two-dimensional, for if

$$aq(\tau)v + b\tau q(\tau)v = 0$$

then $a + b\tau = 0$ on W, which is not the case. Finally, the characteristic polynomial $C_\sigma(x)$ has degree 2 and is divisible by $m_\sigma(x)$, whence $C_\sigma(x) = m_\sigma(x) = p(x)$ is irreducible. Thus, W satisfies condition 2). □

Now we can prove Schur's lemma for real operators.

Theorem 8.9 *(***Schur's lemma: real case***) Every real linear operator $\tau \in \mathcal{L}(V)$ is almost upper triangularizable.*

Proof. As with the complex case, it is simpler to proceed using matrices, by showing that any $n \times n$ real matrix A is similar to an almost upper triangular matrix. The result is clear for $n = 1$ or if A is the zero matrix.

For $n = 2$, the characteristic polynomial $C(x)$ of A has degree 2 and is divisible by the minimal polynomial $m(x)$. If $m(x) = x - \lambda$ is linear then $A = \lambda I_2$ is diagonal. If $m(x) = (x - \lambda)^2$ then A is similar to an upper triangular matrix with diagonal elements λ and if $m(x) = (x - \lambda)(x - \mu)$ with $\lambda \neq \mu$ then A is similar to a diagonal matrix with diagonal entries λ and μ. Finally, if $m(x) = C(x)$ is irreducible then the result still holds.

Assume for the purposes of induction that any square matrix of size less than $n \times n$ is almost upper triangularizable. We wish to show that the same is true for any $n \times n$ matrix A. We may assume that $n \geq 3$.

If A has an eigenvector $v_1 \in \mathbb{R}^n$, then let $W = \langle v_1 \rangle$. If not, then according to Theorem 8.8, there is a pair of vectors $u_1, u_2 \in \mathbb{R}^n$ for which $W = \langle w_1, w_2 \rangle$ is two-dimensional and A-invariant and the characteristic and minimal polynomials of $\tau_A|_W$ are equal and irreducible. Let U be a complement of W. If $\dim(W) = 1$ then let $\mathcal{B} = (v_1, u_1, \ldots, u_{n-1})$ be an ordered basis for $\mathbb{R}^n$ and if $\dim(W) = 2$ then let $\mathcal{B} = (w_1, w_2, u_1, \ldots, u_{n-2})$ be an ordered basis for $\mathbb{R}^n$. In either case, A is similar to a matrix of the form

$$[A]_{\mathcal{B}} = \begin{bmatrix} B_1 & * \\ 0 & A_1 \end{bmatrix}_{\text{block}}$$

where B_1 has size 1×1 or B_1 has size 2×2, with irreducible quadratic minimal polynomial. Also, A_1 has size $k \times k$, where $k = n-1$ or $k = n-2$. Hence, the induction hypothesis applies to A_1 and there exists an invertible matrix $P \in M_k$ for which

$$U = PA_1P^{-1}$$

is almost upper triangular. Hence, if

$$Q = \begin{bmatrix} I_{n-k} & 0 \\ 0 & P \end{bmatrix}_{\text{block}}$$

then Q is invertible and

$$Q[A]_{\mathcal{B}}Q^{-1} = \begin{bmatrix} I_{n-k} & 0 \\ 0 & P \end{bmatrix} \begin{bmatrix} B_1 & * \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} I_{n-k} & 0 \\ 0 & P^{-1} \end{bmatrix} = \begin{bmatrix} B_1 & * \\ 0 & U \end{bmatrix}$$

is almost upper triangular. This completes the proof. □

Unitary Triangularizability

Although we have not yet discussed inner product spaces and orthonormal bases, the reader is no doubt familiar with these concepts. So let us mention that when V is a real or complex inner product space, then if an operator τ on V can be triangularized (or almost triangularized) using an ordered basis $\mathcal{B}$, it can also be triangularized (or almost triangularized) using an *orthonormal* ordered basis $\mathcal{O}$.

To see this, suppose we apply the Gram–Schmidt orthogonalization process to $\mathcal{B} = (v_1, \ldots, v_n)$. The resulting ordered orthonormal basis $\mathcal{O} = (u_1, \ldots, u_n)$ has the property that

$$\langle v_1, \ldots, v_i \rangle = \langle u_1, \ldots, u_i \rangle$$

for all $i \le n$. Since $[\tau]_{\mathcal{B}}$ is upper triangular, that is,

$$\tau(v_i) \in \langle v_1, \ldots, v_i \rangle$$

for all $i \le n$, it follows that

$$\tau(u_i) \in \langle \tau v_1, \ldots, \tau v_i \rangle \subseteq \langle v_1, \ldots, v_i \rangle = \langle u_1, \ldots, u_i \rangle$$

and so the matrix $[\tau]_{\mathcal{O}}$ is also upper triangular. (A similar argument holds in the almost upper triangular case.)

A linear operator τ is **unitarily upper triangularizable** if there is an ordered *orthonormal* basis with respect to which τ is upper triangular. Accordingly, when V is an inner product space, we can replace the term "upper triangularizable" with "unitarily upper triangularizable" in Schur's lemma. (A similar statement holds for almost upper triangular matrices.)

Diagonalizable Operators

A linear operator $\tau \in \mathcal{L}(V)$ is **diagonalizable** if there is an ordered basis $\mathcal{B}$ for which $[\tau]_{\mathcal{B}}$ is diagonal. In the case of an algebraically closed field, we have seen that all operators are upper triangularizable. However, even for such fields, not all operators are diagonalizable.

Our first characterization of diagonalizability amounts to little more than the definitions of the concepts involved.

Theorem 8.10 *An operator $\tau \in \mathcal{L}(V)$ is diagonalizable if and only if there is a basis for V that consists entirely of eigenvectors of τ, that is, if and only if*

$$V = \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_k}$$

where $\lambda_1, \ldots, \lambda_k$ are the distinct *eigenvalues of τ.* □

Diagonalizability can also be characterized via minimal polynomials. Suppose that τ is diagonalizable and that $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V consisting of eigenvectors of τ. Let $\lambda_1, \ldots, \lambda_k$ be a list of the *distinct* eigenvalues of τ. Then each basis vector v_i is an eigenvector for one of these eigenvalues and so

$$\prod_{i=1}^{k} (\tau - \lambda_i) v_j = 0$$

for all basis vectors v_j. Hence, if

$$p(x) = \prod_{i=1}^{k} (x - \lambda_i)$$

then $p(\tau) = 0$ and so $m_\tau(x) \mid p(x)$. But every eigenvalue λ_j is a root of the minimal polynomial of τ and so $p(x) \mid m_\tau(x)$, whence $p(x) = m_\tau(x)$.

Conversely, if the minimal polynomial of τ is a product of distinct linear factors, then the primary decomposition of V looks like

$$V = V_1 \oplus \cdots \oplus V_k$$

where

$$V_i = \{v \in V \mid (\tau - \lambda_i)v = 0\} = \mathcal{E}_{\lambda_i}$$

By Theorem 8.10, τ is diagonalizable. We have established the following result.

Theorem 8.11 *A linear operator $\tau \in \mathcal{L}(V)$ on a finite-dimensional vector space is diagonalizable if and only if its minimal polynomial is the product of* distinct *linear factors.* □

Projections

We have met the following type of operator before.

Definition *Let $V = S \oplus T$. The linear operator $\rho: V \to V$ defined by*

$$\rho(s+t) = s$$

where $s \in S$ and $t \in T$ is called **projection** *on S* **along** *T.* □

The following theorem describes projection operators.

Theorem 8.12

1) *Let ρ be projection on S along T. Then*
   a) $\text{im}(\rho) = S$, $\ker(\rho) = T$
   b) $V = \text{im}(\rho) \oplus \ker(\rho)$
   c) $v \in \text{im}(\rho) \Leftrightarrow \rho(v) = v$

   Note that the last condition says that a vector is in the image of ρ if and only if it is fixed *by ρ.*
2) *Conversely, if $\sigma \in \mathcal{L}(V)$ has the property that*

$$V = \text{im}(\sigma) \oplus \ker(\sigma) \text{ and } \sigma|_{\text{im}(\sigma)} = \iota$$

then σ is projection on $\text{im}(\sigma)$ *along* $\ker(\sigma)$. □

Projection operators play a major role in the spectral theory of linear operators, which we will discuss in Chapter 10. Now we turn to some of the basic properties of these operators.

Theorem 8.13 *A linear operator $\rho \in \mathcal{L}(V)$ is a projection if and only if it is idempotent, that is, if and only if $\rho^2 = \rho$.*
Proof. If ρ is projection on S along T then for any $s \in S$ and $t \in T$,

$$\rho^2(s+t) = \rho(s) = s = \rho(s+t)$$

and so $\rho^2 = \rho$. Conversely, suppose that ρ is idempotent. If $v \in \text{im}(\rho) \cap \ker(\rho)$ then $v = \rho(x)$ and so

$$0 = \rho(v) = \rho^2(x) = \rho(x) = v$$

Hence $\text{im}(\rho) \cap \ker(\rho) = \{0\}$. Moreover, if $v \in V$ then

$$v = [v - \rho(v)] + \rho(v) \in \ker(\rho) \oplus \operatorname{im}(\rho)$$

and so $V = \ker(\rho) \oplus \operatorname{im}(\rho)$. Finally, $\rho[\rho(x)] = \rho(x)$ and so $\rho|_{\operatorname{im}(\rho)} = \iota$. Hence, ρ is projection on $\operatorname{im}(\rho)$ along $\ker(\rho)$. □

The Algebra of Projections

If ρ is projection on S along T then $\iota - \rho$ is idempotent, since

$$(\iota - \rho)^2 = \iota^2 - \iota\rho - \rho\iota + \rho^2 = \iota - \rho$$

Hence, $\iota - \rho$ is also a projection. Since $\ker(\iota - \rho) = \operatorname{im}(\rho)$ and $\operatorname{im}(\iota - \rho) = \ker(\rho)$, it follows that $\iota - \rho$ is projection on T along S.

Orthogonal Projections

Definition *The projections* $\rho, \sigma \in \mathcal{L}(V)$ *are* **orthogonal***, written* $\rho \perp \sigma$*, if* $\rho\sigma = \sigma\rho = 0$. □

Note that $\rho \perp \sigma$ if and only if

$$\operatorname{im}(\rho) \subseteq \ker(\sigma) \text{ and } \operatorname{im}(\sigma) \subseteq \ker(\rho)$$

The following example shows that it is not enough to have $\rho\sigma = 0$ in the definition of orthogonality. In fact, it is possible for $\rho\sigma = 0$ and yet $\sigma\rho$ is not even a projection.

Example 8.1 Let $V = F^2$ and let

$$\begin{aligned} D &= \{(x, x) \mid x \in F\} \\ X &= \{(x, 0) \mid x \in F\} \\ Y &= \{(0, y) \mid y \in F\} \end{aligned}$$

Thus, D is the diagonal, X is the x-axis and Y is the y-axis in F^2. (The reader may wish to draw pictures in $\mathbb{R}^2$.) Using the notation $\rho_{A,B}$ for the projection on A along B, we have

$$\rho_{D,X}\rho_{D,Y} = \rho_{D,Y} \neq \rho_{D,X} = \rho_{D,Y}\rho_{D,X}$$

From this we deduce that if ρ and σ are projections, it may happen that both products $\rho\sigma$ and $\sigma\rho$ are projections, but that they are not equal.

We leave it to the reader to show that $\rho_{Y,X}\rho_{X,D} = 0$ (which is a projection), but that $\rho_{X,D}\rho_{Y,X}$ is not a projection. Thus, it may also happen that $\rho\sigma$ is a projection but that $\sigma\rho$ is not a projection. □

If ρ and σ are projections, it does not necessarily follow that $\rho + \sigma$, $\rho - \sigma$ or $\rho\sigma$ is a projection. Let us consider these one by one.

The Sum of Projections

The sum $\rho + \sigma$ is a projection if and only if

$$(\rho + \sigma)^2 = \rho + \sigma$$

or

$$\rho\sigma + \sigma\rho = 0 \tag{8.5}$$

Of course, this holds if $\rho\sigma = \sigma\rho = 0$, that is, if $\rho \perp \sigma$. We contend that the converse is also true, namely, that (8.5) implies that $\rho \perp \sigma$.

Multiplying (8.5) on the left by ρ and on the right by ρ gives the pair of equations

$$\begin{aligned}\rho\sigma + \rho\sigma\rho &= 0\\ \rho\sigma\rho + \sigma\rho &= 0\end{aligned}$$

Hence $\rho\sigma = \sigma\rho$ which together with (8.5) gives $2\rho\sigma = 0$. Therefore, if $\operatorname{char}(F) \neq 2$ then $\rho\sigma = 0$ and therefore $\sigma\rho = 0$, that is, $\rho \perp \sigma$. We have proven that $\rho + \sigma$ is a projection if and only if $\rho \perp \sigma$ (assuming that F has characteristic different from 2).

Now suppose that $\rho + \sigma$ is a projection. To determine $\ker(\rho + \sigma)$, suppose that

$$(\rho + \sigma)(v) = 0$$

Applying ρ and noting that $\rho^2 = \rho$ and $\rho\sigma = 0$, we get $\rho(v) = 0$. Similarly, $\sigma(v) = 0$ and so $\ker(\rho + \sigma) \subseteq \ker(\rho) \cap \ker(\sigma)$. But the reverse inclusion is obvious and so

$$\ker(\rho + \sigma) = \ker(\rho) \cap \ker(\sigma)$$

As to the image of $\rho + \sigma$, we have

$$v \in \operatorname{im}(\rho + \sigma) \Rightarrow v = (\rho + \sigma)(v) = \rho(v) + \sigma(v) \in \operatorname{im}(\rho) + \operatorname{im}(\sigma)$$

and so $\operatorname{im}(\rho + \sigma) \subseteq \operatorname{im}(\rho) + \operatorname{im}(\sigma)$. But $\rho\sigma = 0$ implies that $\operatorname{im}(\sigma) \subseteq \ker(\rho)$ and so the sum is direct and

$$\operatorname{im}(\rho + \sigma) \subseteq \operatorname{im}(\rho) \oplus \operatorname{im}(\sigma)$$

For the reverse inequality, if $v = r + s$, where $r \in \operatorname{im}(\rho)$ and $s \in \operatorname{im}(\sigma)$ then

$$(\rho + \sigma)(v) = (\rho + \sigma)(r) + (\rho + \sigma)(s) = r + s = v$$

and so $v \in \operatorname{im}(\rho + \sigma)$. Let us summarize.

Theorem 8.14 *Let $\rho, \sigma \in \mathcal{L}(V)$ be projections where V is a vector space over a field of characteristic $\neq 2$. Then $\rho + \sigma$ is a projection if and only if $\rho \perp \sigma$, in which case $\rho + \sigma$ is projection on* $\operatorname{im}(\rho) \oplus \operatorname{im}(\sigma)$ *along* $\ker(\rho) \cap \ker(\sigma)$. □

The Difference of Projections

Let ρ and σ be projections. The difference $\rho-\sigma$ is a projection if and only if

$$\theta=\iota-(\rho-\sigma)=(\iota-\rho)+\sigma$$

is a projection. Hence, we may apply the previous theorem to deduce that $\rho-\sigma$ is a projection if and only if

$$(\iota-\rho)\sigma=\sigma(\iota-\rho)=0$$

or, equivalently,

$$\rho\sigma=\sigma\rho=\sigma$$

Moreover, in this case, $\rho-\sigma=\iota-\theta$ is projection on $\ker(\theta)$ along $\operatorname{im}(\theta)$. Theorem 8.14 also implies that

$$\operatorname{im}(\theta)=\operatorname{im}(\iota-\rho)\oplus\operatorname{im}(\sigma)=\ker(\rho)\oplus\operatorname{im}(\sigma)$$

and

$$\ker(\theta)=\ker(\iota-\rho)\cap\ker(\sigma)=\operatorname{im}(\rho)\cap\ker(\sigma)$$

Theorem 8.15 *Let $\rho,\sigma\in\mathcal{L}(V)$ be projections where V is a vector space over a field of characteristic $\neq 2$. Then $\rho-\sigma$ is a projection if and only if*

$$\rho\sigma=\sigma\rho=\sigma$$

in which case $\rho-\sigma$ is projection on $\operatorname{im}(\rho)\cap\ker(\sigma)$ along $\ker(\rho)\oplus\operatorname{im}(\sigma)$. □

The Product of Projections

Finally, let us consider the product $\rho\sigma$ of two projections.

Theorem 8.16 *Let $\rho,\sigma\in\mathcal{L}(V)$ be projections. If ρ and σ commute, that is, if $\rho\sigma=\sigma\rho$ then $\rho\sigma$ is a projection. In this case, $\rho\sigma$ is projection on* $\operatorname{im}(\rho)\cap\operatorname{im}(\sigma)$ *along* $\ker(\rho)+\ker(\sigma)$. *(Example 8.1 shows that the converse may be false.)*

Proof. If $\rho\sigma=\sigma\rho$ then

$$(\rho\sigma)^2=\rho\sigma\rho\sigma=\rho^2\sigma^2=\rho\sigma$$

and so $\rho\sigma$ is a projection. To find the image of $\rho\sigma$, observe that if $v=\rho\sigma(v)$ then

$$\rho(v)=\rho^2\sigma(v)=\rho\sigma(v)=v$$

and so $v\in\operatorname{im}(\rho)$. Similarly $v\in\operatorname{im}(\sigma)$ and so

$$\operatorname{im}(\rho\sigma)\subseteq\operatorname{im}(\rho)\cap\operatorname{im}(\sigma)$$

For the reverse inclusion, if $x=\rho(v)=\sigma(w)\in\operatorname{im}(\rho)\cap\operatorname{im}(\sigma)$ then

$$\rho\sigma(x) = \rho\sigma^2(w) = \rho\sigma(w) = \rho(x) = \rho^2(v) = \rho(v) = x$$

and so $x \in \operatorname{im}(\rho\sigma)$. Hence,

$$\operatorname{im}(\rho\sigma) = \operatorname{im}(\rho) \cap \operatorname{im}(\sigma)$$

Next, we observe that if $v \in \ker(\rho\sigma)$ then $\rho\sigma(v) = 0$ and so $\sigma(v) \in \ker(\rho)$. Hence,

$$v = \sigma(v) + (v - \sigma(v)) \in \ker(\rho) + \ker(\sigma)$$

Moreover, if $v = r + s \in \ker(\rho) + \ker(\sigma)$ then

$$\rho\sigma(v) = \rho\sigma(r + s) = \sigma\rho(r) + \rho\sigma(s) = 0 + 0 = 0$$

and so $v \in \ker(\rho\sigma)$. Thus,

$$\ker(\rho\sigma) = \ker(\rho) + \ker(\sigma)$$

We should remark that the sum above need not be direct. For example, if $\rho = \sigma$ then $\ker(\rho) = \ker(\sigma)$. □

Resolutions of the Identity

If ρ is a projection then

$$\rho \perp (\iota - \rho) \text{ and } \rho + (\iota - \rho) = \iota$$

Let us generalize this to more than two projections.

Definition *If $\rho_1, \ldots, \rho_k$ are mutually orthogonal projections, that is, $\rho_i \perp \rho_j$ for $i \neq j$ and if*

$$\rho_1 + \cdots + \rho_k = \iota$$

where ι is the identity operator then we refer to this sum as a **resolution of the identity**. □

There is a connection between the resolutions of the identity map on V and the decomposition of V. In general, if the linear operators σ_i on V satisfy

$$\sigma_1 + \cdots + \sigma_k = \iota$$

then for any $v \in V$ we have

$$v = \iota v = \sigma_1(v) + \cdots + \sigma_k(v) \in \operatorname{im}(\sigma_1) + \cdots + \operatorname{im}(\sigma_k)$$

and so

$$V = \operatorname{im}(\sigma_1) + \cdots + \operatorname{im}(\sigma_k)$$

However, the sum need not be direct. The next theorem describes when the sum is direct.

Theorem 8.17 *Resolutions of the identity correspond to direct sum decompositions of* V *in the following sense:*

1) *If* $\rho_1 + \cdots + \rho_k = \iota$ *is a resolution of the identity then*

$$V = \operatorname{im}(\rho_1) \oplus \cdots \oplus \operatorname{im}(\rho_k)$$

and ρ_i *is projection on* $\operatorname{im}(\rho_i)$ *along*

$$\ker(\rho_i) = \bigoplus_{j \neq i} \operatorname{im}(\rho_j)$$

2) *Conversely, suppose that*

$$V = S_1 \oplus \cdots \oplus S_k$$

If ρ_i *is projection on* S_i *along the direct sum of the other subspaces*

$$\bigoplus_{j \neq i} S_j$$

then $\rho_1 + \cdots + \rho_k = \iota$ *is a resolution of the identity.*

Pro**of.** To prove 1) suppose that $\rho_1 + \cdots + \rho_k = \iota$ is a resolution of the identity. Then as we have seen

$$V = \operatorname{im}(\rho_1) + \cdots + \operatorname{im}(\rho_k)$$

To see that the sum is direct, if

$$\rho_1 x_1 + \cdots + \rho_n x_n = 0$$

then applying ρ_i gives $\rho_i x_i = \rho_i^2 x_i = 0$ for all i. Hence, the sum is direct. Finally, we have

$$\operatorname{im}(\rho_i) \oplus \bigoplus_{j \neq i} \operatorname{im}(\rho_j) = V = \operatorname{im}(\rho_i) \oplus \ker(\rho_i)$$

which implies that

$$\ker(\rho_i) = \bigoplus_{j \neq i} \operatorname{im}(\rho_j)$$

To prove part 2), observe that for $i \neq j$,

$$\operatorname{im}(\rho_i) = S_i \subseteq \ker(\rho_j)$$

and similarly $\operatorname{im}(\rho_j) \subseteq \ker(\rho_i)$. Hence, $\rho_i \perp \rho_j$. Also, if $v = s_1 + \cdots + s_k$ where $s_i \in S_i$ then

$$\iota v = s_1 + \cdots + s_k = \rho_1(v) + \cdots + \rho_k(v)$$

and so $\iota = \rho_1 + \cdots + \rho_k$ is a resolution of the identity. □

Spectral Resolutions

Let us try to do something similar to Theorem 8.17 for an arbitrary linear operator τ on V (rather than just the identity ι). Suppose that τ can be resolved as follows

$$\tau = \lambda_1\rho_1 + \cdots + \lambda_k\rho_k$$

where $\rho_1 + \cdots + \rho_k = \iota$ is a resolution of the identity and $\lambda_i \in F$. Then

$$V = \operatorname{im}(\rho_1) \oplus \cdots \oplus \operatorname{im}(\rho_k)$$

Moreover, if $v \in \operatorname{im}(\rho_j)$ then $v = \rho_j(x)$ and so

$$\tau(v) = (\lambda_1\rho_1 + \cdots + \lambda_k\rho_k)\rho_j(x) = \lambda_j\rho_j(x) = \lambda_j v$$

Hence, $\operatorname{im}(\rho_j) \subseteq \mathcal{E}_{\lambda_j}$. But the reverse is also true, since the equation $\tau(v) = \lambda_j v$ is

$$(\lambda_1\rho_1 + \cdots + \lambda_k\rho_k)(v) = \lambda_j(\rho_1 + \cdots + \rho_k)v$$

or

$$(\lambda_1 - \lambda_j)\rho_1(v) + \cdots + (\lambda_k - \lambda_j)\rho_k(v) = 0$$

But since $(\lambda_i - \lambda_j)\rho_i(v) \in \operatorname{im}(\rho_i)$, we deduce that $\rho_i(v) = 0$ for $i \neq j$ and so

$$v = (\rho_1 + \cdots + \rho_k)v = \rho_j v \in \operatorname{im}(\rho_j)$$

Thus, $\operatorname{im}(\rho_j) = \mathcal{E}_{\lambda_j}$ and we can conclude that

$$V = \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_k}$$

that is, τ is diagonalizable. The converse also holds, for if V is the direct sum of the eigenspaces of τ and if ρ_i is projection on $\mathcal{E}_{\lambda_i}$ along the direct sum of the other eigenspaces then

$$\rho_1 + \cdots + \rho_k = \iota$$

But for any $v_i \in \mathcal{E}_{\lambda_i}$, we have

$$\tau(v_i) = \lambda_i v_i = \lambda_i(\rho_1 + \cdots + \rho_k)v_i = (\lambda_1\rho_1 + \cdots + \lambda_k\rho_k)(v_i)$$

and so

$$\tau = \lambda_1\rho_1 + \cdots + \lambda_k\rho_k$$

Theorem 8.18 *A linear operator $\tau \in \mathcal{L}(V)$ is diagonalizable if and only if it can be written in the form*

$$\tau = \lambda_1\rho_1 + \cdots + \lambda_k\rho_k \tag{8.6}$$

where the λ_i's are distinct and $\rho_1 + \cdots + \rho_k = \iota$ is a resolution of the identity. In this case, $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ and the projections ρ_i satisfy

$$\operatorname{im}(\rho_i) = \mathcal{E}_{\lambda_i} \textit{ and } \ker(\rho_i) = \bigoplus_{j \neq i} \mathcal{E}_{\lambda_j}$$

Equation (8.6) is referred to as the **spectral resolution** *of* τ. □

Projections and Invariance

There is a connection between projections and invariant subspaces. Suppose that S is a τ-invariant subspace of V and let ρ be any projection onto S (along any complement of S). Then for any $v \in V$, we have $\rho(v) \in S$ and so $\tau(\rho(v)) \in S$. Hence, $\tau(\rho(v))$ is fixed by ρ, that is

$$\rho\tau\rho(v) = \tau\rho(v)$$

Thus $\rho\tau\rho = \tau\rho$. Conversely, if $\rho\tau\rho = \tau\rho$ then for any $s \in S$, we have $s = \rho(s)$, whence

$$\rho\tau(s) = \rho\tau\rho(s) = \tau\rho(s) = \tau(s)$$

and so $\tau(s)$ is fixed by ρ, from which it follows that $\tau(s) \in S$. In other words, S is τ-invariant.

Theorem 8.19 *Let* $\tau \in \mathcal{L}(V)$. *A subspace* S *of* V *is* τ*-invariant if and only if*

$$\rho\tau\rho = \tau\rho$$

for some projection ρ *on* S. □

We also have the following relationship between projections and reducing pairs.

Theorem 8.20 *Let* $V = S \oplus T$. *Then a linear operator* $\tau \in \mathcal{L}(V)$ *is reduced by the pair* (S,T) *if and only if* $\tau\rho = \rho\tau$, *where* ρ *is projection on* S *along* T.

Proof. Suppose first that $\tau\rho = \rho\tau$, where ρ is projection on S along T. For $s \in S$ we have

$$\rho\tau(s) = \tau\rho(s) = \tau(s)$$

and so ρ fixes $\tau(s)$, which implies that $\tau(s) \in S$. Hence S is invariant under τ. Also, for $t \in T$

$$\rho\tau(t) = \tau\rho(t) = 0$$

and so $\tau(t) \in \ker(\rho) = T$. Hence, T is invariant under τ.

Conversely, suppose that (S,T) reduces τ. The projection operator ρ fixes vectors in S and sends vectors in T to 0. Hence, for $s \in S$ and $t \in T$ we have

$$\rho\tau(s) = \tau(s) = \tau\rho(s)$$

and

$$\rho\tau(t) = 0 = \tau\rho(t)$$

which imply that $\rho\tau = \tau\rho$. □

Exercises

1. Let J be the $n \times n$ matrix all of whose entries are equal to 1. Find the minimal polynomial and characteristic polynomial of J and the eigenvalues.
2. A linear operator $\tau \in \mathcal{L}(V)$ is said to be **nonderogatory** if its minimal polynomial is equal to its characteristic polynomial. Prove that τ is nonderogatory if and only if V is a cyclic module.
3. Prove that the eigenvalues of a matrix do not form a complete set of invariants under similarity.
4. Show that $\tau \in \mathcal{L}(V)$ is invertible if and only if 0 is not an eigenvalue of τ.
5. Let A be an $n \times n$ matrix over a field F that contains all roots of the characteristic polynomial of A. Prove that $\det(A)$ is the product of the eigenvalues of A, counting multiplicity.
6. Show that if λ is an eigenvalue of τ then $p(\lambda)$ is an eigenvalue of $p(\tau)$, for any polynomial $p(x)$. Also, if $\lambda \neq 0$ then λ^{-1} is an eigenvalue for τ^{-1}.
7. An operator $\tau \in \mathcal{L}(V)$ is **nilpotent** if $\tau^n = 0$ for some positive $n \in \mathbb{N}$.
   a) Show that if τ is nilpotent then the spectrum of τ is $\{0\}$.
   b) Find a nonnilpotent operator τ with spectrum $\{0\}$.
8. Show that if $\tau, \sigma \in \mathcal{L}(V)$ then $\tau\sigma$ and $\sigma\tau$ have the same eigenvalues.
9. (Halmos)
   a) Find a linear operator τ that is not idempotent but for which $\tau^2(\iota - \tau) = 0$.
   b) Find a linear operator τ that is not idempotent but for which $\tau(\iota - \tau)^2 = 0$.
   c) Prove that if $\tau^2(\iota - \tau) = \tau(\iota - \tau)^2 = 0$ then τ is idempotent.
10. An **involution** is a linear operator θ for which $\theta^2 = \iota$. If τ is idempotent what can you say about $2\tau - \iota$? Construct a one-to-one correspondence between the set of idempotents on V and the set of involutions.
11. Let $A, B \in M_2(\mathbb{C})$ and suppose that $A^2 = B^3 = I, ABA = B^{-1}$ but $A \neq I$ and $B \neq I$. Show that if $C \in M_2(\mathbb{C})$ commutes with both A and B then $C = rI$ for some scalar $r \in \mathbb{C}$.
12. Suppose that J and K are matrices in Jordan canonical form. Prove that if J and K are similar then they are the same except for the order of the Jordan blocks. Hence, Jordan form is a canonical form for similarity (up to order of the blocks).
13. Fix $\epsilon > 0$. Show that any complex matrix is similar to a matrix that looks just like a Jordan matrix except that the entries that are equal to 1 are replaced by entries with value ϵ, where ϵ is any complex number. Thus, any complex matrix is similar to a matrix that is "almost" diagonal. *Hint*:

consider the fact that

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \epsilon & 0 \\ 0 & 0 & \epsilon^2 \end{bmatrix} \begin{bmatrix} \lambda & 0 & 0 \\ 1 & \lambda & 0 \\ 0 & 1 & \lambda \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \epsilon^{-1} & 0 \\ 0 & 0 & \epsilon^{-2} \end{bmatrix} = \begin{bmatrix} \lambda & 0 & 0 \\ \epsilon & \lambda & 0 \\ 0 & \epsilon & \lambda \end{bmatrix}$$

14. Show that the Jordan canonical form is not very robust in the sense that a small change in the entries of a matrix A may result in a large jump in the entries of the Jordan form J. *Hint*: consider the matrix

$$A_\epsilon = \begin{bmatrix} \epsilon & 0 \\ 1 & 0 \end{bmatrix}$$

What happens to the Jordan form of A_ϵ as $\epsilon \to 0$?

15. Give an example of a complex nonreal matrix all of whose eigenvalues are real. Show that any such matrix is similar to a real matrix. What about the type of the invertible matrices that are used to bring the matrix to Jordan form?
16. Let $J = [\tau]_{\mathcal{B}}$ be the Jordan form of a linear operator $\tau \in \mathcal{L}(V)$. For a given Jordan block of $J(\lambda, e)$ let U be the subspace of V spanned by the basis vectors of $\mathcal{B}$ associated with that block.
    a) Show that $\tau|_U$ has a single eigenvalue λ with geometric multiplicity 1. In other words, there is essentially only one eigenvector (up to scalar multiple) associated with each Jordan block. Hence, the geometric multiplicity of λ for τ is the number of Jordan blocks for λ. Show that the algebraic multiplicity is the sum of the dimensions of the Jordan blocks associated with λ.
    b) Show that the number of Jordan blocks in J is the maximum number of linearly independent eigenvectors of τ.
    c) What can you say about the Jordan blocks if the algebraic multiplicity of every eigenvalue is equal to its geometric multiplicity?
17. Assume that the base field F is algebraically closed. Then assuming that the eigenvalues of A are known, it is possible to determine the Jordan form J of a matrix A by looking at the rank of various matrix powers. A matrix B is **nilpotent** if $B^n = 0$ for some $n > 0$. The smallest such exponent is called the **index of nilpotence**.
    a) Let $J = J(\lambda, n)$ be a single Jordan block of size $n \times n$. Show that $J - \lambda I$ is nilpotent of index n. Thus, n is the smallest integer for which $\text{rk}(J - \lambda I)^n = 0$.

    Now let J be a matrix in Jordan form but possessing only one eigenvalue λ.

    b) Show that $J - \lambda I$ is nilpotent. Let m be its index of nilpotence. Show that m is the maximum size of the Jordan blocks of J and that $\text{rk}(J - \lambda I)^{m-1}$ is the number of Jordan blocks in J of maximum size.
    c) Show that $\text{rk}(J - \lambda I)^{m-2}$ is equal to 2 times the number of Jordan blocks of maximum size plus the number of Jordan blocks of size one less than the maximum.

d) Show that the sequence $\text{rk}(J - \lambda I)^k$ for $k = 1, \ldots, m$ uniquely determines the number and size of all of the Jordan blocks in J, that is, it uniquely determines J up to the order of the blocks.

e) Now let J be an arbitrary Jordan matrix. If λ is an eigenvalue for J show that the sequence $\text{rk}(J - \lambda I)^k$ for $k = 1, \ldots, m$ where m is the first integer for which $\text{rk}(J - \lambda I)^m = \text{rk}(J - \lambda I)^{m+1}$ uniquely determines J up to the order of the blocks.

f) Prove that for any matrix A with spectrum $\{\lambda_1, \ldots, \lambda_s\}$ the sequence $\text{rk}(A - \lambda_i I)^k$ for $i = 1, \ldots, s$ and $k = 1, \ldots, m$ where m is the first integer for which $\text{rk}(A - \lambda I)^m = \text{rk}(A - \lambda I)^{m+1}$ uniquely determines the Jordan matrix J for A up to the order of the blocks.

18. Let $A \in \mathcal{M}_n(F)$.

a) If all the roots of the characteristic polynomial of A lie in F prove that A is similar to its transpose A^t. Hint: Let B be the matrix

$$B = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & 1 & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$

that has 1's on the diagonal that moves up from left to right and 0's elsewhere. Let J be a Jordan block of the same size as B. Show that $BJB^{-1} = J^t$.

b) Let $A, B \in \mathcal{M}_n(F)$. Let K be a field containing F. Show that if A and B are similar over K, that is, if $B = PAP^{-1}$ where $P \in \mathcal{M}_n(K)$ then A and B are also similar over F, that is, there exists $Q \in \mathcal{M}_n(F)$ for which $B = QAQ^{-1}$. *Hint*: consider the equation $XA - BX = 0$ as a homogeneous system of linear equations with coefficients in F. Does it have a solution? Where?

c) Show that any matrix is similar to its transpose.

19. Prove Theorem 8.8 using the complexification of V.

The Trace of a Matrix

20. Let A be an $n \times n$ matrix over a field F. The **trace** of A, denoted by $\text{tr}(A)$, is the sum of the elements on the main diagonal of A. Verify the following statements:

a) $\text{tr}(r\text{A}) = r\,\text{tr}(A)$, for $r \in F$

b) $\text{tr}(A + B) = \text{tr}(A) + \text{tr}(B)$

c) $\text{tr}(AB) = \text{tr}(BA)$

d) Prove that $\text{tr}(ABC) = \text{tr}(CAB) = \text{tr}(BCA)$. Find an example to show that $\text{tr}(ABC)$ may not equal $\text{tr}(ACB)$.

e) The trace is an invariant under similarity

f) If F is algebraically closed then the trace of A is the sum of the eigenvalues of A.

Formulate a definition of the trace of a linear operator, show that it is well-defined and relate this concept to the eigenvalues of the operator.

21. Use the concept of the trace of a matrix, as defined in the previous exercise, to prove that there are no matrices $A, B \in \mathcal{M}_n(\mathbb{C})$ for which
$$AB - BA = I$$
22. Let $T: \mathcal{M}_n(F) \to F$ be a function with the following properties. For all matrices $A,\ B \in \mathcal{M}_n(F)$ and $r \in F$,
    1) $T(r\mathrm{A}) = rT(A)$
    2) $T(A + B) = T(A) + T(B)$
    3) $T(AB) = T(BA)$

    Show that there exists $s \in F$ for which $T(A) = s\,\mathrm{tr}(A)$, for all $A \in \mathcal{M}_n(F)$.

Simultaneous Diagonalizability

23. A pair of linear operators $\sigma, \tau \in \mathcal{L}(V)$ is **simultaneously diagonalizable** if there is an ordered basis $\mathcal{B}$ for V for which $[\tau]_{\mathcal{B}}$ and $[\sigma]_{\mathcal{B}}$ are both diagonal, that is, $\mathcal{B}$ is an ordered basis of eigenvectors for both τ and σ. Prove that two diagonalizable operators σ and τ are simultaneously diagonalizable if and only if they commute, that is, $\sigma\tau = \tau\sigma$. *Hint*: If $\sigma\tau = \tau\sigma$ then the eigenspaces of τ are invariant under σ.

Common Eigenvectors

It is often of interest to know whether a family
$$\mathcal{F} = \{\tau_i \in \mathcal{L}(V) \mid i \in \mathcal{I}\}$$
of linear operators on V has a **common eigenvector**, that is, a single vector $v \in V$ that is an eigenvector for every operator in $\mathcal{F}$ (the corresponding eigenvalues may be different for each operator, however).

A **commuting family** $\mathcal{F}$ of operators is a family in which each pair of operators commutes, that is, $\sigma, \tau \in \mathcal{F}$ implies $\sigma\tau = \tau\sigma$. We say that a subspace U of V is **$\mathcal{F}$-invariant** if it is τ-invariant for every $\tau \in \mathcal{F}$.

24. Let $\sigma, \tau \in \mathcal{L}(V)$. Prove that if σ and τ commute then every eigenspace of σ is τ-invariant. Thus, if $\mathcal{F}$ is a commuting family then every eigenspace of any member of $\mathcal{F}$ is $\mathcal{F}$-invariant.
25. Let $\mathcal{F}$ be a family of operators in $\mathcal{L}(V)$ with the property that each operator in $\mathcal{F}$ has a full set of eigenvalues in the base field F, that is, the characteristic polynomial splits over F. Prove that if $\mathcal{F}$ is a commuting family then $\mathcal{F}$ has a common eigenvector $v \in V$.
26. What do the real matrices
$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \text{ and } B = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$$
have to do with the issue of common eigenvectors?

Geršgorin Disks

It is generally impossible to determine precisely the eigenvalues of a given complex operator or matrix $A \in \mathcal{M}_n(\mathbb{C})$, for if $n \geq 5$ then the characteristic equation has degree 5 and cannot in general be solved. As a result, the approximation of eigenvalues is big business. Here we consider one aspect of this approximation problem, which also has some interesting theoretical consequences.

Let $A \in \mathcal{M}_n(\mathbb{C})$ and suppose that $Av = \lambda v$ where $v = (b_1, \ldots, b_n)^t$. Comparing kth rows gives

$$\sum_{i=1}^{n} A_{ki} b_i = \lambda b_k$$

which can also be written in the form

$$b_k(\lambda - A_{kk}) = \sum_{\substack{i=1 \\ i \neq k}}^{n} A_{ki} b_i$$

If k has the property that $|b_k| \geq |b_i|$ for all i, we have

$$|b_k||\lambda - A_{kk}| \leq \sum_{\substack{i=1 \\ i \neq k}}^{n} |A_{ki}||b_i| \leq |b_k| \sum_{\substack{i=1 \\ i \neq k}}^{n} |A_{ki}|$$

and thus

$$|\lambda - A_{kk}| \leq \sum_{\substack{i=1 \\ i \neq k}}^{n} |A_{ki}| \tag{8.7}$$

The right-hand side is the sum of the absolute values of all entries in the kth row of A *except* the diagonal entry A_{kk}. This sum $R_k(A)$ is the kth **deleted absolute row sum** of A. The inequality (8.7) says that, in the complex plane, the eigenvalue λ lies in the disk centered at the diagonal entry A_{kk} with radius equal to $R_k(A)$. This disk

$$\text{GR}_k(A) = \{z \in \mathbb{C} \mid |z - A_{kk}| \leq R_k(A)\}$$

is called the **Geršgorin row disk** for the kth row of A. The union of all of the Geršgorin row disks is called the **Geršgorin row region** for A.

Since there is no way to know in general which is the index k for which $|b_k| \geq |b_i|$, the best we can say in general is that the eigenvalues of A lie in the *union* of all Geršgorin row disks, that is, in the Geršgorin row region of A.

Similar definitions can be made for columns and since a matrix has the same eigenvalues as its transpose, we can say that the eigenvalues of A lie in the Geršgorin column region of A. The **Geršgorin region** $G(A)$ of a matrix $A \in M_n(F)$ is the intersection of the Geršgorin row region and the Geršgorin column region and we can say that all eigenvalues of A lie in the Geršgorin region of A. In symbols, $\sigma(A) \subseteq G(A)$.

27. Find and sketch the Geršgorin region and the eigenvalues for the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

28. A matrix $A \in M_n(\mathbb{C})$ is **diagonally dominant** if for each $k = 1, \ldots, n$

$$|A_{kk}| \geq R_k(A)$$

and it is **strictly diagonally dominant** if strict inequality holds. Prove that if A is strictly diagonally dominant then it is invertible.

29. Find a matrix $A \in M_n(\mathbb{C})$ that is diagonally dominant but not invertible.
30. Find a matrix $A \in M_n(\mathbb{C})$ that is invertible but not strictly diagonally dominant.

Chapter 9
Real and Complex Inner Product Spaces

We now turn to a discussion of real and complex vector spaces that have an additional function defined on them, called an *inner product*, as described in the upcoming definition. Thus, in this chapter, F will denote either the real or complex field. If r is a complex number then the complex conjugate of r is denoted by $\overline{r}$.

Definition *Let V be a vector space over $F = \mathbb{R}$ or $F = \mathbb{C}$. An* **inner product** *on V is a function $\langle , \rangle : V \times V \to F$ with the following properties:*

1) (**Positive definiteness**) *For all $v \in V$, the inner product $\langle v, v \rangle$ is real and*

$$\langle v, v \rangle \geq 0 \textit{ and } \langle v, v \rangle = 0 \Leftrightarrow v = 0$$

2) *For $F = \mathbb{C}$:* (**Conjugate symmetry**)

$$\langle u, v \rangle = \overline{\langle v, u \rangle}$$

For $F = \mathbb{R}$: (**Symmetry**)

$$\langle u, v \rangle = \langle v, u \rangle$$

3) (**Linearity in the first coordinate**) *For all $u, v \in V$ and $r, s \in F$*

$$\langle ru + sv, w \rangle = r\langle u, w \rangle + s\langle v, w \rangle$$

A real (or complex) vector space V, together with an inner product, is called a **real** *(or* **complex***)* **inner product space**. □

We will study bilinear forms (also called *inner products*) on vector spaces over fields other than $\mathbb{R}$ or $\mathbb{C}$ in Chapter 11. Note that property 1) implies that the quantity $\langle v, v \rangle$ is always real, even if V is a complex vector space.

Combining properties 2) and 3), we get, in the complex case

$$\langle w, ru + sv \rangle = \overline{\langle ru + sv, w \rangle} = \overline{r}\overline{\langle u, w \rangle} + \overline{s}\overline{\langle v, w \rangle} = \overline{r}\langle w, u \rangle + \overline{s}\langle w, v \rangle$$

This is referred to as **conjugate linearity** in the second coordinate. Thus, a complex inner product is linear in its first coordinate and conjugate linear in its second coordinate. This is often described by saying that the inner product is **sesquilinear**. (Sesqui means "one and a half times.") In the real case ($F=\mathbb{R}$), the inner product is linear in both coordinates—a property referred to as **bilinearity**.

Example 9.1

1) The vector space $\mathbb{R}^n$ is an inner product space under the **standard inner product**, or **dot product**, defined by
$$\langle (r_1,\ldots,r_n),(s_1,\ldots,s_n)\rangle = r_1s_1+\cdots+r_ns_n$$
The inner product space $\mathbb{R}^n$ is often called **n-dimensional Euclidean space**.
2) The vector space $\mathbb{C}^n$ is an inner product space under the **standard inner product** defined by
$$\langle (r_1,\ldots,r_n),(s_1,\ldots,s_n)\rangle = r_1\overline{s}_1+\cdots+r_n\overline{s}_n$$
This inner product space is often called **n-dimensional unitary space**.
3) The vector space $C[a,b]$ of all continuous complex-valued functions on the closed interval $[a,b]$ is a complex inner product space under the inner product
$$\langle f,g\rangle = \int_{\mathrm{a}}^{\mathrm{b}} f(x)\overline{g(x)}\,\mathrm{dx}$$ □

Example 9.2 One of the most important inner product spaces is the vector space ℓ^2 of all real (or complex) sequences (s_n) with the property that $\sum|s_n|^2<\infty$, under the inner product
$$\langle (s_n),(t_n)\rangle = \sum_{n=0}^{\infty} s_n\overline{t_n}$$

Of course, for this inner product to make sense, the sum on the right must converge. To see this, note that if $(s_n),(t_n)\in\ell^2$ then
$$0\le(|s_n|-|t_n|)^2=|s_n|^2-2|s_n||t_n|+|t_n|^2$$

and so
$$2|s_nt_n|\le|s_n|^2+|t_n|^2$$

which gives

$$2\left|\sum_{n=0}^{\infty} s_n \overline{t_n}\right| \le 2\sum_{n=0}^{\infty} |s_n t_n| \le \sum_{n=0}^{\infty} |s_n|^2 + \sum_{n=0}^{\infty} |t_n|^2 < \infty$$

We leave it to the reader to verify that ℓ^2 is an inner product space. □

The following simple result is quite useful and easy to prove.

Lemma 9.1 *If V is an inner product space and $\langle u, x\rangle = \langle v, x\rangle$ for all $x \in V$ then $u = v$.* □

Note that a vector subspace S of an inner product space V is also an inner product space under the restriction of the inner product of V to S.

Norm and Distance

If V is an inner product space, the **norm**, or **length** of $v \in V$ is defined by

$$\|v\| = \sqrt{\langle v, v\rangle} \tag{9.1}$$

A vector v is a **unit vector** if $\|v\| = 1$. Here are the basic properties of the norm.

Theorem 9.2

1) $\|v\| \ge 0$ *and* $\|v\| = 0$ *if and only if* $v = 0$.

2) $\|rv\| = |r|\|v\|$ *for all* $r \in F,\ v \in V$

3) (**The Cauchy-Schwarz inequality**) *For all* $u, v \in V$,

$$|\langle u, v\rangle| \le \|u\|\|v\|$$

with equality if and only if one of u and v is a scalar multiple of the other.

4) (**The triangle inequality**) *For all* $u, v \in V$

$$\|u + v\| \le \|u\| + \|v\|$$

with equality if and only if one of u and v is a scalar multiple of the other.

5) *For all* $u, v, x \in V$

$$\|u - v\| \le \|u - x\| + \|x - v\|$$

6) *For all* $u, v \in V$

$$|\|u\| - \|v\|| \le \|u - v\|$$

7) (**The parallelogram law**) *For all* $u, v \in V$

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

Proof. We prove only Cauchy-Schwarz and the triangle inequality. For Cauchy-Schwarz, if either u or v is zero the result follows, so assume that $u, v \neq 0$. Then, for any scalar $r \in F$,

$$\begin{aligned}0 &\le \|u-rv\|^2\\ &= \langle u-rv, u-rv\rangle\\ &= \langle u,u\rangle - \overline{r}\langle u,v\rangle - r[\langle v,u\rangle - \overline{r}\langle v,v\rangle]\end{aligned}$$

Choosing $\overline{r} = \langle v,u\rangle/\langle v,v\rangle$ makes the value in the square brackets equal to 0 and so

$$0 \le \langle u,u\rangle - \frac{\langle v,u\rangle\langle u,v\rangle}{\langle v,v\rangle} = \|u\|^2 - \frac{|\langle u,v\rangle|^2}{\|v\|^2}$$

which is equivalent to the Cauchy-Schwarz inequality. Furthermore, equality holds if and only if $\|u-rv\|^2 = 0$, that is, if and only if $u - rv = 0$, which is equivalent to u and v being scalar multiples of one another.

To prove the triangle inequality, the Cauchy-Schwarz inequality gives

$$\begin{aligned}\|u+v\|^2 &= \langle u+v, u+v\rangle\\ &= \langle u,u\rangle + \langle u,v\rangle + \langle v,u\rangle + \langle v,v\rangle\\ &\le \|u\|^2 + 2\|u\|\|v\| + \|v\|^2\\ &= (\|u\| + \|v\|)^2\end{aligned}$$

from which the triangle inequality follows. The proof of the statement concerning equality is left to the reader. □

Any vector space V, together with a function $\|\cdot\|: V \to \mathbb{R}$ that satisfies properties 1), 2) and 4) of Theorem 9.2, is called a **normed linear space**. (And the function $\|\cdot\|$ is called a **norm**.) Thus, any inner product space is a normed linear space, under the norm given by (9.1).

It is interesting to observe that the inner product on V can be recovered from the norm.

Theorem 9.3 *(***The polarization identities***)*

1) *If V is a real inner product space, then*

$$\langle u,v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2)$$

2) *If V is a complex inner product space, then*

$$\langle u,v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2) + \frac{1}{4}i(\|u+iv\|^2 - \|u-iv\|^2)$$

The formulas in Theorem 9.3 are known as the **polarization identities**.

The norm can be used to define the distance between any two vectors in an inner product space.

Definition Let V be an inner product space. The **distance** $d(u,v)$ between any two vectors u and v in V is

$$d(u,v) = \|u-v\| \tag{9.2}$$ □

Here are the basic properties of distance.

Theorem 9.4
1) $d(u,v) \geq 0$ *and* $d(u,v) = 0$ *if and only if* $u = v$
2) (**Symmetry**)

$$d(u,v) = d(v,u)$$

3) (**The triangle inequality**)

$$d(u,v) \leq d(u,w) + d(w,v)$$ □

Any nonempty set V, together with a function $d\colon V \times V \to \mathbb{R}$ that satisfies the properties of Theorem 9.4, is called a **metric space** and the function d is called a **metric** on V. Thus, any inner product space is a metric space under the metric (9.2).

Before continuing, we should make a few remarks about our goals in this and the next chapter. The presence of an inner product (and hence a metric) raises a host of topological issues related to the notion of convergence. We say that a sequence (v_n) of vectors in an inner product space **converges** to $v \in V$ if

$$\lim_{n\to\infty} d(v_n, v) = 0$$

that is, if

$$\lim_{n\to\infty} \|v_n - v\| = 0$$

Some of the more important concepts related to convergence are closedness and closures, completeness and the continuity of linear operators and linear functionals.

In the finite-dimensional case, the situation is very straightforward: all subspaces are closed, all inner product spaces are complete and all linear operators and functionals are continuous. However, in the infinite-dimensional case, things are not as simple.

Our goals in this chapter and the next are to describe some of the basic properties of inner product spaces—both finite and infinite-dimensional—and then discuss certain special types of operators (normal, unitary and self-adjoint) in the finite-dimensional case only. To achieve the latter goal as rapidly as possible, we will postpone a discussion of topological properties until Chapter

13. This means that we must state some results only for the finite-dimensional case in this chapter, deferring the infinite-dimensional case to Chapter 13.

Isometries

An isomorphism of vector spaces preserves the vector space operations. The corresponding concept for inner product spaces is the following.

Definition *Let V and W be inner product spaces and let $\tau \in \mathcal{L}(V, W)$.*

1) *τ is an* **isometry** *if it preserves the inner product, that is, if*

$$\langle \tau(u), \tau(v) \rangle = \langle u, v \rangle$$

for all $u, v \in V$.

2) *A bijective isometry is called an* **isometric isomorphism**. *When $\tau: V \to W$ is a bijective isometry, we say that V and W are* **isometrically isomorphic**. □

It is not hard to show that an isometry is injective and so it is an isometric isomorphism provided it is also surjective. Moreover, if

$$\dim(V) = \dim(W) < \infty$$

injectivity implies surjectivity and so the concepts of isometry and isometric isomorphism are equivalent. On the other hand, the following example shows that this is not the case for infinite-dimensional inner product spaces.

Example 9.3 Let $\tau: \ell^2 \to \ell^2$ be defined by

$$\tau(x_1, x_2, x_3 \ldots) = (0, x_1, x_2, \ldots)$$

(This is the *right shift operator*.) Then τ is an isometry, but it is clearly not surjective. □

Theorem 9.5 *A linear transformation $\tau \in \mathcal{L}(V, W)$ is an isometry if and only if it preserves the norm, that is, if and only if*

$$\|\tau(v)\| = \|v\|$$

for all $v \in V$.

Proof. Clearly, an isometry preserves the norm. The converse follows from the polarization identities. In the real case, we have

$$\begin{aligned}\langle\tau(u),\tau(v)\rangle &= \frac{1}{4}(\|\tau(u)+\tau(v)\|^2-\|\tau(u)-\tau(v)\|^2)\\ &= \frac{1}{4}(\|\tau(u+v)\|^2-\|\tau(u-v)\|^2)\\ &= \frac{1}{4}(\|u+v\|^2-\|u-v\|^2)\\ &= \langle u,v\rangle\end{aligned}$$

and so τ is an isometry. The complex case is similar. □

The next result points out one of the main differences between real and complex inner product spaces.

Theorem 9.6 *Let V be an inner product space and let $\tau\in\mathcal{L}(V)$.*
1) If $\langle\tau(v),w\rangle=0$ for all $v,\ w\in V$ then $\tau=0$.
2) If V is a complex inner product space and $\langle\tau(v),v\rangle=0$ for all $v\in V$ then $\tau=0$.
3) Part 2) does not hold in general for real inner product spaces.

Proof. Part 1) follows directly from Lemma 9.1. As for part 2), let $v=rx+y$, for $x,y\in V$ and $r\in F$. Then

$$\begin{aligned}0 &= \langle\tau(rx+y),rx+y\rangle\\ &= |r|^2\langle\tau(x),x\rangle+\langle\tau(y),y\rangle+r\langle\tau(x),y\rangle+\overline{r}\langle\tau(y),x\rangle\\ &= r\langle\tau(x),y\rangle+\overline{r}\langle\tau(y),x\rangle\end{aligned}$$

Setting $r=1$ gives

$$\langle\tau(x),y\rangle+\langle\tau(y),x\rangle=0$$

and setting $r=i$ gives

$$\langle\tau(x),y\rangle-\langle\tau(y),x\rangle=0$$

These two equations imply that $\langle\tau(x),y\rangle=0$ for all $x,y\in V$ and so $\tau=0$ by part 1). As for part 3), rotation by $\pi/2$ in the real plane $\mathbb{R}^2$ has the property that $\langle\tau(v),v\rangle=0$ for all v, yet τ is not zero. □

Orthogonality

The presence of an inner product allows us to define the concept of orthogonality.

Definition *Let V be an inner product space.*
1) Two vectors $u,v\in V$ are **orthogonal***, written $u\perp v$, if $\langle u,v\rangle=0$.*
2) Two subsets $X,Y\subseteq V$ are **orthogonal***, written $X\perp Y$, if $x\perp y$ for all $x\in X$ and $y\in Y$.*

3) *The* **orthogonal complement** *of a subset $X \subseteq V$ is the set*

$$X^{\perp} = \{v \in V \mid \{v\} \perp X\} \qquad \square$$

The following result is easily proved.

Theorem 9.7 *Let V be an inner product space.*
1) *The orthogonal complement $X^{\perp}$ of any subset $X \subseteq V$ is a subspace of V.*
2) *For any subspace S of V, $S \cap S^{\perp} = \{0\}$.* □

Orthogonal and Orthonormal Sets

Definition *A nonempty set $\mathcal{O} = \{u_i \mid i \in K\}$ of vectors in an inner product space is said to be an* **orthogonal set** *if $u_i \perp u_j$ for all $i \neq j \in K$. If, in addition, each vector u_i is a unit vector, the set $\mathcal{O}$ is an* **orthonormal set**. *Thus, a set is orthonormal if*

$$\langle u_i, u_j \rangle = \delta_{i,j}$$

for all $i, j \in K$, where $\delta_{i,j}$ is the Kronecker delta function. □

Of course, given any nonzero vector $v \in V$, we may obtain a unit vector u by multiplying v by the reciprocal of its norm

$$u = \frac{1}{\|v\|} v$$

Thus, it is a simple matter to construct an orthonormal set from an orthogonal set of nonzero vectors.

Note that if $u \perp v$ then

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2$$

and the converse holds if $F = \mathbb{R}$.

Theorem 9.8 *Any orthogonal set of nonzero vectors in V is linearly independent.*
Proof. Let $\mathcal{O} = \{u_i \mid i \in K\}$ be an orthogonal set of nonzero vectors and suppose that

$$r_1 u_1 + \cdots + r_n u_n = 0$$

Then, for any $k = 1, \ldots, n$,

$$0 = \langle r_1 u_1 + \cdots + r_n u_n, u_k \rangle = r_k \langle u_k, u_k \rangle$$

and so $r_k = 0$, for all k. Hence, $\mathcal{O}$ is linearly independent. □

Definition *A* **maximal orthonormal set** *in an inner product space V is called a* **Hilbert basis** *for V.* □

Zorn's lemma can be used to show that any nontrivial inner product space has a Hilbert basis. We leave the details to the reader.

Extreme care must be taken here not to confuse the concepts of a basis for a vector space and a Hilbert basis for an inner product space. To avoid confusion, a vector space basis, that is, a maximal linearly independent set of vectors, is referred to as a **Hamel basis**. An orthonormal *Hamel* basis will be called an **orthonormal basis**, to distinguish it from a Hilbert basis.

The following example shows that, in general, the two concepts of basis are not the same.

Example 9.4 Let $V = \ell^2$ and let M be the set of all vectors of the form

$$e_i = (0, \dots, 0, 1, 0, \dots)$$

where e_i has a 1 in the ith coordinate and 0's elsewhere. Clearly, M is an orthonormal set. Moreover, it is maximal. For if $v = (x_n) \in \ell^2$ has the property that $v \perp M$ then

$$x_i = \langle v, e_i \rangle = 0$$

for all i and so $v = 0$. Hence, no nonzero vector $v \notin M$ is orthogonal to M. This shows that M is a Hilbert basis for the inner product space ℓ^2.

On the other hand, the vector space span of M is the subspace S of all sequences in ℓ^2 that have finite support, that is, have only a finite number of nonzero terms and since $\text{span}(M) = S \neq \ell^2$, we see that M is not a Hamel basis for the vector space ℓ^2. □

We will show in Chapter 13 that all Hilbert bases for an inner product space have the same cardinality and so we can define the **Hilbert dimension** of an inner product space to be that cardinality. Once again, to avoid confusion, the cardinality of any Hamel basis for V is referred to as the **Hamel dimension** of V. The Hamel dimension is, in general, not equal to the Hilbert dimension. However, as we will now show, they are equal when either dimension is finite.

Theorem 9.9 *Let V be an inner product space.*

1) (**Gram–Schmidt orthogonalization**) *If $\mathcal{B} = (v_1, v_2, \dots)$ is a linearly independent sequence in V, then there is an orthogonal sequence $\mathcal{O} = (u_1, u_2, \dots)$ in V for which*

$$\text{span}(u_1, \dots, u_n) = \text{span}(v_1, \dots, v_n)$$

for all $n > 0$.

2) *If* $\dim(V) = n$ *is finite then* V *has a Hilbert basis of size* n *and all Hilbert bases for* V *have size* n *.*
3) *If* V *has a finite Hilbert basis of size* n*, then* $\dim(V) = n$*.*

Proof. To prove part 1), first let $u_1 = v_1$. Once the orthogonal set $\{u_1, \dots, u_k\}$ of nonzero vectors has been chosen so that

$$\text{span}(u_1, \dots, u_k) = \text{span}(v_1, \dots, v_k)$$

the next vector u_{k+1} is chosen by setting

$$u_k = v_k + r_1 u_1 + \cdots + r_{k-1} u_{k-1}$$

and requiring that u_{k+1} be orthogonal to each u_i for $i < k$, that is,

$$0 = \langle u_k, u_i \rangle = \langle v_k + r_1 u_1 + \cdots + r_{k-1} u_{k-1}, u_i \rangle = \langle v_k, u_i \rangle + r_i \langle u_i, u_i \rangle$$

or, finally,

$$r_i = -\frac{\langle v_k, u_i \rangle}{\langle u_i, u_i \rangle}$$

for all $i = 1, \dots, k$.

For part 2), applying the Gram–Schmidt orthogonalization process to a Hamel basis gives a Hilbert basis of the same size n. Moreover, if V has a Hilbert basis of size greater than n, it must also have a Hamel basis of size greater than n, which is not possible. Finally, if V has a Hilbert basis $\mathcal{B}$ of size less than n then $\mathcal{B}$ can be extended to a proper superset $\mathcal{C}$ that is also linearly independent. The Gram–Schmidt process applied to $\mathcal{C}$ gives a proper superset of $\mathcal{B}$ that is orthonormal, which is not possible. Hence, all Hilbert bases have size n.

For part 3), suppose that $\dim(V) > n$. Since a Hilbert basis $\mathcal{H}$ of size n is linearly independent, we can adjoin a new vector to $\mathcal{H}$ to get a linearly independent set of size $n + 1$. Applying the Gram–Schmidt process to this set gives an orthonormal set that properly contains $\mathcal{H}$, which is not possible. □

For reference, let us state the Gram–Schmidt orthogonalization process separately and give an example of its use.

Theorem 9.10 *(***The Gram–Schmidt orthogonalization process***) If* $\mathcal{B} = (v_1, v_2, \dots)$ *is a sequence of linearly independent vectors in an inner product space* V*, then the sequence* $\mathcal{O} = (u_1, u_2, \dots)$ *defined by*

$$u_k = v_k - \sum_{i=1}^{k-1} \frac{\langle v_k, u_i \rangle}{\langle u_i, u_i \rangle} u_i$$

is an orthogonal sequence in V with the property that

$$\operatorname{span}(u_1, \ldots, u_k) = \operatorname{span}(v_1, \ldots, v_k)$$

for all $k > 0$. □

Of course, from the orthogonal sequence (u_i), we get the orthonormal sequence (w_i), where $w_i = u_i / \|u_i\|$.

Example 9.5 Consider the inner product space $\mathbb{R}[x]$ of real polynomials, with inner product defined by

$$\langle p(x), q(x) \rangle = \int_{-1}^{1} p(x)q(x)dx$$

Applying the Gram–Schmidt process to the sequence $\mathcal{B} = (1, x, x^2, x^3, \ldots)$ gives

$$\begin{aligned} u_1(x) &= 1 \\ u_2(x) &= x - \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} dx} \cdot 1 = x \\ u_3(x) &= x^2 - \frac{\int_{-1}^{1} x^2\,dx}{\int_{-1}^{1} dx} \cdot 1 - \frac{\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} x\,dx} \cdot x = x^2 - \frac{1}{3} \\ u_4(x) &= x^3 - \frac{\int_{-1}^{1} x^3\,dx}{\int_{-1}^{1} dx} \cdot 1 - \frac{\int_{-1}^{1} x^4\,dx}{\int_{-1}^{1} x\,dx} \cdot x - \frac{\int_{-1}^{1} x^3(x^2 - \frac{1}{3})dx}{\int_{-1}^{1} (x^2 - \frac{1}{3})^2 dx} \cdot \left(x^2 - \frac{1}{3}\right) \\ &= x^3 - \frac{3}{5}x \end{aligned}$$

and so on. The polynomials in this sequence are (at least up to multiplicative constants) the **Legendre polynomials**. □

Orthonormal bases have a great advantage over arbitrary bases. From a computational point of view, if $\mathcal{B} = \{v_1, \ldots, v_n\}$ is a basis for V then each $v \in V$ has the form

$$v = r_1 v_1 + \cdots + r_n v_n$$

In general, however, determining the coordinates r_i requires solving a system of linear equations of size $n \times n$.

On the other hand, if $\mathcal{O} = \{u_1, \ldots, u_n\}$ is an orthonormal basis for $V$and

$$v = r_1 u_1 + \cdots + r_n u_n$$

then the coefficients are quite easily computed:

$$\langle v, u_i \rangle = \langle r_1 u_1 + \cdots + r_n u_n, u_i \rangle = r_i \langle u_i, u_i \rangle = r_i$$

Even if $\mathcal{O} = \{u_1, \ldots, u_n\}$ is not a basis (but just an orthonormal set), we can still consider the expansion

$$\widehat{v} = \langle v, u_1 \rangle u_1 + \cdots + \langle v, u_n \rangle u_n$$

Proof of the following characterization of orthonormal (Hamel) bases is left to the reader.

Theorem 9.11 *Let $\mathcal{O} = \{u_1, \ldots, u_k\}$ be an orthonormal set of vectors in a finite-dimensional inner product space V. For any $v \in V$, the* **Fourier expansion** *of v* **with respect to** $\mathcal{O}$ *is*

$$\widehat{v} = \langle v, u_1 \rangle u_1 + \cdots + \langle v, u_k \rangle u_k$$

In this case, **Bessel's inequality** *holds for all $v \in V$, that is*

$$\|\widehat{v}\| \le \|v\|$$

Moreover, the following are equivalent:

1) *The set $\mathcal{O}$ is an orthonormal basis for V.*
2) *Every vector is equal to its Fourier expansion, that is, for all $v \in V$*

$$\widehat{v} = v$$

3) **Bessel's identity** *holds for all $v \in V$, that is*

$$\|\widehat{v}\| = \|v\|$$

4) **Parseval's identity** *holds for all $v, w \in V$, that is*

$$\langle v, w \rangle = \langle v, u_1 \rangle \overline{\langle w, u_1 \rangle} + \cdots + \langle v, u_k \rangle \overline{\langle w, u_k \rangle}$$ □

The Projection Theorem and Best Approximations

We have seen that if S is a subspace of an inner product space V then $S \cap S^\perp = \{0\}$. This raises the question of whether or not the orthogonal complement $S^\perp$ is a vector space complement of S, that is, whether or not $V = S \oplus S^\perp$.

If S is a finite-dimensional subspace of V, the answer is yes, but for infinite-dimensional subspaces, S must have the topological property of being *complete*. Hence, in accordance with our goals in this chapter, we will postpone a discussion of the general case to Chapter 13, contenting ourselves here with an example to show that, in general, $V \ne S \oplus S^\perp$.

Example 9.6 As in Example 9.4, let $V = \ell^2$ and let S be the subspace of all sequences of finite support, that is, S is spanned by the vectors

$$e_i = (0, \ldots, 0, 1, 0, \ldots)$$

where e_i has a 1 in the ith coordinate and 0s elsewhere. If $x = (x_n) \in S^\perp$ then

$x_i = \langle x, e_i \rangle = 0$ for all i and so $x = 0$. Therefore, $S^\perp = \{0\}$. However,

$$S \oplus S^\perp = S \neq \ell^2 \qquad \square$$

As the next theorem shows, in the finite-dimensional case, orthogonal complements are also vector space complements. This theorem is often called the *projection theorem*, for reasons that will become apparent when we discuss projection operators. (We will discuss the projection theorem in the infinite-dimensional case in Chapter 13.)

Theorem 9.12 *(***The projection theorem***) If S is a finite-dimensional subspace of an inner product space V (which need not be finite-dimensional) then*

$$V = S \oplus S^\perp$$

Proof. Let $\mathcal{O} = \{u_1, \ldots, u_k\}$ be an orthonormal basis for S. For each $v \in V$, consider the Fourier expansion

$$\widehat{v} = \langle v, u_1 \rangle u_1 + \cdots + \langle v, u_k \rangle u_k$$

with respect to $\mathcal{O}$. We may write

$$v = \widehat{v} + (v - \widehat{v})$$

where $\widehat{v} \in S$. Moreover, $v - \widehat{v} \in S^\perp$, since

$$\langle v - \widehat{v}, u_i \rangle = \langle v, u_i \rangle - \langle \widehat{v}, u_i \rangle = 0$$

Hence $V = S + S^\perp$. We have already observed that $S \cap S^\perp = \{0\}$ and so $V = S \oplus S^\perp$. $\square$

According to the proof of the projection theorem, the component of v that lies in S is just the Fourier expansion of v with respect to any orthonormal basis $\mathcal{O}$ for S.

Best Approximations

The projection theorem implies that if $v = \widehat{v} + s^\perp$ where $\widehat{v} \in S$ and $s^\perp \in S^\perp$ then $\widehat{v}$ is the element of S that is *closest* to v, that is, $\widehat{v}$ is the **best approximation** to v from within S. For if $t \in S$ then since $v - \widehat{v} \in S^\perp$ we have $(v - \widehat{v}) \perp (\widehat{v} - t)$ and so

$$\|v - t\|^2 = \|v - \widehat{v} + \widehat{v} - t\|^2 = \|v - \widehat{v}\|^2 + \|\widehat{v} - t\|^2$$

It follows that $\|v - t\|$ is smallest when $t = \widehat{v}$. Also, note that $\widehat{v}$ is the *unique* vector in S for which $v - \widehat{v} \perp S$. Thus, we can say that the best approximation to v from within S is the unique vector $s \in S$ for which $(v - s) \perp S$ and that this vector is the Fourier expansion $\widehat{v}$ of v.

Orthogonal Direct Sums

Definition *Let V be an inner product space and let $S_1,\dots,S_n$ be subspaces of V. Then V is the* **orthogonal direct sum** *of $S_1,\dots,S_n$, written*

$$S = S_1 \odot \cdots \odot S_n$$

if
1) $V = S_1 \oplus \cdots \oplus S_n$
2) $S_i \perp S_j$ for $i \neq j$
In general, to say that the orthogonal direct sum $S_1 \odot \cdots \odot S_n$ of subspaces **exists** *is to say that the direct sum $S_1 \oplus \cdots \oplus S_n$ exists and that 2) holds.* □

Theorem 9.12 states that $V = S \odot S^{\perp}$, for any finite-dimensional subspace S of a vector space V. The following simple result is very useful.

Theorem 9.13 *Let V be an inner product space. The following are equivalent.*
1) $V = S \odot T$
2) $V = S \oplus T$ and $T = S^{\perp}$
3) $V = S \oplus T$ and $T \subseteq S^{\perp}$
Proof. Suppose 1) holds. Then $V = S \oplus T$ and $S \perp T$, which implies that $T \subseteq S^{\perp}$. But if $w \in S^{\perp}$ then $w = s + t$ for $s \in S,\ t \in T$ and so

$$0 = \langle s, w\rangle = \langle s, s\rangle + \langle s, t\rangle = \langle s, s\rangle$$

Hence $s = 0$ and $w \in T$, which implies that $S^{\perp} \subseteq T$. Hence, $S^{\perp} = T$, which gives 2). Of course, 2) implies 3). Finally, if 3) holds then $T \subseteq S^{\perp}$, which implies that $S \perp T$ and so 1) holds. □

Theorem 9.14 *Let V be an inner product space.*
1) If $\dim(V) < \infty$ and S is a subspace of V then

$$\dim(S^{\perp}) = \dim(V) - \dim(S)$$

2) If S is a finite-dimensional subspace of V then

$$S^{\perp\perp} = S$$

3) If X is a subset *of V and $\dim(\text{span}(X)) < \infty$ then*

$$X^{\perp\perp} = \text{span}(X)$$

Proof. Since $V = S \oplus S^{\perp}$, we have $\dim(V) = \dim(S) + \dim(S^{\perp})$, which proves part 1). As for part 2), it is clear that $S \subseteq S^{\perp\perp}$. On the other hand, if $v \in S^{\perp\perp}$ then by the projection theorem

$$v = s + s'$$

where $s \in S$ and $s' \in S^{\perp}$. But $v \in S^{\perp\perp}$ implies that $0 = \langle v, s' \rangle = \langle s', s' \rangle$ and so $s' = 0$, showing that $v \in S$. Therefore, $S^{\perp\perp} \subseteq S$ and $S^{\perp\perp} = S$. We leave the proof of part 3) as an exercise. □

The Riesz Representation Theorem

If x is a vector in an inner product space V then the function $\phi_x \colon V \to F$ defined by

$$\phi_x(v) = \langle v, x \rangle$$

is easily seen to be a linear functional on V. The following theorem shows that all linear functionals on a finite-dimensional inner product space V have this form. (We will see in Chapter 13 that, in the infinite-dimensional case, all *continuous* linear functionals on V have this form.)

Theorem 9.15 (**The Riesz representation theorem**) Let V be a finite-dimensional inner product space and let $f \in V^*$ be a linear functional on V. Then there exists a unique vector $x \in V$ for which

$$f(v) = \langle v, x \rangle \tag{9.3}$$

for all $v \in V$. Let us call x the **Riesz vector** for f and denote it by R_f. (This term and notation are not standard.)

Proof. If f is the zero functional, we may take $x = 0$, so let us assume that $f \neq 0$. Then $K = \ker(f)$ has codimension 1 and so

$$V = \langle w \rangle \odot K$$

for $w \in K^{\perp}$. If $x = \alpha w$ for some $\alpha \in F$, then (9.3) holds if and only if

$$f(v) = \langle v, \alpha w \rangle$$

and since any $v \in V$ has the form $v = \beta w + k$ for $\beta \in F$ and $k \in K$, this is equivalent to

$$f(\beta w) = \langle \beta w, \alpha w \rangle$$

or

$$f(w) = \overline{\alpha} \langle w, w \rangle = \overline{\alpha} \|w\|^2$$

Hence, we may take $\alpha = \overline{f(w)}/\|w\|^2$ and

$$x = \frac{\overline{f(w)}}{\|w\|^2} w$$

Proof of uniqueness is left as an exercise. □

If $V = \mathbb{R}^n$, then it is easy to see that $R_f = (f(e_1), \ldots, f(e_n))$ where $(e_1, \ldots, e_n)$ is the standard basis for $\mathbb{R}^n$.

Using the Riesz representation theorem, we can define a map $\phi: V^* \to V$ by setting $\phi(f) = R_f$, where R_f is the Riesz vector for f. Since

$$\begin{aligned}\langle v, \phi(rf+sg)\rangle &= (rf+sg)(v)\\ &= rf(v)+sg(v)\\ &= \langle v, \overline{r}\phi(f)\rangle + \langle v, \overline{s}\phi(g)\rangle\\ &= \langle v, \overline{r}\phi(f)+\overline{s}\phi(g)\rangle\end{aligned}$$

for all $v \in V$, we have

$$\phi(rf+sg) = \overline{r}\phi(f)+\overline{s}\phi(g)$$

and so ϕ is **conjugate linear**. Since ϕ is bijective, the map $\phi: V^* \to V$ is a "conjugate isomorphism."

Exercises

1. Verify the statement concerning equality in the triangle inequality.
2. Prove the parallelogram law.
3. Prove the **Appolonius identity**

$$\|w-u\|^2 + \|w-v\|^2 = \frac{1}{2}\|u-v\|^2 + 2\left\|w - \frac{1}{2}(u+v)\right\|^2$$

4. Let V be an inner product space with basis $\mathcal{B}$. Show that the inner product is uniquely defined by the values $\langle u, v\rangle$, for all $u, v \in \mathcal{B}$.
5. Prove that two vectors u and v in a real inner product space V are orthogonal if and only if

$$\|u+v\|^2 = \|u\|^2 + \|v\|^2$$

6. Show that an isometry is injective.
7. Use Zorn's lemma to show that any nontrivial inner product space has a Hilbert basis.
8. Prove Bessel's inequality.
9. Prove that an orthonormal set $\mathcal{O}$ is a basis for V if and only if $\widehat{v} = v$, for all $v \in V$.
10. Prove that an orthonormal set $\mathcal{O}$ is a basis for V if and only if Bessel's identity holds for all $v \in V$, that is, if and only if

$$\|\widehat{v}\| = \|v\|$$

for all $v \in V$.

11. Prove that an orthonormal set $\mathcal{O}$ is a basis for V if and only if Parseval's identity holds for all $v, w \in V$, that is, if and only if

$$\langle v, w\rangle = \langle v, u_1\rangle\langle w, u_1\rangle + \cdots + \langle v, u_k\rangle\langle w, u_k\rangle$$

for all $v, w \in V$.

12. Let $u = (r_1, \ldots, r_n)$ and $v = (s_1, \ldots, s_n)$ be in $\mathbb{R}^n$. The Cauchy-Schwarz inequality states that

$$|r_1 s_1 + \cdots + r_n s_n|^2 \leq (r_1^2 + \cdots + r_n^2)(s_1^2 + \cdots + s_n^2)$$

Prove that we can do better:

$$(|r_1 s_1| + \cdots + |r_n s_n|)^2 \leq (r_1^2 + \cdots + r_n^2)(s_1^2 + \cdots + s_n^2)$$

13. Let V be a finite-dimensional inner product space. Prove that for any subset X of V, we have $X^{\perp\perp} = \text{span}(X)$.
14. Let $\mathcal{P}_3$ be the inner product of all polynomials of degree at most 3, under the inner product

$$\langle p(x), q(x) \rangle = \int_{-\infty}^{\infty} p(x)q(x)e^{-x^2} dx$$

Apply the Gram–Schmidt process to the basis $\{1, x, x^2, x^3\}$, thereby computing the first four **Hermite polynomials** (at least up to a multiplicative constant).
15. Verify uniqueness in the Riesz representation theorem.
16. Let V be a complex inner product space and let S be a subspace of V. Suppose that $v \in V$ is a vector for which $\langle v, s \rangle + \langle s, v \rangle \leq \langle s, s \rangle$ for all $s \in S$. Prove that $v \in S^{\perp}$.
17. If V and W are inner product spaces, consider the function on $V \boxplus W$ defined by

$$\langle (v_1, w_1), (v_2, w_2) \rangle = \langle v_1, v_2 \rangle + \langle w_1, w_2 \rangle$$

Is this an inner product on $V \boxplus W$?
18. A **normed vector space** over $\mathbb{R}$ or $\mathbb{C}$ is a vector space (over $\mathbb{R}$ or $\mathbb{C}$) together with a function $\|\,\|\colon V \to \mathbb{R}$ for which for all $u, v \in V$ and scalars r we have
a) $\|rv\| = |r|\|v\|$
b) $\|u + v\| \leq \|u\| + \|v\|$
c) $\|v\| = 0$ if and only if $v = 0$
If V is a real normed space (over $\mathbb{R}$) and if the norm satisfies the parallelogram law

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$$

prove that the polarization identity

$$\langle u, v \rangle = \frac{1}{4}(\|u + v\|^2 - \|u - v\|^2)$$

defines an inner product on V.
19. Let S be a subspace of an inner product space V. Prove that each coset in V/S contains *exactly one* vector that is orthogonal to S.

Extensions of Linear Functionals

20. Let f be a linear functional on a subspace S of a finite-dimensional inner product space V. Let $f(v) = \langle v, R_f \rangle$. Suppose that $g \in V^*$ is an extension of f, that is, $g|_S = f$. What is the relationship between the Riesz vectors R_f and R_g?
21. Let f be a linear functional on a subspace S of a finite-dimensional inner product space V and let $K = \ker(f)$. Show that if $g \in V^*$ is an extension of f then $R_g \in K^\perp \setminus \mathcal{S}^\perp$. Moreover, for each vector $u \in K^\perp \setminus S^\perp$ there is exactly one scalar λ for which the linear functional $g(X) = \langle X, \lambda u \rangle$ is an extension of f.

Positive Linear Functionals on $\mathbb{R}^n$

A vector $v = (a_1, \ldots, a_n)$ in $\mathbb{R}^n$ is **nonnegative** (also called **positive**), written $v \geq 0$, if $a_i \geq 0$ for all i. The vector v is **strictly positive**, written $v > 0$, if v is nonnegative but not 0. The set $\mathbb{R}^n_+$ of all strictly positive vectors in $\mathbb{R}^n$ is called the **nonnegative orthant** in $\mathbb{R}^n$. The vector v is **strongly positive**, written $v \gg 0$, if $a_i > 0$ for all i. The set $\mathbb{R}^n_{++}$, of all strongly positive vectors in $\mathbb{R}^n$ is the **strongly positive orthant** in $\mathbb{R}^n$.

Let $f\colon S \to \mathbb{R}$ be a linear functional on a subspace S of $\mathbb{R}^n$. Then f is **nonnegative** (also called **positive**), written $f \geq 0$, if

$$v > 0 \Rightarrow f(v) \geq 0$$

for all $v \in S$ and f is **strictly positive**, written $f > 0$, if

$$v > 0 \Rightarrow f(v) > 0$$

for all $v \in S$.

22. Prove that a linear functional f on $\mathbb{R}^n$ is positive if and only if $R_f > 0$ and strictly positive if and only if $R_f \gg 0$. If S is a subspace of $\mathbb{R}^n$ is it true that a linear functional f on S is nonnegative if and only if $R_f > 0$?
23. Let $f\colon S \to \mathbb{R}$ be a strictly positive linear functional on a subspace S of $\mathbb{R}^n$. Prove that f has a strictly positive extension to $\mathbb{R}^n$. Use the fact that if $U \cap \mathbb{R}^m_+ = \{0\}$, where

$$\mathbb{R}^n_+ = \{(a_1, \ldots, a_n) \mid a_i \geq 0 \text{ all } i\}$$

and U is a subspace of $\mathbb{R}^n$ then $U^\perp$ contains a strongly positive vector.
24. If V is a real inner product space, then we can define an inner product on its complexification $V^{\mathbb{C}}$ as follows (this is the same formula as for the ordinary inner product on a complex vector space)

$$\langle u + vi, x + yi \rangle = \langle u, x \rangle + \langle v, y \rangle + (\langle v, x \rangle - \langle u, y \rangle)i$$

Show that

$$\|(u + vi)\|^2 = \|u\|^2 + \|v\|^2$$

where the norm on the left is induced by the inner product on $V^{\mathbb{C}}$ and the norm on the right is induced by the inner product on V.

Chapter 10
Structure Theory for Normal Operators

Throughout this chapter, all vector spaces are assumed to be finite-dimensional unless otherwise noted.

The Adjoint of a Linear Operator

The purpose of this chapter is to study the structure of certain special types of linear operators on finite-dimensional inner product spaces. In order to define these operators, we introduce another type of adjoint (different from the operator adjoint of Chapter 3).

Theorem 10.1 *Let V and W be finite-dimensional inner product spaces over F and let $\tau \in \mathcal{L}(V,W)$. Then there is a unique function $\tau^*: W \to V$, defined by the condition*

$$\langle \tau(v), w\rangle = \langle v, \tau^*(w)\rangle$$

for all $v \in V$ and $w \in W$. This function is in $\mathcal{L}(W,V)$ and is called the **adjoint** *of τ.*

Proof. For a fixed $w \in W$, consider the function $\theta_w: V \to F$ defined by

$$\theta_w(v) = \langle \tau(v), w\rangle$$

It is easy to verify that θ_w is a linear functional on V and so, by the Riesz representation theorem, there exists a unique vector $x \in V$ for which

$$\theta_w(v) = \langle v, x\rangle$$

for all $v \in V$. Hence, if $\tau^*(w) = x$ then

$$\langle \tau(v), w\rangle = \langle v, \tau^*(w)\rangle$$

for all $v \in V$. This establishes the existence and uniqueness of τ^*. To show that τ^* is linear, observe that

$$\begin{aligned}\langle v,\tau^*(rw+su)\rangle &= \langle \tau(v),rw+su\rangle\\ &= \overline{r}\langle\tau(v),w\rangle+\overline{s}\langle\tau(v),u\rangle\\ &= \overline{r}\langle v,\tau^*(w)\rangle+\overline{s}\langle v,\tau^*(u)\rangle\\ &= \langle v,r\tau^*(v)\rangle+\langle v,s\tau^*(u)\rangle\\ &= \langle v,r\tau^*(w)+s\tau^*(u)\rangle\end{aligned}$$

for all $v\in V$ and so

$$\tau^*(rw+su)=r\tau^*(w)+s\tau^*(u)$$

Hence $\tau^*\in\mathcal{L}(V,W)$. □

Here are some of the basic properties of the adjoint. Proof is left to the reader.

Theorem 10.2 *Let V and W be finite-dimensional inner product spaces. For every $\sigma,\tau\in\mathcal{L}(V,W)$ and $r\in F$*

1) $(\sigma+\tau)^*=\sigma^*+\tau^*$

2) $(r\tau)^*=\overline{r}\tau^*$

3) $\tau^{**}=\tau$

4) If $V=W$ then $(\sigma\tau)^*=\tau^*\sigma^*$

5) If τ is invertible then $(\tau^{-1})^*=(\tau^*)^{-1}$

6) If $V=W$ and $p(x)\in\mathbb{R}[x]$ then $p(\tau)^*=p(\tau^*)$. □

Now let us relate the kernel and image of a linear transformation to those of its adjoint.

Theorem 10.3 *Let $\tau\in\mathcal{L}(V,W)$ where V and W are finite-dimensional inner product spaces. Then*

1) $\ker(\tau^*)=\operatorname{im}(\tau)^\perp$

2) $\operatorname{im}(\tau^*)=\ker(\tau)^\perp$

3) τ is injective if and only if τ^* is surjective.

4) τ is surjective if and only if τ^* is injective.

5) $\ker(\tau^*\tau)=\ker(\tau)$

6) $\ker(\tau\tau^*)=\ker(\tau^*)$

7) $\operatorname{im}(\tau^*\tau)=\operatorname{im}(\tau^*)$

8) $\operatorname{im}(\tau\tau^*)=\operatorname{im}(\tau)$

9) If ρ is projection onto $\operatorname{im}(\rho)$ *along* $\ker(\rho)$ *then ρ^* is projection onto* $\ker(\rho)^\perp$ *along* $\operatorname{im}(\rho)^\perp$.

Proof. For part 1),

$$\begin{aligned}u\in\ker(\tau^*) &\Leftrightarrow \tau^*(u)=0\\ &\Leftrightarrow \langle\tau^*(u),v\rangle=0 \text{ for all } v\\ &\Leftrightarrow \langle u,\tau(v)\rangle=0 \text{ for all } v\\ &\Leftrightarrow u\in\operatorname{im}(\tau)^\perp\end{aligned}$$

Part 2) follows from part 1) by replacing τ by τ^* and taking orthogonal complements. Parts 3) and 4) follow from parts 1) and 2). For part 5), it is clear that $\tau(u)=0$ implies that $\tau^*\tau(u)=0$. For the converse, we have

$$\begin{aligned}\tau^*\tau(u)=0 &\Rightarrow \langle\tau^*\tau(u),u\rangle=0\\ &\Rightarrow \langle\tau(u),\tau(u)\rangle=0\\ &\Rightarrow \tau(u)=0\end{aligned}$$

Part 6) follows from part 5) by replacing τ with τ^*. We leave parts 7)–9) for the reader. □

The Operator Adjoint and the Hilbert Space Adjoint

We should make some remarks about the relationship between the operator adjoint $\tau^\times$ of τ, as defined in Chapter 3 and the adjoint τ^* that we have just defined, which is sometimes called the **Hilbert space adjoint**. In the first place, if $\tau\colon V\to W$ then $\tau^\times$ and τ^* have different domains and ranges

$$\tau^\times\colon W^*\to V^*$$

but

$$\tau^*\colon W\to V$$

These maps are shown in Figure 10.1, where $\phi_1\colon V^*\to V$ and $\phi_2\colon W^*\to W$ are the conjugate linear maps defined by the conditions

$$f(v)=\langle v,\phi_1(f)\rangle$$

for all $f\in V^*$ and $v\in V$ and

$$g(w)=\langle w,\phi_2(g)\rangle$$

for all $g\in W^*$ and $w\in W$, and whose existence is guaranteed by the Riesz representation theorem.

$$\begin{tikzcd}
V^* \arrow[d, "\phi_1"'] & W^* \arrow[l, "\tau^\times"'] \arrow[d, "\phi_2"] \\
V \arrow[r, shift right, "\tau"'] & W \arrow[l, shift right, "\tau^*"']
\end{tikzcd}$$

Figure 10.1

The map $\sigma\colon W^*\to V^*$ defined by

$$\sigma=(\phi_1)^{-1}\tau^*\phi_2$$

is linear. Moreover, for all $f \in W^*$ and $v \in V$

$$\begin{aligned}[\sigma(f)](v) &= [(\phi_1)^{-1}\tau^*\phi_2(f)](v) \\ &= \{(\phi_1)^{-1}[\tau^*\phi_2(f)]\}(v) \\ &= \langle v, \tau^*\phi_2(f)\rangle \\ &= \langle \tau(v), \phi_2(f)\rangle \\ &= f(\tau(v)) \\ &= \tau^\times(f)(v)\end{aligned}$$

and so $\sigma = \tau^\times$. Hence, the relationship between $\tau^\times$ and τ^* is

$$\tau^\times = (\phi_1)^{-1}\tau^*\phi_2$$

The functions ϕ are like "change of variables" functions from linear functionals to vectors, and we can say, loosely speaking, that τ^* does to Riesz vectors what $\tau^\times$ does to the corresponding linear functional.

In Chapter 3, we showed that the matrix of the operator adjoint $\tau^\times$ is the transpose of the matrix of the map τ. For Hilbert space adjoints, the situation is slightly different (due to the conjugate linearity of the inner product). Suppose that $\mathcal{B} = (b_1, \ldots, b_n)$ is an ordered orthonormal basis for V and $\mathcal{C} = (c_1, \ldots, c_m)$ is an ordered orthonormal basis for W. Then

$$([\tau^*]_{\mathcal{C},\mathcal{B}})_{i,j} = \langle \tau^*(c_i), b_j\rangle = \langle c_i, \tau(b_j)\rangle = \overline{\langle \tau(b_j), c_i\rangle} = \overline{([\tau]_{\mathcal{B},\mathcal{C}})_{j,i}}$$

and so $[\tau^*]_{\mathcal{C},\mathcal{B}}$ and $[\tau]_{\mathcal{B},\mathcal{C}}$ are conjugate transposes. If $A = (a_{i,j})$ is a matrix over F, let us write the conjugate transpose as

$$A^* = (\overline{a}_{i,j})^t$$

Theorem 10.4 *Let $\tau \in \mathcal{L}(V, W)$, where V and W are finite-dimensional inner product spaces.*

1) *The operator adjoint $\tau^\times$ and the Hilbert space adjoint τ^* are related by*

$$\tau^\times = (\phi_1)^{-1}\tau^*\phi_2$$

where ϕ_i maps a linear functional f to its Riesz vector R_f.

2) *If $\mathcal{B}$ and $\mathcal{C}$ are ordered* orthonormal bases *for V and W, respectively, then*

$$[\tau^*]_{\mathcal{C},\mathcal{B}} = ([\tau]_{\mathcal{B},\mathcal{C}})^*$$

In words, the matrix of the adjoint τ^* is the conjugate transpose of the matrix of τ. □

Unitary Diagonalizability

Recall that a linear operator $\tau \in \mathcal{L}(V)$ on a finite-dimensional vector space V is diagonalizable if and only if V has a basis consisting entirely of eigenvectors of τ, or equivalently, V can be written as a direct sum of the eigenspaces of τ

$$V = \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_k}$$

Of course, each eigenspace $\mathcal{E}_{\lambda_i}$ has an orthonormal basis $\mathcal{O}_i$, but the union of these bases, while certainly a basis for V, need not be orthonormal.

Definition *Let V be a finite-dimensional inner product space and let $\tau \in \mathcal{L}(V)$. If there is an orthonormal basis $\mathcal{O}$ for V for which $[\tau]_{\mathcal{O}}$ is a diagonal matrix, we say that τ is* **unitarily diagonalizable** *when V is complex and* **orthogonally diagonalizable** *when V is real.* □

For simplicity in exposition, we will tend to use the term unitarily diagonalizable for both cases. It is clear that the following statements are equivalent:

1) τ is unitarily diagonalizable
2) There is an orthonormal basis for V consisting entirely of eigenvectors of τ
3) V can be written as an orthogonal direct sum of eigenspaces

$$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

Since unitarily diagonalizable operators are so well behaved, it is natural to seek a characterization of such operators. Remarkably, there is a simple one.

Let us first suppose that τ is unitarily diagonalizable and that $\mathcal{O}$ is an ordered orthonormal basis of eigenvectors for τ. Then the matrix $[\tau]_{\mathcal{O}}$ is diagonal

$$[\tau]_{\mathcal{O}} = \operatorname{diag}(\lambda_{i_1}, \ldots, \lambda_{i_n})$$

and so

$$[\tau^*]_{\mathcal{O}} = \operatorname{diag}(\overline{\lambda}_{i_1}, \ldots, \overline{\lambda}_{i_n})$$

Clearly, $[\tau]_{\mathcal{O}}$ and $[\tau^*]_{\mathcal{O}}$ commute. Hence, τ and τ^* also commute, that is

$$\tau\tau^* = \tau^*\tau$$

It is a surprising fact that the converse also holds on a *complex* inner product space, that is, if τ and τ^* commute then τ is unitarily diagonalizable. (Something similar holds for real inner product spaces, as we will see.)

Normal Operators

It is clear from the preceding discussion that the following concept is key.

Definition A linear operator τ on an inner product space V is **normal** if it commutes with its adjoint

$$\tau\tau^* = \tau^*\tau$$ □

Normal operators have some very nice properties.

Theorem 10.5 *Let $\mathcal{N}$ be the set of normal operators on a finite-dimensional inner product space V. Then $\mathcal{N}$ satisfies the following properties:*

1) (**Closure under linear combinations**)

$$r,s\in F,\sigma,\tau\in\mathcal{N}\Rightarrow r\sigma+s\tau\in\mathcal{N}$$

2) (**Closure under multiplication, under a commutivity requirement**)

$$\sigma,\tau\in\mathcal{N},\sigma^*\tau=\tau\sigma^*\Rightarrow\sigma\tau\in\mathcal{N}$$

3) (**Closure under inverses**)

$$\tau\in\mathcal{N},\ \tau\ \textit{invertible}\Rightarrow\tau^{-1}\in\mathcal{N}$$

4) (**Closure under polynomials**)

$$\tau\in\mathcal{N}\Rightarrow p(\tau)\in\mathcal{N}\ \textit{for any}\ p(x)\in F[x]$$

Moreover, if $\tau\in\mathcal{N}$ then

5) $\tau(v)=0\Leftrightarrow\tau^*(v)=0$

6) $\tau^k(v)=0\Leftrightarrow\tau(v)=0$

7) *The minimal polynomial of τ is a product of distinct irreducible monic polynomials.*

8) $\tau(v)=\lambda v\Leftrightarrow\tau^*(v)=\overline{\lambda}v$

9) *Let S and T be submodules of V_τ, whose orders are relatively prime real polynomials. Then $S\perp T$.*

10) *If λ and μ are distinct eigenvalues of τ then $\mathcal{E}_\lambda\perp\mathcal{E}_\mu$.*

Proof. We leave parts 1)–4) for the reader. Normality implies that

$$\|\tau(v)\|^2=\langle\tau(v),\tau(v)\rangle=\langle\tau^*(v),\tau^*(v)\rangle=\|\tau^*(v)\|^2$$

and so part 5) follows. For part 6) let $\sigma=\tau^*\tau$. Then σ has the property that

$$\langle\sigma(v),w\rangle=\langle\tau^*\tau(v),w\rangle=\langle\tau(v),\tau(w)\rangle=\langle v,\tau^*\tau(w)\rangle=\langle v,\sigma(w)\rangle$$

and so $\sigma^*=\sigma$. (We will discuss this property of being *self-adjoint* in detail later.) Now we can easily prove part 6) for σ. For if $\sigma^k(v)=0$ for $k>1$ then

$$0=\langle\sigma^k(v),\sigma^{k-2}(v)\rangle=\langle\sigma^{k-1}(v),\sigma^{k-1}(v)\rangle$$

and so $\sigma^{k-1}(v)=0$. Continuing in this way gives $\sigma(v)=0$. Now, if $\tau^k(v)=0$ for $k>1$, then the normality of τ implies that

$$\sigma^k(v)=(\tau^*)^k\tau^k(v)=0$$

and so $\sigma(v)=0$. Hence

$$0=\langle\sigma(v),v\rangle=\langle\tau^*\tau(v),v\rangle=\langle\tau(v),\tau(v)\rangle$$

and so $\tau(v)=0$.

For part 7), suppose that

$$m_\tau(x) = p_1^{e_1}(x)\cdots p_k^{e_k}(x)$$

where $p_1(x), \ldots, p_k(x)$ are distinct, irreducible and monic. If $e_i > 1$ then for any $v \in V$

$$p_i^{e_i}(\tau)[m_\tau^{(i)}(\tau)v] = 0$$

where $m_\tau^{(i)}(x) = m_\tau(x)/p_i^{e_i}(x)$. Hence, since $p_i(\tau)$ is also normal, part 6) implies that

$$p_i(\tau)[m_\tau^{(i)}(\tau)v] = 0$$

for all $v \in V$ and so $p_i(\tau)m_\tau^{(i)}(\tau) = 0$, which is false since the polynomial $p_i(x)m_\tau^{(i)}(x)$ has degree less than that of $m_\tau(x)$. Hence, $e_i = 1$ for all i.

For part 8), using part 5) we have

$$\begin{aligned} \tau(v) = \lambda v &\Leftrightarrow (\tau - \lambda\iota)(v) = 0 \\ &\Leftrightarrow (\tau - \lambda\iota)^*(v) = 0 \\ &\Leftrightarrow \tau^*(v) = \overline{\lambda}(v) \end{aligned}$$

For part 9), let $\operatorname{ann}(S) = \langle p(x) \rangle$ and $\operatorname{ann}(T) = \langle q(x) \rangle$. Then there are real polynomials $a(x)$ and $b(x)$ for which $a(x)p(x) + b(x)q(x) = 1$. If $u \in S$ and $v \in T$ then since $p(\tau)u = 0$ implies that $p(\tau^*)u = 0$, we have

$$\begin{aligned} \langle u, v \rangle &= \langle [a(\tau^*)p(\tau^*) + b(\tau^*)q(\tau^*)]u, v \rangle \\ &= \langle b(\tau^*)q(\tau^*)u, v \rangle \\ &= \langle u, q(\tau)b(\tau)v \rangle \\ &= 0 \end{aligned}$$

Hence, $S \perp T$. For part 10), we have for $v \in \mathcal{E}_\lambda$ and $w \in \mathcal{E}_\mu$

$$\lambda\langle v, w \rangle = \langle \tau(v), w \rangle = \langle v, \tau^*(w) \rangle = \langle v, \overline{\mu}w \rangle = \mu\langle v, w \rangle$$

and since $\lambda \neq \mu$ we get $\langle v, w \rangle = 0$. □

Special Types of Normal Operators

Before discussing the structure of normal operators, we want to introduce some special types of normal operators that will play an important role in the theory.

Definition *Let V be an inner product space and let $\tau \in \mathcal{L}(V)$.*

1) *τ is* **self-adjoint** *(also called* **Hermitian** *in the complex case and* **symmetric** *in the real case), if*

$$\tau^* = \tau$$

2) *τ is called* **skew-Hermitian** *in the complex case and* **skew-symmetric** *in the real case, if*

$$\tau^* = -\tau$$

3) *τ is called* **unitary** *in the complex case and* **orthogonal** *in the real case if τ is invertible and*

$$\tau^* = \tau^{-1}$$ □

There are also matrix versions of these definitions, obtained simply by replacing the operator τ by a matrix A. In the finite-dimensional case, we have seen that

$$[\tau^*]_{\mathcal{O}} = [\tau]^*_{\mathcal{O}}$$

for any ordered orthonormal basis $\mathcal{O}$ of V and so if τ is normal then

$$[\tau]_{\mathcal{O}}[\tau]^*_{\mathcal{O}} = [\tau]_{\mathcal{O}}[\tau^*]_{\mathcal{O}} = [\tau\tau^*]_{\mathcal{O}} = [\tau^*\tau]_{\mathcal{O}} = [\tau^*]_{\mathcal{O}}[\tau]_{\mathcal{O}} = [\tau]^*_{\mathcal{O}}[\tau]_{\mathcal{O}}$$

which implies that the matrix $[\tau]_{\mathcal{O}}$ of τ is normal. The converse holds as well. In fact, we can say that τ is normal (respectively: Hermitian, symmetric, skew-Hermitian, unitary, orthogonal) if and only if any matrix that represents τ, with respect to an ordered *orthonormal* basis $\mathcal{O}$, is normal (respectively: Hermitian, symmetric, skew-Hermitian, unitary, orthogonal).

In some sense, square complex matrices are a generalization of complex numbers. Also, the adjoint (conjugate transpose) of a matrix seems to be a generalization of the complex conjugate. In looking for a tighter analogy—one that will lead to some useful mnemonics, we could consider just the diagonal matrices, but this is a bit too tight. The next logical choice is the normal operators.

Among the complex numbers, there are some special subsets: the real numbers, the positive numbers and the numbers on the unit circle. We will soon see that a normal operator is self-adjoint if and only if its complex eigenvalues are all real. This would suggest that the analog of the set of real numbers is the set of self-adjoint operators. Also, we will see that a normal operator is unitary if and only if all of its eigenvalues have norm 1, so numbers on the unit circle seem to correspond to the set of unitary operators. Of course, this is just an analogy.

Self-Adjoint Operators

Let us consider some properties of self-adjoint operators. The **quadratic form** associated with the linear operator τ is the function $Q_\tau\colon V \to F$ defined by

$$Q_\tau(v) = \langle \tau(v), v\rangle$$

We have seen that in a *complex* inner product space $\tau = 0$ if and only if $Q_\tau = 0$ but this does not hold, in general, for real inner product spaces. However, it does hold for symmetric operators on a real inner product space.

Theorem 10.6 *Let $\mathcal{H}$ be the set of self-adjoint operators on a finite-dimensional inner product space V. Then $\mathcal{H}$ satisfies the following properties:*

1) **(Closure under addition)**

$$\sigma, \tau \in \mathcal{H} \Rightarrow \sigma + \tau \in \mathcal{H}$$

2) **(Closure under real scalar multiplication)**

$$r \in \mathbb{R}, \tau \in \mathcal{H} \Rightarrow r\tau \in \mathcal{H}$$

3) **(Closure under multiplication if the factors commute)**

$$\sigma, \tau \in \mathcal{H}, \sigma\tau = \tau\sigma \Rightarrow \sigma\tau \in \mathcal{H}$$

4) **(Closure under inverses)**

$$\tau \in \mathcal{H},\ \tau \text{ invertible} \Rightarrow \tau^{-1} \in \mathcal{H}$$

5) **(Closure under real polynomials)**

$$\tau \in \mathcal{H} \Rightarrow p(\tau) \in \mathcal{H} \text{ for any } p(x) \in \mathbb{R}[x]$$

6) A complex operator τ is Hermitian if and only if $Q_\tau(v)$ is real for all $v \in V$.

7) If $F = \mathbb{C}$ or if $F = \mathbb{R}$ and τ is symmetric then $\tau = 0$ if and only if $Q_\tau = 0$

8) If τ is self-adjoint, then the characteristic polynomial of τ splits over $\mathbb{R}$ and so all complex eigenvalues are real.

Proof. For part 6), if τ is Hermitian then

$$\langle \tau(v), v \rangle = \langle v, \tau(v) \rangle = \overline{\langle \tau(v), v \rangle}$$

and so $Q_\tau(v) = \langle \tau(v), v \rangle$ is real. Conversely, if $\langle \tau(v), v \rangle \in \mathbb{R}$ then

$$\langle v, \tau(v) \rangle = \langle \tau(v), v \rangle = \langle v, \tau^*(v) \rangle$$

and so $\langle v, (\tau - \tau^*)(v) \rangle = 0$ for all $v \in V$, whence $\tau - \tau^* = 0$, which shows that τ is Hermitian.

As for part 7), the first case ($F = \mathbb{C}$) is just Theorem 9.6 so we need only consider the real case, for which

$$\begin{aligned}
0 &= \langle \tau(x+y), x+y \rangle \\
&= \langle \tau(x), x \rangle + \langle \tau(y), y \rangle + \langle \tau(x), y \rangle + \langle \tau(y), x \rangle \\
&= \langle \tau(x), y \rangle + \langle \tau(y), x \rangle \\
&= \langle \tau(x), y \rangle + \langle x, \tau(y) \rangle \\
&= \langle \tau(x), y \rangle + \langle \tau(x), y \rangle \\
&= 2\langle \tau(x), y \rangle
\end{aligned}$$

and so $\tau = 0$.

For part 8), if τ is Hermitian ($F = \mathbb{C}$) and $\tau(v) = \lambda v$ then

$$\lambda\langle v, v\rangle = \langle \tau(v), v\rangle = Q_\tau(v)$$

is real by part 5) and so λ must be real. If τ is symmetric ($F = \mathbb{R}$), we must be a bit careful, since if λ is a complex root of $C_\tau(x)$, it does not follow that $\tau(v) = \lambda v$ for some $0 \neq v \in V$. However, we can proceed as follows. Let τ be represented by the matrix A, with respect to some ordered basis for V. Then $C_\tau(x) = C_A(x)$. Now, A is a real symmetric matrix, but can be thought of as a complex Hermitian matrix that happens to have real entries. As such, it represents a Hermitian linear operator on the complex space $\mathbb{C}^n$ and so, by what we have just shown, all (complex) roots of its characteristic polynomial are real. But the characteristic polynomial of A is the same, whether we think of A as a real or a complex matrix and so the result follows. □

Unitary Operators and Isometries

We now turn to the basic properties of unitary operators. These are the workhorse operators, in that a unitary operator is precisely a normal operator that maps orthonormal bases to orthonormal bases.

Note that τ is unitary if and only if

$$\langle \tau(v), w\rangle = \langle v, \tau^{-1}(w)\rangle$$

for all $v, w \in V$.

Theorem 10.7 *Let $\mathcal{U}$ be the set of unitary operators on a finite-dimensional inner product space V. Then $\mathcal{U}$ satisfies the following properties:*

1) (**Closure under scalar multiplication by complex numbers of norm 1**)

$$r \in \mathbb{C}, |r| = 1 \textit{ and } \tau \in \mathcal{U} \Rightarrow r\tau \in \mathcal{U}$$

2) (**Closure under multiplication**)

$$\sigma, \tau \in \mathcal{U} \Rightarrow \sigma\tau \in \mathcal{U}$$

3) (**Closure under inverses**)

$$\tau \in \mathcal{U} \Rightarrow \tau^{-1} \in \mathcal{U}$$

4) τ is unitary/orthogonal if and only it is an isometric isomorphism.

5) τ is unitary/orthogonal if and only if it takes an orthonormal basis to an orthonormal basis.

6) If τ is unitary/orthogonal then the eigenvalues of τ have absolute value 1.

Proof. We leave the proofs of 1)–3) to the reader. For part 4), a unitary/orthogonal map is injective and since the range and domain have the same finite dimension, it is also surjective. Moreover, for a bijective linear map τ, we have

$$\begin{aligned}\tau \text{ is an isometry} &\Leftrightarrow \langle\tau(v),\tau(w)\rangle=\langle v,w\rangle \text{ for all } v,w\in V\\ &\Leftrightarrow \langle v,\tau^*\tau(w)\rangle=\langle v,w\rangle \text{ for all } v,w\in V\\ &\Leftrightarrow \tau^*\tau(w)=w \text{ for all } w\in V\\ &\Leftrightarrow \tau^*\tau=\iota\\ &\Leftrightarrow \tau^*=\tau^{-1}\\ &\Leftrightarrow \tau \text{ is unitary/orthogonal}\end{aligned}$$

For part 5), suppose that τ is unitary/orthogonal and that $\mathcal{O}=\{u_1,\dots,u_n\}$ is an orthonormal basis for V. Then

$$\langle\tau(u_i),\tau(u_j)\rangle=\langle u_i,u_j\rangle=\delta_{i,j}$$

and so $\tau(\mathcal{O})$ is an orthonormal basis for V. Conversely, suppose that $\mathcal{O}$ and $\tau(\mathcal{O})$ are orthonormal bases for V. Then

$$\langle\tau(u_i),\tau(u_j)\rangle=\delta_{i,j}=\langle u_i,u_j\rangle$$

Using the conjugate linearity/bilinearity of the inner product, we get $\langle\tau(v),\tau(w)\rangle=\langle v,w\rangle$ and so τ is unitary/orthogonal.

For part 6), if τ is unitary and $\tau(v)=\lambda v$ then

$$\lambda\overline{\lambda}\langle v,v\rangle=\langle\lambda v,\lambda v\rangle=\langle\tau(v),\tau(v)\rangle=\langle v,v\rangle$$

and so $|\lambda|^2=\lambda\overline{\lambda}=1$, which implies that $|\lambda|=1$. □

We also have the following theorem concerning unitary (and orthogonal) matrices.

Theorem 10.8 *Let A be an $n\times n$ matrix.*

1) *The following are equivalent:*
   a) *A is unitary*
   b) *The columns of A form an orthonormal set in $\mathbb{C}^n$.*
   c) *The rows of A form an orthonormal set in $\mathbb{C}^n$.*
2) *If A is unitary then $|\det(A)|=1$. In particular, if A is orthogonal then* $\det(A)=\pm 1$.

Proof. The matrix A is unitary if and only if $AA^*=I$, which is equivalent to saying that the rows of A are orthonormal. Similarly, A is unitary if and only if $A^*A=I$, which is equivalent to saying that the columns of A are orthonormal. As for part 2), we have

$$AA^*=I\Rightarrow \det(A)\det(A^*)=1\Rightarrow \det(A)\overline{\det(A)}=1$$

from which the result follows. □

Unitary/orthogonal matrices play the role of change of basis matrices when we restrict attention to orthonormal bases. Let us first note that if $\mathcal{B}=(u_1,\dots,u_n)$

is an ordered orthonormal basis and

$$\begin{aligned} v &= a_1u_1 + \cdots + a_nu_n \\ w &= b_1u_1 + \cdots + b_nu_n \end{aligned}$$

then

$$\langle v, w\rangle = a_1b_1 + \cdots + a_nb_n = [v]_{\mathcal{B}} \cdot [w]_{\mathcal{B}}$$

and so $v \perp w$ if and only if $[v]_{\mathcal{B}} \perp [w]_{\mathcal{B}}$.

We can now state the analog of Theorem 2.13.

Theorem 10.9 *If we are given any two of the following:*
1) A unitary/orthogonal $n \times n$ matrix A.
2) An ordered orthonormal basis $\mathcal{B}$ for F^n.
3) An ordered orthonormal basis $\mathcal{C}$ for F^n.
then the third is uniquely determined by the equation

$$A = M_{\mathcal{B},\mathcal{C}}$$ □

Unitary Similarity

We have seen that the change of basis formula for operators is given by

$$[\tau]_{\mathcal{B}'} = P[\tau]_{\mathcal{B}}P^{-1}$$

where P is an invertible matrix. What happens when the bases are orthonormal?

Definition *Two complex matrices A and B are* **unitarily similar** *(also called* **unitarily equivalent***) if there exists a unitary matrix U for which*

$$B = UAU^{-1} = UAU^*$$

The equivalence classes associated with unitary similarity are called **unitary similarity classes**. *Similarly, two real matrices A and B are* **orthogonally similar** *(also called* **orthogonally equivalent***) if there exists an orthogonal matrix O for which*

$$B = OAO^{-1} = OAO^t$$

The equivalence classes associated with orthogonal similarity are called **orthogonal similarity classes**. □

The analog of Theorem 2.19 is the following.

Theorem 10.10 *Let V be an inner product space of dimension n. Then two $n \times n$ matrices A and B are unitarily/orthogonally similar if and only if they represent the same linear operator $\tau \in \mathcal{L}(V)$, but possibly with respect to different ordered orthonormal bases. In this case, A and B represent exactly the* *same set of linear operators in $\mathcal{L}(V)$, when we restrict attention to orthonormal bases.*

Proof. If A and B represent $\tau \in \mathcal{L}(V)$, that is, if

$$A = [\tau]_{\mathcal{B}} \text{ and } B = [\tau]_{\mathcal{C}}$$

for ordered orthonormal bases $\mathcal{B}$ and $\mathcal{C}$ then

$$B = M_{\mathcal{B},\mathcal{C}} A M_{\mathcal{C},\mathcal{B}}$$

and according to Theorem 10.9, $M_{\mathcal{B},\mathcal{C}}$ is unitary/orthogonal. Hence, A and B are unitarily/orthogonally similar.

Now suppose that A and B are unitarily/orthogonally similar, say

$$B = UAU^{-1}$$

where U is unitary/orthogonal. Suppose also that A represents a linear operator $\tau \in \mathcal{L}(V)$ for some ordered orthonormal basis $\mathcal{B}$, that is,

$$A = [\tau]_{\mathcal{B}}$$

Theorem 10.9 implies that there is a unique ordered orthonormal basis $\mathcal{C}$ for V for which $U = M_{\mathcal{B},\mathcal{C}}$. Hence

$$B = M_{\mathcal{B},\mathcal{C}}[\tau]_{\mathcal{B}} M_{\mathcal{B},\mathcal{C}}^{-1} = [\tau]_{\mathcal{C}}$$

Hence, B also represents τ. By symmetry, we see that A and B represent the same set of linear operators, under all possible ordered orthonormal bases. □

Unfortunately, canonical forms for unitary similarity are rather complicated and not well discussed. We have shown in Chapter 8 that any complex matrix A is unitarily similar to an upper triangular matrix, that is, that A is unitarily upper triangularizable. (This is Schur's lemma.) However, just as in the nonunitary case, upper triangular matrices do not form a canonical form for unitary similarity. We will soon show that every complex normal matrix is unitarily diagonalizable. However, we will not discuss canonical forms for unitary similarity in this book, but instead refer the reader to the survey article [Sh].

Reflections

The following defines a very special type of unitary operator.

Definition *For a nonzero $v \in V$, the unique operator H_v for which*

$$H_v v = -v,\ (H_v)|_{\langle v \rangle^{\perp}} = \iota$$

is called a **reflection** *or a* **Householder transformation**. □

It is easy to verify that

$$H_v(x) = x - \frac{2\langle x, v\rangle}{\langle v, v\rangle} v$$

Note also that if τ is a reflection then $\tau = H_x$ if and only if $\tau(x) = -x$. For if $\tau = H_v$ and $\tau(x) = -x$ then we can write $x = av + z$ where $z \perp v$ and so

$$-(av + z) = -x = \tau(x) = H_v(av + z) = -av + z$$

which implies that $z = 0$, whence $H_v = H_{a^{-1}x} = H_x$.

If H_v is reflection and we extend v to an ordered orthonormal basis $\mathcal{B}$ for V then $[H_v]_{\mathcal{B}}$ is the matrix obtained from the identity matrix by replacing the $(1,1)$ entry by -1. Thus, we see that a reflection is unitary, Hermitian and idempotent ($H_v^2 = \iota$).

Theorem 10.11 *Let $v, w \in V$ with $\|v\| = \|w\| \neq 0$. Then H_{v-w} is the unique reflection sending v to w, that is, $H_{v-w}(v) = w$.*
Proof. If $\|v\| = \|w\|$ then $(v - w) \perp (v + w)$ and so

$$\begin{aligned} H_{v-w}(v - w) &= w - v \\ H_{v-w}(v + w) &= v + w \end{aligned}$$

from which it follows that $H_{v-w}(v) = w$. As to uniqueness, suppose H_x is a reflection for which $H_x(v) = w$. Since $H_x^{-1} = H_x$, we have $H_x(w) = v$ and so

$$H_x(v - w) = -(v - w)$$

which implies that $H_x = H_{v-w}$. □

Reflections can be used to characterize unitary operators.

Theorem 10.12 *An operator τ on a finite-dimensional inner product space V is unitary (for $F = \mathbb{C}$) or orthogonal (for $F = \mathbb{R}$) if and only if it is a product of reflections.*
Proof. Since reflections are unitary (orthogonal) and the product of unitary (orthogonal) operators is unitary, one direction is easy.

For the converse, let τ be unitary. Let $\mathcal{B} = (u_1, \ldots, u_n)$ be an orthonormal basis for V. Hence $\tau(\mathcal{B})$ is also an orthonormal basis for V. We make repeated use of the fact that $H_{x-y}(x) = y$. For example, if

$$x_1 = \tau(u_1) - u_1$$

then

$$(H_{x_1}\tau)(u_1) = u_1$$

and so $H_{x_1}\tau$ is the identity on $\langle u_1 \rangle$. Next, if

$$x_2 = (H_{x_1}\tau)(u_2) - u_2$$

then

$$(H_{x_2}H_{x_1}\tau)(u_2) = u_2$$

Also, we claim that $x_2 \perp u_1$. Since $H_{x_1}u_1 = H_{x_1}^{-1}u_1 = \tau(u_1)$, it follows that

$$\begin{aligned}\langle (H_{x_1}\tau)(u_2) - u_2, u_1\rangle &= \langle (H_{x_1}\tau)(u_2), u_1\rangle \\ &= \langle \tau(u_2), H_{x_1}u_1\rangle \\ &= \langle \tau(u_2), \tau(u_1)\rangle \\ &= \langle u_2, u_1\rangle \\ &= 0\end{aligned}$$

Hence

$$(H_{x_2}H_{x_1}\tau)(u_1) = H_{x_2}(u_1) = u_1$$

and so $H_{x_2}H_{x_1}\tau$ is the identity on $\langle u_1, u_2\rangle$. Now let us generalize.

Assume that for $k > 1$, we have found reflections $H_{x_{k-1}}, \ldots, H_{x_1}$ for which $H_{x_{k-1}}\cdots H_{x_1}\tau$ is the identity on $\langle$u$_1, \ldots, u_{k-1}\rangle$. If

$$x_k = (H_{x_{k-1}}\cdots H_{x_1}\tau)(u_k) - u_k$$

then

$$(H_{x_k}\cdots H_{x_1}\tau)(u_k) = u_k$$

Also, we claim that $x_k \perp u_i$ for all $i < k$. Since $H_{x_{k-1}}\cdots H_{x_1}u_i = \tau(u_i)$ it follows that

$$\begin{aligned}\langle (H_{x_{k-1}}\cdots H_{x_1}\tau)(u_k) - u_k, u_i\rangle &= \langle (H_{x_{k-1}}\cdots H_{x_1}\tau)(u_k), u_i\rangle \\ &= \langle \tau(u_k), H_{x_{k-1}}\cdots H_{x_1}u_i\rangle \\ &= \langle \tau(u_k), \tau(u_i)\rangle \\ &= \langle u_k, u_i\rangle \\ &= 0\end{aligned}$$

Hence

$$(H_{x_k}\cdots H_{x_1}\tau)(u_i) = H_{x_k}(u_i) = u_k$$

and so $H_{x_k}\cdots H_{x_1}\tau$ is the identity on $\langle u_1, \ldots, u_k\rangle$.

Thus, for $k = n$ we have $H_{x_n}\cdots H_{x_1}\tau = \iota$ and so $\tau = H_{x_n}\cdots H_{x_1}$, as desired. □

The Structure of Normal Operators

We are now ready to consider the structure of a normal operator τ on a finite-dimensional inner product space. According to Theorem 10.5, the minimal polynomial of τ has the form

$$m_r(x) = p_1(x)\cdots p_n(x)$$

where the p_i's are distinct monic irreducible polynomials.

If $F = \mathbb{C}$, then each $p_i(x)$ is linear. Theorem 8.11 then implies that τ is diagonalizable and

$$V = \mathcal{E}_{\lambda_1} \oplus \cdots \oplus \mathcal{E}_{\lambda_k}$$

where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of τ. Theorem 10.5 also tells us that if $\lambda_i \neq \lambda_j$ then $\mathcal{E}_{\lambda_i} \perp \mathcal{E}_{\lambda_j}$ and so

$$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

This is equivalent to saying that V has an orthonormal basis of eigenvectors of τ.

The converse is also true, that is, if $\mathcal{B} = (u_1, \ldots, u_n)$ is an ordered orthonormal basis of eigenvectors of τ then

$$\langle \tau^* u_i, u_j \rangle = \langle u_i, \tau u_j \rangle = \langle u_i, \lambda_j u_j \rangle = \overline{\lambda}_j \delta_{i,j} = \langle \overline{\lambda}_i u_i, u_j \rangle$$

and so $\tau^* u_i = \overline{\lambda}_i u_i$. It follows that

$$\tau\tau^* u_i = \overline{\lambda}_i \lambda_i u_i = \tau^* \tau u_i$$

and so τ is normal.

Theorem 10.13 *(***The structure theorem for normal operators: complex case***) Let V be a finite-dimensional complex inner product space. Then a linear operator τ on V is normal if and only if V has an orthonormal basis $\mathcal{B}$ consisting entirely of eigenvectors of τ, that is*

$$V_\tau = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ. Put another way, τ is normal if and only if it is unitarily diagonalizable. □

Now let us consider the real case, which is more complex than the complex case. However, we can take advantage of the corresponding result for $F = \mathbb{C}$, by using the complexification process (which we will review in a moment).

First, let us observe that when $F = \mathbb{R}$, the minimal polynomial of τ is a product of distinct real linear and real quadratic factors, say

$$m_\tau(x) = (x - r_1)\cdots(x - r_k) p_1(x)\cdots p_d(x)$$

where the r_i's are distinct and the $p_i(x)$'s are distinct real irreducible quadratics.

Hence, according to part 9) of Theorem 10.5, the primary decomposition of V_τ has the form

$$V_\tau = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k} \odot W_1 \odot \cdots \odot W_s$$

where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ and where

$$W_i = \{v \in V \mid p_i(\tau)v = 0\}$$

are τ-invariant subspaces. Accordingly, we can focus on the subspace $W = W_1 \odot \cdots \odot W_s$, upon which τ is normal, with a minimal polynomial that is the product of distinct irreducible quadratic factors.

Let us briefly review the complexification process. Recall that if V is a real vector space, then the set

$$V^{\mathbb{C}} = \{u + vi \mid u, v \in V\}$$

is a complex vector space under addition and scalar multiplication "borrowed" from the field of complex numbers, that is,

$$\begin{aligned} (u+vi)+(x+yi) &= (u+x)+(v+y)i \\ (a+bi)(u+vi) &= (au-bv)+(av+bu)i \end{aligned}$$

Recall also that the complexification map cpx: $V \to V^{\mathbb{C}}$ defined by

$$\text{cpx}(v) = v + 0i$$

is an injective linear transformation from the real vector space V to the real version $(V^{\mathbb{C}})_{\mathbb{R}}$ of the complexification $V^{\mathbb{C}}$.

If $\mathcal{B} = \{v_j \mid j \in I\}$ is a basis for V over $\mathbb{R}$, then the complexification of $\mathcal{B}$

$$\text{cpx}(\mathcal{B}) = \{v_j + 0i \mid v_j \in \mathcal{B}\}$$

is a basis for the vector space $V^{\mathbb{C}}$ over $\mathbb{C}$ and so

$$\dim(V^{\mathbb{C}}) = \dim(V)$$

For any linear operator τ on V, we can define a linear operator $\tau^{\mathbb{C}}$ on $V^{\mathbb{C}}$ by

$$\tau^{\mathbb{C}}(u + vi) = \tau(u) + \tau(v)i$$

Note that

$$(\sigma\tau)^{\mathbb{C}} = \sigma^{\mathbb{C}}\tau^{\mathbb{C}}$$

Also, if $p(x)$ is a real polynomial then $p(\tau^{\mathbb{C}}) = [p(\tau)]^{\mathbb{C}}$.

For any ordered basis $\mathcal{B}$ of V, we have

$$[\tau^{\mathbb{C}}]_{\mathrm{cpx}(\mathcal{B})} = [\tau]_{\mathcal{B}}$$

Hence, if a real matrix A represents a linear operator τ on V then A also represents the complexification of τ on $V^{\mathbb{C}}$. In particular, the polynomial $c(x) = \det(xI - A)$ is the characteristic polynomial of both τ and $\tau^{\mathbb{C}}$.

If V is a real inner product space, then we can define an inner product on the complexification $V^{\mathbb{C}}$ as follows (this is the same formula as for the ordinary inner product on a complex vector space)

$$\langle u+vi, x+yi\rangle = \langle u,x\rangle + \langle v,y\rangle + (\langle v,x\rangle - \langle u,y\rangle)i$$

From this, it follows that if $u, x \in V$ then

$$\langle u^{\mathbb{C}}, x^{\mathbb{C}}\rangle = \langle u,x\rangle$$

and, in particular, $u \perp x$ in V if and only if $u^{\mathbb{C}} \perp x^{\mathbb{C}}$ in $V^{\mathbb{C}}$.

Next, we have $(\tau^*)^{\mathbb{C}} = (\tau^{\mathbb{C}})^*$, since

$$\begin{aligned}\langle u+vi, (\tau^*)^{\mathbb{C}}(x+yi)\rangle &= \langle u+vi, \tau^*(x)+\tau^*(y)i\rangle\\ &= \langle u,\tau^*(x)\rangle + \langle v,\tau^*(y)\rangle + (\langle v,\tau^*(x)\rangle - \langle u,\tau^*(y)\rangle)i\\ &= \langle \tau(u),x\rangle + \langle \tau(v),y\rangle + (\langle \tau(v),x\rangle - \langle \tau(u),y\rangle)i\\ &= \langle \tau(u)+\tau(v)i, x+yi\rangle\\ &= \langle \tau^{\mathbb{C}}(u+vi), x+yi\rangle\\ &= \langle u+vi, (\tau^{\mathbb{C}})^*(x+yi)\rangle\end{aligned}$$

It follows that τ is normal if and only if $\tau^{\mathbb{C}}$ is normal.

Now consider a normal linear operator τ on a real vector space V and suppose that the minimal polynomial $m_\tau(x)$ of τ is the product of distinct irreducible quadratic factors

$$m_\tau(x) = p_1(x)\cdots p_d(x)$$

Hence, $m_\tau(x)$ has distinct roots, all of which are nonreal, say

$$\lambda_1, \overline{\lambda}_1, \ldots, \lambda_d, \overline{\lambda}_d$$

and since the characteristic polynomial $c(x)$ of τ and $\tau^{\mathbb{C}}$ is a multiple of $m_\tau(x)$, these scalars are characteristic roots of $\tau^{\mathbb{C}}$.

Also, since $m_\tau(x)$ is real, it follows that

$$m_\tau(\tau^{\mathbb{C}}) = [m_\tau(\tau)]^{\mathbb{C}} = 0$$

and so $m_{\tau^{\mathbb{C}}}(x) \mid m_\tau(x)$. However, the eigenvalues $\lambda_i, \overline{\lambda}_i$ of $\tau^{\mathbb{C}}$ are roots of $m_{\tau^{\mathbb{C}}}(x)$ and so $m_\tau(x) \mid m_{\tau^{\mathbb{C}}}(x)$. Thus, $m_{\tau^{\mathbb{C}}}(x) = m_\tau(x)$.

Since $m_{\tau^{\mathbb{C}}}(x)$ is the product of distinct *linear* factors over $\mathbb{C}$, we deduce immediately that $\tau^{\mathbb{C}}$ is diagonalizable. Hence, $V^{\mathbb{C}}$ has a basis of eigenvectors of $\tau^{\mathbb{C}}$, that is,

$$V^{\mathbb{C}} = \mathcal{E}_{\lambda_1} \odot \mathcal{E}_{\overline{\lambda}_1} \odot \cdots \odot \mathcal{E}_{\lambda_d} \odot \mathcal{E}_{\overline{\lambda}_d}$$

Note also that since $\tau^{\mathbb{C}}$ is normal, these eigenspaces are orthogonal under the inner product on $V^{\mathbb{C}}$.

Let us consider a particular eigenvalue pair λ and $\overline{\lambda}$ and the subspace $\mathcal{E}_\lambda \odot \mathcal{E}_{\overline{\lambda}}$. (For convenience, we have dropped the subscript.) Suppose that $\lambda = a + bi$ and that

$$\mathcal{O} = (u_1 + v_1 i, \ldots, u_m + v_m i)$$

is an ordered orthonormal basis for $\mathcal{E}_\lambda$. Then for any $j = 1, \ldots, m$,

$$\tau^{\mathbb{C}}(u_j + v_j i) = (a + bi)(u_j + v_j i)$$

and so

$$\begin{aligned} \tau(u_j) &= au_j - bv_j \\ \tau(v_j) &= bu_j + av_j \end{aligned}$$

It follows that

$$\begin{aligned} \tau^{\mathbb{C}}(u_j - v_j i) &= \tau(u_j) - \tau(v_j) i \\ &= au_j - bv_j - (bu_j + av_j) i \\ &= (a - bi)(u_j - v_j i) \\ &= \overline{\lambda}(u_j - v_j i) \end{aligned}$$

which shows that $u_j - v_j i$ is an eigenvector for $\tau^{\mathbb{C}}$ associated with $\overline{\lambda}$ and so

$$\overline{\mathcal{O}} = (u_1 - v_1 i, \ldots, u_m - v_m i) \subseteq \mathcal{E}_{\overline{\lambda}}$$

But the set $\overline{\mathcal{O}}$ is easily seen to be linearly independent and so $\dim(\mathcal{E}_{\overline{\lambda}}) \geq \dim(\mathcal{E}_\lambda)$. Using the same argument with λ replaced by $\overline{\lambda}$, we see that this inequality is an equality. Hence $\overline{\mathcal{O}}$ is an ordered orthonormal basis for $\mathcal{E}_{\overline{\lambda}}$.

It follows that

$$\mathcal{E}_\lambda \odot \mathcal{E}_{\overline{\lambda}} = U_1 \odot \cdots \odot U_m$$

where

$$U_j = \operatorname{span}(u_j + v_j i, u_j - v_j i)$$

is two-dimensional, because the eigenvectors $u_j + v_j i$ and $u_j - v_j i$ are associated with distinct eigenvalues and are therefore linearly independent.

Hence, $V^{\mathbb{C}}$ is the orthogonal direct sum of τ-invariant two-dimensional subspaces

$$V^{\mathbb{C}} = U_1 \odot \cdots \odot U_n$$

where $2n = \dim(V^{\mathbb{C}}) = \dim(V)$ and where each subspace U_j has the property that

$$\begin{aligned}\tau(u_j) &= a_j u_j - b_j v_j \\ \tau(v_j) &= b_j u_j + a_j v_j\end{aligned}$$

and where the scalars $\lambda = a_j + b_j i$ range over the distinct eigenvalues $\lambda_1, \overline{\lambda}_1, \ldots, \lambda_d, \overline{\lambda}_d$ of $\tau^{\mathbb{C}}$.

Now we wish to drop down to V. For each $j = 1, \ldots, n$, let $S_j = \text{span}(u_j, v_j)$ be the subspace of V spanned by the real and imaginary parts of the eigenvectors $u_j + v_j i$ and $u_j - v_j i$ that span U_j. To see that S_j is two-dimensional, consider its complexification

$$S_j^{\mathbb{C}} = \{x + yi \mid x, y \in S_j\}$$

Since $U_j \subseteq S_j^{\mathbb{C}}$, we have

$$2 = \dim(U_j) \leq \dim(S_j^{\mathbb{C}}) = \dim(S_j) \leq 2$$

(This can also be seen directly by applying $\tau^{\mathbb{C}}$ to the equation $ru_j + sv_j = 0$ and solving the resulting pair of equations in u_j and v_j.)

Next, we observe that if $x \in S_j$ and $y \in S_k$ with $j \neq k$, then since $x^{\mathbb{C}} \in U_j$ and $y^{\mathbb{C}} \in U_k$, we have $x^{\mathbb{C}} \perp y^{\mathbb{C}}$ and so $x \perp y$. Thus, $S_j \perp S_k$.

In summary, if $\mathcal{B}_j = (u_j, v_j)$, then the subspaces S_j are two-dimensional, τ-invariant and pairwise orthogonal, with matrix

$$[\tau]_{\mathcal{B}_j} = \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}$$

It follows that $S_1 \odot \cdots \odot S_n \subseteq V$ but since the dimensions of both sides are equal, we have equality

$$V = S_1 \odot \cdots \odot S_n$$

Theorem 10.14 (The structure theorem for normal operators: real case) Let V be a finite-dimensional real inner product space. A linear operator τ on V is normal if and only if

$$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k} \odot S_1 \odot \cdots \odot S_m$$

where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ and each S_j is a two-dimensional τ-invariant subspace for which there exists an ordered basis $\mathcal{B}_j = (u_j, v_j)$ for which

$$[\tau]_{\mathcal{B}_j} = \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}$$

for $a_j, b_j \in \mathbb{R}$.

Proof. We need only show that if V has such a decomposition, then τ is normal. But since $[\tau]_{\mathcal{B}_j}([\tau]_{\mathcal{B}_j})^t = (a_j^2 + b_j^2)I_2$, it is clear that $[\tau]_{\mathcal{B}_j}$ is real normal. it follows easily that τ is real normal. □

The following theorem includes the structure theorems stated above for the real and complex cases, along with some further refinements related to self-adjoint and unitary/orthogonal operators.

Theorem 10.15 *(***The structure theorem for normal operators***)*

1) (**Complex case**) *Let V be a finite-dimensional complex inner product space. Then*
   a) *An operator τ on V is normal if and only if V has an orthonormal basis $\mathcal{B}$ consisting entirely of eigenvectors of τ, that is*

   $$V_\tau = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

   where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ. Put another way, τ is normal if and only if it is unitarily diagonalizable.
   b) *Among the normal operators, the Hermitian operators are precisely those for which all complex eigenvalues are real.*
   c) *Among the normal operators, the unitary operators are precisely those for which all eigenvalues have norm* 1.
2) (**Real case**) *Let V be a finite-dimensional real inner product space. Then*
   a) *A linear operator τ on V is normal if and only if*

   $$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k} \odot S_1 \odot \cdots \odot S_m$$

   where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ and each S_j is a two-dimensional τ-invariant subspace for which there exists an ordered basis $\mathcal{B}_j = (u_j, v_j)$ for which

   $$[\tau]_{\mathcal{B}_j} = \begin{bmatrix} a_j & -b_j \\ b_j & a_j \end{bmatrix}$$

   for $a_j, b_j \in \mathbb{R}$.
   b) *Among the real normal operators, the symmetric operators are precisely those for which there are no subspaces U_i in the*

decomposition of part 2b). Hence, an operator is symmetric if and only if it is orthogonally diagonalizable.

c) *Among the real normal operators, the orthogonal operators are precisely those for which the eigenvalues are equal to ± 1 and the matrices $[\tau]_{\mathcal{B}_i}$ described in part 2a) have rows (and columns) of norm 1, that is,*

$$[\tau]_{\mathcal{B}_i} = \begin{bmatrix} \sin\theta & -\cos\theta \\ \cos\theta & \sin\theta \end{bmatrix}$$

for some $\theta \in \mathbb{R}$.

Proof. We have proved part 1a). As to part 1b), it is only necessary to look at the matrix A of τ with respect to a basis $\mathcal{B}$ consisting of eigenvectors for τ. This matrix is diagonal and so it is Hermitian ($A = A^*$) if and only if the diagonal entries are real. Similarly, A is unitary ($A^{-1} = A^*$) if and only if the diagonal entries have absolute value equal to 1.

We have proved part 2a). Parts 2b) and 2c) are seen to be true by looking at the matrix $A = [\tau]_{\mathcal{B}}$, which is symmetric ($A = A^t$) if and only if A is diagonal and A is orthogonal if and only if $\lambda_i = \pm 1$ and the matrices M have orthonormal rows. □

Matrix Versions

We can formulate matrix versions of the structure theorem for normal operators.

Theorem 10.16 *(***The structure theorem for normal matrices***)*

1) (**Complex case**)

a) *A complex matrix A is normal if and only if there is a unitary matrix U for which*

$$UAU^{-1} = \text{diag}(\lambda_1, \ldots, \lambda_k)$$

where $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of τ. Put another way, A is normal if and only if it is unitarily diagonalizable .

b) *A complex matrix A is Hermitian if and only if 1a) holds where all eigenvalues λ_i are real.*

c) *A complex matrix A is unitary if and only if 1a) holds where all eigenvalues λ_i have norm 1.*

2) (**Real case**)

a) *A real matrix A is real normal if and only if there is an orthogonal matrix O for which OAO^{-1} has the block diagonal form*

$$OAO^{-1} = \text{diag}\left(\lambda_1, \ldots, \lambda_k, \begin{bmatrix} a_1 & -b_1 \\ b_1 & a_1 \end{bmatrix}, \ldots, \begin{bmatrix} a_m & -b_m \\ b_m & a_m \end{bmatrix}\right)$$

b) *A real matrix A is symmetric if and only if there is an orthogonal matrix O for which OAO^{-1} has the block diagonal form*

$$OAO^{-1} = \text{diag}(\lambda_1, \ldots, \lambda_k)$$

That is, a real matrix A is symmetric if and only if it is orthogonally diagonalizable.

c) A real matrix A is orthogonal if and only if there is an orthogonal matrix O for which OAO^{-1} has the block diagonal form

$$OAO^{-1} = \text{diag}\left(\lambda_1, \ldots, \lambda_k, \begin{bmatrix} \sin\theta_1 & -\cos\theta_1 \\ \cos\theta_1 & \sin\theta_1 \end{bmatrix}, \ldots, \begin{bmatrix} \sin\theta_m & -\cos\theta_m \\ \cos\theta_m & \sin\theta_m \end{bmatrix}\right)$$

for some $\theta_1, \ldots, \theta_m \in \mathbb{R}$. □

Orthogonal Projections

We now wish to characterize unitary diagonalizability in terms of projection operators.

Definition *Let $V = S \odot S^{\perp}$. The projection map $\rho: V \to S$ on S along $S^{\perp}$ is called* **orthogonal projection** *onto S. Put another way, a projection map ρ is an orthogonal projection if* $\text{im}(\rho) \perp \text{ker}(\rho)$. □

Note that some care must be taken to avoid confusion between the term orthogonal projection and the concept of projections that are orthogonal to each other, that is, for which $\rho\sigma = \sigma\rho = 0$.

We saw in Chapter 8 that an operator ρ is a projection operator if and only if it is idempotent. Here is the analogous characterization of orthogonal projections.

Theorem 10.17 *Let V be a finite-dimensional inner product space. The following are equivalent for an operator ρ on V:*

1) ρ is an orthogonal projection

2) ρ is idempotent and self-adjoint

3) ρ is idempotent and does not expand lengths, that is

$$\|\rho(v)\| \le \|v\|$$

for all $v \in V$.

Proof. To see that 1) and 2) are equivalent, we have

$$\begin{aligned} \rho = \rho^* &\Leftrightarrow \text{im}(\rho) = \text{im}(\rho^*) \text{ and } \ker(\rho) = \ker(\rho^*) \\ &\Leftrightarrow \text{im}(\rho) = \ker(\rho)^{\perp} \text{ and } \ker(\rho) = \text{im}(\rho)^{\perp} \\ &\Leftrightarrow \text{im}(\rho) \perp \ker(\rho) \end{aligned}$$

To prove that 1) implies 3), note that $v = \rho(v) + z$ where $z \in \ker(\rho)$ and since $\rho(v) \perp z$ we have

$$\|v\|^2 = \|\rho(v)\|^2 + \|z\|^2 \geq \|\rho(v)\|^2$$

from which the result follows.

Now suppose that 3) holds. We know that $V = \operatorname{im}(\rho) \oplus \ker(\rho)$ and wish to show that this sum is orthogonal. According to Theorem 9.13, it is sufficient to show that $\operatorname{im}(\rho) \subseteq \ker(\rho)^{\perp}$. Let $w \in \operatorname{im}(\rho)$. Then since $V = \ker(\rho) \odot \ker(\rho)^{\perp}$, we have $w = x + y$, where $x \in \ker(\rho)$ and $y \in \ker(\rho)^{\perp}$ and so

$$w = \rho(w) = \rho(x) + \rho(y) = \rho(y)$$

Hence

$$\|x\|^2 + \|y\|^2 = \|w\|^2 = \|\rho(y)\|^2 \leq \|y\|^2$$

which implies that $\|x\| = 0$ and hence that $x = 0$. Thus, $w = y \in \ker(\rho)^{\perp}$ and so $\operatorname{im}(\rho) \subseteq \ker(\rho)^{\perp}$, as desired. □

Note that for an orthogonal projection ρ, we have

$$\langle v, \rho(v)\rangle = \langle v, \rho^2(v)\rangle = \langle \rho(v), \rho(v)\rangle = \|\rho(v)\|^2$$

Next we give some additional properties of orthogonal projections.

Theorem 10.18 *Let V be an inner product space over a field of characteristic $\neq 2$. Let ρ, $\rho_1, \dots, \rho_k$ and σ be projections, each of which is orthogonal. Then*

1) $\rho\sigma = 0$ *if and only if* $\sigma\rho = 0$.
2) $\rho + \sigma$ *is an orthogonal projection if and only if* $\rho \perp \sigma$, *in which case* $\rho + \sigma$ *is projection on* $\operatorname{im}(\rho) \odot \operatorname{im}(\sigma)$ *along* $\ker(\rho) \cap \ker(\sigma)$.
3) $\rho = \rho_1 + \cdots + \rho_k$ *is an orthogonal projection if and only if* $\rho_i \perp \rho_j$ *for all* $i \neq j$.
4) $\rho - \sigma$ *is an orthogonal projection if and only if*
$$\rho\sigma = \sigma\rho = \sigma$$
in which case $\rho - \sigma$ *is projection on* $\operatorname{im}(\rho) \cap \ker(\sigma)$ *along* $\ker(\rho) \odot \operatorname{im}(\sigma)$.
5) *If* $\rho\sigma = \sigma\rho$ *then* $\rho\sigma$ *is an orthogonal projection. In this case,* $\rho\sigma$ *is projection on* $\operatorname{im}(\rho) \cap \operatorname{im}(\sigma)$ *along* $\ker(\rho) \odot \ker(\sigma)$.
6) *a)* ρ^* *is orthogonal projection onto* $\ker(\rho)^{\perp}$ *along* $\operatorname{im}(\rho)^{\perp}$
   b) $\rho^*\rho$ *is orthogonal projection onto* $\ker(\rho)$ *along* $\ker(\rho)^{\perp}$
   c) $\rho\rho^*$ *is orthogonal projection onto* $\operatorname{im}(\rho)$ *along* $\operatorname{im}(\rho)^{\perp}$

Proof. We prove only part 3). If the ρ_i's are orthogonal projections and if $\rho_i \perp \rho_j$ for all $i \neq j$ then $\rho_i\rho_j = 0$ for all $i \neq j$ and so it is straightforward to check that $\rho^2 = \rho$ and that $\rho^* = \rho$. Hence, ρ is an orthogonal projection. Conversely, suppose that ρ is an orthogonal projection and that $x \in \operatorname{im}(\rho_i)$ for a fixed i. Then $\rho_i(x) = x$ and so

$$\begin{aligned}\|x\|^2 \geq \|\rho(x)\|^2 &= \langle \rho(x), \rho(x)\rangle = \langle \rho(x), x\rangle \\ &= \sum_j \langle \rho_j(x), x\rangle = \sum_j \|\rho_j(x)\|^2 \geq \|\rho_i(x)\|^2 = \|x\|^2\end{aligned}$$

which implies that $\rho_j(x) = 0$ for $j \neq i$. In other words,

$$\mathrm{im}(\rho_i) \subseteq \ker(\rho_j) = \mathrm{im}(\rho_j)^{\perp}$$

Therefore,

$$0 = \langle \rho_j(v), \rho_i(w)\rangle = \langle \rho_i\rho_j(v), w\rangle$$

for all $v, w \in V$, which shows that $\rho_i\rho_j = 0$, that is, $\rho_i \perp \rho_j$. □

For orthogonal projections σ and ρ, we can define a partial order by defining $\sigma \leq \rho$ to be $\mathrm{im}(\sigma) \subseteq \mathrm{im}(\rho)$. it is easy to verify that this is a reflexive, antisymmetric, transitive relation on the set of all orthogonal projections. Furthermore, we have the following characterizations.

Theorem 10.19 *The following statements are equivalent for orthogonal projections ρ and σ:*

1) $\sigma \leq \rho$
2) $\rho\sigma = \sigma$
3) $\sigma\rho = \sigma$
4) $\|\sigma(v)\| \leq \|\rho(v)\|$ *for all* $v \in V$.
5) $Q_{\rho-\sigma}(v) \geq 0$, *for all* $v \in V$.

Proof. First, we show that 2) and 3) are equivalent. If 2) holds then

$$\sigma\rho = \sigma^*\rho^* = (\rho\sigma)^* = \sigma^* = \sigma$$

and so 3) holds. Similarly, 3) implies 2). Next, note that 4) and 5) are equivalent, since

$$\begin{aligned}Q_{\rho-\sigma}(v) &= \langle \rho(v), v\rangle - \langle \sigma(v), v\rangle \\ &= \langle \rho(v), \rho(v)\rangle - \langle \sigma(v), \sigma(v)\rangle \\ &= \|\rho(v)\|^2 - \|\sigma(v)\|^2\end{aligned}$$

Now, 2) is equivalent to the statement that ρ fixes each element of $\mathrm{im}(\sigma)$, which is equivalent to the statement that $\mathrm{im}(\sigma) \subseteq \mathrm{im}(\rho)$, which is 1). Hence, 1)–3) are equivalent.

Finally, if 2) holds, then 3) also holds and so by Theorem 10.18, the difference $\rho - \sigma$ is an orthogonal projection, from which it follows that

$$Q_{\rho-\sigma}(v) = \langle (\rho-\sigma)(v), v\rangle = \langle (\rho-\sigma)(v), (\rho-\sigma)(v)\rangle \geq 0$$

which is 5). Also, if 4) holds, then for any $v \in \mathrm{im}(\sigma)$, we have $v = x + y$, where

$x \in \operatorname{im}(\rho), y \in \ker(\rho)$ and $x \perp y$. Then,

$$\|x\|^2 + \|y\|^2 = \|v\|^2 = \|\sigma(v)\|^2 \leq \|\rho(v)\|^2 = \|x\|^2$$

and so $y = 0$, that is, $v \in \operatorname{im}(\rho)$. Hence, 1) holds. □

Orthogonal Resolutions of the Identity

Recall that resolutions of the identity

$$\rho_1 + \cdots + \rho_k = \iota$$

correspond to direct sum decompositions of the space V. It is the mutual orthogonality of the projections that gives the directness of the sum. If, in addition, the projections are themselves orthogonal, then the direct sum is an orthogonal sum.

Definition If $\rho_1 + \cdots + \rho_k = \iota$ is a resolution of the identity and if each ρ_i is orthogonal, then we say that $\rho_1 + \cdots + \rho_k = \iota$ is an **orthogonal resolution of the identity**. □

The following theorem displays a correspondence between orthogonal direct sum decompositions of V and orthogonal resolutions of the identity.

Theorem 10.20

1) If $\rho_1 + \cdots + \rho_k = \iota$ is an orthogonal resolution of the identity then

$$V = \operatorname{im}(\rho_1) \odot \cdots \odot \operatorname{im}(\rho_k)$$

2) Conversely, if $V = S_1 \odot \cdots \odot S_k$ and ρ_i is projection on S_i along $S_1 \odot \cdots \odot \widehat{S}_i \odot \cdots \odot S_k$, where the hat ^ means that the corresponding term is missing from the direct sum, then

$$\rho_1 + \cdots + \rho_k = \iota$$

is an orthogonal resolution of the identity.

Proof. To prove 1) suppose that $\rho_1 + \cdots + \rho_k = \iota$ is an orthogonal resolution of the identity. According to Theorem 8.17, we have

$$V = \operatorname{im}(\rho_1) \oplus \cdots \oplus \operatorname{im}(\rho_k)$$

However, since the ρ_i's are orthogonal, they are self-adjoint and so for $i \neq j$,

$$\langle \rho_i(v), \rho_j(w) \rangle = \langle v, \rho_i \rho_j(w) \rangle = \langle v, 0 \rangle = 0$$

Hence

$$V = \operatorname{im}(\rho_1) \odot \cdots \odot \operatorname{im}(\rho_k)$$

For the converse, we know from Theorem 8.17 that $\rho_1 + \cdots + \rho_k = \iota$ is a resolution of the identity and we need only show that each ρ_i is an orthogonal projection. But this follows from the fact that

$$\text{im}(\rho_i)^{\perp} = \text{im}(\rho_1) \odot \cdots \odot \text{im}(\rho_{i-1}) \odot \text{im}(\rho_{i+1}) \odot \cdots \odot \text{im}(\rho_k) = \ker(\rho_i) \quad \square$$

The Spectral Theorem

We can now characterize the normal (unitarily diagonalizable) operators on a finite-dimensional complex inner product space using projections.

Theorem 10.21 *(***The spectral theorem for normal operators***) Let τ be an operator on a finite-dimensional complex inner product space V. The following statements are equivalent:*

1) τ is normal

2) τ is unitarily diagonalizable, that is,

$$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

3) τ has the **orthogonal spectral resolution**

$$\tau = \lambda_1\rho_1 + \cdots + \lambda_k\rho_k \tag{10.1}$$

where $\lambda_i \in \mathbb{C}$ and where $\rho_1 + \cdots + \rho_k = \iota$ is an orthogonal resolution of the identity.

Moreover, if τ has the form (10.1), where the λ_i's are distinct and the ρ_i's are nonzero then the λ_i's are the eigenvalues of τ and $\text{im}(\rho_i)$ is the eigenspace associated with λ_i.

Proof. We have seen that 1) and 2) are equivalent. Suppose that τ is unitarily diagonalizable. Let ρ_i be orthogonal projection onto $\mathcal{E}_{\lambda_i}$. Then any $v \in V$ can be written as a sum of orthogonal eigenvectors

$$v = v_1 + \cdots + v_k$$

and so

$$\tau(v) = \lambda_1 v_1 + \cdots + \lambda_k v_k = (\lambda_1\rho_1 + \cdots + \lambda_k\rho_k)(v)$$

Hence, 3) holds. Conversely, if (10.1) holds, we have

$$V = \text{im}(\rho_1) \odot \cdots \odot \text{im}(\rho_k)$$

But Theorem 8.18 implies that $\text{im}(\rho_i) = \mathcal{E}_{\lambda_i}$ and so τ is unitarily diagonalizable. $\square$

In the real case, we have the following.

Theorem 10.22 *(***The spectral theorem for self-adjoint operators***) Let τ be an operator on a finite-dimensional real inner product space V. The following statements are equivalent:*

1) τ is self-adjoint

2) τ is orthogonally diagonalizable, that is,

$$V = \mathcal{E}_{\lambda_1} \odot \cdots \odot \mathcal{E}_{\lambda_k}$$

3) τ has the **orthogonal spectral resolution**

$$\tau = \lambda_1 \rho_1 + \cdots + \lambda_k \rho_k \tag{10.2}$$

where $\lambda_i \in \mathbb{R}$ and $\rho_1 + \cdots + \rho_k = \iota$ is an orthogonal resolution of the identity.

Moreover, if τ has the form (10.2), where the λ_i's are distinct and the ρ_i's are nonzero then the λ_i's are the eigenvalues of τ and $\operatorname{im}(\rho_i)$ is the eigenspace associated with λ_i. □

Spectral Resolutions and Functional Calculus

Let τ be a linear operator on a finite-dimensional inner product space V, and let τ have spectral resolution

$$\tau = \lambda_1 \rho_1 + \cdots + \lambda_k \rho_k$$

Since ρ_i is idempotent, we have $\rho_i^m = \rho_i$ for all $m \ge 1$. The mutual orthogonality of the projections means that $\rho_i \rho_j = 0$ for all $i \ne j$ and so

$$\tau^n = (\lambda_1 \rho_1 + \cdots + \lambda_k \rho_k)^n = \lambda_1^n \rho_1 + \cdots + \lambda_k^n \rho_k$$

More generally, for any polynomial $p(x)$ over F, we have

$$p(\tau) = p(\lambda_1)\rho_1 + \cdots + p(\lambda_k)\rho_k$$

Now, we can extend this further by defining, for *any* function

$$f\colon \{\lambda_1, \ldots, \lambda_k\} \to F$$

the linear operator $f(\tau)$ by setting

$$f(\tau) = f(\lambda_1)\rho_1 + \cdots + f(\lambda_k)\rho_k$$

For example, we may define $\sqrt{\tau}$, τ^{-1}, e^{τ} and so on. Notice, however, that since the spectral resolution of τ is a *finite* sum, we gain nothing (but convenience) by using functions other than polynomials, for we can always find a polynomial $p(x)$ for which $p(\lambda_i) = f(\lambda_i)$ for $i = 1, \ldots, k$ and so

$$f(\tau) = f(\lambda_1)\rho_1 + \cdots + f(\lambda_k)\rho_k = p(\lambda_1)\rho_1 + \cdots + p(\lambda_k)\rho_k = p(\tau)$$

The study of the properties of functions of an operator τ is referred to as the **functional calculus** of τ.

According to the spectral theorem, if V is complex and τ is normal then $f(\tau)$ is a normal operator whose eigenvalues are $f(\lambda_i)$. Similarly, if V is real and τ is self-adjoint then $f(\tau)$ is self-adjoint, with eigenvalues $f(\lambda_i)$. Let us consider some special cases of this construction.

If $p_j(x)$ is a polynomial for which

$$p_j(\lambda_i) = \delta_{i,j}$$

for $i = 1, \ldots, k$, then

$$p_j(\tau) = \rho_j$$

and so we see that each projection ρ_j in the spectral resolution is a polynomial function of τ.

If τ is invertible then $\lambda_i \neq 0$ for all i and so we may take $f(x) = x^{-1}$, giving

$$\tau^{-1} = \lambda_1^{-1}\rho_1 + \cdots + \lambda_k^{-1}\rho_k$$

as can easily be verified by direct calculation.

If $f(\lambda_i) = \overline{\lambda}_i$ and if τ is normal then each ρ_i is self-adjoint and so

$$f(\tau) = \overline{\lambda}_1\rho_1 + \cdots + \overline{\lambda}_k\rho_k = \tau^*$$

Commutativity

The functional calculus can be applied to the study of the commutativity properties of operators. Here are two simple examples.

Theorem 10.23 *Let V be a finite-dimensional complex inner product space. For $\tau, \sigma \in \mathcal{L}(V)$, let us write $\tau \sim \sigma$ to denote the fact that τ and σ commute. Let τ and σ have spectral resolutions*

$$\begin{aligned}\tau &= \lambda_1\rho_1 + \cdots + \lambda_k\rho_k \\ \sigma &= \mu_1\nu_1 + \cdots + \mu_m\nu_m\end{aligned}$$

Then

1) *For any $\phi \in \mathcal{L}(V)$, we have $\phi \sim \tau$ if and only if $\phi \sim \rho_i$ for all i.*
2) *$\tau \sim \sigma$ if and only if $\rho_i \sim \nu_j$, for all i, j.*
3) *If $f: \{\lambda_1, \ldots, \lambda_k\} \to F$ and $g: \{\mu_1, \ldots, \mu_m\} \to F$ are injective functions, then $f(\tau) \sim g(\sigma)$ if and only if $\tau \sim \sigma$.*

Proof. The proof is based on the fact that if α and β are operators then $\alpha \sim \beta$ implies that $p(\alpha) \sim q(\beta)$ for any polynomials $p(x)$ and $q(x)$, and hence $f(\alpha) \sim g(\beta)$ for any functions f and g.

For 1), it is clear that $\phi \sim \rho_i$ for all i implies that $\phi \sim \tau$. The converse follows from the fact that ρ_i is a polynomial in τ. Part 2) is similar. For part 3), $\tau \sim \sigma$ clearly implies $f(\tau) \sim g(\sigma)$. For the converse, let $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$. Since f is injective, the inverse function $f^{-1}: f(\Lambda) \to \Lambda$ is well-defined and

$f^{-1}(f(\tau)) = \tau$. Thus, τ is a function of $f(\tau)$. Similarly, σ is a function of $g(\sigma)$. it follows that $f(\tau) \sim g(\sigma)$ implies $\tau \sim \sigma$. □

Theorem 10.24 *Let V be a finite-dimensional complex inner product space and let τ and σ be normal operators on V. Then τ and σ commute if and only if they have the form*

$$\tau = p(r(\tau, \sigma))$$
$$\sigma = q(r(\tau, \sigma))$$

where $p(x), q(x)$ and $r(x, y)$ are polynomials.

Proof. If τ and σ are polynomials in θ then they clearly commute. For the converse, suppose that $\tau\sigma = \sigma\tau$ and let

$$\tau = \lambda_1 \rho_1 + \cdots + \lambda_k \rho_k$$

and

$$\sigma = \mu_1 \nu_1 + \cdots + \mu_m \nu_m$$

be the orthogonal spectral resolutions of τ and σ. Then according to Theorem 10.23, $\rho_i \nu_j = \nu_j \rho_i$. Now, let us choose any polynomial $r(x, y)$ with the property that $\alpha_{i,j} = r(\lambda_i, \mu_j)$ are distinct. Since each ρ_i and ν_j is self-adjoint, we may set $\theta = r(\tau, \sigma)$ and deduce (after some algebra) that

$$\theta = r(\tau, \sigma) = \sum_{i,j} \alpha_{i,j} \rho_i \nu_j$$

We also choose $p(x)$ and $q(x)$ so that $p(\alpha_{i,j}) = \lambda_i$ for all j and $q(\alpha_{i,j}) = \mu_j$ for all i. Then

$$p(\theta) = \sum_{i,j} p(\alpha_{i,j}) \rho_i \nu_j = \sum_{i,j} \lambda_i \rho_i \nu_j = (\sum_i \lambda_i \rho_i)(\sum_j \nu_j) = \sum_i \lambda_i \rho_i = \tau$$

and similarly, $q(\theta) = \sigma$. □

Positive Operators

One of the most important cases of the functional calculus is when $f(x) = \sqrt{x}$. First, we need some definitions. Recall that the quadratic form associated with a linear operator τ is

$$Q_\tau(v) = \langle \tau(v), v \rangle$$

Definition A self-adjoint linear operator $\tau \in \mathcal{L}(V)$ is

1) **positive** *if $Q_\tau(v) \geq 0$ for all $v \in V$*
2) **positive definite** *if $Q_\tau(v) > 0$ for all $v \neq 0$.* □

Theorem 10.25 *A self-adjoint operator τ on a finite-dimensional inner product space is*

1) positive if and only if all of its eigenvalues are nonnegative
2) positive definite if and only if all of its eigenvalues are positive.

Proof. If $Q_\tau(v) \geq 0$ and $\tau(v) = \lambda v$ then

$$0 \leq \langle \tau(v), v \rangle = \lambda \langle v, v \rangle$$

and so $\lambda \geq 0$. Conversely, if all eigenvalues of τ are nonnegative then we have

$$\tau = \lambda_1 \rho_1 + \cdots + \lambda_k \rho_k, \ \lambda_i \geq 0$$

and since $\iota = \rho_1 + \cdots + \rho_k$,

$$\langle \tau(v), v \rangle = \sum_{i,j} \lambda_i \langle \rho_i(v), \rho_j(v) \rangle = \sum_i \lambda_i \|\rho_i(v)\|^2 \geq 0$$

and so τ is positive. Part 2) is proved similarly. □

If τ is a positive operator, with spectral resolution

$$\tau = \lambda_1 \rho_1 + \cdots + \lambda_k \rho_k, \ \lambda_i \geq 0$$

then we may take the **positive square root** of τ,

$$\sqrt{\tau} = \sqrt{\lambda_1} \rho_1 + \cdots + \sqrt{\lambda_k} \rho_k$$

where $\sqrt{\lambda_i}$ is the nonnegative square root of λ_i.

It is clear that

$$(\sqrt{\tau})^2 = \tau$$

and it is not hard to see that $\sqrt{\tau}$ is the only positive operator whose square is τ. In other words, every positive operator has a unique positive square root. Conversely, if τ has a positive square root, that is, if $\tau = \sigma^2$, for some positive operator σ then τ is positive. Hence, an operator τ is positive if and only if it has a positive square root.

If τ is positive then $\sqrt{\tau}$ is self-adjoint and so

$$(\sqrt{\tau})^* \sqrt{\tau} = \tau$$

Conversely, if $\tau = \sigma^* \sigma$ for some operator σ then τ is positive, since it is clearly self-adjoint and

$$\langle \tau(v), v \rangle = \langle \sigma^* \sigma(v), v \rangle = \langle \sigma(v), \sigma(v) \rangle \geq 0$$

Thus, τ is positive if and only if it has the form $\tau = \sigma^* \sigma$ for some operator σ.

Here is an application of square roots.

Theorem 10.26 *If τ and σ are positive operators and $\tau\sigma = \sigma\tau$ then $\tau\sigma$ is positive.*

Proof. Since τ is a positive operator, it has a positive square root $\sqrt{\tau}$, which is a polynomial in τ. A similar statement holds for σ. Therefore, since τ and σ commute, so do $\sqrt{\tau}$ and $\sqrt{\sigma}$. Hence,

$$(\sqrt{\tau}\sqrt{\sigma})^2 = (\sqrt{\tau})^2(\sqrt{\sigma})^2 = \tau\sigma$$

Since $\sqrt{\tau}$ and $\sqrt{\sigma}$ are self-adjoint and commute, their product is self-adjoint and so $\tau\sigma$ is positive. □

The Polar Decomposition of an Operator

It is well known that any nonzero complex number z can be written in the *polar form* $z = re^{i\theta}$, where r is a positive number and θ is real. We can do the same for any nonzero linear operator τ on a finite-dimensional complex inner product space.

Theorem 10.27 *Let τ be a nonzero linear operator on a finite-dimensional complex inner product space V. Then*

1) *There exists a positive operator ρ and a unitary operator ν for which $\tau = \nu\rho$. Moreover, ρ is unique and if τ is invertible then ν is also unique.*
2) *There exists a positive operator σ and a unitary operator μ for which $\tau = \sigma\mu$. Moreover, σ is unique and if τ is invertible then μ is also unique.*

Proof. Let us suppose for a moment that $\tau = \nu\rho$. Then

$$\tau^* = (\nu\rho)^* = \rho^*\nu^* = \rho\nu^{-1}$$

and so

$$\tau^*\tau = \rho\nu^{-1}\nu\rho = \rho^2$$

Also, if $v \in V$ then

$$\tau(v) = \nu(\rho(v))$$

These equations give us a clue as to how to define ρ and ν.

Let us define ρ to be the unique positive square root of the positive operator $\tau^*\tau$. Then

$$\|\rho(v)\|^2 = \langle\rho(v), \rho(v)\rangle = \langle\rho^2(v), v\rangle = \langle\tau^*\tau(v), v\rangle = \|\tau(v)\|^2 \qquad (10.3)$$

Let us define ν on $\mathrm{im}(\rho)$ by

$$\nu(\rho(v)) = \tau(v) \qquad (10.4)$$

for all $v \in V$. Equation (10.3) shows that $\rho(x) = \rho(y)$ implies that $\tau(x) = \tau(y)$ and so this definition of ν on $\mathrm{im}(\rho)$ is well-defined.

Moreover, ν is an isometry on $\mathrm{im}(\rho)$, since (10.3) gives

$$\|\nu(\rho(v))\| = \|\tau(v)\| = \|\rho(v)\|$$

Thus, if $\mathcal{B} = \{b_1, \ldots, b_k\}$ is an orthonormal basis for $\mathrm{im}(\rho)$, then $\nu(\mathcal{B}) = \{\nu(b_1), \ldots, \nu(b_k)\}$ is an orthonormal basis for $\nu(\mathrm{im}(\rho)) = \mathrm{im}(\tau)$. Finally, we may extend both orthonormal bases to orthonormal bases for V and then extend the definition of ν to an isometry on V, for which $\tau = \nu\rho$.

As for the uniqueness, we have seen that ρ must satisfy $\rho^2 = \tau^*\tau$ and since ρ^2 has a unique positive square root, we deduce that ρ is uniquely defined. Finally, if τ is invertible then so is ρ since $\ker(\rho) \subseteq \ker(\tau)$. Hence, $\nu = \tau\rho^{-1}$ is uniquely determined by τ.

Part 2) can be proved by applying the previous theorem to the map τ^*, to get

$$\tau = (\tau^*)^* = (\nu\rho)^* = \rho\nu^{-1} = \rho\mu$$

where μ is unitary. □

We leave it as an exercise to show that any unitary operator μ has the form $\mu = e^{i\sigma}$, where σ is a self-adjoint operator. This gives the following corollary.

Corollary 10.27 (**Polar decomposition**) Let τ be a nonzero linear operator on a finite-dimensional complex inner product space. Then there is a positive operator ρ and a self-adjoint operator σ for which τ has the **polar decomposition**

$$\tau = \rho e^{i\sigma}$$

Moreover, ρ is unique and if τ is invertible then σ is also unique. □

Normal operators can be characterized using the polar decomposition.

Theorem 10.28 *Let $\tau = \rho e^{i\sigma}$ be a polar decomposition of a nonzero linear operator τ. Then τ is normal if and only if $\rho\sigma = \sigma\rho$.*

Proof. Since

$$\tau\tau^* = \rho e^{i\sigma} e^{-i\sigma}\rho = \rho^2$$

and

$$\tau^*\tau = e^{-i\sigma}\rho\rho e^{i\sigma} = e^{-i\sigma}\rho^2 e^{i\sigma}$$

we see that τ is normal if and only if

$$e^{-i\sigma}\rho^2 e^{i\sigma} = \rho^2$$

or equivalently,

$$\rho^2 e^{i\sigma} = e^{i\sigma}\rho^2 \tag{10.5}$$

Now, ρ is a polynomial in ρ^2 and σ is a polynomial in $e^{i\sigma}$ and so (10.5) holds if and only if $\rho\sigma = \sigma\rho$. □

Exercises

1. Let $\tau \in \mathcal{L}(U,V)$. If τ is surjective, find a formula for the right inverse of τ in terms of τ^*. If τ is injective, find a formula for the left inverse of τ in terms of τ^*. *Hint*: Consider $\ker(\tau\tau^*)$ and $\ker(\tau^*\tau)$.
2. Let $\tau \in \mathcal{L}(V)$ where V is a complex vector space and let

   $$\tau_1 = \frac{1}{2}(\tau + \tau^*) \text{ and } \tau_2 = \frac{1}{2i}(\tau - \tau^*)$$

   Show that τ_1 and τ_2 are self-adjoint and that

   $$\tau = \tau_1 + i\tau_2 \text{ and } \tau^* = \tau_1 - i\tau_2$$

   What can you say about the uniqueness of these representations of τ and τ^*?
3. Prove that all of the roots of the characteristic polynomial of a skew-Hermitian matrix are pure imaginary.
4. Give an example of a normal operator that is neither self-adjoint nor unitary.
5. Prove that if $\|\tau(v)\| = \|\tau^*(v)\|$ for all $v \in V$, where V is complex then τ is normal.
6. a) Show that if τ is a normal operator on a finite-dimensional inner product space then $\tau^* = p(\tau)$, for some polynomial $p(x) \in F[x]$.
   b) Show that if τ is normal and $\sigma\tau = \tau\sigma$ then $\sigma\tau^* = \tau^*\sigma$. In other words, τ^* commutes with all operators that commute with τ.
7. Show that a linear operator τ on a finite-dimensional complex inner product space V is normal if and only if whenever S is an invariant subspace under τ, so is $S^\perp$.
8. Let V be a finite-dimensional inner product space and let τ be a normal operator on V.
   a) Prove that if τ is idempotent then it is also self-adjoint.
   b) Prove that if τ is nilpotent then $\tau = 0$.
   c) Prove that if $\tau^2 = \tau^3$ then τ is idempotent.
9. Show that if τ is a normal operator on a finite-dimensional complex inner product space then the algebraic multiplicity is equal to the geometric multiplicity for all eigenvalues of τ.
10. Show that two orthogonal projections σ and ρ are orthogonal to each other if and only if $\operatorname{im}(\sigma) \perp \operatorname{im}(\rho)$.
11. Let τ be a normal operator and let σ be any operator on V. If the eigenspaces of τ are σ-invariant, show that τ and σ commute.

12. Prove that if τ and σ are normal operators on a finite-dimensional inner complex product space and if $\tau\theta = \theta\sigma$ for some operator θ then $\tau^*\theta = \theta\sigma^*$.
13. Prove that if two normal $n \times n$ complex matrices are similar then they are *unitarily similar*, that is, similar via a unitary matrix.
14. If ν is a unitary operator on a complex inner product space, show that there exists a self-adjoint operator σ for which $\nu = e^{i\sigma}$.
15. Show that a positive operator has a unique positive square root.
16. Let α_i, β_i be complex numbers, for $i = 1, \ldots, k$. Construct a polynomial $p(x)$ for which $p(\alpha_i) = \beta_i$ for all i.
17. Prove that if τ has a square root, that is, if $\tau = \sigma^2$, for some positive operator σ then τ is positive.
18. Prove that if $\sigma \le \tau$ and if θ is a positive operator that commutes with both σ and τ then $\sigma\theta \le \tau\theta$.
19. Does every self-adjoint operator on a finite-dimensional real inner product space have a square root?
20. Let τ be a linear operator on $\mathbb{C}^n$ and let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of τ, each one written a number of times equal to its algebraic multiplicity. Show that
$$\sum_i |\lambda_i|^2 \le \operatorname{tr}(\tau^*\tau)$$
where tr is the trace. Show also that equality holds if and only if τ is normal.
21. If $\tau \in \mathcal{L}(V)$ where V is a real inner product space, show that the Hilbert space adjoint satisfies $(\tau^*)^{\mathbb{C}} = (\tau^{\mathbb{C}})^*$.

Part II—Topics

Chapter 11
Metric Vector Spaces: The Theory of Bilinear Forms

In this chapter, we study vector spaces over arbitrary fields that have a bilinear form defined upon them.

Unless otherwise mentioned, all vector spaces are assumed to be finite-dimensional. The symbol F denotes an arbitrary field and F_q denotes a finite field of size q.

Symmetric, Skew-Symmetric and Alternate Forms

We begin with the basic definition.

Definition *Let V be a vector space over F. A mapping $\langle,\rangle\colon V \times V \to F$ is called a* **bilinear form** *if it is a linear function of each coordinate, that is, if*

$$\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$$

and

$$\langle z, \alpha x + \beta y\rangle = \alpha\langle z, x\rangle + \beta\langle z, y\rangle$$

A bilinear form is

1) **symmetric** *if*

$$\langle x, y\rangle = \langle y, x\rangle$$

for all $x,\ y \in V$.

2) **skew-symmetric** *(or* **antisymmetric***) if*

$$\langle x, y\rangle = -\langle y, x\rangle$$

for all $x, y \in V$.

3) **alternate** *(or* **alternating***) if*

$$\langle x, x\rangle = 0$$

for all $x \in V$.

A bilinear form that is either symmetric, skew-symmetric, or alternate is referred to as an **inner product** *and a pair $(V, \langle , \rangle)$, where V is a vector space and $\langle , \rangle$ is an inner product on V, is called a* **metric vector space** *or* **inner product space***. If $\langle , \rangle$ is symmetric then $(V, \langle , \rangle)$ (or just V) is called an* **orthogonal geometry** *over F and if $\langle , \rangle$ is alternate then $(V, \langle , \rangle)$ (or just V) is called a* **symplectic geometry** *over F.* □

As an aside, the term *symplectic*, from the Greek for "intertwined" was introduced in 1939 by the famous mathematician Hermann Weyl in his book *The Classical Groups*, as a substitute for the term *complex*. According to the dictionary, symplectic means "relating to or being an intergrowth of two different minerals." An example is *ophicalcite*, which is marble spotted with green serpentine.

Example 11.1 Minkowski space M_4 is the four-dimensional real orthogonal geometry $\mathbb{R}^4$ with inner product defined by

$$\begin{aligned}
&\langle e_1, e_1\rangle = \langle e_2, e_2\rangle = \langle e_3, e_3\rangle = 1\\
&\langle e_4, e_4\rangle = -1\\
&\langle e_i, e_j\rangle = 0 \text{ for } i \neq j
\end{aligned}$$

where $e_1, \ldots, e_4$ is the standard basis for $\mathbb{R}^4$. □

As is traditional, when the inner product is understood, we will use the phrase "let V be a metric vector space."

The real inner products discussed in Chapter 9 are inner products in the present sense and have the additional property of being *positive definite*—a notion that does not even make sense if the base field is not ordered. Thus, a real inner product space is an orthogonal geometry. On the other hand, the complex inner products of Chapter 9, being sesquilinear, are not inner products in the present sense. For this reason, we prefer to use the term *metric vector space* rather than *inner product space*.

If S is a vector subspace of a metric vector space V then S inherits the metric structure from V. With this structure, we refer to S as a **subspace** of V.

The concepts of being symmetric, skew-symmetric and alternate are not independent. However, their relationship depends on the characteristic of the base field F, as do many other properties of metric vector spaces. In fact, the next theorem tells us that we do not need to consider skew-symmetric forms per se, since skew-symmetry is always equivalent to either symmetry or alternateness.

Theorem 11.1 *Let V be a vector space over a field F.*

1) *If* $\operatorname{char}(F)=2$ *then*

$$\textit{symmetric} \Leftrightarrow \textit{skew-symmetric}$$
$$\textit{alternate} \Rightarrow \textit{skew-symmetric}$$

2) *If* $\operatorname{char}(F)\neq 2$ *then*

$$\textit{alternate} \Leftrightarrow \textit{skew-symmetric}$$

Also, the only form that is both alternate and symmetric is the zero form: $\langle x,y\rangle=0$ *for all* $x,y\in V$.

Proof. First note that for any base field, if $\langle,\rangle$ is alternate then

$$0=\langle x+y,x+y\rangle=\langle x,x\rangle+\langle x,y\rangle+\langle y,x\rangle+\langle y,y\rangle=\langle x,y\rangle+\langle y,x\rangle$$

Thus,

$$\langle x,y\rangle+\langle y,x\rangle=0$$

or

$$\langle x,y\rangle=-\langle y,x\rangle$$

which shows that $\langle,\rangle$ is skew-symmetric. Thus, alternate always implies skew-symmetric.

If $\operatorname{char}(F)=2$ then $-1=1$ and so the definitions of symmetric and skew-symmetric are equivalent. This proves 1). If $\operatorname{char}(F)\neq 2$ and $\langle,\rangle$ is skew-symmetric, then for any $x\in V$, we have $\langle x,x\rangle=-\langle x,x\rangle$ or $2\langle x,x\rangle=0$, which implies that $\langle x,x\rangle=0$. Hence, $\langle,\rangle$ is alternate. Finally, if the form is alternate and symmetric, then it is also skew-symmetric and so $\langle u,v\rangle=-\langle u,v\rangle$ for all $u,v\in V$ and so $\langle u,v\rangle=0$ for all $u,v\in V$. □

Example 11.2 The standard inner product on $V(n,q)$, defined by

$$(x_1,\ldots,x_n)\cdot(y_1,\ldots,y_n)=x_1y_1+\cdots+x_ny_n$$

is symmetric, but not alternate, since

$$(1,0,\ldots,0)\cdot(1,0,\ldots,0)=1\neq 0$$ □

The Matrix of a Bilinear Form

If $\mathcal{B} = (b_1, \ldots, b_n)$ is an ordered basis for a metric vector space V then the form $\langle , \rangle$ is completely determined by the $n \times n$ matrix of values

$$M_{\mathcal{B}} = (a_{i,j}) = (\langle b_i, b_j \rangle)$$

This is referred to as the **matrix of the form** $\langle , \rangle$ with respect to the ordered basis $\mathcal{B}$. We also refer to $M_{\mathcal{B}}$ as the **matrix of** V with respect to $\mathcal{B}$ and write $M_{\mathcal{B}}(V)$ when the space needs emphasis.

Observe that multiplication of the coordinate matrix of a vector by $M_{\mathcal{B}}$ produces a vector of inner products, to wit, if $x = \Sigma r_i b_i$ then

$$M_{\mathcal{B}}[x]_{\mathcal{B}} = \begin{bmatrix} \langle b_1, x \rangle \\ \vdots \\ \langle b_n, x \rangle \end{bmatrix}$$

and

$$[x]_{\mathcal{B}}^t M_{\mathcal{B}} = (\, \langle x, b_1 \rangle \quad \cdots \quad \langle x, b_n \rangle \,)$$

It follows that if $y = \sum s_i b_i$ then

$$[x]_{\mathcal{B}}^t M_{\mathcal{B}}[y]_{\mathcal{B}} = (\, \langle x, b_1 \rangle \quad \cdots \quad \langle x, b_n \rangle \,) \begin{bmatrix} s_1 \\ \vdots \\ s_n \end{bmatrix} = \langle x, y \rangle$$

and this uniquely defines the matrix $M_{\mathcal{B}}$, that is, if $[x]_{\mathcal{B}}^t A[y]_{\mathcal{B}} = \langle x, y \rangle$ then $A = M_{\mathcal{B}}$.

Notice also that a form is symmetric if and only if the matrix $M_{\mathcal{B}}$ is symmetric, skew-symmetric if and only if $M_{\mathcal{B}}$ is skew-symmetric and alternate if and only if $M_{\mathcal{B}}$ is skew-symmetric and has 0's on the main diagonal. The latter type of matrix is referred to as **alternate**.

Now let us see how the matrix of a form behaves with respect to a change of basis. Let $\mathcal{C} = (c_1, \ldots, c_n)$ be an ordered basis for V. Recall from Chapter 2 that the change of basis matrix $M_{\mathcal{C},\mathcal{B}}$, whose ith column is $[c_i]_{\mathcal{B}}$, satisfies

$$[v]_{\mathcal{B}} = M_{\mathcal{C},\mathcal{B}}[v]_{\mathcal{C}}$$

Hence,

$$\begin{aligned} \langle x, y \rangle &= [x]_{\mathcal{B}}^t \, M_{\mathcal{B}}[y]_{\mathcal{B}} \\ &= ([x]_{\mathcal{C}}^t M_{\mathcal{C},\mathcal{B}}^t \;) M_{\mathcal{B}}(M_{\mathcal{C},\mathcal{B}}[y]_{\mathcal{C}} \;) \\ &= [x]_{\mathcal{C}}^t (M_{\mathcal{C},\mathcal{B}}^t \;\; M_{\mathcal{B}} M_{\mathcal{C},\mathcal{B}})[y]_{\mathcal{C}} \end{aligned}$$

and so

$$M_{\mathcal{C}} = M_{\mathcal{C},\mathcal{B}}^t \, M_{\mathcal{B}} M_{\mathcal{C},\mathcal{B}}$$

This prompts the following definition.

Definition Two matrices $A, B \in \mathcal{M}_n(F)$ are said to be **congruent** if there exists an invertible matrix P for which

$$A = P^t B P$$

The equivalence classes under congruence are called **congruence classes**. □

Let us summarize.

Theorem 11.2 *If the matrix of a bilinear form on V with respect to an ordered basis $\mathcal{B} = (b_1, \ldots, b_n)$ is*

$$M_{\mathcal{B}} = (\langle b_i, b_j \rangle)$$

then

$$\langle x, y \rangle = [x]_{\mathcal{B}}^t \, M_{\mathcal{B}} [y]_{\mathcal{B}}$$

Furthermore, if $\mathcal{C} = (c_1, \ldots, c_n)$ is an ordered basis for V then

$$M_{\mathcal{C}} = M_{\mathcal{C},\mathcal{B}}^t \, M_{\mathcal{B}} M_{\mathcal{C},\mathcal{B}}$$

where $M_{\mathcal{C},\mathcal{B}}$ is the change of basis matrix from $\mathcal{C}$ to $\mathcal{B}$. □

We have shown that if two matrices represent the same bilinear form on V, they must be congruent. Conversely, congruent matrices represent the same bilinear form on V. For suppose that $B = M_{\mathcal{B}}$ represents a bilinear form on V, with respect to the ordered basis $\mathcal{B}$ and that

$$A = P^t B P$$

where P is nonsingular. We saw in Chapter 2 that there is an ordered basis $\mathcal{C}$ for V with the property that

$$P = M_{\mathcal{C},\mathcal{B}}$$

and so

$$A = M_{\mathcal{C},\mathcal{B}}^t \, M_{\mathcal{B}} M_{\mathcal{C},\mathcal{B}}$$

Thus, $A = M_{\mathcal{C}}$ represents the same form with respect to $\mathcal{C}$.

Theorem 11.3 *Two matrices A and B represent a bilinear form $\langle , \rangle$ on V if and only if they are congruent, in which case they represent the same set of bilinear forms on V.* □

In view of the fact that congruent matrices have the same rank, we may define the **rank** of a bilinear form (or of V) to be the rank of any matrix that represents that form.

The Discriminant of a Form

If A and B are congruent matrices then

$$\det(A) = \det(P^t AP) = \det(P)^2 \det(B)$$

and so $\det(A)$ and $\det(B)$ differ by a square factor. The **discriminant** of a bilinear form is the set of all determinants of the matrices that represent the form under all choices of ordered bases. Thus, if $\det(A) = d$ for some matrix A representing the form then the discriminant of the form is the set

$$\Delta = F^2 d = \{r^2 d \mid 0 \neq r \in F\}$$

Quadratic Forms

There is a close link between symmetric bilinear forms and another important type of function defined on a vector space.

Definition *A* **quadratic form** *on a vector space V is a map $Q{:}\,V \to F$ with the following properties:*

1) For all $r \in F,\ v \in V$

$$Q(rv) = r^2 Q(v)$$

2) The map

$$\langle u, v \rangle_Q = Q(u+v) - Q(u) - Q(v)$$

is a (symmetric) bilinear form. □

Thus, every quadratic form Q defines a symmetric bilinear form $\langle u, v \rangle_Q$. On the other hand, if $\operatorname{char}(F) \neq 2$ and if $\langle , \rangle$ is a symmetric bilinear form on V then we can define a quadratic form Q by

$$Q(x) = \frac{1}{2}\langle x, x \rangle$$

We leave it to the reader to verify that this is indeed a quadratic form. Moreover, if Q is defined from a bilinear form in this way then the bilinear form associated with Q is

$$\begin{aligned}\langle u,v\rangle_Q &= Q(u+v)-Q(u)-Q(v)\\ &= \frac{1}{2}\langle u+v,u+v\rangle-\frac{1}{2}\langle u,u\rangle-\frac{1}{2}\langle v,v\rangle\\ &= \frac{1}{2}\langle u,v\rangle+\frac{1}{2}\langle v,u\rangle=\langle u,v\rangle\end{aligned}$$

which is the original bilinear form. In other words, the maps $\langle,\rangle\to Q$ and $Q\to\langle,\rangle_Q$ are inverses and so there is a one-to-one correspondence between symmetric bilinear forms on V and quadratic forms on V. Put another way, knowing the quadratic form is equivalent to knowing the corresponding bilinear form.

Again assuming that $\mathrm{char}(F)\neq 2$, if $\mathcal{B}=(v_1,\dots,v_n)$ is an ordered basis for an orthogonal geometry V and if the matrix of the symmetric form on V is $M_{\mathcal{B}}=(a_{i,j})$ then for $x=\Sigma x_iv_i$,

$$Q(x)=\frac{1}{2}\langle x,x\rangle=\frac{1}{2}[x]_{\mathcal{B}}^t\,M_{\mathcal{B}}[x]_{\mathcal{B}}=\sum_{i,j}\frac{1}{2}a_{i,j}x_ix_j$$

and so $Q(x)$ is a homogeneous polynomial of degree 2 in the coordinates x_i. (The term "form" means *homogeneous polynomial*—hence the term quadratic *form*.)

Orthogonality

As we will see, not all metric vector spaces behave as nicely as real inner product spaces and this necessitates the introduction of a new set of terminology to cover various types of behavior. (The base field F is the culprit, of course.) The most striking differences stem from the possibility that $\langle x,x\rangle=0$ for a nonzero vector $x\in V$.

The following terminology should be familiar.

Definition *A vector x is* **orthogonal** *to a vector y, written $x\perp y$, if $\langle x,y\rangle=0$. A vector $x\in V$ is* **orthogonal** *to a subset S of V, written $x\perp S$, if $\langle x,s\rangle=0$ for all $s\in S$. A subset S of V is* **orthogonal** *to a subset T of V, written $S\perp T$, if $\langle s,t\rangle=0$ for all $s\in S$ and $t\in T$. The* **orthogonal complement** *of a subset X of a metric vector space V, denoted by $X^\perp$, is the subspace*

$$X^\perp=\{v\in V\mid v\perp X\}\qquad\square$$

Note that regardless of whether the form is symmetric or alternate (and hence skew-symmetric), orthogonality is a symmetric relation, that is, $x\perp y$ implies $y\perp x$. Indeed, this is precisely why we restrict attention to these two types of bilinear forms. We will have more to say about this issue momentarily.

There are two types of degenerate behaviors that a vector may possess: It may be orthogonal to itself or, worse yet, it may be orthogonal to *every* vector in V. With respect to the former, we have the following terminology.

Definition *Let* $(V,\langle,\rangle)$ *be a metric vector space.*

1) *A nonzero* $x\in V$ *is* **isotropic** *(or* **null***) if* $\langle x,x\rangle=0$*; otherwise it is* **nonisotropic**.
2) V *is* **nonisotropic** *(also called* **anisotropic***) if it contains no isotropic vectors.*
3) V *is* **isotropic** *if it contains at least one isotropic vector.*
4) V *is* **totally isotropic** *(that is, symplectic) if all vectors in* V *are isotropic.* □

With respect to the latter (and more severe) form of degeneracy, we have the following terminology.

Definition *Let* $(V,\langle,\rangle)$ *be a metric vector space.*

1) *The set* $V^\perp$ *of all degenerate vectors is called the* **radical** *of* V *and written*
$$\mathrm{rad}(V)=V^\perp$$
2) V *is* **nonsingular***, or* **nondegenerate***, if* $\mathrm{rad}(V)=\{0\}$.
3) V *is* **singular***, or* **degenerate***, if* $\mathrm{rad}(V)\neq\{0\}$.
4) V *is* **totally singular** *or* **totally degenerate** *if* $\mathrm{rad}(V)=V$. □

Let us make a few remarks about these terms. Some of the above terminology is not entirely standard, so care should be exercised in reading the literature. Also, it is not hard to see that a metric vector space V is nonsingular if and only if the matrix $M_{\mathcal{B}}$ is nonsingular, for any ordered basis $\mathcal{B}$.

If v is an isotropic vector then so is av for all $a\in F$. This can be expressed by saying that the set I of isotropic vectors in V is a **cone** in V.

With respect to subspaces, to say that a subspace S of V is totally degenerate, for example, means that S is totally degenerate as a metric vector space in its own right and so each vector in S is orthogonal *to all other vectors in* S, not necessarily in V. In fact, we have
$$\mathrm{rad}(S)=S^{\perp_S}=S\cap S^{\perp_V}$$
where the symbols $\perp_S$ and $\perp_V$ refer to the orthogonal complements with respect to S and V, respectively.

Example 11.3 Recall that $V(n,q)$ is the set of all ordered n-tuples, whose components come from the finite field F_q. It is easy to see that the subspace

$$S=\{0000,1100,0011,1111\}$$

of $V(4,2)$ has the property that $S=S^{\perp}$. Note also that $V(4,2)$ is nonsingular and yet the subspace S is *totally* singular. □

The following result explains why we restrict attention to symmetric or alternate forms (which includes skew-symmetric forms).

Theorem 11.4 *Let $\langle,\rangle$ be a bilinear form on V. Then orthogonality is a symmetric relation, that is,*

$$x\perp y \;\Rightarrow\; y\perp x \tag{11.1}$$

if and only if $\langle,\rangle$ is either symmetric or alternate, that is, if and only if V is a metric vector space.

Proof. It is clear that (11.1) holds if $\langle,\rangle$ is symmetric. If $\langle,\rangle$ is alternate then it is skew-symmetric and so (11.1) also holds. For the converse, assume that (11.1) holds.

Let us introduce the notation $x\bowtie y$ to mean that $\langle x,y\rangle=\langle y,x\rangle$ and the notation $x\bowtie V$ to mean that $\langle x,v\rangle=\langle v,x\rangle$ for all $v\in V$. For $x,y,z\in V$, let

$$w=\langle x,y\rangle z-\langle x,z\rangle y$$

Then $x\perp w$ and so by (11.1) we have $w\perp x$, that is,

$$\langle x,y\rangle\langle z,x\rangle-\langle x,z\rangle\langle y,x\rangle=0$$

From this we deduce that

$$\begin{aligned} x\bowtie y &\Rightarrow \langle x,y\rangle(\langle z,x\rangle-\langle x,z\rangle)=0\\ &\Rightarrow \langle x,y\rangle=0 \text{ or } x\bowtie z \text{ for all } z\in V \end{aligned}$$

It follows that

$$x\bowtie y\Rightarrow x\perp y \text{ or } (x\bowtie V \text{ and } y\bowtie V) \tag{11.2}$$

Of course, $x\bowtie x$ and so for all $x\in V$

$$x\perp x \text{ or } x\bowtie V \tag{11.3}$$

Now we can show that if V is not symmetric, it must be alternate. If V is not symmetric, there exist $u,v\in V$ for which $u\not\bowtie v$. Thus, $u\not\bowtie V$ and $v\not\bowtie V$, which implies by (11.3) that u and v are both isotropic. We wish to show that all vectors in V are isotropic.

According to (11.3), if $w \not\bowtie V$, then w is isotropic. On the other hand, if $w \bowtie V$ then to see that w is also isotropic, we use the fact that if x and y are *orthogonal* isotropic vectors, then $x - y$ is also isotropic.

In particular, consider the vectors $w + u$ and u. We have seen that u is isotropic. The fact that $w \bowtie V$ implies $w \bowtie u$ and $w \bowtie v$ and so (11.2) gives $w \perp u$ and $w \perp v$. Hence, $(w + u) \perp u$. To see that $w + u$ is isotropic, note that

$$\langle w + u, v\rangle = \langle u, v\rangle \neq \langle v, u\rangle = \langle v, w + u\rangle$$

and so $(w + u) \not\bowtie V$. Hence, (11.3) implies that $w + u$ is isotropic. Thus, u and $w + u$ are orthogonal isotropic vectors, and so $w = (w + u) - u$ is also isotropic. □

Linear Functionals

Recall that the Riesz representation theorem says that for any linear functional f on a finite-dimensional real or complex inner product space V, there is a vector $R_f \in V$, which we called the *Riesz vector* for f, that represents f, in the sense that

$$f(v) = \langle v, R_f\rangle$$

for all $v \in V$. A similar result holds for *nonsingular* metric vector spaces.

Let V be a metric vector space over F. Let $x \in V$ and consider the "inner product on the right" map $\phi_x: V \to F$ defined by

$$\phi_x(v) = \langle v, x\rangle$$

This is easily seen to be a linear functional and so we can define a function $\tau: V \to V^*$ by

$$\tau(x) = \phi_x$$

The bilinearity of the form insures that τ is linear and the kernel of τ is

$$\ker(\tau) = \{x \in V \mid \phi_x = 0\} = \{x \in V \mid \langle v, x\rangle = 0 \text{ for all } v \in V\} = V^\perp$$

Hence, if V is nonsingular then $\ker(\tau) = V^\perp = \{0\}$ and so τ is injective. Moreover, since $\dim(V) = \dim(V^*)$, it follows that τ is surjective and so τ is an isomorphism from V onto V^*. This implies that every linear functional on V has the form ϕ_x, for a unique $x \in V$. We have proved the Riesz representation theorem for finite-dimensional nonsingular metric vector spaces.

Theorem 11.5 *(***The Riesz representation theorem***) Let V be a finite-dimensional nonsingular metric vector space. The linear functional $\tau: V \to V^*$ defined by*

$$\tau(x) = \phi_x$$

where $\phi_x(v) = \langle v, x\rangle$ for all $v \in V$, is an isomorphism from V to V^*. It follows that for each $f \in V^*$ there exists a unique vector $x \in V$ for which $f = \phi_x$, that is,

$$f(v) = \langle v, x\rangle$$

for all $v \in V$. □

The requirement that V be nonsingular is necessary. As a simple example, if V is totally singular, then no nonzero linear functional could possibly be represented by an inner product.

We would like to extend the Riesz representation theorem to the case of subspaces of a metric vector space. The Riesz representation theorem applies to nonsingular metric vector spaces. Thus, if S is a nonsingular subspace of V, the Riesz representation theorem applies to S and so all linear functionals on S have the form of an inner product by a (unique) element of S. This is nothing new.

As long as V is nonsingular, even if S is singular, we can still say something very useful. The reason is that any linear functional $f \in S^*$ can be extended to a linear functional g on V (perhaps in many ways) and since V is nonsingular, the extension g has the form of inner product by a vector in V, that is,

$$g(v) = \langle v, x\rangle$$

for some $x \in V$. Hence, f also has this form, where its "Riesz vector" is an element of V, not necessarily S. Here is the formal statement.

Theorem 11.6 *Let V be a metric vector space and let S be a subspace of V. If either V or S is nonsingular, the linear transformation $\tau\colon V \to S^*$ defined by*

$$\tau(x) = \phi_x|_S$$

where $\phi_x(v) = \langle v, x\rangle$, is surjective. Hence, for any linear functional $f \in S^*$ there is a (not necessarily unique) vector $x \in V$ for which $f(s) = \langle s, x\rangle$. Moreover, if S is nonsingular then x can be taken from S, in which case it is unique. □

Orthogonal Complements and Orthogonal Direct Sums

If S is a subspace of a real inner product space, then the projection theorem says that the orthogonal complement $S^\perp$ of S is a true vector space complement of S, that is,

$$V = S \odot S^\perp$$

Hence, the term orthogonal *complement* is justified. However, in general metric vector spaces, an orthogonal complement may not be a vector space complement. In fact, Example 11.3 shows that we may have the opposite extreme, that is, $S^{\perp} = S$. As we will see, the orthogonal complement of S is a true complement if and only if S is nonsingular.

Definition *A metric vector space V is the* **orthogonal direct sum** *of the subspaces S and T, written*

$$V = S \odot T$$

if $V = S \oplus T$ and $S \perp T$. □

In a real inner product space, if $V = S \odot T$ then $T = S^{\perp}$. However, in a metric vector space in general, we may have a proper inclusion $T \subset S^{\perp}$. (In fact, $S^{\perp}$ may be all of V.)

Many nice properties of orthogonality in real inner product spaces do carry over to *nonsingular* metric vector spaces. The next result shows that the restriction to nonsingular spaces is not that severe.

Theorem 11.7 *Let V be a metric vector space. Then*

$$V = \text{rad}(V) \odot S$$

where S is nonsingular and $\text{rad}(V)$ *is totally singular.*

Proof. Ignoring the metric structure for a moment, we know that all subspaces of a vector space, including $\text{rad}(V)$, have a complement, say $V = \text{rad}(V) \oplus S$. But $\text{rad}(V) \perp S$ and so $V = \text{rad}(V) \odot S$. To see that S is nonsingular, if $x \in \text{rad}(S)$ then $x \perp S$ and so $x \perp V$, which implies that $x \in \text{rad}(V) \cap S = \{0\}$, that is, $x = 0$. Hence, $\text{rad}(S) = \{0\}$ and S is nonsingular. □

Under the assumption of nonsingularity of V, we get many nice properties, just short of the projection theorem. The first property in the next theorem is key: It says that if S is a subspace of a *nonsingular* space V, then the orthogonal complement of S always has the "correct" dimension, even if it is not well behaved with respect to its intersection with S, that is,

$$\dim(S^{\perp}) = \dim(V) - \dim(S)$$

just as in the case of a real inner product space.

Theorem 11.8 *Let V be a nonsingular metric vector space V and let S be any subspace of V. Then*

1) $\dim(S) + \dim(S^{\perp}) = \dim(V)$

2) if $V = S + S^{\perp}$ then $V = S \odot S^{\perp}$

3) $S^{\perp\perp} = S$

4) $\text{rad}(S) = \text{rad}(S^{\perp})$

5) S is nonsingular if and only if $S^{\perp}$ is nonsingular

Proof. For part 1), the map $\tau\colon V \to S^*$ of Theorem 11.6 is surjective and

$$\ker(\tau) = \{x \in V \mid \phi_x|_S = 0\} = \{x \in V \mid \langle s, x\rangle = 0 \text{ for all } s \in S\} = S^\perp$$

Thus, the rank-plus-nullity theorem implies that

$$\dim(S^*) + \dim(S^\perp) = \dim(V)$$

However, $\dim(S^*) = \dim(S)$ and so part 1) follows.

For part 2), we have using part 1)

$$\begin{aligned}\dim(V) &= \dim(S + S^\perp) \\ &= \dim(S) + \dim(S^\perp) - \dim(S \cap S^\perp) \\ &= \dim(V) - \dim(S \cap S^\perp)\end{aligned}$$

and so $S \cap S^\perp = \{0\}$.

For part 3), part 1) implies that

$$\dim(S) + \dim(S^\perp) = \dim(V)$$

and

$$\dim(S^\perp) + \dim(S^{\perp\perp}) = \dim(V)$$

and so $\dim(S^{\perp\perp}) = \dim(S)$. But $S \subseteq S^{\perp\perp}$ and so equality holds.

For part 4), we have

$$\text{rad}(S) = S \cap S^\perp = S^\perp \cap S^{\perp\perp} = \text{rad}(S^\perp)$$

and part 5) follows from part 4). □

The previous theorem cannot in general be strengthened. Consider the two-dimensional metric vector space $V = \text{span}(u, v)$ where

$$\langle u, u\rangle = 1, \langle u, v\rangle = 0, \langle v, v\rangle = 0$$

Let $S = \text{span}(u)$. Then $S^\perp = \text{span}(v)$. Now, S is nonsingular but $S^\perp$ is singular and so 5) does not hold. Also, $\text{rad}(S) = \{0\}$ and $\text{rad}(S^\perp) = S^\perp$ and so 4) fails. Finally, $S^{\perp\perp} = V \neq S$ and so 3) fails.

On the other hand, we should note that if S is singular, then so is $S^\perp$, regardless of whether V is singular or nonsingular. To see this, note that if S is singular then there is a nonzero $s \in \text{rad}(S) = S \cap S^\perp$. Hence, $s \in S \subseteq S^{\perp\perp}$ and so $s \in S^\perp \cap S^{\perp\perp} = \text{rad}(S^\perp)$, which implies that $S^\perp$ is singular.

Now let us state the projection theorem for arbitrary metric vector spaces.

Theorem 11.9 Let S be a subspace of a finite-dimensional metric vector space V. Then

$$V = S \odot S^{\perp}$$

if and only if S is nonsingular, that is, if and only if $S \cap S^{\perp} = \{0\}$.
Proof. If $V = S \odot S^{\perp}$ then by definition of orthogonal direct sum, we have

$$\text{rad}(S) = S \cap S^{\perp} = \{0\}$$

and so S is nonsingular. Conversely, if S is nonsingular, then $S \cap S^{\perp} = \{0\}$ and so $S \odot S^{\perp}$ exists. Now, the same proof used in part 1) of the previous theorem works if S is nonsingular (even if V is singular). To wit, the map $\tau: V \to S^*$ of Theorem 11.6 is surjective and

$$\ker(\tau) = \{x \in V \mid \phi_x|_S = 0\} = \{x \in V \mid \langle s, x\rangle = 0 \text{ for all } s \in S\} = S^{\perp}$$

Thus, the rank-plus-nullity theorem gives

$$\dim(S^*) + \dim(S^{\perp}) = \dim(V)$$

But $\dim(S^*) = \dim(S)$ and so

$$\dim(S \odot S^{\perp}) = \dim(S) + \dim(S^{\perp}) = \dim(V)$$

It follows that $V = S \odot S^{\perp}$. □

Isometries

We now turn to a discussion of structure-preserving maps on metric vector spaces.

Definition *Let V and W be metric vector spaces. We use the same notation $\langle , \rangle$ for the bilinear form on each space. A* bijective *linear map $\tau: V \to W$ is called an* **isometry** *if*

$$\langle \tau u, \tau v\rangle = \langle u, v\rangle$$

for all vectors u and v in V. If an isometry exists from V to W, we say that V and W are **isometric** *and write $V \approx W$. It is evident that the set of all isometries from V to V forms a group under composition.*

If V is a nonsingular orthogonal geometry, an isometry of V is called an **orthogonal transformation**. *The set $\mathcal{O}(V)$ of all orthogonal transformations on V is a group under composition, known as the* **orthogonal group** *of V.*

If V is a nonsingular symplectic geometry, an isometry of V is called a **symplectic transformation**. *The set* Sp(V) *of all symplectic transformations on V is a group under composition, known as the* **symplectic group** *of V.* □

Here are a few of the basic properties of isometries.

Theorem 11.10 *Let $\tau \in \mathcal{L}(V,W)$ be a linear transformation between finite-dimensional metric vector spaces V and W.*

1) *Let $\mathcal{B} = \{v_1,\dots,v_n\}$ be a basis for V. Then τ is an isometry if and only if τ is bijective and*
$$\langle \tau v_i, \tau v_j\rangle = \langle v_i, v_j\rangle$$
for all i,j.
2) *If V is orthogonal and* $\text{char}(F) \neq 2$ *then τ is an isometry if and only if it is bijective and*
$$\langle \tau(v), \tau(v)\rangle = \langle v, v\rangle$$
for all $v \in V$.
3) *Suppose that τ is an isometry and $V = S \odot S^{\perp}$ and $W = T \odot T^{\perp}$. If $\tau(S) = T$ then $\tau(S^{\perp}) = T^{\perp}$. In particular, if $\tau \in \mathcal{L}(V)$ is an isometry and $V = S \odot S^{\perp}$ then if S is τ-invariant, so is $S^{\perp}$.*

Proof. We prove part 3) only. To see that $\tau(S^{\perp}) = T^{\perp}$, if $z \in S^{\perp}$ and $t \in T$ then since $T = \tau(S)$, we can write $t = \tau(s)$ for some $s \in S$ and so
$$\langle \tau(z), t\rangle = \langle \tau(z), \tau(s)\rangle = \langle z, s\rangle = 0$$
whence $\tau(S^{\perp}) \subseteq T^{\perp}$. But since $\dim(\tau(S^{\perp})) = \dim(T^{\perp})$, we deduce that $\tau(S^{\perp}) = T^{\perp}$. □

Hyperbolic Spaces

A special type of two-dimensional metric vector space plays an important role in the structure theory of metric vector spaces.

Definition Let V be a metric vector space. If $u, v \in V$ have the property that
$$\langle u,u\rangle = \langle v,v\rangle = 0,\ \langle u,v\rangle = 1$$
the ordered pair (u,v) is called a **hyperbolic pair**. Note that $\langle v,u\rangle = 1$ if V is an orthogonal geometry and $\langle v,u\rangle = -1$ if V is symplectic. In either case, the subspace $H = \text{span}(u,v)$ is called a **hyperbolic plane** and any space of the form
$$\mathcal{H} = H_1 \odot \cdots \odot H_k$$
where each H_i is a hyperbolic plane, is called a **hyperbolic space**. If (u_i, v_i) is a hyperbolic pair for H_i then we refer to the basis $(u_1, v_1, \dots, u_k, v_k)$ for $\mathcal{H}$ as a **hyperbolic basis**. (In the symplectic case, the usual term is **symplectic basis**.)□

Note that any hyperbolic space $\mathcal{H}$ is nonsingular.

In the orthogonal case, hyperbolic planes can be characterized by their degree of isotropy, so-to-speak. (In the symplectic case, all spaces are totally isotropic by definition.) Indeed, we leave it as an exercise to prove that a two-dimensional nonsingular orthogonal geometry V is a hyperbolic plane if and only if V contains exactly two one-dimensional totally isotropic (equivalently: totally degenerate) subspaces. Put another way, the cone of isotropic vectors is the union of two one-dimensional subspaces of V.

Nonsingular Completions of a Subspace

Let U be a subspace of a nonsingular metric vector space V. If U is singular, it is of interest to find a *minimal* nonsingular extension of U, that is, minimal nonsingular subspace of V containing U. Such extensions of U are called **nonsingular completions** of U.

Theorem 11.11 *(***Nonsingular extension theorem***) Let V be a nonsingular metric vector space over F. We assume that* $\operatorname{char}(F) \neq 2$ *when V is orthogonal.*

1) *Let S be a subspace of V. For each isotropic vector $v \in S^{\perp} \setminus S$, there is a hyperbolic plane $H = \operatorname{span}(v, z)$ contained in $S^{\perp}$. Hence, $H \odot S$ is an extension of S containing v.*
2) *Let U be a subspace of V and write $U = \operatorname{rad}(U) \odot W$ where W is nonsingular and $\{v_1, \ldots, v_k\}$ is a basis for $\operatorname{rad}(U)$. Then there is a hyperbolic space $H_1 \odot \cdots \odot H_k$ with hyperbolic basis $(v_1, z_1, \ldots, v_k, z_k)$ for which*

$$\overline{U} = H_1 \odot \cdots \odot H_k \odot W$$

is a nonsingular extension of U, called a **nonsingular completion** *of U.*

Proof. For part 1), the nonsingularity of V implies that $S^{\perp\perp} = S$ and so $v \notin S$ is equivalent to $v \notin S^{\perp\perp}$. Hence, there is a vector $x \in S^{\perp}$ for which $\langle v, x \rangle \neq 0$. If V is symplectic then we can take $z = (1/\langle v, x \rangle)x$. If V is orthogonal, let $z = rv + sx$. The conditions defining (v, z) as a hyperbolic pair are (since v is isotropic)

$$1 = \langle v, z \rangle = \langle v, rv + sx \rangle = s\langle v, x \rangle$$

and

$$0 = \langle z, z \rangle = \langle rv + sx, rv + sx \rangle = 2rs\langle v, x \rangle + s^2\langle x, x \rangle$$

Since $\langle v, x \rangle \neq 0$, the first of these equations can be solved for s and since $\operatorname{char}(F) \neq 2$, the second can then be solved for r. Thus, in either case, a vector $z \in S^{\perp}$ exists for which (v, z) is a hyperbolic pair and $\operatorname{span}(v, z) \subseteq S^{\perp}$.

For part 2), we proceed by indution on $k = \dim(\operatorname{rad}(U))$. If $k = 1$ then v_1 is isotropic and $v_1 \in W^{\perp} \setminus W$. Hence, part 1) applied to $S = W$ implies that there is a hyperbolic plane $H = \text{span}(v_1, z)$ for which $H \odot W$ is an extension of W containing $\text{span}(v_1) = \text{rad}(U)$. Hence, part 2) holds when $k = 1$.

Let us assume that the result is true when $\dim(\text{rad}(U)) < k$ and assume that $\dim(\text{rad}(U)) = k$. Let

$$U_1 = \text{span}(v_2, \ldots, v_k) \odot W$$

Then v_1 is (still) isotropic and $v_1 \in U_1^\perp \setminus U_1$. Hence, we may apply part 1) to the subspace $S = U_1$ to deduce the existence of a hyperbolic plane $H_1 = \text{span}(v_1, z_1)$ contained in $U_1^\perp$.

Now, since H_1 is nonsingular, we have

$$V = H_1 \odot H_1^\perp$$

Since $H_1 \subseteq U_1^\perp$, it follows that $U_1 = U_1^{\perp\perp} \subseteq H_1^\perp$. Thus, we may apply the induction hypothesis to U_1 as a subspace of the nonsingular $H_1^\perp$, giving a space

$$H_2 \odot \cdots \odot H_k \odot W \subseteq H_1^\perp$$

containing U_1. It follows that $H_1 \odot H_2 \odot \cdots \odot H_k \odot W$ is the desired extension of U. □

Note that if $\overline{U}$ is a nonsingular extension of U then

$$\dim(\overline{U}) = \dim(U) + \dim(\text{rad}(U))$$

Theorem 11.12 *Let V be a nonsingular metric vector space and let U be a subspace of V. The following are equivalent:*
1) W is a nonsingular completion of U
2) W is a minimal nonsingular extension of U
3) W is a nonsingular extension of U and

$$\dim(W) = \dim(U) + \dim(\text{rad}(U))$$

Moreover, any two nonsingular completions of U are isomorphic.
Proof. If 1) holds and if $U \subseteq X \subseteq W$ where X is nonsingular, then we may apply the nonsingular extension theorem to U as a subspace of X, to obtain a nonsingular extension U' of U for which

$$U \subseteq U' \subseteq X \subseteq W$$

But U' and W have the same dimension and so must be equal. Hence, $X = W$. Thus W is a minimal nonsingular extention of U and 2) holds. If 2) holds then we have $U \subseteq \overline{U} \subseteq W$ where $\overline{U}$ is a nonsingular completion of U. But the minimality of W implies that $\overline{U} = W$ and so 3) holds. If 3) holds then again we have $U \subseteq \overline{U} \subseteq W$. But $\dim(\overline{U}) = \dim(W)$ and so $W = \overline{U}$ is a nonsingular completion of U and 1) holds.

If $X = \mathcal{H} \odot W$ and $Y = \mathcal{H}' \odot W$ are nonsingular completions of $U = W \odot \text{rad}(U)$ then $\mathcal{H}$ and $\mathcal{H}'$ are hyperbolic spaces of the same dimension and are therefore isometric. It follows that X and Y are isometric. □

Extending Isometries to Nonsingular Completions

Let V and V' be isometric nonsingular metric vector spaces and let $U = W \odot \text{rad}(U)$ be a subspace of V, with nonsingular completion $\overline{U}$. If $\tau: U \to \tau(U)$ is an isometry, then it is a simple matter to extend τ to an isometry $\overline{\tau}$ from $\overline{U}$ onto a nonsingular completion of $\tau(U)$. To see this, let

$$\overline{U} = \mathcal{H} \odot W$$

where W is nonsingular and $(u_1, v_1, \ldots, u_k, v_k)$ is a hyperbolic basis for $\mathcal{H}$. Since $(u_1, \ldots, u_k)$ is a basis for rad(U), it follows that $(\tau(u_1), \ldots, \tau(u_k))$ is a basis for rad$(\tau(U))$.

Now we complete $\tau(U) = \tau(W) \odot \text{rad}(\tau(W))$ to get

$$\tau(\overline{U}) = \mathcal{H}' \odot \tau(W)$$

where $\mathcal{H}'$ has hyperbolic basis $(\tau(u_1), z_1, \ldots, \tau(u_k), z_k)$. To extend τ, simply set $\overline{\tau}(v_i) = z_i$ for all $i = 1, \ldots, k$.

Theorem 11.13 *Let V and V' be isometric nonsingular metric vector spaces and let U be a subspace of V, with nonsingular completion $\overline{U}$. Any isometry $\tau: U \to \tau(U)$ can be extended to an isometry from $\overline{U}$ onto a nonsingular completion of $\tau(U)$.* □

The Witt Theorems: A Preview

There are two important theorems that are quite easy to prove in the case of real inner product spaces, but require more work in the case of metric vector spaces in general. Let V and V' be nonsingular isometric metric vector spaces over a field F. We assume that char$(F) \neq 2$ if V is orthogonal.

The *Witt extension theorem* says that if S is a subspace of V and

$$\tau: S \to \tau(S) \subseteq V'$$

is an isometry, then τ can be extended to an isometry from V to V'. The *Witt cancellation theorem* says that if

$$V = S \odot S^{\perp} \quad \text{and} \quad V' = T \odot T^{\perp}$$

then

$$S \approx T \Rightarrow S^{\perp} \approx T^{\perp}$$

We will prove these theorems in both the orthogonal and symplectic cases later in the chapter. For now, we simply want to show that it is easy to prove one Witt theorem from the other.

Suppose that the Witt extension theorem holds and assume that

$$V = S \odot S^{\perp} \quad \text{and} \quad V' = T \odot T^{\perp}$$

and $S \approx T$. Then any isometry $\tau\colon S \to T$ can be extended to an isometry $\overline{\tau}$ from V to V'. According to Theorem 11.10, we have $\overline{\tau}(S^{\perp}) = T^{\perp}$ and so $S^{\perp} \approx T^{\perp}$. Hence, the Witt cancellation theorem holds.

Conversely, suppose that the Witt cancellation theorem holds and let $\tau\colon S \to \tau(S) \subseteq V'$ be an isometry. Then we may extend τ to a nonsingular completion of S. Hence, we may assume that S is nonsingular. Then

$$V = S \odot S^{\perp}$$

Since τ is an isometry, $\tau(S)$ is also nonsingular and we can write

$$V' = \tau(S) \odot \tau(S)^{\perp}$$

Since $S \approx \tau(S)$, Witt's cancellation theorem implies that $S^{\perp} \approx \tau(S)^{\perp}$. If $\mu\colon S^{\perp} \to \tau(S)^{\perp}$ is an isometry then the map $\sigma\colon V \to V'$ defined by

$$\sigma(u + v) = \tau(u) + \mu(v)$$

for $u \in S$ and $v \in S^{\perp}$ is an isometry that extends τ. Hence Witt's extension theorem holds.

The Classification Problem for Metric Vector Spaces

The **classification problem** for a class of metric vector spaces (such as the orthogonal or symplectic spaces) is the problem of determining when two metric vector spaces in the class are isometric. The classification problem is considered "solved," at least in a theoretical sense, by finding a set of canonical forms or a complete set of invariants for matrices under congruence.

To see why, suppose that $\tau\colon V \to W$ is an isometry and $\mathcal{B} = (v_1, \ldots, v_n)$ is an ordered basis for V. Then $\mathcal{C} = (\tau(v_1), \ldots, \tau(v_n))$ is an ordered basis for W and

$$M_{\mathcal{B}}(V) = (\langle v_i, v_j \rangle) = (\langle \tau(v_i), \tau(v_j) \rangle) = M_{\mathcal{C}}(W)$$

Thus, the congruence class of matrices representing V is identical to the congruence class of matrices representing W.

Conversely, suppose that V and W are metric vector spaces with the same congruence class of representing matrices. Then if $\mathcal{B} = (v_1, \ldots, v_n)$ is an ordered basis for V, there is an ordered basis $\mathcal{C} = (w_1, \ldots, w_n)$ for W for which

$$(\langle v_i, v_j \rangle) = M_{\mathcal{B}}(V) = M_{\mathcal{C}}(W) = (\langle w_i, w_j \rangle)$$

Hence, the map $\tau\colon V \to W$ defined by $\tau(v_i) = w_i$ is an isometry from V to W.

We have shown that two metric vector spaces are isometric if and only if they have the same congruence class of representing matrices. Thus, we can determine whether any two metric vector spaces are isometric by representing each space with a matrix and determining if these matrices are congruent, using a set of canonical forms or a set of complete invariants.

Symplectic Geometry

We now turn to a study of the structure of orthogonal and symplectic geometries and their isometries. Since the study of the structure (and the structure itself) of symplectic geometries is simpler than that of orthogonal geometries, we begin with the symplectic case. The reader who is interested only in the orthogonal case may omit this section.

Throughout this section, let V be a nonsingular symplectic geometry.

The Classification of Symplectic Geometries

Among the simplest types of metric vector spaces are those that possess an orthogonal basis, that is, a basis $\mathcal{B} = \{u_1, \ldots, u_n\}$ for which $\langle u_i, u_j \rangle = 0$ when $i \neq j$. For in this case, we may write V as an orthogonal direct sum of one-dimensional subspaces

$$V = \text{span}(u_1) \odot \cdots \odot \text{span}(u_n)$$

However, it is easy to see that a symplectic geometry V has an orthogonal basis if and only if it is totally degenerate. For if $\mathcal{B}$ is an orthogonal basis for V then $\langle u_i, u_j \rangle = 0$ for $i = j$ since the form is alternate and for $i \neq j$ since the basis is orthogonal. It follows that V is totally degenerate. Thus, no "interesting" symplectic geometries have orthogonal bases.

Thus, in searching for an orthogonal decomposition of V, we must look not at one-dimensional subspaces, but at two-dimensional subspaces. Given a nonzero $u \in V$, the nonsingularity of V implies that there must exist a vector $v \in V$ for which $\langle u, v \rangle = a \neq 0$. Replacing v by $a^{-1}v$, we have

$$\langle u, u \rangle = \langle v, v \rangle = 0, \quad \langle u, v \rangle = 1 \quad \text{and} \quad \langle v, u \rangle = -1$$

and so $H = \text{span}(u, v)$ is a hyperbolic plane and the matrix of H with respect to the hyperbolic pair $\mathcal{B} = (u, v)$ is

$$M_{\mathcal{B}} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

Since H is nonsingular, we can write

$$V = H \odot H^{\perp}$$

where $H^{\perp}$ is also nonsingular. Hence, we may repeat the preceding decomposition in $H^{\perp}$, eventually obtaining an orthogonal decomposition of V of the form

$$V = H_1 \odot H_2 \odot \cdots \odot H_k$$

where each H_i is a hyperbolic plane. This proves the following structure theorem for symplectic geometries.

Theorem 11.14

1) *A symplectic geometry has an orthogonal basis if and only if it is totally degenerate.*
2) *Any nonsingular symplectic geometry V is a hyperbolic space, that is,*

$$V = H_1 \odot H_2 \odot \cdots \odot H_k$$

where each H_i is a hyperbolic plane. Thus, there is a basis for V for which the matrix of the form is

$$Y_{2k} = \begin{bmatrix} 0 & 1 & & & & & \\ -1 & 0 & & & & & \\ & & 0 & 1 & & & \\ & & -1 & 0 & & & \\ & & & & \ddots & & \\ & & & & & 0 & 1 \\ & & & & & -1 & 0 \end{bmatrix}$$

In particular, the dimenison of V is even.

3) *Any symplectic geometry V has the form*

$$V = \operatorname{rad}(V) \odot \mathcal{H}$$

where $\mathcal{H}$ is a hyperbolic space and $\operatorname{rad}(V)$ *is a totally degenerate space. The rank of the form is* $\dim(\mathcal{H})$ *and V is uniquely determined up to isometry by its rank and its dimension. Put another way, up to isometry, there is precisely one symplectic geometry of each rank and dimension.* □

Symplectic forms are represented by alternate matrices, that is, skew-symmetric matrices with zero diagonal. Moreover, according to Theorem 11.14, each $n \times n$ alternate matrix is congruent to a matrix of the form

$$X_{2k,n-2k} = \begin{bmatrix} Y_{2k} & 0 \\ 0 & I_{n-2k} \end{bmatrix}_{\text{block}}$$

Since the rank of $X_{2k,n-2k}$ is $2k$, no two such matrices are congruent.

Theorem 11.15 *The set of $n \times n$ matrices of the form $X_{2k,n-2k}$ is a set of canonical forms for alternate matrices under congruence.* □

The previous theorems solve the classification problem for symplectic geometries by stating that the rank and dimension of V form a complete set of invariants under congruence and that the set of all matrices of the form $X_{2k,n-2k}$ is a set of canonical forms.

Witt's Extension and Cancellation Theorems

We now prove the Witt theorems for symplectic geometries.

Theorem 11.16 *(***Witt's extension theorem***) Let V and V' be nonsingular isometric symplectic geometries over a field F. Suppose that S is a subspace of V and*

$$\tau: S \to \tau(S) \subseteq V'$$

is an isometry. Then τ can be extended to an isometry from V to V'.
Proof. According to Theorem 11.13, we can extend τ to a nonsingular completion of S, so we may simply assume that S and $\tau(S)$ are nonsingular. Hence,

$$V = S \odot S^{\perp}$$

and

$$V' = \tau(S) \odot \tau(S)^{\perp}$$

To complete the extension of τ to V, we need only choose a hyperbolic basis

$$(e_1, f_1, \ldots, e_p, f_p)$$

for $S^{\perp}$ and a hyperbolic basis

$$(e_1', f_1', \ldots, e_p', f_p')$$

for $\tau(S)^{\perp}$ and define the extension by setting $\tau(e_i) = e_i'$ and $\tau(f_i) = f_i'$. □

As a corollary to Witt's extension theorem, we have Witt's cancellation theorem.

Theorem 11.17 *(***Witt's cancellation theorem***) Let V and V' be isometric nonsingular symplectic geometries over a field F. If*

$$V = S \odot S^{\perp} \quad \textit{and} \quad V' = T \odot T^{\perp}$$

then

$$S \approx T \Rightarrow S^{\perp} \approx T^{\perp}$$ □

The Structure of the Symplectic Group: Symplectic Transvections

To understand the nature of symplectic transformations on a nonsingular symplectic geometry V, we begin with the following definition.

Definition *Let V be a nonsingular symplectic geometry over F. Let $v \in V$ be nonzero and let $a \in F$. The map $\tau_{v,a}: V \to V$ defined by*

$$\tau_{v,a}(x) = x + a\langle x, v\rangle v$$

is called the **symplectic transvection** *determined by v and a.* □

The first thing to notice about a symplectic transvection $\tau_{v,a}$ is that if $a = 0$ then $\tau_{v,a}$ is the identity and if $a \neq 0$ then $\tau_{v,a}$ is the identity precisely on the subspace $\text{span}(v)^\perp$, which is very large, in the sense of having codimension 1. Thus, despite the name, symplectic transvections are not highly complex maps. On the other hand, we should point out that since v is isotropic, the subspace $\text{span}(v)$ is singular and $\text{span}(v) \cap \text{span}(v)^\perp = \text{span}(v)$. Hence, $\text{span}(v)$ is *not* a vector space complement of the space $\text{span}(v)^\perp$ upon which $\tau_{v,a}$ is the identity. In other words, while we can write

$$V = \text{span}(v)^\perp \oplus U$$

where $\dim(U) = 1$ and $\tau|_{\text{span}(v)^\perp} = \iota$, we cannot say that $U = \text{span}(v)$.

Here are the basic properties of symplectic transvections.

Theorem 11.18 *Let $\tau_{v,a}$ be a symplectic transvection on V. Then*
1) *$\tau_{v,a}$ is a symplectic transformation.*
2) *$\tau_{v,a} = \iota$ if and only if $a = 0$.*
3) *If $x \perp v$ then $\tau_{v,a}(x) = x$. For $a \neq 0$, $x \perp v$ if and only if $\tau_{v,a}(x) = x$*
4) $\tau_{v,a}\tau_{v,b} = \tau_{v,a+b}$
5) $\tau_{v,a}^{-1} = \tau_{v,-a}$
6) *For any symplectic transformation σ,*

$$\sigma\tau_{v,a}\sigma^{-1} = \tau_{\sigma(v),a}$$

7) *For $b \in F^*$,*

$$\tau_{bv,a} = \tau_{v,ab^2}$$ □

Note that if U is a subspace of V and if $\tau_{u,a}$ is a symplectic transvection on U then, by definition, $u \in U$. However, the map $\tau_{u,a}$ can also be thought of as a symplectic transvection on V, defined by the same formula

$$\tau_{u,a}(x) = x + a\langle x, u\rangle u$$

where x can be any vector in V. Moreover, for any $z \in U^{\perp}$ we have $\tau_{u,a}(z) = z$ and so $\tau_{u,a}$ is the identity on $U^{\perp}$.

We now wish to prove that any symplectic transformation on a nonsingular symplectic geometry V is the product of symplectic transvections. The proof is not difficult, but it is a bit lengthy, so we break it up into parts. Our first goal is to show that we can get from any hyperbolic pair to any other hyperbolic pair using a product of symplectic transvections.

Let us say that two *hyperbolic pairs* (x,y) and (w,z) are **connected** if there is a product μ of symplectic transvections that carries x to w and y to z and write

$$\mu\colon (x,y) \leftrightarrow (w,z)$$

or just $(x,y) \leftrightarrow (w,z)$. It is clear that connectedness is an equivalence relation on the set of hyperbolic pairs.

Theorem 11.19 *Let V be a nonsingular symplectic geometry.*

1) *For every hyperbolic pair (u,v) and nonzero vector $w \in V$, there is a vector x for which $(u,v) \leftrightarrow (w,x)$.*
2) *Any two hyperbolic pairs (u,v) and (u,w) with the same first coordinate are connected.*
3) *Every pair (u,v) and (w,z) of hyperbolic pairs is connected.*

Proof. For part 1), all we need to do is find a product μ of symplectic transvections for which $\mu(u) = w$, because an isometry maps hyperbolic pairs to hyperbolic pairs and so we can simply set $x = \mu(v)$.

If $\langle u,w\rangle \neq 0$ then $u \neq w$ and

$$\tau_{u-w,a}(u) = u + a\langle u, u-w\rangle(u-w) = u - a\langle u,w\rangle(u-w)$$

Taking $a = 1/\langle u,w\rangle$ gives $\tau_{u-w,a}(u) = w$, as desired.

Now suppose that $\langle u,w\rangle = 0$. If there is a vector y that is *not* orthogonal to either u or w, then by what we have just proved, there is a vector x_1 such that $(u,v) \leftrightarrow (y,x_1)$ and a vector x for which $(y,x_1) \leftrightarrow (w,x)$. Then transitivity implies that $(u,v) \leftrightarrow (w,x)$.

But there is a nonzero vector $y \in V$ that is *not* orthogonal to either u or w since there is a linear functional f on V for which $f(u) \neq 0$ and $f(w) \neq 0$. But the Riesz representation theorem implies that there is a nonzero vector y such that $f(x) = \langle x,y\rangle$ for all $x \in V$.

For part 2), suppose first that $\langle v,w\rangle \neq 0$. Then $v \neq w$ and since $\langle u, v-w\rangle = 0$, we know that $\tau_{v-w,a}(u) = u$. Also,

$$\tau_{v-w,a}(v) = v + a\langle v, v-w\rangle(v-w) = v - a\langle v, w\rangle(v-w)$$

Taking $a = 1/\langle v, w\rangle$ gives $\tau_{v-w,a}(v) = w$, as desired. If $\langle v, w\rangle = 0$, then $\langle v, u+v\rangle \neq 0$ implies that $(u, v) \leftrightarrow (u, u+v)$ and $\langle u+v, w\rangle \neq 0$ implies that $(u, u+v) \leftrightarrow (u, w)$. It follows by transitivity that $(u, v) \leftrightarrow (u, w)$.

For part 3), parts 1) and 2) imply that there is a vector x for which

$$(u, v) \leftrightarrow (u, x) \leftrightarrow (w, z)$$

as desired. □

We can now show that the symplectic transvections generate the symplectic group.

Theorem 11.20 *Every symplectic transformation on a nonsingular symplectic geometry V is the product of symplectic transvections.*
Proof. Let μ be a symplectic transformation on V. We proceed by induction on $d = \dim(V)$, which must be even.

If $d = 2$ then $V = H = \text{span}(u, v)$ is a hyperbolic plane and by the previous theorem, there is a product τ of symplectic transvections on V for which

$$\tau\colon (u, v) \leftrightarrow (\mu(u), \mu(v))$$

Hence $\mu = \tau$. This proves the result if $d = 2$. Assume that the result holds for all dimensions less than d and let $\dim(V) = d$.

Let $H = \text{span}(u, v)$ be a hyperbolic plane in V and write

$$V = H \odot H^{\perp}$$

where $H^{\perp}$ is a nonsingular symplectic geometry of degree less than d.

Since $(\mu(u), \mu(v))$ is a hyperbolic pair, we again have a product τ of symplectic transvections on V for which

$$\tau\colon (u, v) \leftrightarrow (\mu(u), \mu(v))$$

Thus $\mu = \tau$ on the subspace H. Also, since H is invariant under $\tau^{-1}\mu$, so is $H^{\perp}$.

If we restrict $\tau^{-1}\mu$ to $H^{\perp}$, we may apply the induction hypotheses to get a product π of symplectic transvections on $H^{\perp}$ for which $\tau^{-1}\mu = \pi$ on $H^{\perp}$. Hence, $\mu = \tau\pi$ on $H^{\perp}$ and $\mu = \tau$ on H. But since the vectors that define the symplectic transvections making up π belong to $H^{\perp}$, we may extend π to V and $\pi = \iota$ on H. Thus, $\mu = \tau\pi$ on H as well, and we have $\mu = \tau\pi$ on V.

The Structure of Orthogonal Geometries: Orthogonal Bases

We have seen that no interesting (not totally degenerate) symplectic geometries have orthgonal bases. In contradistinction to the symplectic case, almost all interesting orthogonal geometries V have orthogonal bases. The only problem arises when V is also symplectic and $\operatorname{char}(F) = 2$.

In particular, if V is orthogonal and symplectic, then it cannot possess an orthogonal basis unless it is totally degenerate. When $\operatorname{char}(F) \neq 2$, the only orthogonal, symplectic geometries are the totally degenerate ones, since the matrix of V with respect to any basis is both symmetric and skew-symmetric, with zeros on the main diagonal and so must be the zero matrix. However, when $\operatorname{char}(F) = 2$, such a nonzero matrix exists, for example

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Thus, there are orthogonal, symplectic geometries that are not totally degenerate when $\operatorname{char}(F) = 2$. These geometries do not have orthogonal bases and we will not consider them further.

Once we have an orthogonal basis for V, the natural question is: “How close can we come to obtaining an orthonormal basis?” Clearly, this is possible only if V is nonsingular. As we will see, the answer to this question depends on the nature of the base field, and is different for algebraically closed fields, the real field and finite fields—the three cases that we will consider in this book.

We should mention that, even when V has an orthogonal basis, the Gram–Schmidt orthogonalization process may not apply, because even nonsingular orthogonal geometries may have isotropic vectors, and so division by $\langle u, u \rangle$ is problematic.

For example, consider an orthogonal hyperbolic plane $H = \operatorname{span}(u, v)$ and assume that $\operatorname{char}(F) \neq 2$. Thus, u and v are isotropic and $\langle u, v \rangle = \langle v, u \rangle = 1$. The vector u cannot be extended to an orthogonal basis, as would be possible for a real inner product space, using the Gram–Schmidt process, for it is easy to see that the set $\{u, au + bv\}$ cannot be an orthogonal basis for any $a, b \in F$. However, H has an orthogonal basis, namely, $\{u + v, u - v\}$.

Orthogonal Bases

Let V be an orthogonal geometry. If V is also symplectic, then V has an orthogonal basis if and only if it is totally degenerate. Moreover, when $\operatorname{char}(F) \neq 2$, these are the only types of orthogonal, symplectic geometries.

Now, let V be an orthogonal geometry that is not symplectic. Hence, V contains a nonisotropic vector u_1, the subspace $\text{span}(u_1)$ is nonsingular and

$$V = \text{span}(u_1) \odot V_1$$

where $V_1 = \text{span}(u_1)^{\perp}$. If V_1 is not symplectic, then we may decompose V_1 to get

$$V = \text{span}(u_1) \odot \text{span}(u_2) \odot V_2$$

This process may be continued until we reach a decomposition

$$V = \text{span}(u_1) \odot \cdots \odot \text{span}(u_k) \odot U$$

where U is symplectic as well as orthogonal. (This includes the case $U = \{0\}$.)

If $\text{char}(F) \neq 2$, then U is totally degenerate. Thus, if $\mathcal{B} = (u_1, \ldots, u_k)$ and $\mathcal{C}$ is any basis for U, the union $\mathcal{B} \cup \mathcal{C}$ is an orthogonal basis for V. Hence, when $\text{char}(F) \neq 2$, any orthogonal geometry has an orthogonal basis.

When $\text{char}(F) = 2$, we must work a bit harder. Since U is symplectic, it has the form $U = \mathcal{H} \odot \text{rad}(U)$ where $\mathcal{H}$ is a hyperbolic space and so

$$V = \text{span}(u_1) \odot \cdots \odot \text{span}(u_k) \odot \mathcal{H} \odot \mathcal{N}$$

where $\mathcal{N}$ is totally degenerate and the u_i are nonisotropic. If $\mathcal{B} = (u_1, \ldots, u_k)$ and $\mathcal{C} = (x_1, y_1, \ldots, x_m, y_m)$ is a hyperbolic basis for $\mathcal{H}$ and $\mathcal{D} = (z_1, \ldots, z_m)$ is an ordered basis for $\mathcal{N}$ then the union

$$\mathcal{E} = \mathcal{B} \cup \mathcal{C} \cup \mathcal{D} = (u_1, \ldots, u_k, x_1, y_1, \ldots, x_m, y_m, z_1, \ldots, z_m)$$

is an ordered basis for V. However, we can do better.

The following lemma says that, when $\text{char}(F) = 2$, a pair of isotropic basis vectors (such as x_i, y_i) can be replaced by a pair of nonisotropic basis vectors, in the presence of a nonisotropic basis vector (such as u_k).

Lemma 11.21 *Suppose that* $\text{char}(F) = 2$*. Let* W *be a three-dimensional orthogonal geometry with ordered basis* $\mathcal{B} = (u, x, y)$ *for which the matrix of the form with respect to* $\mathcal{B}$ *is*

$$M_{\mathcal{B}} = \begin{bmatrix} a & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

where $a \neq 0$*. Then the vectors*

$$\begin{aligned} v_1 &= u + x + y \\ v_2 &= u + ax \\ v_3 &= u + (1-a)x + y \end{aligned}$$

form an orthogonal basis of W consisting of nonisotropic vectors.

Proof. It is straightforward to check that the vectors v_1, v_2 and v_3 are linearly independent and mutually orthogonal. Details are left to the reader. □

Using the previous lemma, we can replace the vectors $\{u_k, x_1, y_1\}$ with the nonisotropic vectors $\{v_k, v_{k+1}, v_{k+2}\}$ and still have an ordered basis

$$\mathcal{E}_1 = (u_1, \ldots, u_{k-1}, v_k, v_{k+1}, v_{k+2}, x_2, y_2, \ldots, x_m, y_m)$$

for V. The replacement process can be repeated until the isotropic vectors are absorbed, leaving an orthogonal basis of nonisotropic vectors.

Let us summarize.

Theorem 11.22 *Let V be an orthogonal geometry.*

1) *If V is also symplectic, then V has an orthogonal basis if and only if it is totally degenerate. (When* $\operatorname{char}(F) \neq 2$*, these are the only types of orthogonal, symplectic geometries. When* $\operatorname{char}(F) = 2$*, orthogonal, symplectic geometries that are not totally degenerate do exist.)*
2) *If V is not symplectic, then V has an ordered orthogonal basis $\mathcal{B} = (u_1, \ldots, u_k, z_1, \ldots, z_m)$ for which $\langle u_i, u_i \rangle = a_i \neq 0$ and $\langle z_i, z_i \rangle = 0$. Hence, $M_{\mathcal{B}}$ has the diagonal form*

$$M_{\mathcal{B}} = \begin{bmatrix} a_1 & & & & & \\ & \ddots & & & & \\ & & a_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

with $k = \operatorname{rk}(M_{\mathcal{B}})$ nonzero entries and m zeros on the diagonal. □

As a corollary, we get a nice theorem about symmetric matrices.

Corollary 11.23 *Let M be a symmetric matrix that is not alternate if* $\operatorname{char}(F) = 2$*. Then M is congruent to a diagonal matrix.* □

The Classification of Orthogonal Geometries: Canonical Forms

We now want to consider the question of improving upon Theorem 11.22. The diagonal matrices of this theorem do not form a set of canonical forms for congruence. In fact, if $r_1, \ldots, r_k$ are nonzero scalars, then the matrix of V with respect to the basis $\mathcal{C} = (r_1u_1, \ldots, r_ku_n, z_1, \ldots, z_m)$ is

$$M_{\mathcal{C}} = \begin{bmatrix} r_1^2a_1 & & & & & \\ & \ddots & & & & \\ & & r_k^2a_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix} \tag{11.5}$$

Hence, $M_{\mathcal{B}}$ and $M_{\mathcal{C}}$ are congruent diagonal matrices, and by a simple change of basis, we can multiply any diagonal entry by a nonzero square in F.

The determination of a set of canonical forms for symmetric (nonalternate when $\text{char}(F) = 2$) matrices under congruence depends on the properties of the base field. Our plan is to consider three types of base fields: algebraically closed fields, the real field $\mathbb{R}$ and finite fields. Here is a preview of the forthcoming results.

1) When the base field F is algebraically closed, there is an ordered basis $\mathcal{B}$ for which

$$M_{\mathcal{B}} = Z_{k,m} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

If V is nonsingular, then $M_{\mathcal{B}}$ is an identity matrix and V has an orthonormal basis.

2) Over the real base field, there is an ordered basis $\mathcal{B}$ for which

$$M_{\mathcal{B}} = \mathrm{Z}_{p,m,k} = \begin{bmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & -1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & -1 & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{bmatrix}$$

3) If F is a finite field, there is an ordered basis $\mathcal{B}$ for which
$$M_{\mathcal{B}} = \mathrm{Z}_{k,m}(d) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & d & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$
where d is unique up to multiplication by a square and if $\mathrm{char}(F) = 2$ then we can take $d = 1$.

Now let us turn to the details.

Algebraically Closed Fields

If F is algebraically closed then for every $r \in F$, the polynomial $x^2 - r$ has a root in F, that is, every element of F has a square root in F. Therefore, we may choose $r_i = 1/\sqrt{a_i}$ in (11.5), which leads to the following result.

Theorem 11.24 *Let V be an orthogonal geometry over an algebraically closed field F. Provided that V is not symplectic as well when* $\mathrm{char}(F) = 2$, *then V has an ordered orthogonal basis $\mathcal{B} = (u_1, \ldots, u_k, z_1, \ldots, z_m)$ for which $\langle u_i, u_i \rangle = 1$ and $\langle z_i, z_i \rangle = 0$. Hence, $M_{\mathcal{B}}$ has the diagonal form*
$$M_{\mathcal{B}} = Z_{k,m} = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$
with k ones and m zeros on the diagonal. In particular, if V is nonsingular then V has an orthonormal basis. □

The matrix version of Theorem 11.24 follows.

Theorem 11.25 Let $\mathcal{S}_n$ be the set of all $n \times n$ symmetric matrices over an algebraically closed field F. If $\mathrm{char}(F) = 2$, we restrict $\mathcal{S}_n$ to the set of all symmetric matrices with at least one nonzero entry on the main diagonal.

1) *Any matrix M in $\mathcal{S}_n$ is congruent to a unique matrix of the form $Z_{k,m}$, in fact, $k = \mathrm{rk}(M)$ and $m = n - \mathrm{rk}(M)$.*
2) *The set of all matrices of the form $Z_{k,m}$ for $k + m = n$, is a set of canonical forms for congruence on $\mathcal{S}_n$.*
3) *The rank of a matrix is a complete invariant for congruence on $\mathcal{S}_n$.* □

The Real Field $\mathbb{R}$

If $F = \mathbb{R}$, we can choose $r_i = 1/\sqrt{|a_i|}$, so that all nonzero diagonal elements in (11.5) will be either 0, 1 or -1.

Theorem 11.26 *(***Sylvester's law of inertia***) Any orthogonal geometry V over the real field $\mathbb{R}$ has an ordered orthogonal basis*

$$\mathcal{B} = (u_1, \ldots, u_p, v_1, \ldots, v_m, z_1, \ldots, z_k)$$

for which $\langle u_i, u_i \rangle = 1$, $\langle v_i, v_i \rangle = -1$ and $\langle z_i, z_i \rangle = 0$. Hence, the matrix $M_\mathcal{B}$ has the diagonal form

$$M_\mathcal{B} = Z_{p,m,k} = \begin{bmatrix} 1 &&&&&&& \\ & \ddots &&&&&& \\ && 1 &&&&& \\ &&& -1 &&&& \\ &&&& \ddots &&& \\ &&&&& -1 && \\ &&&&&& 0 & \\ &&&&&&& \ddots & \\ &&&&&&&& 0 \end{bmatrix}$$

with p ones, m negative ones and k zeros on the diagonal. □

Here is the matrix version of Theorem 11.26.

Theorem 11.27 *Let $\mathcal{S}_n$ be the set of all $n \times n$ symmetric matrices over the real field $\mathbb{R}$.*

1) *Any matrix in $\mathcal{S}_n$ is congruent to a unique matrix of the form $Z_{p,m,k}$, for some p, m and $k = n - p - m$.*
2) *The set of all matrices of the form $Z_{p,m,k}$ for $p + m + k = n$ is a set of canonical forms for congruence on $\mathcal{S}_n$.*
3) *Let $M \in \mathcal{S}_n$ and let M be congruent to $Z_{p,m,k}$. The number $p + m$ is the rank of M, the number $p - m$ is the* **signature** *of M and the triple (p, m, k) is the* **inertia** *of M. The pair (p, m), or equivalently the pair $(p + m, p - m)$, is a complete invariant under congruence on $\mathcal{S}_n$.*

Proof. We need only prove the uniqueness statement in part 1). Let

$$\mathcal{B} = (u_1, \ldots, u_p, v_1, \ldots, v_m, z_1, \ldots, z_k)$$

and

$$\mathcal{C} = (u'_1, \ldots, u'_{p'}, v'_1, \ldots, v'_{m'} z'_1, \ldots, z'_{k'})$$

be ordered bases for which the matrices $M_\mathcal{B}$ and $M_\mathcal{C}$ have the form shown in Theorem 11.26. Since the rank of these matrices must be equal, we have $p + m = p' + m'$ and so $k = k'$.

If $x \in \operatorname{span}(u_1, \ldots, u_p)$ and $x \neq 0$ then

$$\langle x, x\rangle = \left\langle \sum r_i u_i, \sum r_j u_j \right\rangle = \sum_{i,j} r_i r_j \langle u_i, u_j\rangle = \sum_{i,j} r_i r_j \delta_{i,j} = \sum r_i^2 > 0$$

On the other hand, if $y \in \operatorname{span}(v'_1, \ldots, v'_{m'})$ and $y \neq 0$ then

$$\langle y, y\rangle = \left\langle \sum s_i v'_i, \sum s_j v'_j \right\rangle = \sum_{i,j} s_i s_j \langle v'_i, v'_j\rangle = -\sum_{i,j} s_i s_j \delta_{i,j} = -\sum s_i^2 < 0$$

Hence, if $y \in \operatorname{span}(v'_1, \ldots, v'_{m'}, z'_1, \ldots, z'_{k'})$ then $\langle y, y\rangle \le 0$. It follows that

$$\operatorname{span}(u_1, \ldots, u_p) \cap \operatorname{span}(v'_1, \ldots, v'_{m'}, z'_1, \ldots, z'_{k'}) = \{0\}$$

and so

$$p + (n - p') \le n$$

that is, $p \le p'$. By symmetry, $p' \le p$ and so $p = p'$. Finally, since $k = k'$, it follows that $m = m'$. □

Finite Fields

To deal with the case of finite fields, we must know something about the distribution of squares in finite fields, as well as the possible values of the scalars $\langle v, v\rangle$.

Theorem 11.28 *Let F_q be a finite field with q elements.*

1) *If* $\operatorname{char}(F_q) = 2$ *then every element of F_q is a square.*
2) *If* $\operatorname{char}(F_q) \neq 2$ *then exactly half of the nonzero elements of F_q are squares, that is, there are $(q-1)/2$ nonzero squares in F_q. Moreover, if x is any nonsquare in F_q then all nonsquares have the form $r^2 x$, for some $r \in F_q$.*

Proof. Write $F = F_q$, let F^* be the subgroup of all nonzero elements in F and let

$$(F^*)^2 = \{a^2 \mid a \in F^*\}$$

be the subgroup of all nonzero squares in F. The *Frobenius map* $\phi\colon F^* \to (F^*)^2$ defined by $\phi(a) = a^2$ is a surjective group homomorphism, with kernel

$$\ker(\phi) = \{a \in F \mid a^2 = 1\} = \{-1, 1\}$$

If $\operatorname{char}(F) = 2$, then $\ker(\phi) = \{1\}$ and so ϕ is bijective and $|F^*| = |(F^*)^2|$, which proves part 1). If $\operatorname{char}(F) \neq 2$, then $|\ker(\phi)| = 2$ and so $|F^*| = 2|(F^*)^2|$, which proves the first part of part 2). We leave proof of the last statement to the reader. □

Definition *A bilinear form on V is* **universal** *if for any nonzero $c \in F$ there exists a vector $v \in V$ for which $\langle v, v \rangle = c$.* □

Theorem 11.29 *Let V be an orthogonal geometry over a finite field F with* $\operatorname{char}(F) \neq 2$ *and assume that V has a nonsingular subspace of dimension at least* 2. *Then the bilinear form of V is universal.*

Proof. Theorem 11.22 implies that V contains two linearly independent vectors u and v for which

$$\langle u, u \rangle = a \neq 0,\ \langle v, v \rangle = b \neq 0,\ \langle u, v \rangle = 0$$

Given any $c \in F$, we want to find α and β for which

$$c = \langle \alpha u + \beta v, \alpha u + \beta v \rangle = a\alpha^2 + b\beta^2$$

or

$$a\alpha^2 = c - b\beta^2$$

If $A = \{a\alpha^2 \mid \alpha \in F\}$ then $|A| = (q+1)/2$, since there are $(q-1)/2$ nonzero squares α^2 and also we must consider $\alpha = 0$. Also, if $B = \{c - b\beta^2 \mid \beta \in F\}$ then for the same reasons $|B| = (q+1)/2$. It follows that $A \cap B$ cannot be the empty set and so there are α and β for which $a\alpha^2 = c - b\beta^2$, as desired. □

Now we can proceed with the business at hand.

Theorem 11.30 *Let V be an orthogonal geometry over a finite field F and assume that V is not symplectic if* $\operatorname{char}(F) = 2$. *If* $\operatorname{char}(F) \neq 2$ *then let d be a fixed nonsquare in F. For any nonzero $a \in F$, write*

$$X_k(a) = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a & & & \\ & & & & 0 & & \\ & & & & & \ddots & \\ & & & & & & 0 \end{bmatrix}$$

where $\operatorname{rk}(X_k(a)) = k$.

1) *If* $\operatorname{char}(F) = 2$ *then there is an ordered basis $\mathcal{B}$ for which $M_{\mathcal{B}} = X_k(1)$.*
2) *If* $\operatorname{char}(F) \neq 2$, *then there is an ordered basis $\mathcal{B}$ for which $M_{\mathcal{B}}$ equals $X_k(1)$ or $X_k(d)$.*

Proof. We can dispose of the case $\operatorname{char}(F) = 2$ quite easily: Referring to (11.5), since every element of F has a square root, we may take $r_i = (\sqrt{a_i})^{-1}$.

If $\operatorname{char}(F) \neq 2$, then Theorem 11.22 implies that there is an ordered orthogonal basis

$$\mathcal{B} = (u_1, \ldots, u_k, z_1, \ldots, z_m)$$

for which $\langle u_i, u_i \rangle = a_i \neq 0$ and $\langle z_i, z_i \rangle = 0$. Hence, $M_{\mathcal{B}}$ has the diagonal form

$$M_{\mathcal{B}} = \begin{bmatrix} a_1 & & & & & \\ & \ddots & & & & \\ & & a_k & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}$$

Now, consider the nonsingular orthogonal geometry $V_1 = \text{span}(u_1, u_2)$. According to Theorem 11.29, the form is universal when restricted to V_1. Hence, there exists a $v_1 \in V_1$ for which $\langle v_1, v_1 \rangle = 1$.

Now, $v_1 = ru_1 + su_2$ for $r, s \in F$ not both 0, and we may swap u_1 and u_2 if necessary to ensure that $r \neq 0$. Hence,

$$\mathcal{B}_1 = (v_1, u_2, \ldots, u_k, z_1, \ldots, z_m)$$

is an ordered basis for V for which the matrix $M_{\mathcal{B}_1}$ is diagonal and has a 1 in the upper left entry. We can repeat the process with the subspace $V_2 = \text{span}(v_2, v_3)$. Continuing in this way, we can find an ordered basis

$$\mathcal{C} = (v_1, v_2, \ldots, v_k, z_1, \ldots, z_m)$$

for which $M_{\mathcal{C}} = X_k(a)$ for some nonzero $a \in F$. Now, if a is a square in F then we can replace v_k by $(1/\sqrt{a})v_k$ to get a basis $\mathcal{D}$ for which $M_{\mathcal{D}} = X_k(1)$. If a is not a square in F, then $a = r^2 d$ for some $r \in F$ and so replacing v_k by $(1/r)v_k$ gives a basis $\mathcal{D}$ for which $M_{\Delta} = X_k(d)$. $\square$

Theorem 11.31 *Let $\mathcal{S}_n$ be the set of all $n \times n$ symmetric matrices over a finite field F. If* char$(F) = 2$, *we restrict $\mathcal{S}_n$ to the set of all symmetric matrices with at least one nonzero entry on the main diagonal.*

1) *If* char$(F) = 2$ *then any matrix in $\mathcal{S}_n$ is congruent to a unique matrix of the form $X_k(1)$ and the matrices $\{X_k(1) \mid k = 0, \ldots, n\}$ form a set of canonical forms for $\mathcal{S}_n$ under congruence. Also, the rank is a complete invariant.*
2) *If* char$(F) \neq 2$, *let d be a fixed nonsquare in F. Then any matrix $\mathcal{S}_n$ is congruent to a unique matrix of the form $X_k(1)$ or $X_k(d)$. The set $\{X_k(1), X_k(d) \mid k = 0, \ldots, n\}$ is a set of canonical forms for congruence on $\mathcal{S}_n$. (Thus, there are exactly two congruence classes for each rank k.)*$\square$

The Orthogonal Group

Having "settled" the classification question for orthogonal geometries over certain types of fields, let us turn to a discussion of the structure-preserving maps, that is, the isometries.

Rotations and Reflections

We have seen that if $\mathcal{B}$ is an ordered basis for V, then for any $x, y \in V$

$$\langle x, y \rangle = [x]_{\mathcal{B}}^t M_{\mathcal{B}} [y]_{\mathcal{B}}$$

Now, for any $\tau \in \mathcal{L}(V)$, we have

$$\langle \tau x, \tau y \rangle = [\tau x]_{\mathcal{B}}^t M_{\mathcal{B}} [\tau y]_{\mathcal{B}} = [x]_{\mathcal{B}}^t ([\tau]_{\mathcal{B}}^t M_{\mathcal{B}} [\tau]_{\mathcal{B}}) [y]_{\mathcal{B}}$$

and so τ is an isometry if and only if

$$[\tau]_{\mathcal{B}}^t M_{\mathcal{B}} [\tau]_{\mathcal{B}} = M_{\mathcal{B}}$$

Taking determinants gives

$$\det(M_{\mathcal{B}}) = \det([\tau]_{\mathcal{B}})^2 \det(M_{\mathcal{B}})$$

Therefore, if V is nonsingular then

$$\det([\tau]_{\mathcal{B}}) = \pm 1$$

Since the determinant is an invariant under similarity, we can make the following definition.

Definition *Let τ be an isometry on a nonsingular orthogonal geometry V. The* **determinant** *of τ is the determinant of any matrix $[\tau]_{\mathcal{B}}$ representing τ. If* $\det(\tau) = 1$ *then τ is called a* **rotation** *and if* $\det(\tau) = -1$ *then τ is called a* **reflection**. □

The set $\mathcal{O}^+(V)$ of rotations forms a subgroup of the orthogonal group $\mathcal{O}(V)$ and the surjective determinant map $\det\colon \mathcal{O}(V) \to \{-1, 1\}$ has kernel $\mathcal{O}^+(V)$. Hence, if $\operatorname{char}(F) \neq 2$, then $\mathcal{O}^+(V)$ is a normal subgroup of $\mathcal{O}(V)$ of index 2.

Symmetries

Recall that for a real (or complex) inner product space V, we defined a *reflection* to be a linear map H_v for which

$$H_v v = -v, \; (H_v)|_{\langle v \rangle^\perp} = \iota$$

The term *symmetry* is often used in the context of general orthogonal geometries.

In particular, suppose that V is a nonsingular orthogonal geometry over F, where $\operatorname{char}(F) \neq 2$ and let $u \in V$ be nonisotropic. Write

$$V = \operatorname{span}(u) \odot \operatorname{span}(u)^\perp$$

Then there is a unique isometry σ_u with the properties

1) $\sigma_u(u) = -u$
2) $\sigma_u(x) = x$ for all $x \in \operatorname{span}(u)^\perp$

We can also write $\sigma_u = -\iota \odot \iota$, that is

$$\sigma_u(x+y) = -x+y$$

for all $x \in \operatorname{span}(u)$ and $y \in \operatorname{span}(u)^\perp$. It is easy to see that

$$\sigma_u(v) = v - \frac{2\langle v,u\rangle}{\langle u,u\rangle}u$$

The map σ_u is called the **symmetry** determined by u.

Note that the requirement that u be nonisotropic is required, since otherwise we would have $u \in \operatorname{span}(u)^\perp$ and so $-u = \sigma_u(u) = u$, which implies that $u = 0$. (Thus, symplectic geometries do not have symmetries.)

In the context of real inner product spaces, Theorem 10.11 says that if $\|v\| = \|w\| \neq 0$, then H_{v-w} is the unique reflection sending v to w, that is, $H_{v-w}(v) = w$. In the present context, we must be careful, since symmetries are defined for nonisotropic vectors only. Here is what we can say.

Theorem 11.32 *Let V be a nonsingular orthogonal geometry over a field F, with* $\operatorname{char}(F) \neq 2$. *If $u,v \in V$ have the same nonzero "length," that is, if*

$$\langle u,u\rangle = \langle v,v\rangle \neq 0$$

then there exists a symmetry σ for which

$$\sigma(u) = v \quad \textit{or} \quad \sigma(u) = -v$$

Proof. In general, if x and y are orthogonal isotropic vectors, then $x+y$ and $x-y$ are also isotropic. Hence, since u and v are not isotropic, it follows that one of $u-v$ and $u+v$ must be nonisotropic. If $u+v$ is nonisotropic, then

$$\sigma_{u+v}(u+v) = -(u+v)$$

and

$$\sigma_{u+v}(u-v) = u-v$$

Combining these two gives $\sigma_{u+v}(u) = -v$. On the other hand, if $u-v$ is nonisotropic, then

$$\sigma_{u-v}(u-v) = -(u-v)$$

and

$$\sigma_{u-v}(u+v) = u+v$$

These equations give $\sigma_{u-v}(u) = v$. □

Recall that an operator on a real inner product space is unitary if and only if it is a product of reflections. Here is the generalization to nonsingular orthogonal geometries.

Theorem 11.33 *Let V be a nonsingular orthogonal geometry over a field F with* $\text{char}(F) \neq 2$. *A linear transformation τ on V is an orthogonal transformation (an isometry) if and only if τ is the product of symmetries on V.*
Proof. We proceed by induction on $d = \dim(V)$. If $d = 1$ then $V = \text{span}(v)$ where $\langle v, v\rangle \neq 0$. Let $\tau(v) = \alpha v$ where $\alpha \in F$. Since τ is unitary

$$\alpha^2\langle v, v\rangle = \langle \alpha v, \alpha v\rangle = \langle \tau(v), \tau(v)\rangle = \langle v, v\rangle$$

and so $\alpha = \pm 1$. If $\alpha = 1$ then τ is the identity, which is equal to σ_v^2. On the other hand, if $\alpha = -1$ then $\tau = \sigma_v$. In either case, τ is a product of symmetries.

Assume now that the theorem is true for dimensions less than d and let $\dim(V) = d$. Let $v \in V$ be nonisotropic. Since $\langle \tau(v), \tau(v)\rangle = \langle v, v\rangle \neq 0$, Theorem 11.32 implies the existence of a symmetry σ on V for which

$$\sigma(\tau(v)) = \epsilon v$$

where $\epsilon = \pm 1$. Thus, $\sigma\tau = \pm\iota$ on $\text{span}(v)$. Since Theorem 11.15 implies that $\text{span}(v)^\perp$ is $\sigma\tau$-invariant, we may apply the induction hypothesis to $\sigma\tau$ on $\text{span}(v)^\perp$ to get

$$\sigma\tau|_{\text{span}(v)^\perp} = \sigma_{w_1}\cdots\sigma_{w_k} = \rho$$

where $w_i \in \text{span}(v)^\perp$ and σ_{w_i} is a symmetry on $\text{span}(v)^\perp$. But each σ_{w_i} can be extended to a symmetry on V by setting $\sigma_{w_i}(v) = v$. Assume that $\overline{\rho}$ is the extension of ρ to V, where $\rho = \iota$ on $\text{span}(v)$. Hence, $\sigma\tau = \overline{\rho}$ on $\text{span}(v)^\perp$ and $\sigma\tau = \epsilon\overline{\rho}$ on $\text{span}(v)$.

If $\epsilon = 1$ then $\sigma\tau = \overline{\rho}$ on V and so $\tau = \sigma\overline{\rho}$, which completes the proof. If $\epsilon = -1$ then $\sigma\tau = \sigma_v\overline{\rho}$ on $\text{span}(v)^\perp$ since σ_v is the identity on $\text{span}(v)^\perp$ and $\sigma\tau = \sigma_v\overline{\rho}$ on $\text{span}(v)$. Hence, $\sigma\tau = \sigma_v\overline{\rho}$ on V and so $\tau = \sigma\sigma_v\overline{\rho}$ on V. □

The Witt's Theorems for Orthogonal Geometries

We are now ready to consider the Witt theorems for orthogonal geometries.

Theorem 11.34 *(***Witt's cancellation theorem***) Let V and W be isometric nonsingular orthogonal geometries over a field F with* $\text{char}(F) \neq 2$. *Suppose that*

$$V = S \odot S^\perp \quad \textit{and} \quad W = T \odot T^\perp$$

Then

$$S \approx T \Rightarrow S^{\perp} \approx T^{\perp}$$

Proof. First, we prove that it is sufficient to consider the case $V = W$. Suppose that the result holds when $V = W$ and that $\mu: V \to W$ is an isometry. Then

$$\mu(S) \odot \mu(S^{\perp}) = \mu(S \odot S^{\perp}) = \mu(V) = W = T \odot T^{\perp}$$

Furthermore, $\mu(S) \approx S \approx T$. We can therefore apply the theorem to W to get

$$S^{\perp} \approx \mu(S^{\perp}) \approx T^{\perp}$$

as desired.

To prove the theorem when $V = W$, assume that

$$V = S \odot S^{\perp} = T \odot T^{\perp}$$

where S and T are nonsingular and $S \approx T$. Let $\tau: S \to T$ be an isometry. We proceed by induction on dim(S).

Suppose first that dim$(S) = 1$ and that $S =$ span(s). Since

$$\langle \tau(s), \tau(s) \rangle = \langle s, s \rangle \neq 0$$

Theorem 11.32 implies that there is a symmetry σ for which $\sigma(s) = \epsilon\tau(s)$ where $\epsilon = \pm 1$. Hence, σ is an isometry of V for which $T = \sigma(S)$ and Theorem 11.10 implies that $T^{\perp} = \sigma(S^{\perp})$. Thus, $\sigma|_{S^{\perp}}$ is the desired isometry.

Now suppose the theorem is true for dim$(S) < k$ and let dim$(S) = k$. Let $\tau: S \to T$ be an isometry. Since S is nonsingular, we can choose a nonisotropic vector $s \in S$ and write $S =$ span$(s) \odot U$, where U is nonsingular. It follows that

$$V = S \odot S^{\perp} = \text{span}(s) \odot U \odot S^{\perp}$$

and

$$V = T \odot T^{\perp} = \tau(\text{span}(s)) \odot \tau(U) \odot T^{\perp}$$

Now we may apply the one-dimensional case to deduce that

$$U \odot S^{\perp} \approx \tau(U) \odot T^{\perp}$$

If $\sigma: U \odot S^{\perp} \to \tau(U) \odot T^{\perp}$ is an isometry then

$$\sigma(U) \odot \sigma(S^{\perp}) = \sigma(U \odot S^{\perp}) = \tau(U) \odot T^{\perp}$$

But $\sigma(U) \approx \tau(U)$ and since dim$(\sigma(U)) =$ dim$(U) < k$, the induction hypothesis implies that $S^{\perp} \approx \sigma(S^{\perp}) \approx T^{\perp}$. □

As we have seen, Witt's extension theorem is a corollary of Witt's cancellation theorem.

Theorem 11.35 *(***Witt's extension theorem***) Let V and V' be isometric nonsingular orthogonal geometries over a field F, with* $\text{char}(F) \neq 2$. *Suppose that U is a subspace of V and*

$$\tau: U \to \tau(U) \subseteq V'$$

is an isometry. Then τ can be extended to an isometry from V to V'. □

Maximal Hyperbolic Subspaces of an Orthogonal Geometry

We have seen that any orthogonal geometry V can be written in the form

$$V = U \odot \text{rad}(V)$$

where U is nonsingular. Nonsingular spaces are better behaved than singular ones, but they can still possess isotropic vectors.

We can improve upon the preceding decomposition by noticing that if $u \in U$ is isotropic, then $\text{span}(u)$ is totally degenerate and so it can be "captured" in a hyperbolic plane $H = \text{span}(u, x)$, namely, the nonsingular extension of $\text{span}(u)$. Then we can write

$$V = H \odot H^{\perp_U} \odot \text{rad}(V)$$

where $H^{\perp_U}$ is the orthogonal complement of H in U and has "one fewer" isotropic vector.

In order to generalize this process, we first discuss maximal totally degenerate subspaces.

Maximal Totally Degenerate Subspaces

Let V be a nonsingular orthogonal geometry over a field F, with $\text{char}(F) \neq 2$. Suppose that U and U' are maximal totally degenerate subspaces of V. We claim that $\dim(U) = \dim(U')$. For if $\dim(U) \leq \dim(U')$, then there is a vector space isomorphism $\tau: U \to \tau(U) \subseteq U'$, which is also an isometry, since U and U' are totally degenerate. Thus, Witt's extension theorem implies the existence of an isometry $\overline{\tau}: V \to V$ that extends τ. In particular, $\overline{\tau}^{-1}(U')$ is a totally degenerate space that contains U and so $\overline{\tau}^{-1}(U') = U$, which shows that $\dim(U) = \dim(U')$.

We have proved the following.

Theorem 11.36 *Let V be a nonsingular orthogonal geometry over a field F, with* $\text{char}(F) \neq 2$.

1) *All maximal totally degenerate subspaces of V have the same dimension, which is called the* **Witt index** *of V and is denoted by $w(V)$.*
2) *Any totally degenerate subspace of V of dimension $w(V)$ is maximal.* □

Maximal Hyperbolic Subspaces

We can prove by a similar argument that all maximal hyperbolic subspaces of V have the same dimension. Let

$$\mathcal{H}_{2k} = H_1 \odot \cdots \odot H_k$$

and

$$\mathcal{K}_{2m} = K_1 \odot \cdots \odot K_m$$

be maximal hyperbolic subspaces of V and suppose that $H_i = \text{span}(u_i, v_i)$ and $K_i = \text{span}(x_i, y_i)$. We may assume that $\dim(\mathcal{H}) \le \dim(\mathcal{K})$.

The linear map $\tau: \mathcal{H} \to \mathcal{K}$ defined by

$$\tau(u_i) = x_i,\ \tau(v_i) = y_i$$

is clearly an isometry from $\mathcal{H}$ to $\tau(\mathcal{H})$. Thus, Witt's extension theorem implies the existence of an isometry $\overline{\tau}: V \to V$ that extends τ. In particular, $\overline{\tau}^{-1}(\mathcal{K})$ is a hyperbolic space that contains $\mathcal{H}$ and so $\overline{\tau}^{-1}(\mathcal{K}) = \mathcal{H}$. It follows that $\dim(\mathcal{K}) = \dim(\mathcal{H})$.

It is not hard to see that the maximum dimension $h(V)$ of a hyperbolic subspace of V is $2w(V)$, where $w(V)$ is the Witt index of V. First, the nonsingular extension of a maximal totally degenerate subspace U_w of V is a hyperbolic space of dimension $2w(V)$ and so $h(V) \ge 2w(V)$. On the other hand, there is a totally degenerate subspace U_k contained in any hyperbolic space $\mathcal{H}_{2k}$ and so $k \le w(V)$, that is, $\dim(\mathcal{H}_{2k}) \le 2w(V)$. Hence $h(V) \le 2w(V)$ and so $h(V) = 2w(V)$.

Theorem 11.37 *Let V be a nonsingular orthogonal geometry over a field F, with* $\text{char}(F) \ne 2$.

1) *All maximal hyperbolic subspaces of V have dimension $2w(V)$.*
2) *Any hyperbolic subspace of dimenison $2w(V)$ must be maximal.*
3) *The Witt index of a hyperbolic space $\mathcal{H}_{2k}$ is k.* □

The Anisotropic Decomposition of an Orthogonal Geometry

If $\mathcal{H}$ is a maximal hyperbolic subspace of V then

$$V = \mathcal{H} \odot \mathcal{H}^{\perp}$$

Since $\mathcal{H}$ is maximal, $\mathcal{H}^{\perp}$ is anisotropic, for if $u \in \mathcal{H}^{\perp}$ is isotropic then the nonsingular extension of $\mathcal{H} \odot \text{span}(u)$ would be a hyperbolic space strictly larger than $\mathcal{H}$.

Thus, we arrive at the following decomposition theorem for orthogonal geometries.

Theorem 11.38 *(***The anisotropic decomposition of an orthogonal geometry***)* *Let* $V = U \odot \mathrm{rad}(V)$ *be an orthogonal geometry over* F*, with* $\mathrm{char}(F) \neq 2$*. Let* $\mathcal{H}$ *be a maximal hyperbolic subspace of* U*, where* $\mathcal{H} = \{0\}$ *if* U *has no isotropic vectors. Then*

$$V = S \odot \mathcal{H} \odot \mathrm{rad}(V)$$

where S *is anisotropic,* $\mathcal{H}$ *is hyperbolic of dimension* $2w(V)$ *and* $\mathrm{rad}(V)$ *is totally degenerate.* □

Exercises

1. Let U, W be subspaces of a metric vector space V. Show that
   a) $U \subseteq W \Rightarrow W^{\perp} \subseteq U^{\perp}$
   b) $U \subseteq U^{\perp\perp}$
   c) $U^{\perp} = U^{\perp\perp\perp}$
2. Let U, W be subspaces of a metric vector space V. Show that
   a) $(U + W)^{\perp} = U^{\perp} \cap W^{\perp}$
   b) $(U \cap W)^{\perp} = U^{\perp} + W^{\perp}$
3. Prove that the following are equivalent:
   a) V is nonsingular
   b) $\langle u, x\rangle = \langle v, x\rangle$ for all $x \in V$ implies $u = v$
4. Show that a metric vector space V is nonsingular if and only if the matrix $M_{\mathcal{B}}$ of the form is nonsingular, for every ordered basis $\mathcal{B}$.
5. Let V be a finite-dimensional vector space with a bilinear form $\langle , \rangle$. We do *not* assume that the form is symmetric or alternate. Show that the following are equivalent:
   a) $\{v \in V \mid \langle v, w\rangle = 0 \text{ for all } w \in V\} = 0$
   b) $\{v \in V \mid \langle w, v\rangle = 0 \text{ for all } w \in V\} = 0$
   Hint: Consider the singularity of the matrix of the form.
6. Find a diagonal matrix congruent to
$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 0 & 1 \\ 3 & 1 & -1 \end{bmatrix}$$
7. Prove that the matrices
$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } M = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$$
   are congruent over the base field $F = \mathbb{Q}$ of rational numbers. Find an invertible matrix P such that $P^t I_2 P = M$.

8. Let V be an orthogonal geometry over a field F with $\mathrm{char}(F) \neq 2$. We wish to construct an orthogonal basis $\mathcal{O} = (u_1, \ldots, u_n)$ for V, starting with any generating set $\mathcal{G} = (v_1, \ldots, v_n)$. Justify the following steps, essentially due to Lagrange. We may assume that V is not totally degenerate.
   a) If $\langle v_i, v_i \rangle \neq 0$ for some i then let $u_1 = v_i$. Otherwise, there are indices $i \neq j$ for which $\langle v_i, v_j \rangle \neq 0$. Let $u_1 = v_i + v_j$.
   b) Assume we have found an ordered set of vectors $\mathcal{O}_k = (u_1, \ldots, u_k)$ that form an orthogonal basis for a subspace V_k of V and that none of the u_i's are isotropic. Then $V = V_k \odot V_k^{\perp}$.
   c) For each $v_i \in \mathcal{G}$, let

$$w_i = v_i - \sum_{j=1}^{k} \frac{\langle v_i, u_j \rangle}{\langle u_j, u_j \rangle} u_j$$

   Then the vectors w_i span $V_k^{\perp}$. If $V_k^{\perp}$ is totally degenerate, take any basis for $V_k^{\perp}$ and append it to $\mathcal{O}_k$. Otherwise, repeat step a) on $V_k^{\perp}$ to get another vector u_{k+1} and let $\mathcal{O}_{k+1} = (u_1, \ldots, u_{k+1})$. Eventually, we arrive at an orthogonal basis $\mathcal{O}_n$ for V.
9. Prove that orthogonal hyperbolic planes may be characterized as two-dimensional nonsingular orthogonal geometries that have exactly two one-dimensional totally isotropic (equivalently: totally degenerate) subspaces.
10. Prove that a two-dimensional nonsingular orthogonal geometry is a hyperbolic plane if and only if its discriminant is $F^2(-1)$.
11. Does Minkowski space contain any isotropic vectors? If so, find them.
12. Is Minkowski space isometric to Euclidean space $\mathbb{R}^4$?
13. If $\langle , \rangle$ is a symmetric bilinear form on V and $\mathrm{char}(F) \neq 2$, show that $Q(x) = \langle x, x \rangle / 2$ is a quadratic form.
14. Let V be a vector space over a field F, with ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$. Let $p(x_1, \ldots, x_n)$ be a *homogeneous* polynomial of degree d over F, that is, a polynomial each of whose terms has degree d. The ***d*-form** defined by p is the function from V to F defined as follows. If $v = \Sigma a_i v_i$ then

$$p(v) = p(a_1, \ldots, a_n)$$

   (We use the same notation for the form and the polynomial.) Prove that 2-forms are the same as quadratic forms.
15. Show that τ is an isometry on V if and only if $Q(\tau(v)) = Q(v)$ where Q is the quadratic form associated with the bilinear form on V. (Assume that $\mathrm{char}(F) \neq 2$.)
16. Show that a quadratic form Q on V satisfies the parallelogram law:

$$Q(x+y) + Q(x-y) = 2[Q(x) + Q(y)]$$

17. Show that if V is a nonsingular orthogonal geometry over a field F, with $\mathrm{char}(F) \neq 2$ then any totally isotropic subspace of V is also a totally degenerate space.
18. Is it true that $V = \mathrm{rad}(V) \odot \mathrm{rad}(V)^{\perp}$?

19. Let V be a nonsingular symplectic geometry and let $\tau_{v,a}$ be a symplectic transvection. Prove that
    a) $\tau_{v,a}\tau_{v,b} = \tau_{v,a+b}$
    b) For any symplectic transformation σ,
    $$\sigma\tau_{v,a}\sigma^{-1} = \tau_{\sigma(v),a}$$
    c) For $b \in F^*$,
    $$\tau_{bv,a} = \tau_{v,ab^2}$$
    d) For a fixed $v \neq 0$, the map $a \mapsto \tau_{v,a}$ is an isomorphism from the additive group of F onto the group $\{\tau_{v,a} \mid a \in F\} \subseteq \mathrm{Sp}(V)$.
20. Prove that if x is any nonsquare in a finite field F_q then all nonsquares have the form r^2x, for some $r \in F$. Hence, the product of any two nonsquares in F_q is a square.
21. Formulate Sylvester's law of inertia in terms of quadratic forms on V.
22. Show that a two-dimensional space is a hyperbolic plane if and only if it is nonsingular and contains an isotropic vector. Assume that $\mathrm{char}(F) \neq 2$.
23. Prove directly that a hyperbolic plane in an orthogonal geometry cannot have an orthogonal basis when $\mathrm{char}(F) = 2$.
24. a) Let U be a subspace of V. Show that the inner product $\langle x+U, y+U\rangle = \langle x, y\rangle$ on the quotient space V/U is well-defined if and only if $U \subseteq \mathrm{rad}(V)$.
    b) If $U \subseteq \mathrm{rad}(V)$, when is V/U nonsingular?
25. Let $V = N \odot S$, where N is a totally degenerate space.
    a) Prove that $N = \mathrm{rad}(V)$ if and only if S is nonsingular.
    b) If S is nonsingular, prove that $S \approx V/\mathrm{rad}(V)$.
26. Let $\dim(V) = \dim(W)$. Prove that $V/\mathrm{rad}(V) \approx W/\mathrm{rad}(W)$ implies $V \approx W$.
27. Let $V = S \odot T$. Prove that
    a) $\mathrm{rad}(V) = \mathrm{rad}(S) \odot \mathrm{rad}(T)$
    b) $V/\mathrm{rad}(V) \approx S/\mathrm{rad}(S) \odot T/\mathrm{rad}(T)$
    c) $\dim(\mathrm{rad}(V)) = \dim(\mathrm{rad}(S)) + \dim(\mathrm{rad}(T))$
    d) V is nonsingular if and only if S and T are both nonsingular.
28. Let V be a nonsingular metric vector space. Because the Riesz representation theorem is valid in V, we can define the adjoint τ^* of a linear map $\tau \in \mathcal{L}(V)$ exactly as in the case of real inner product spaces. Prove that τ is an isometry if and only if it is bijective and unitary (that is, $\tau\tau^* = \iota$).
29. If $\mathrm{char}(F) \neq 2$, prove that $\tau \in \mathcal{L}(V, W)$ is an isometry if and only if it is bijective and $\langle \tau(v), \tau(v)\rangle = \langle v, v\rangle$ for all $v \in V$.
30. Let $\mathcal{B} = \{v_1, \ldots, v_n\}$ be a basis for V. Prove that $\tau \in \mathcal{L}(V, W)$ is an isometry if and only if it is bijective and $\langle \tau v_i, \tau v_j\rangle = \langle v_i, v_j\rangle$ for all i, j.
31. Let τ be a linear operator on a metric vector space V. Let $\mathcal{B} = (v_1, \ldots, v_n)$ be an ordered basis for V and let $M_{\mathcal{B}}$ be the matrix of the form relative to

$\mathcal{B}$. Prove that τ is an isometry if and only if

$$[\tau]_{\mathcal{B}}^t\, M_{\mathcal{B}}[\tau]_{\mathcal{B}} = M_{\mathcal{B}}$$

32. Let V be a nonsingular orthogonal geometry and let $\tau \in \mathcal{L}(V)$ be an isometry.
    a) Show that $\dim(\ker(\iota - \tau)) = \dim(\operatorname{im}(\iota - \tau)^{\perp})$.
    b) Show that $\ker(\iota - \tau) = \operatorname{im}(\iota - \tau)^{\perp}$. How would you describe $\ker(\iota - \tau)$ in words?
    c) If τ is a symmetry, what is $\dim(\ker(\iota - \tau))$?
    d) Can you characterize symmetries by means of $\dim(\ker(\iota - \tau))$?
33. A linear transformation $\tau \in \mathcal{L}(V)$ is called **unipotent** if $\tau - \iota$ is nilpotent. Suppose that V is a nonisotropic metric vector space and that τ is unipotent and isometric. Show that $\tau = \iota$.
34. Let V be a hyperbolic space of dimension $2m$ and let U be a hyperbolic subspace of V of dimension $2k$. Show that for each $k \le j \le m$, there is a hyperbolic subspace $\mathcal{H}_{2j}$ of V for which $U \subseteq \mathcal{H}_{2j} \subseteq V$.
35. Let $\operatorname{char}(F) \ne 2$. Prove that if X is a totally degenerate subspace of an orthgonal geometry V then $\dim(X) \le \dim(V)/2$.
36. Prove that an orthogonal geometry V of dimension n is a hyperbolic space if and only if V is nonsingular, n is even and V contains a totally degenerate subspace of dimension $n/2$.
37. Prove that a symplectic transformation has determinant equal to 1.

Chapter 12
Metric Spaces

The Definition

In Chapter 9, we studied the basic properties of real and complex inner product spaces. Much of what we did does not depend on whether the space in question is finite or infinite-dimensional. However, as we discussed in Chapter 9, the presence of an inner product and hence a metric, on a vector space, raises a host of new issues related to convergence. In this chapter, we discuss briefly the concept of a metric space. This will enable us to study the convergence properties of real and complex inner product spaces.

A metric space is not an algebraic structure. Rather it is designed to model the abstract properties of distance.

Definition *A* **metric space** *is a pair* (M,d)*, where* M *is a nonempty set and* $d\colon M\times M\to\mathbb{R}$ *is a real-valued function, called a* **metric** *on* M*, with the following properties. The expression* $d(x,y)$ *is read "the distance from* x *to* y*."*

1) (**Positive definiteness**) *For all* $x,y\in M$*,*

$$d(x,y)\geq 0$$

and $d(x,y)=0$ *if and only if* $x=y$*.*

2) (**Symmetry**) *For all* $x,y\in M$*,*

$$d(x,y)=d(y,x)$$

3) (**Triangle inequality**) *For all* $x,y,z\in M$*,*

$$d(x,y)\leq d(x,z)+d(z,y) \qquad \square$$

As is customary, when there is no cause for confusion, we simply say "let M be a metric space."

Example 12.1 Any nonempty set M is a metric space under the **discrete metric**, defined by

$$d(x,y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases} \qquad \square$$

Example 12.2

1) The set $\mathbb{R}^n$ is a metric space, under the metric defined for $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ by

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$$

This is called the **Euclidean metric** on $\mathbb{R}^n$. We note that $\mathbb{R}^n$ is also a metric space under the metric

$$d_1(x,y) = |x_1 - y_1| + \cdots + |x_n - y_n|$$

Of course, $(\mathbb{R}^n, d)$ and $(\mathbb{R}^n, d_1)$ are different metric spaces.

2) The set $\mathbb{C}^n$ is a metric space under the **unitary metric**

$$d(x,y) = \sqrt{|x_1 - y_1|^2 + \cdots + |x_n - y_n|^2}$$

where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ are in $\mathbb{C}^n$. $\square$

Example 12.3

1) The set $C[a,b]$ of all real-valued (or complex-valued) continuous functions on $[a,b]$ is a metric space, under the metric

$$d(f,g) = \sup_{x \in [a,b]} |f(x) - g(x)|$$

We refer to this metric as the **sup metric**.

2) The set $C[a,b]$ of all real-valued (or complex-valued) continuous functions on $[a,b]$ is a metric space, under the metric

$$d_1(f(x), g(x)) = \int_{\mathrm{a}}^{b} |f(x) - g(x)| \, \mathrm{dx} \qquad \square$$

Example 12.4 Many important sequence spaces are metric spaces. We will often use boldface roman letters to denote sequences, as in $\boldsymbol{x} = (x_n)$ and $\boldsymbol{y} = (y_n)$.

1) The set $\ell^\infty_{\mathbb{R}}$ of all bounded sequences of real numbers is a metric space under the metric defined by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sup_n |x_n - y_n|$$

The set $\ell^\infty_{\mathbb{C}}$ of all bounded complex sequences, with the same metric, is also a metric space. As is customary, we will usually denote both of these spaces by ℓ^∞.

2) For $p \geq 1$, let ℓ^p be the set of all sequences $\boldsymbol{x} = (x_n)$ of real (or complex) numbers for which

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

We define the p**-norm** of $\boldsymbol{x}$ by

$$\|\boldsymbol{x}\|_p = \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p}$$

Then ℓ^p is a metric space, under the metric

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p = \left(\sum_{n=1}^{\infty} |x_n - y_n|^p\right)^{1/p}$$

The fact that ℓ^p is a metric follows from some rather famous results about sequences of real or complex numbers, whose proofs we leave as (well-hinted) exercises.

Hőlder's inequality Let $p, q \geq 1$ and $p + q = pq$. If $\boldsymbol{x} \in \ell^p$ and $\boldsymbol{y} \in \ell^q$ then the product sequence $\boldsymbol{xy} = (x_n y_n)$ is in ℓ^1 and

$$\|\boldsymbol{xy}\|_1 \leq \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q$$

that is,

$$\sum_{n=1}^{\infty} |x_n y_n| \leq \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p} \left(\sum_{n=1}^{\infty} |y_n|^q\right)^{1/q}$$

A special case of this (with $p = q = 2$) is the **Cauchy-Schwarz inequality**

$$\sum_{n=1}^{\infty} |x_n y_n| \leq \sqrt{\sum_{n=1}^{\infty} |x_n|^2} \sqrt{\sum_{n=1}^{\infty} |y_n|^2}$$

Minkowski's inequality For $p \geq 1$, if $\boldsymbol{x}, \boldsymbol{y} \in \ell^p$ then the sum $\boldsymbol{x} + \boldsymbol{y} = (x_n + y_n)$ is in ℓ^p and

$$\|\boldsymbol{x} + \boldsymbol{y}\|_p \leq \|\boldsymbol{x}\|_p + \|\boldsymbol{y}\|_p$$

that is,

$$\left(\sum_{n=1}^{\infty} |x_n + y_n|^p\right)^{1/p} \leq \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p} + \left(\sum_{n=1}^{\infty} |y_n|^p\right)^{1/p} \qquad \square$$

If M is a metric space under a metric d then any nonempty subset S of M is also a metric under the restriction of d to $S \times S$. The metric space S thus obtained is called a **subspace** of M.

Open and Closed Sets

Definition *Let M be a metric space. Let $x_0 \in M$ and let r be a positive real number.*

1) The **open ball** *centered at x_0, with radius r, is*

$$B(x_0, r) = \{x \in M \mid d(x, x_0) < r\}$$

2) The **closed ball** *centered at x_0, with radius r, is*

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

3) The **sphere** *centered at x_0, with radius r, is*

$$S(x_0, r) = \{x \in M \mid d(x, x_0) = r\} \qquad \square$$

Definition *A subset S of a metric space M is said to be* **open** *if each point of S is the center of an open ball that is contained completely in S. More specifically, S is open if for all $x \in S$, there exists an $r > 0$ such that $B(x, r) \subseteq S$. Note that the empty set is open. A set $T \subseteq M$ is* **closed** *if its complement T^c in M is open.* □

It is easy to show that an open ball is an open set and a closed ball is a closed set. If $x \in M$, we refer to any open set S containing x as an **open neighborhood** of x. It is also easy to see that a set is open if and only if it contains an open neighborhood of each of its points.

The next example shows that it is possible for a set to be both open and closed, or neither open nor closed.

Example 12.5 In the metric space $\mathbb{R}$ with the usual Euclidean metric, the open balls are just the open intervals

$$B(x_0, r) = (x_0 - r, x_0 + r)$$

and the closed balls are the closed intervals

$$\overline{B}(x_0, r) = [x_0 - r, x_0 + r]$$

Consider the half-open interval $S = (a, b]$, for a $< b$. This set is not open, since it contains no open ball centered at $b \in S$ and it is not closed, since its complement $S^c = (-\infty, a] \cup (b, \infty)$ is not open, since it contains no open ball about a.

Observe also that the empty set is both open and closed, as is the entire space $\mathbb{R}$. (Although we will not do so, it is possible to show that these are the only two sets that are both open and closed in $\mathbb{R}$.) □

It is not our intention to enter into a detailed discussion of open and closed sets, the subject of which belongs to the branch of mathematics known as *topology*. In order to put these concepts in perspective, however, we have the following result, whose proof is left to the reader.

Theorem 12.1 *The collection $\mathcal{O}$ of all open subsets of a metric space M has the following properties:*

1) $\emptyset \in \mathcal{O}$, $M \in \mathcal{O}$

2) If $S, T \in \mathcal{O}$ then $S \cap T \in \mathcal{O}$

3) If $\{S_i \mid i \in K\}$ is any collection of open sets then $\bigcup_{i \in K} S_i \in \mathcal{O}$. □

These three properties form the basis for an axiom system that is designed to generalize notions such as convergence and continuity and leads to the following definition.

Definition *Let X be a nonempty set. A collection $\mathcal{O}$ of subsets of X is called a* **topology** *for X if it has the following properties:*

1) $\emptyset \in \mathcal{O}$, $X \in \mathcal{O}$

2) If $S, T \in \mathcal{O}$ then $S \cap T \in \mathcal{O}$

3) If $\{S_i \mid i \in K\}$ is any collection of sets in $\mathcal{O}$ then $\bigcup\limits_{i \in K} S_i \in \mathcal{O}$.

We refer to subsets in $\mathcal{O}$ as **open sets** *and the pair $(X, \mathcal{O})$ as a* **topological space**. □

According to Theorem 12.1, the open sets (as we defined them earlier) in a metric space M form a topology for M, called the topology **induced** by the metric.

Topological spaces are the most general setting in which we can define concepts such as convergence and continuity, which is why these concepts are called topological concepts. However, since the topologies with which we will be dealing are induced by a metric, we will generally phrase the definitions of the topological properties that we will need directly in terms of the metric.

Convergence in a Metric Space

Convergence of sequences in a metric space is defined as follows.

Definition *A sequence (x_n) in a metric space M* **converges** *to $x \in M$, written $(x_n) \to x$, if*

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

Equivalently, $(x_n) \to x$ if for any $\epsilon > 0$, there exists an $N > 0$ such that

$$n > N \Rightarrow d(x_n, x) < \epsilon$$

or, equivalently

$$n > N \Rightarrow x_n \in B(x, \epsilon)$$

In this case, x is called the **limit** *of the sequence (x_n).* □

If M is a metric space and S is a subset of M, by a *sequence in S*, we mean a sequence whose terms all lie in S. We next characterize closed sets and therefore also open sets, using convergence.

Theorem 12.2 *Let M be a metric space. A subset $S \subseteq M$ is closed if and only if whenever (x_n) is a sequence in S and $(x_n) \rightarrow x$ then $x \in S$. In loose terms, a subset S is closed if it is closed under the taking of sequential limits.*
Proof. Suppose that S is closed and let $(x_n) \rightarrow x$, where $x_n \in S$ for all n. Suppose that $x \notin S$. Then since $x \in S^c$ and S^c is open, there exists an $\epsilon > 0$ for which $x \in B(x, \epsilon) \subseteq S^c$. But this implies that

$$B(x, \epsilon) \cap \{x_n\} = \emptyset$$

which contradicts the fact that $(x_n) \rightarrow x$. Hence, $x \in S$.

Conversely, suppose that S is closed under the taking of limits. We show that S^c is open. Let $x \in S^c$ and suppose to the contrary that no open ball about x is contained in S^c. Consider the open balls $B(x, 1/n)$, for all $n \geq 1$. Since none of these balls is contained in S^c, for each n, there is an $x_n \in S \cap B(x, 1/n)$. It is clear that $(x_n) \rightarrow x$ and so $x \in S$. But x cannot be in both S and S^c. This contradiction implies that S^c is open. Thus, S is closed. □

The Closure of a Set

Definition *Let S be any subset of a metric space M. The* **closure** *of S, denoted by* cl(S)*, is the smallest closed set containing S.* □

We should hasten to add that, since the entire space M is closed and since the intersection of any collection of closed sets is closed (exercise), the closure of any set S does exist and is the intersection of all closed sets containing S. The following definition will allow us to characterize the closure in another way.

Definition Let S be a nonempty subset of a metric space M. An element $x \in M$ is said to be a **limit point**, or **accumulation point** of S if every open ball centered at x meets S at a point other than x itself. Let us denote the set of all limit points of S by $\ell(S)$. □

Here are some key facts concerning limit points and closures.

Theorem 12.3 *Let S be a nonempty subset of a metric space M.*

1) $x \in \ell(S)$ if and only if there is a sequence (x_n) in S for which $x_n \neq x$ for all n and $(x_n) \to x$.

2) S is closed if and only if $\ell(S) \subseteq S$. In words, S is closed if and only if it contains all of its limit points.

3) $\mathrm{cl}(S) = S \cup \ell(S)$.

4) $x \in \mathrm{cl}(S)$ if and only if there is a sequence (x_n) in S for which $(x_n) \to x$.

Proof. For part 1), assume first that $x \in \ell(S)$. For each n, there exists a point $x_n \neq x$ such that $x_n \in B(x, 1/n) \cap S$. Thus, we have

$$d(x_n, x) < 1/n$$

and so $(x_n) \to x$. For the converse, suppose that $(x_n) \to x$, where $x \neq x_n \in S$. If $B(x,r)$ is any ball centered at x then there is some N such that $n > N$ implies $x_n \in B(x,r)$. Hence, for any ball $B(x,r)$ centered at x, there is a point $x_n \neq x$, such that $x_n \in S \cap B(x,r)$. Thus, x is a limit point of S.

As for part 2), if S is closed then by part 1), any $x \in \ell(S)$ is the limit of a sequence (x_n) in S and so must be in S. Hence, $\ell(S) \subseteq S$. Conversely, if $\ell(S) \subseteq S$ then S is closed. For if (x_n) is any sequence in S and $(x_n) \to x$ then there are two possibilities. First, we might have $x_n = x$ for some n, in which case $x = x_n \in S$. Second, we might have $x_n \neq x$ for all n, in which case $(x_n) \to x$ implies that $x \in \ell(S) \subseteq S$. In either case, $x \in S$ and so S is closed under the taking of limits, which implies that S is closed.

For part 3), let $T = S \cup \ell(S)$. Clearly, $S \subseteq T$. To show that T is closed, we show that it contains all of its limit points. So let $x \in \ell(T)$. Hence, there is a sequence $(x_n) \in T$ for which $x_n \neq x$ and $(x_n) \to x$. Of course, each x_n is either in S, or is a limit point of S. We must show that $x \in T$, that is, that x is either in S or is a limit point of S.

Suppose for the purposes of contradiction that $x \notin S$ and $x \notin \ell(S)$. Then there is a ball $B(x,r)$ for which $B(x,r) \cap S \neq \emptyset$. However, since $(x_n) \to x$, there must be an $x_n \in B(x,r)$. Since x_n cannot be in S, it must be a limit point of S. Referring to Figure 12.1, if $d(x_n, x) = d < r$ then consider the ball $B(x_n, (r-d)/2)$. This ball is completely contained in $B(x,r)$ and must contain an element y of S, since its center x_n is a limit point of S. But then $y \in S \cap B(x,r)$, a contradiction. Hence, $x \in S$ or $x \in \ell(S)$. In either case, $x \in T = S \cup \ell(S)$ and so T is closed.

Thus, T is closed and contains S and so $\mathrm{cl}(S) \subseteq T$. On the other hand, $T = S \cup \ell(S) \subseteq \mathrm{cl}(S)$ and so $\mathrm{cl}(S) = T$.

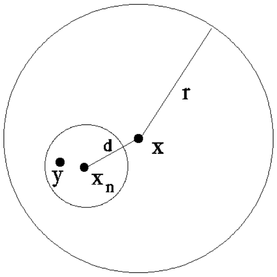

Figure 12.1

For part 4), if $x \in \mathrm{cl}(S)$ then there are two possibilities. If $x \in S$ then the constant sequence (x_n), with $x_n = x$ for all x, is a sequence in S that converges to x. If $x \notin S$ then $x \in \ell(S)$ and so there is a sequence (x_n) in S for which $x_n \neq x$ and $(x_n) \to x$. In either case, there is a sequence in S converging to x. Conversely, if there is a sequence (x_n) in S for which $(x_n) \to x$ then either $x_n = x$ for some n, in which case $x \in S \subseteq \mathrm{cl}(S)$, or else $x_n \neq x$ for all n, in which case $x \in \ell(S) \subseteq \mathrm{cl}(S)$. □

Dense Subsets

The following concept is meant to convey the idea of a subset $S \subseteq M$ being "arbitrarily close" to every point in M.

Definition *A subset S of a metric space M is* **dense** *in M if* $\mathrm{cl}(S) = M$. *A metric space is said to be* **separable** *if it contains a* countable *dense subset.* □

Thus, a subset S of M is dense if every open ball about any point $x \in M$ contains at least one point of S.

Certainly, any metric space contains a dense subset, namely, the space itself. However, as the next examples show, not every metric space contains a countable dense subset.

Example 12.6

1) The real line $\mathbb{R}$ is separable, since the rational numbers $\mathbb{Q}$ form a countable dense subset. Similarly, $\mathbb{R}^n$ is separable, since the set $\mathbb{Q}^n$ is countable and dense.
2) The complex plane $\mathbb{C}$ is separable, as is $\mathbb{C}^n$ for all n.
3) A discrete metric space is separable if and only if it is countable. We leave proof of this as an exercise. □

Example 12.7 The space ℓ^∞ is not separable. Recall that ℓ^∞ is the set of all bounded sequences of real numbers (or complex numbers), with metric

$$d(\boldsymbol{x},\boldsymbol{y}) = \sup_n |x_n - y_n|$$

To see that this space is not separable, consider the set S of all binary sequences

$$S = \{(x_n) \mid x_i = 0 \text{ or } 1 \text{ for all } i\}$$

This set is in one-to-one correspondence with the set of all subsets of $\mathbb{N}$ and so is uncountable. (It has cardinality $2^{\aleph_0} > \aleph_0$.) Now, each sequence in S is certainly bounded and so lies in ℓ^∞. Moreover, if $\boldsymbol{x} \neq \boldsymbol{y} \in \ell^\infty$ then the two sequences must differ in at least one position and so $d(x,y) = 1$.

In other words, we have a subset S of ℓ^∞ that is uncountable and for which the distance between any two distinct elements is 1. This implies that the uncountable collection of balls $\{B(s,1/3) \mid s \in S\}$ is mutually disjoint. Hence, no countable set can meet every ball, which implies that no countable set can be dense in ℓ^∞. □

Example 12.8 The metric spaces ℓ^p are separable, for $p \geq 1$. The set S of all sequences of the form

$$s = (q_1, \ldots, q_n, 0, \ldots)$$

for all $n > 0$, where the q_i's are rational, is a countable set. Let us show that it is dense in ℓ^p. Any $x \in \ell^p$ satisfies

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

Hence, for any $\epsilon > 0$, there exists an N such that

$$\sum_{n=N+1}^{\infty} |x_n|^p < \frac{\epsilon}{2}$$

Since the rational numbers are dense in $\mathbb{R}$, we can find rational numbers q_i for which

$$|x_i - q_i|^p < \frac{\epsilon}{2N}$$

for all $i = 1, \ldots, N$. Hence, if $s = (q_1, \ldots, q_N, 0, \ldots)$ then

$$d(x,s)^p = \sum_{n=1}^{N} |x_n - q_n|^p + \sum_{n=N+1}^{\infty} |x_n|^p < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

which shows that there is an element of S arbitrarily close to any element of ℓ^p. Thus, S is dense in ℓ^p and so ℓ^p is separable. □

Continuity

Continuity plays a central role in the study of linear operators on infinite-dimensional inner product spaces.

Definition *Let $f\colon M \to M'$ be a function from the metric space (M,d) to the metric space (M',d'). We say that f is* **continuous at** $x_0 \in M$ *if for any $\epsilon > 0$, there exists a $\delta > 0$ such that*

$$d(x,x_0) < \delta \Rightarrow d'(f(x),f(x_0)) < \epsilon$$

or, equivalently,

$$f\left(B(x_0,\delta)\right) \subseteq B(f(x_0),\epsilon)$$

(See Figure 12.2.) A function is **continuous** *if it is continuous at every $x_0 \in M$.* □

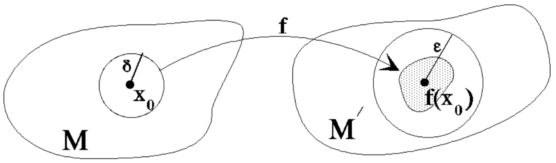

Figure 12.2

We can use the notion of convergence to characterize continuity for functions between metric spaces.

Theorem 12.4 *A function $f\colon M \to M'$ is continuous if and only if whenever (x_n) is a sequence in M that converges to $x_0 \in M$ then the sequence $(f(x_n))$ converges to $f(x_0)$, in short,*

$$(x_n) \to x_0 \Rightarrow (f(x_n)) \to f(x_0)$$

Proof. Suppose first that f is continuous at x_0 and let $(x_n) \to x_0$. Then, given $\epsilon > 0$, the continuity of f implies the existence of a $\delta > 0$ such that

$$f\left(B(x_0,\delta)\right) \subseteq B(f(x_0),\epsilon)$$

Since $(x_n) \to x$, there exists an $N > 0$ such that $x_n \in B(x_0,\delta)$ for $n > N$ and so

$$n > N \Rightarrow f(x_n) \in B(f(x_0),\epsilon)$$

Thus, $f(x_n) \to f(x_0)$.

Conversely, suppose that $(x_n) \to x_0$ implies $(f(x_n)) \to f(x_0)$. Suppose, for the purposes of contradiction, that f is not continuous at x_0. Then there exists an

$\epsilon > 0$ such that, for all $\delta > 0$

$$f\Big(B(x_0,\delta)\Big) \not\subseteq B(f(x_0),\epsilon)$$

Thus, for all $n > 0$,

$$f\left(B\left(x_0,\frac{1}{n}\right)\right) \not\subseteq B(f(x_0),\epsilon)$$

and so we may construct a sequence (x_n) by choosing each term x_n with the property that

$$x_n \in B\left(x_0,\frac{1}{n}\right), \text{ but } f(x_n) \notin B(f(x_0),\epsilon)$$

Hence, $(x_n) \to x_0$, but $f(x_n)$ does not converge to $f(x_0)$. This contradiction implies that f must be continuous at x_0. □

The next theorem says that the distance function is a continuous function in both variables.

Theorem 12.5 *Let (M,d) be a metric space. If $(x_n) \to x$ and $(y_n) \to y$ then $d(x_n,y_n) \to d(x,y)$.*
Proof. We leave it as an exercise to show that

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

But the right side tends to 0 as $n \to \infty$ and so $d(x_n,y_n) \to d(x,y)$. □

Completeness

The reader who has studied analysis will recognize the following definitions.

Definition *A sequence (x_n) in a metric space M is a* **Cauchy sequence** *if, for any $\epsilon > 0$, there exists an $N > 0$ for which*

$$n,m > N \Rightarrow d(x_n,x_m) < \epsilon \qquad \square$$

We leave it to the reader to show that any convergent sequence is a Cauchy sequence. When the converse holds, the space is said to be *complete*.

Definition *Let M be a metric space.*
1) *M is said to be* **complete** *if every Cauchy sequence in M converges in M.*
2) *A subspace S of M is* **complete** *if it is complete as a metric space. Thus, S is complete if every Cauchy sequence (s_n) in S converges to an element in S.* □

Before considering examples, we prove a very useful result about completeness of subspaces.

Theorem 12.6 *Let M be a metric space.*

1) *Any complete subspace of M is closed.*
2) *If M is complete then a subspace S of M is complete if and only if it is closed.*

Proof. To prove 1), assume that S is a complete subspace of M. Let (x_n) be a sequence in S for which $(x_n) \to x \in M$. Then (x_n) is a Cauchy sequence in S and since S is complete, (x_n) must converge to an element of S. Since limits of sequences are unique, we have $x \in S$. Hence, S is closed.

To prove part 2), first assume that S is complete. Then part 1) shows that S is closed. Conversely, suppose that S is closed and let (x_n) be a Cauchy sequence in S. Since (x_n) is also a Cauchy sequence in the complete space M, it must converge to some $x \in M$. But since S is closed, we have $(x_n) \to x \in S$. Hence, S is complete. □

Now let us consider some examples of complete (and incomplete) metric spaces.

Example 12.9 It is well known that the metric space $\mathbb{R}$ is complete. (However, a proof of this fact would lead us outside the scope of this book.) Similarly, the complex numbers $\mathbb{C}$ are complete. □

Example 12.10 The Euclidean space $\mathbb{R}^n$ and the unitary space $\mathbb{C}^n$ are complete. Let us prove this for $\mathbb{R}^n$. Suppose that (x_k) is a Cauchy sequence in $\mathbb{R}^n$, where

$$x_k = (x_{k,1}, \ldots, x_{k,n})$$

Thus,

$$d(x_k, x_m)^2 = \sum_{i=1}^{n} (x_{k,i} - x_{m,i})^2 \to 0 \text{ as } k, m \to \infty$$

and so, for each coordinate position i,

$$(x_{k,i} - x_{m,i})^2 \le d(x_k, x_m)^2 \to 0$$

which shows that the sequence $(x_{k,i})_{k=1,2,\ldots}$ of ith coordinates is a Cauchy sequence in $\mathbb{R}$. Since $\mathbb{R}$ is complete, we must have

$$(x_{k,i}) \to y_i \text{ as } k \to \infty$$

If $y = (y_1, \ldots, y_n)$ then

$$d(x_k, y)^2 = \sum_{i=1}^{n} (x_{k,i} - y_i)^2 \to 0 \text{ as } k \to \infty$$

and so $(x_n) \to y \in \mathbb{R}^n$. This proves that $\mathbb{R}^n$ is complete. □

Example 12.11 The metric space $(C[a,b],d)$ of all real-valued (or complex-valued) continuous functions on $[a,b]$, with metric

$$d(f,g) = \sup_{x\in[a,b]} |f(x)-g(x)|$$

is complete. To see this, we first observe that the limit with respect to d is the uniform limit on $[a,b]$, that is $d(f_n,f)\to 0$ if and only if for any $\epsilon > 0$, there is an $N > 0$ for which

$$n > N \ \Rightarrow |f_n(x)-f(x)| \le \epsilon \text{ for all } x \in [a,b]$$

Now, let (f_n) be a Cauchy sequence in $(C[a,b],d)$. Thus, for any $\epsilon > 0$, there is an N for which

$$m,n > N \Rightarrow |f_n(x)-f_m(x)| \le \epsilon \text{ for all } x \in [a,b] \tag{12.1}$$

This implies that, for each $x \in [a,b]$, the sequence $(f_n(x))$ is a Cauchy sequence of real (or complex) numbers and so it converges. We can therefore define a function f on $[a,b]$ by

$$f(x) = \lim_{n\to\infty} f_n(x)$$

Letting $m \to \infty$ in (12.1), we get

$$n > N \Rightarrow |f_n(x)-f(x)| \le \epsilon \text{ for all } x \in [a,b]$$

Thus, $f_n(x)$ converges to $f(x)$ uniformly. It is well known that the uniform limit of continuous functions is continuous and so $f(x) \in C[a,b]$. Thus, $(f_n(x)) \to f(x) \in C[a,b]$ and so $(C[a,b],d)$ is complete. □

Example 12.12 The metric space $(C[a,b],d_1)$ of all real-valued (or complex-valued) continuous functions on $[a,b]$, with metric

$$d_1(f(x),g(x)) = \int_{\mathrm{a}}^{b} |f(x)-g(x)|dx$$

is not complete. For convenience, we take $[a,b]=[0,1]$ and leave the general case for the reader. Consider the sequence of functions $f_n(x)$ whose graphs are shown in Figure 12.3. (The definition of $f_n(x)$ should be clear from the graph.)

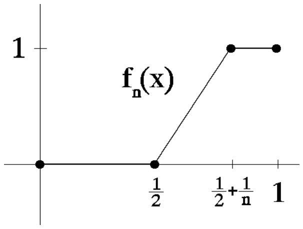

Figure 12.3

We leave it to the reader to show that the sequence $(f_n(x))$ is Cauchy, but does not converge in $(C[0,1],d_1)$. (The sequence converges to a function that is not continuous.)□

Example 12.13 The metric space ℓ^∞ is complete. To see this, suppose that (x_n) is a Cauchy sequence in ℓ^∞, where

$$x_n = (x_{n,1}, x_{n,2}, \dots)$$

Then, for each coordinate position i, we have

$$|x_{n,i} - x_{m,i}| \le \sup_j |x_{n,j} - x_{m,j}| \to 0 \text{ as } n, m \to \infty \tag{12.2}$$

Hence, for each i, the sequence $(x_{n,i})$ of ith coordinates is a Cauchy sequence in $\mathbb{R}$ (or $\mathbb{C}$). Since $\mathbb{R}$ (or $\mathbb{C}$) is complete, we have

$$(x_{n,i}) \to y_i \text{ as } n \to \infty$$

for each coordinate position i. We want to show that $y = (y_i) \in \ell^\infty$ and that $(x_n) \to y$.

Letting $m \to \infty$ in (12.2) gives

$$\sup_j |x_{n,j} - y_j| \to 0 \text{ as } n \to \infty \tag{12.3}$$

and so, for some n,

$$|x_{n,j} - y_j| < 1 \text{ for all } j$$

and so

$$|y_j| < 1 + |x_{n,j}| \text{ for all } j$$

But since $x_n \in \ell^\infty$, it is a bounded sequence and therefore so is (y_j). That is, $y = (y_j) \in \ell^\infty$. Since (12.3) implies that $(x_n) \to y$, we see that ℓ^∞ is complete. □

Example 12.14 The metric space ℓ^p is complete. To prove this, let (x_n) be a Cauchy sequence in ℓ^p, where

$$x_n = (x_{n,1}, x_{n,2}, \dots)$$

Then, for each coordinate position i,

$$|x_{n,i} - x_{m,i}|^p \le \sum_{j=1}^{\infty} |x_{n,j} - x_{m,j}|^p = d(x_n, x_m)^p \to 0$$

which shows that the sequence $(x_{n,i})$ of ith coordinates is a Cauchy sequence in $\mathbb{R}$ (or $\mathbb{C}$). Since $\mathbb{R}$ (or $\mathbb{C}$) is complete, we have

$$(x_{n,i}) \to y_i \text{ as } n \to \infty$$

We want to show that $y = (y_i) \in \ell^p$ and that $(x_n) \to y$.

To this end, observe that for any $\epsilon > 0$, there is an N for which

$$n, m > N \Rightarrow \sum_{i=1}^{r} |x_{n,i} - x_{m,i}|^p \le \epsilon$$

for all $r > 0$. Now, we let $m \to \infty$, to get

$$n > N \Rightarrow \sum_{i=1}^{r} |x_{n,i} - y_i|^p \le \epsilon$$

for all $r > 0$. Letting $r \to \infty$, we get, for any $n > N$,

$$\sum_{i=1}^{\infty} |x_{n,i} - y_i|^p < \epsilon$$

which implies that $(x_n) - y \in \ell^p$ and so $y = y - (x_n) + (x_n) \in \ell^p$ and in addition, $(x_n) \to y$. □

As we will see in the next chapter, the property of completeness plays a major role in the theory of inner product spaces. Inner product spaces for which the induced metric space is complete are called **Hilbert spaces**.

Isometries

A function between two metric spaces that preserves distance is called an isometry. Here is the formal definition.

Definition Let (M, d) and (M', d') be metric spaces. A function $f\colon M \to M'$ is called an **isometry** if

$$d'(f(x), f(y)) = d(x, y)$$

for all $x, y \in M$. If $f: M \to M'$ is a bijective isometry from M to M', we say that M and M' are **isometric** and write $M \approx M'$. □

Theorem 12.7 *Let $f: (M, d) \to (M', d')$ be an isometry. Then*
1) f is injective
2) f is continuous
3) $f^{-1}: f(M) \to M$ is also an isometry and hence also continuous.
Proof. To prove 1), we observe that

$$f(x) = f(y) \Leftrightarrow d'(f(x), f(y)) = 0 \Leftrightarrow d(x, y) = 0 \Leftrightarrow x = y$$

To prove 2), let $(x_n) \to x$ in M then

$$d'(f(x_n), f(x)) = d(x_n, x) \to 0 \text{ as } n \to \infty$$

and so $(f(x_n)) \to f(x)$, which proves that f is continuous. Finally, we have

$$d(f^{-1}(f(x)), f^{-1}(f(y)) = d(x, y) = d'(f(x), f(y))$$

and so $f^{-1}: f(M) \to M$ is an isometry. □

The Completion of a Metric Space

While not all metric spaces are complete, any metric space can be embedded in a complete metric space. To be more specific, we have the following important theorem.

Theorem 12.8 *Let (M, d) be any metric space. Then there is a complete metric space (M', d') and an isometry $\tau: M \to \tau(M) \subseteq M'$ for which $\tau(M)$ is dense in M'. The metric space (M', d') is called a* **completion** *of (M, d). Moreover, (M', d') is unique, up to bijective isometry.*
Proof. The proof is a bit lengthy, so we divide it into various parts. We can simplify the notation considerably by thinking of sequences (x_n) in M as functions $f: \mathbb{N} \to M$, where $f(n) = x_n$.

Cauchy Sequences in M

The basic idea is to let the elements of M' be equivalence classes of Cauchy sequences in M. So let CS(M) denote the set of all Cauchy sequences in M. If $f, g \in$ CS(M) then, intuitively speaking, the terms $f(n)$ get closer together as $n \to \infty$ and so do the terms $g(n)$. Therefore, it seems reasonable that $d(f(n), g(n))$ should approach a finite limit as $n \to \infty$. Indeed, since

$$|d(f(n), g(n)) - d(f(m), g(m))| \le d(f(n), f(m)) + d(g(n), g(m)) \to 0$$

as $n, m \to \infty$ it follows that $d(f(n), g(n))$ is a Cauchy sequence of real numbers, which implies that

$$\lim_{n\to\infty} d(f(n),g(n)) < \infty \tag{12.4}$$

(That is, the limit exists and is finite.)

Equivalence Classes of Cauchy Sequences in M

We would like to define a metric d' on the set CS(M) by

$$d'(f,g) = \lim_{n\to\infty} d(f(n),g(n))$$

However, it is possible that

$$\lim_{n\to\infty} d(f(n),g(n)) = 0$$

for distinct sequences f and g, so this does not define a metric. Thus, we are led to define an equivalence relation on CS(M) by

$$f \sim g \Leftrightarrow \lim_{n\to\infty} d(f(n),g(n)) = 0$$

Let $\overline{\text{CS}(M)}$ be the set of all equivalence classes of Cauchy sequences and define, for $\overline{f},\ \overline{g} \in \overline{\text{CS}(M)}$

$$d'(\overline{f},\overline{g}) = \lim_{n\to\infty} d(f(n),g(n)) \tag{12.5}$$

where $f \in \overline{f}$ and $g \in \overline{g}$.

To see that d' is well-defined, suppose that $f' \in \overline{f}$ and $g' \in \overline{g}$. Then since $f' \sim f$ and $g' \sim g$, we have

$$|d(f'(n),g'(n)) - d(f(n),g(n))| \le d(f'(n),f(n)) + d(g'(n),g(n)) \to 0$$

as $n \to \infty$. Thus,

$$\begin{aligned} f' \sim f \text{ and } g' \sim g &\Rightarrow \lim_{n\to\infty} d(f'(n),g'(n)) = \lim_{n\to\infty} d(f(n),g(n)) \\ &\Rightarrow\ d'(f',g') = d'(f,g) \end{aligned}$$

which shows that d' is well-defined. To see that d' is a metric, we verify the triangle inequality, leaving the rest to the reader. If f,g and h are Cauchy sequences then

$$d(f(n),g(n)) \le d(f(n),h(n)) + d(h(n),g(n))$$

Taking limits gives

$$\lim_{n\to\infty} d(f(n),g(n)) \le \lim_{n\to\infty} d(f(n),h(n)) + \lim_{n\to\infty} d(h(n),g(n))$$

and so

$$d'(\overline{f},\overline{g}) \le d'(\overline{f},\overline{h}) + d'(\overline{h},\overline{g})$$

Embedding (M,d) in (M',d')

For each $x \in M$, consider the constant Cauchy sequence $[x]$, where $[x](n) = x$ for all n. The map $\tau: M \to M'$ defined by

$$\tau(x) = \overline{[x]}$$

is an isometry, since

$$d'(\tau(x),\tau(y)) = d'(\overline{[x]},\overline{[y]}) = \lim_{n\to\infty} d([x](n),[y](n)) = d(x,y)$$

Moreover, $\tau(M)$ is dense in M'. This follows from the fact that we can approximate any Cauchy sequence in M by a constant sequence. In particular, let $\overline{f} \in M'$. Since $f \in \overline{f}$ is a Cauchy sequence, for any $\epsilon > 0$, there exists an N such that

$$n,m \ge N \Rightarrow d(f(n),f(m)) < \epsilon$$

Now, for the constant sequence $[f(N)]$ we have

$$d'\left(\overline{[f(N)]},\overline{f}\right) = \lim_{n\to\infty} d(f(N),f(n)) \le \epsilon$$

and so $\tau(M)$ is dense in M'.

(M',d') Is Complete

Suppose that

$$\overline{f_1},\overline{f_2},\overline{f_3},\ \ldots$$

is a Cauchy sequence in M'. We wish to find a Cauchy sequence g in M for which

$$d'(\overline{f_k},\overline{g}) = \lim_{n\to\infty} d(f_k(n),g(n)) \to 0 \text{ as } k \to \infty$$

Since $\overline{f_k} \in M'$ and since $\tau(M)$ is dense in M', there is a constant sequence

$$[c_k] = (c_k,c_k,\ldots)$$

for which

$$d'(\overline{f_k},\overline{[c_k]}) < \frac{1}{k}$$

We can think of c_k as a constant approximation to f_k, with error at most $1/k$. Let g be the sequence of these constant approximations

$$g(k) = c_k$$

This is a Cauchy sequence in M. Intuitively speaking, since the f_k's get closer to each other as $k \to \infty$, so do the constant approximations. In particular, we have

$$\begin{aligned} d(c_k, c_j) &= d'(\overline{[c_k]}, \overline{[c_j]}) \\ &\le d'(\overline{[c_k]}, \overline{f_k}) + d'(\overline{f_k}, \overline{f_j}) + d'(\overline{f_j}, \overline{[c_j]}) \\ &\le \frac{1}{k} + d'(\overline{f_k}, \overline{f_j}) + \frac{1}{j} \to 0 \end{aligned}$$

as $k, j \to \infty$. To see that $\overline{f_k}$ converges to $\overline{g}$, observe that

$$\begin{aligned} d'(\overline{f_k}, \overline{g}) &\le d'(\overline{f_k}, \overline{[c_k]}) + d'(\overline{[c_k]}, \overline{g}) < \frac{1}{k} + \lim_{n\to\infty} d(c_k, g(n)) \\ &= \frac{1}{k} + \lim_{n\to\infty} d(c_k, c_n) \end{aligned}$$

Now, since g is a Cauchy sequence, for any $\epsilon > 0$, there is an N such that

$$k, n \ge N \Rightarrow d(c_k, c_n) < \epsilon$$

In particular,

$$k \ge N \Rightarrow \lim_{n\to\infty} d(c_k, c_n) \le \epsilon$$

and so

$$k \ge N \Rightarrow d'(\overline{f_k}, \overline{g}) \le \frac{1}{k} + \epsilon$$

which implies that $\overline{f_k} \to g$, as desired.

Uniqueness

Finally, we must show that if (M', d') and (M'', d'') are both completions of (M, d) then $M' \approx M''$. Note that we have bijective isometries

$$\tau\colon M \to \tau(M) \subseteq M' \text{ and } \sigma\colon M \to \sigma(M) \subseteq M''$$

Hence, the map

$$\rho = \sigma\tau^{-1}\colon \tau(M) \to \sigma(M)$$

is a bijective isometry from $\tau(M)$ onto $\sigma(M)$, where $\tau(M)$ is dense in M'. (See Figure 12.4.)

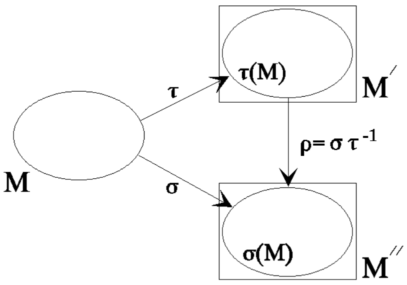

Figure 12.4

Our goal is to show that ρ can be extended to a bijective isometry $\overline{\rho}$ from M' to M''.

Let $x \in M'$. Then there is a sequence (a_n) in $\tau(M)$ for which $(a_n) \to x$. Since (a_n) is a Cauchy sequence in $\tau(M)$, $(\rho(a_n))$ is a Cauchy sequence in $\sigma(M) \subseteq M''$ and since M'' is complete, we have $(\rho(a_n)) \to y$ for some $y \in M''$. Let us define $\overline{\rho}(x) = y$.

To see that $\overline{\rho}$ is well-defined, suppose that $(a_n) \to x$ and $(b_n) \to x$, where both sequences lie in $\tau(M)$. Then

$$d''(\rho(a_n), \rho(b_n)) = d'(a_n, b_n) \to 0 \text{ as } n \to \infty$$

and so $(\rho(a_n))$ and $(\rho(b_n))$ converge to the same element of M'', which implies that $\overline{\rho}(x)$ does not depend on the choice of sequence in $\tau(M)$ converging to x. Thus, $\overline{\rho}$ is well-defined. Moreover, if $a \in \tau(M)$ then the constant sequence $[a]$ converges to a and so $\overline{\rho}(a) = \lim \rho(a) = \rho(a)$, which shows that $\overline{\rho}$ is an extension of ρ.

To see that $\overline{\rho}$ is an isometry, suppose that $(a_n) \to x$ and $(b_n) \to y$. Then $(\rho(a_n)) \to \overline{\rho}(x)$ and $(\rho(b_n)) \to \overline{\rho}(y)$ and since d'' is continuous, we have

$$d''(\overline{\rho}(x), \overline{\rho}(y)) = \lim_{n\to\infty} d''(\rho(a_n), \rho(b_n)) = \lim_{n\to\infty} d'(a_n, b_n) = d'(x, y)$$

Thus, we need only show that $\overline{\rho}$ is surjective. Note first that $\sigma(M) = \text{im}(\rho) \subseteq \text{im}(\overline{\rho})$. Thus, if $\text{im}(\overline{\rho})$ is closed, we can deduce from the fact that $\sigma(M)$ is dense in M'' that $\text{im}(\overline{\rho}) = M''$. So, suppose that $(\overline{\rho}(x_n))$ is a sequence in $\text{im}(\overline{\rho})$ and $(\overline{\rho}(x_n)) \to z$. Then $(\overline{\rho}(x_n))$ is a Cauchy sequence and therefore so is (x_n). Thus, $(x_n) \to x \in M'$. But $\overline{\rho}$ is continuous and so $(\overline{\rho}(x_n)) \to \overline{\rho}(x)$, which implies that $\overline{\rho}(x) = z$ and so $z \in \text{im}(\overline{\rho})$. Hence, $\overline{\rho}$ is surjective and $M' \approx M''$. □

Exercises

1. Prove the generalized triangle inequality

$$d(x_1, x_n) \leq d(x_1, x_2) + d(x_2, x_3) + \cdots + d(x_{n-1}, x_n)$$

2. a) Use the triangle inequality to prove that

$$|d(x, y) - d(a, b)| \leq d(x, a) + d(y, b)$$

   b) Prove that

$$|d(x, z) - d(y, z)| \leq d(x, y)$$

3. Let $S \subseteq \ell^\infty$ be the subspace of all binary sequences (sequences of 0's and 1's). Describe the metric on S.
4. Let $M = \{0,1\}^n$ be the set of all binary n-tuples. Define a function $h: S \times S \to \mathbb{R}$ by letting $h(x, y)$ be the number of positions in which x and y differ. For example, $h[(11010), (01001)] = 3$. Prove that h is a metric. (It is called the **Hamming distance function** and plays an important role in the theory of error-correcting codes.)
5. Let $1 \leq p < \infty$.
   a) If $\boldsymbol{x} = (x_n) \in \ell^p$ show that $x_n \to 0$
   b) Find a sequence that converges to 0 but is not an element of any ℓ^p for $1 \leq p < \infty$.
6. a) Show that if $\boldsymbol{x} = (x_n) \in \ell^p$ then $\boldsymbol{x} \in \ell^q$ for all $q > p$.
   b) Find a sequence $\boldsymbol{x} = (x_n)$ that is in ℓ^p for $p > 1$, but is not in ℓ^1.
7. Show that a subset S of a metric space M is open if and only if S contains an open neighborhood of each of its points.
8. Show that the intersection of any collection of closed sets in a metric space is closed.
9. Let (M, d) be a metric space. The **diameter** of a nonempty subset $S \subseteq M$ is

$$\delta(S) = \sup_{x,y \in S} d(x, y)$$

   A set S is **bounded** if $\delta(S) < \infty$.
   a) Prove that S is bounded if and only if there is some $x \in M$ and $r \in \mathbb{R}$ for which $S \subseteq B(x, r)$.
   b) Prove that $\delta(S) = 0$ if and only if S consists of a single point.
   c) Prove that $S \subseteq T$ implies $\delta(S) \leq \delta(T)$.
   d) If S and T are bounded, show that $S \cup T$ is also bounded.
10. Let (M, d) be a metric space. Let d' be the function defined by

$$d'(x, y) = \frac{d(x, y)}{1 + d(x, y)}$$

a) Show that (M,d') is a metric space and that M is bounded under this metric, even if it is not bounded under the metric d.

b) Show that the metric spaces (M,d) and (M,d') have the same open sets.

11. If S and T are subsets of a metric space (M,d), we define the **distance** between S and T by

$$\rho(S,T) = \inf_{x \in S, t \in T} d(x,y)$$

a) Is it true that $\rho(S,T) = 0$ if and only if $S = T$? Is ρ a metric?

b) Show that $x \in \mathrm{cl}(S)$ if and only if $\rho(\{x\},S) = 0$.

12. Prove that $x \in M$ is a limit point of $S \subseteq M$ if and only if every neighborhood of x meets S in a point other than x itself.
13. Prove that $x \in M$ is a limit point of $S \subseteq M$ if and only if every open ball $B(x,r)$ contains infinitely many points of S.
14. Prove that limits are unique, that is, $(x_n) \to x$, $(x_n) \to y$ implies that $x = y$.
15. Let S be a subset of a metric space M. Prove that $x \in \mathrm{cl}(S)$ if and only if there exists a sequence (x_n) in S that converges to x.
16. Prove that the closure has the following properties:

a) $S \subseteq \mathrm{cl}(S)$

b) $\mathrm{cl}(\mathrm{cl}(S)) = S$

c) $\mathrm{cl}(S \cup T) = \mathrm{cl}(S) \cup \mathrm{cl}(T)$

d) $\mathrm{cl}(S \cap T) \subseteq \mathrm{cl}(S) \cap \mathrm{cl}(T)$

Can the last part be strengthened to equality?

17. a) Prove that the closed ball $\overline{B}(x,r)$ is always a closed subset.

b) Find an example of a metric space in which the closure of an open ball $B(x,r)$ is not equal to the closed ball $\overline{B}(x,r)$.

18. Provide the details to show that $\mathbb{R}^n$ is separable.
19. Prove that $\mathbb{C}^n$ is separable.
20. Prove that a discrete metric space is separable if and only if it is countable.
21. Prove that the metric space $\mathcal{B}[a,b]$ of all bounded functions on $[a,b]$, with metric

$$d(f,g) = \sup_{x \in [a,b]} |f(x) - g(x)|$$

is not separable.

22. Show that a function $f\colon (M,d) \to (M',d')$ is continuous if and only if the inverse image of any open set is open, that is, if and only if $f^{-1}(U) = \{x \in M \mid f(x) \in U\}$ is open in M whenever U is an open set in M'.
23. Repeat the previous exercise, replacing the word open by the word closed.
24. Give an example to show that if $f\colon (M,d) \to (M',d')$ is a continuous function and U is an open set in M, it need not be the case that $f(U)$ is open in M'.

25. Show that any convergent sequence is a Cauchy sequence.
26. If $(x_n) \to x$ in a metric space M, show that any subsequence (x_{n_k}) of (x_n) also converges to x.
27. Suppose that (x_n) is a Cauchy sequence in a metric space M and that some subsequence (x_{n_k}) of (x_n) converges. Prove that (x_n) converges to the same limit as the subsequence.
28. Prove that if (x_n) is a Cauchy sequence then the set $\{x_n\}$ is bounded. What about the converse? Is a bounded sequence necessarily a Cauchy sequence?
29. Let (x_n) and (y_n) be Cauchy sequences in a metric space M. Prove that the sequence $d_n = d(x_n, y_n)$ converges.
30. Show that the space of all convergent sequences of real numbers (or complex numbers) is complete as a subspace of ℓ^∞.
31. Let $\mathcal{P}$ denote the metric space of all polynomials over $\mathbb{C}$, with metric

$$d(p,q) = \sup_{x \in [a,b]} |p(x) - q(x)|$$

Is $\mathcal{P}$ complete?
32. Let $S \subseteq \ell^\infty$ be the subspace of all sequences with finite support (that is, with a finite number of nonzero terms). Is S complete?
33. Prove that the metric space $\mathbb{Z}$ of all integers, with metric $d(n,m) = |n - m|$, is complete.
34. Show that the subspace S of the metric space $C[a,b]$ (under the sup metric) consisting of all functions $f \in C[a,b]$ for which $f(a) = f(b)$ is complete.
35. If $M \approx M'$ and M is complete, show that M' is also complete.
36. Show that the metric spaces $C[a,b]$ and $C[c,d]$, under the sup metric, are isometric.
37. Prove Hölder's inequality

$$\sum_{n=1}^{\infty} |x_n y_n| \le \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{1/p} \left(\sum_{n=1}^{\infty} |y_n|^q\right)^{1/q}$$

as follows.

a) Show that $s = t^{p-1} \Rightarrow t = s^{q-1}$

b) Let u and v be positive real numbers and consider the rectangle R in $\mathbb{R}^2$ with corners $(0,0)$, $(u,0)$, $(0,v)$ and (u,v), with area uv. Argue geometrically (that is, draw a picture) to show that

$$uv \le \int_0^u t^{p-1} dt + \int_0^v s^{q-1} ds$$

and so

$$uv \le \frac{u^p}{p} + \frac{v^q}{q}$$

c) Now let $X = \Sigma |x_n|^p < \infty$ and $Y = \Sigma |y_n|^q < \infty$. Apply the results of part b), to

$$u = \frac{|x_n|}{X^{1/p}}, \quad v = \frac{|y_n|}{Y^{1/q}}$$

and then sum on n to deduce Hölder's inequality.

38. Prove Minkowski's inequality

$$\left(\sum_{n=1}^{\infty}|x_n + y_n|^p\right)^{1/p} \le \left(\sum_{n=1}^{\infty}|x_n|^p\right)^{1/p} + \left(\sum_{n=1}^{\infty}|y_n|^p\right)^{1/p}$$

as follows.

a) Prove it for $p = 1$ first.

b) Assume $p > 1$. Show that

$$|x_n + y_n|^p \le |x_n||x_n + y_n|^{p-1} + |y_n||x_n + y_n|^{p-1}$$

c) Sum this from $n = 1$ to k and apply Hölder's inequality to each sum on the right, to get

$$\begin{aligned}&\sum_{n=1}^{k}|x_n + y_n|^p \\ &\quad\le \left\{\left(\sum_{n=1}^{k}|x_n|^p\right)^{1/p} + \left(\sum_{n=1}^{k}|y_n|^p\right)^{1/p}\right\}\left(\sum_{n=1}^{k}|x_n + y_n|^p\right)^{1/q}\end{aligned}$$

Divide both sides of this by the last factor on the right and let $n \to \infty$ to deduce Minkowski's inequality.

39. Prove that ℓ^p is a metric space.

Chapter 13
Hilbert Spaces

Now that we have the necessary background on the topological properties of metric spaces, we can resume our study of inner product spaces without qualification as to dimension. As in Chapter 9, we restrict attention to real and complex inner product spaces. Hence F will denote either $\mathbb{R}$ or $\mathbb{C}$.

A Brief Review

Let us begin by reviewing some of the results from Chapter 9. Recall that an inner product space V over F is a vector space V, together with an inner product $\langle , \rangle : V \times V \to F$. If $F = \mathbb{R}$ then the inner product is bilinear and if $F = \mathbb{C}$, the inner product is sesquilinear.

An inner product induces a norm on V, defined by

$$\|v\| = \sqrt{\langle v, v \rangle}$$

We recall in particular the following properties of the norm.

Theorem 13.1

1) (**The Cauchy-Schwarz inequality**) *For all $u, v \in V$,*

$$|\langle u, v \rangle| \leq \|u\| \; \|v\|$$

with equality if and only if $u = rv$ for some $r \in F$.

2) (**The triangle inequality**) *For all $u, v \in V$,*

$$\|u + v\| \leq \|u\| + \|v\|$$

with equality if and only if $u = rv$ for some $r \in F$.

3) (**The parallelogram law**)

$$\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2 \qquad \square$$

We have seen that the inner product can be recovered from the norm, as follows.

Theorem 13.2

1) *If* V *is a real inner product space then*

$$\langle u,v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2)$$

2) *If* V *is a complex inner product space then*

$$\langle u,v\rangle = \frac{1}{4}(\|u+v\|^2 - \|u-v\|^2) + \frac{1}{4}i(\|u+iv\|^2 - \|u-iv\|^2) \quad \square$$

The inner product also induces a metric on V defined by

$$d(u,v) = \|u-v\|$$

Thus, any inner product space is a metric space.

Definition *Let* V *and* W *be inner product spaces and let* $\tau \in \mathcal{L}(V,W)$.

1) τ *is an* **isometry** *if it preserves the inner product, that is, if*

$$\langle \tau(u), \tau(v)\rangle = \langle u,v\rangle$$

for all $u,v \in V$.

2) *A bijective isometry is called an* **isometric isomorphism**. *When* $\tau: V \to W$ *is an isometric isomorphism, we say that* V *and* W *are* **isometrically isomorphic**. □

It is easy to see that an isometry is always injective but need not be surjective, even if $V = W$. (See Example 10.3.)

Theorem 13.3 *A linear transformation* $\tau \in \mathcal{L}(V,W)$ *is an isometry if and only if it preserves the norm, that is, if and only if*

$$\|\tau(v)\| = \|v\|$$

for all $v \in V$. □

The following result points out one of the main differences between real and complex inner product spaces.

Theorem 13.4 *Let* V *be an inner product space and let* $\tau \in \mathcal{L}(V)$.

1) *If* $\langle \tau(v), w\rangle = 0$ *for all* $v,\ w \in V$ *then* $\tau = 0$.
2) *If* V *is a complex inner product space and* $Q_\tau(v) = \langle \tau(v), v\rangle = 0$ *for all* $v \in V$ *then* $\tau = 0$.
3) *Part 2) does not hold in general for real inner product spaces.* □

Hilbert Spaces

Since an inner product space is a metric space, all that we learned about metric spaces applies to inner product spaces. In particular, if (x_n) is a sequence of vectors in an inner product space V then

$$(x_n) \to x \text{ if and only if } \|x_n - x\| \to 0 \text{ as } n \to \infty$$

The fact that the inner product is continuous as a function of either of its coordinates is extremely useful.

Theorem 13.5 *Let V be an inner product space. Then*
1) $(x_n) \to x,\ (y_n) \to y \Rightarrow \langle x_n, y_n \rangle \to \langle x, y \rangle$
2) $(x_n) \to x \Rightarrow \|x_n\| \to \|x\|$ □

Complete inner product spaces play an especially important role in both theory and practice.

Definition *An inner product space that is complete under the metric induced by the inner product is said to be a* **Hilbert space***.* □

Example 13.1 One of the most important examples of a Hilbert space is the space ℓ^2 of Example 10.2. Recall that the inner product is defined by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{n=1}^{\infty} x_n \overline{y}_n$$

(In the real case, the conjugate is unnecessary.) The metric induced by this inner product is

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2 = \left(\sum_{n=1}^{\infty} |x_n - y_n|^2 \right)^{1/2}$$

which agrees with the definition of the metric space ℓ^2 given in Chapter 12. In other words, the metric in Chapter 12 is induced by this inner product. As we saw in Chapter 12, this inner product space is complete and so it is a Hilbert space. (In fact, it is the prototype of all Hilbert spaces, introduced by David Hilbert in 1912, even before the axiomatic definition of Hilbert space was given by John von Neumann in 1927.) □

The previous example raises the question of whether or not the other metric spaces ℓ^p $(p \neq 2)$, with distance given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_p = \left(\sum_{n=1}^{\infty} |x_n - y_n|^p \right)^{1/p} \tag{13.1}$$

are complete inner product spaces. The fact is that they are not even inner product spaces! More specifically, there is no inner product whose induced metric is given by (13.1). To see this, observe that, according to Theorem 13.1, any norm that comes from an inner product must satisfy the parallelogram law

$$\|\boldsymbol{x}+\boldsymbol{y}\|^2+\|\boldsymbol{x}-\boldsymbol{y}\|^2=2\|\boldsymbol{x}\|^2+2\|\boldsymbol{y}\|^2$$

But the norm in (13.1) does not satisfy this law. To see this, take $\boldsymbol{x}=(1,1,0\ldots)$ and $\boldsymbol{y}=(1,-1,0\ldots)$. Then

$$\|\boldsymbol{x}+\boldsymbol{y}\|_p=2,\ \|\boldsymbol{x}-\boldsymbol{y}\|_p=2$$

and

$$\|\boldsymbol{x}\|_p=2^{1/p},\ \|\boldsymbol{y}\|_p=2^{1/p}$$

Thus, the left side of the parallelogram law is 8 and the right side is $4\cdot 2^{2/p}$, which equals 8 if and only if $p=2$.

Just as any metric space has a completion, so does any inner product space.

Theorem 13.6 *Let V be an inner product space. Then there exists a Hilbert space H and an isometry $\tau\colon V\to H$ for which $\tau(V)$ is dense in H. Moreover, H is unique up to isometric isomorphism.*

Proof. We know that the metric space (V,d), where d is induced by the inner product, has a unique completion (V',d'), which consists of equivalence classes of Cauchy sequences in V. If $(x_n)\in\overline{(x_n)}\in V'$ and $(y_n)\in\overline{(y_n)}\in V'$ then we set

$$\overline{(x_n)}+\overline{(y_n)}=\overline{(x_n+y_n)},\ r\overline{(x_n)}=\overline{(rx_n)}$$

and

$$\langle\overline{(x_n)},\overline{(y_n)}\rangle=\lim_{n\to\infty}\langle x_n,y_n\rangle$$

It is easy to see that, since (x_n) and (y_n) are Cauchy sequences, so are (x_n+y_n) and (rx_n). In addition, these definitions are well-defined, that is, they are independent of the choice of representative from each equivalence class. For instance, if $(\widehat{x}_n)\in\overline{(x_n)}$ then

$$\lim_{n\to\infty}\|x_n-\widehat{x}_n\|=0$$

and so

$$|\langle x_n,y_n\rangle-\langle\widehat{x}_n,y_n\rangle|=|\langle x_n-\widehat{x}_n,y_n\rangle|\le\|x_n-\widehat{x}_n\|\|y_n\|\to 0$$

(The Cauchy sequence (y_n) is bounded.) Hence,

$$\langle\overline{(x_n)},\overline{(y_n)}\rangle=\lim_{n\to\infty}\langle x_n,y_n\rangle=\lim_{n\to\infty}\langle\widehat{x}_n,y_n\rangle=\langle\overline{(\widehat{x}_n)},\overline{(y_n)}\rangle$$

We leave it to the reader to show that V' is an inner product space under these operations.

Moreover, the inner product on V' induces the metric d', since

$$\begin{aligned}\langle(\overline{x_n - y_n}), (\overline{x_n - y_n})\rangle &= \lim_{n\to\infty}\langle x_n - y_n, x_n - y_n\rangle \\ &= \lim_{n\to\infty} d(x_n, y_n)^2 \\ &= d'((x_n), (y_n))^2\end{aligned}$$

Hence, the metric space isometry $\tau\colon V \to V'$ is an isometry of inner product spaces, since

$$\langle\tau(x), \tau(y)\rangle = d'(\tau(x), \tau(y))^2 = d(x, y)^2 = \langle x, y\rangle$$

Thus, V' is a complete inner product space and $\tau(V)$ is a dense subspace of V' that is isometrically isomorphic to V. We leave the issue of uniqueness to the reader. □

The next result concerns subspaces of inner product spaces.

Theorem 13.7

1) *Any complete subspace of an inner product space is closed.*
2) *A subspace of a Hilbert space is a Hilbert space if and only if it is closed.*
3) *Any finite-dimensional subspace of an inner product space is closed and complete.*

Proof. Parts 1) and 2) follow from Theorem 12.6. Let us prove that a finite-dimensional subspace S of an inner product space V is closed. Suppose that (x_n) is a sequence in S, $(x_n) \to x$ and $x \notin S$. Let $\mathcal{B} = \{b_1, \ldots, b_m\}$ be an orthonormal Hamel basis for S. The Fourier expansion

$$s = \sum_{i=1}^{m}\langle x, b_i\rangle b_i$$

in S has the property that $x - s \neq 0$ but

$$\langle x - s, b_j\rangle = \langle x, b_j\rangle - \langle s, b_j\rangle = 0$$

Thus, if we write $y = x - s$ and $y_n = x_n - s \in S$, the sequence (y_n), which is in S, converges to a vector y that is orthogonal to S. But this is impossible, because $y_n \perp y$ implies that

$$\|y_n - y\|^2 = \|y_n\|^2 + \|y\|^2 \geq \|y\|^2 \not\to 0$$

This proves that S is closed.

To see that any finite-dimensional subspace S of an inner product space is complete, let us embed S (as an inner product space in its own right) in its completion S'. Then S (or rather an isometric copy of S) is a finite-dimensional subspace of a complete inner product space S' and as such it is closed. However, S is dense in S' and so $S = S'$, which shows that S is complete. □

Infinite Series

Since an inner product space allows both addition of vectors and convergence of sequences, we can define the concept of infinite sums, or infinite series.

Definition Let V be an inner product space. The **nth partial sum** of the sequence (x_k) in V is

$$s_n = x_1 + \cdots + x_n$$

If the sequence (s_n) of partial sums converges to a vector $s \in V$, that is, if

$$\|s_n - s\| \to 0 \text{ as } n \to \infty$$

then we say that the series $\sum x_n$ **converges** to s and write

$$\sum_{n=1}^{\infty} x_n = s$$ □

We can also define absolute convergence.

Definition *A series* $\sum x_k$ *is said to be* **absolutely convergent** *if the series*

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

converges. □

The key relationship between convergence and absolute convergence is given in the next theorem. Note that completeness is required to guarantee that absolute convergence implies convergence.

Theorem 13.8 *Let V be an inner product space. Then V is complete if and only if absolute convergence of a series implies convergence.*
Proof. Suppose that V is complete and that $\sum \|x_k\| < \infty$. Then the sequence s_n of partial sums is a Cauchy sequence, for if $n > m$, we have

$$\|s_n - s_m\| = \left\| \sum_{k=m+1}^{n} x_k \right\| \le \sum_{k=m+1}^{n} \|x_k\| \to 0$$

Hence, the sequence (s_n) converges, that is, the series $\sum x_k$ converges.

Conversely, suppose that absolute convergence implies convergence and let (x_n) be a Cauchy sequence in V. We wish to show that this sequence converges. Since (x_n) is a Cauchy sequence, for each $k > 0$, there exists an N_k

with the property that

$$i, j \geq N_k \Rightarrow \|x_i - x_j\| < \frac{1}{2^k}$$

Clearly, we can choose $N_1 < N_2 < \cdots$, in which case

$$\|x_{N_{k+1}} - x_{N_k}\| < \frac{1}{2^k}$$

and so

$$\sum_{k=1}^{\infty} \|x_{N_{k+1}} - x_{N_k}\| \leq \sum_{k=1}^{\infty} \frac{1}{2^k} < \infty$$

Thus, according to hypothesis, the series

$$\sum_{k=1}^{\infty} (x_{N_{k+1}} - x_{N_k})$$

converges. But this is a telescoping series, whose nth partial sum is

$$x_{N_{n+1}} - x_{N_1}$$

and so the subsequence (x_{N_k}) converges. Since any Cauchy sequence that has a convergent subsequence must itself converge, the sequence (x_k) converges and so V is complete. □

An Approximation Problem

Suppose that V is an inner product space and that S is a subset of V. It is of considerable interest to be able to find, for any $x \in V$, a vector in S that is *closest* to x in the metric induced by the inner product, should such a vector exist. This is the **approximation problem** for V.

Suppose that $x \in V$ and let

$$\delta = \inf_{s \in S} \|x - s\|$$

Then there is a sequence s_n for which

$$\delta_n = \|x - s_n\| \to \delta$$

as shown in Figure 13.1.

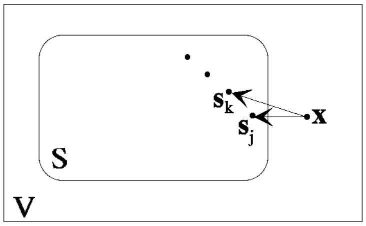

Figure 13.1

Let us see what we can learn about this sequence. First, if we let $y_k = x - s_k$ then according to the parallelogram law

$$\|y_k + y_j\|^2 + \|y_k - y_j\|^2 = 2(\|y_k\|^2 + \|y_j\|^2)$$

or

$$\|y_k - y_j\|^2 = 2(\|y_k\|^2 + \|y_j\|^2) - 4\left\|\frac{y_k + y_j}{2}\right\|^2 \tag{13.2}$$

Now, if the set S is **convex**, that is, if

$$x, y \in S \Rightarrow rx + (1-r)y \in S \text{ for all } 0 \leq r \leq 1$$

(in words S contains the line segment between any two of its points) then $(s_k + s_j)/2 \in S$ and so

$$\left\|\frac{y_k + y_j}{2}\right\| = \left\|x - \frac{s_k + s_j}{2}\right\| \geq \delta$$

Thus, (13.2) gives

$$\|y_k - y_j\|^2 \leq 2(\|y_k\|^2 + \|y_j\|^2) - 4\delta^2 \to 0$$

as $k, j \to \infty$. Hence, if S is convex then the sequence $(y_n) = (x - s_n)$ is a Cauchy sequence and therefore so is (s_n).

If we also require that S be complete then the Cauchy sequence (s_n) converges to a vector $\widehat{x} \in S$ and by the continuity of the norm, we must have $\|x - \widehat{x}\| = \delta$. Let us summarize and add a remark about uniqueness.

Theorem 13.9 *Let V be an inner product space and let S be a complete convex subset of V. Then for any $x \in V$, there exists a unique $\widehat{x} \in S$ for which*

$$\|x - \widehat{x}\| = \inf_{s \in S}\|x - s\|$$

The vector $\widehat{x}$ is called the **best approximation** *to x in S.*

Proof. Only the uniqueness remains to be established. Suppose that

$$\|x-\widehat{x}\| = \delta = \|x-x'\|$$

Then, by the parallelogram law,

$$\begin{aligned}
\|\widehat{x}-x'\|^2 &= \|(x-x')-(x-\widehat{x})\|^2 \\
&= 2\|x-\widehat{x}\|^2 + 2\|x-x'\|^2 - \|2x-\widehat{x}-x'\|^2 \\
&= 2\|x-\widehat{x}\|^2 + 2\|x-x'\|^2 - 4\left\|x-\frac{\widehat{x}+x'}{2}\right\|^2 \\
&\le 2\delta^2 + 2\delta^2 - 4\delta^2 = 0
\end{aligned}$$

and so $\widehat{x} = x'$. □

Since any subspace S of an inner product space V is convex, Theorem 13.9 applies to complete subspaces. However, in this case, we can say more.

Theorem 13.10 *Let V be an inner product space and let S be a complete* subspace *of V. Then for any $x \in V$, the best approximation to x in S is the unique vector $x' \in S$ for which $x - x' \perp S$.*

Proof. Suppose that $x-x' \perp S$, where $x' \in S$. Then for any $s \in S$, we have $x - x' \perp s - x'$ and so

$$\|x-s\|^2 = \|x-x'\|^2 + \|x'-s\|^2 \ge \|x-x'\|^2$$

Hence $x' = \widehat{x}$ is the best approximation to x in S. Now we need only show that $x-\widehat{x} \perp S$, where $\widehat{x}$ is the best approximation to x in S. For any $s \in S$, a little computation reminiscent of completing the square gives

$$\begin{aligned}
\|x-rs\|^2 &= \langle x-rs, x-rs\rangle \\
&= \|x\|^2 - \overline{r}\langle x,s\rangle - r\langle s,x\rangle + r\overline{r}\|s\|^2 \\
&= \|x\|^2 + \|s\|^2\left(r\overline{r} - \overline{r}\frac{\langle x,s\rangle}{\|s\|^2} - r\frac{\overline{\langle x,s\rangle}}{\|s\|^2}\right) \\
&= \|x\|^2 + \|s\|^2\left(r - \frac{\langle x,s\rangle}{\|s\|^2}\right)\left(\overline{r} - \frac{\overline{\langle x,s\rangle}}{\|s\|^2}\right) - \frac{|\langle x,s\rangle|^2}{\|s\|^2} \\
&= \|x\|^2 + \|s\|^2\left|r - \frac{\langle x,s\rangle}{\|s\|^2}\right|^2 - \frac{|\langle x,s\rangle|^2}{\|s\|^2}
\end{aligned}$$

Now, this is smallest when

$$r = r_0 := \frac{\langle x,s\rangle}{\|s\|^2}$$

in which case

$$\|x - r_0 s\|^2 = \|x\|^2 - \frac{|\langle x, s\rangle|^2}{\|s\|^2}$$

Replacing x by $x - \widehat{x}$ gives

$$\|x - \widehat{x} - r_0 s\|^2 = \|x - \widehat{x}\|^2 - \frac{|\langle x - \widehat{x}, s\rangle|^2}{\|s\|^2}$$

But $\widehat{x}$ is the best approximation to x in S and since $\widehat{x} - r_0 s \in S$ we must have

$$\|x - \widehat{x} - r_0 s\|^2 \geq \|x - \widehat{x}\|^2$$

Hence,

$$\frac{|\langle x - \widehat{x}, s\rangle|^2}{\|s\|^2} = 0$$

or, equivalently,

$$\langle x - \widehat{x}, s\rangle = 0$$

Hence, $x - \widehat{x} \perp S$. □

According to Theorem 13.10, if S is a complete subspace of an inner product space V then for any $x \in V$, we may write

$$x = \widehat{x} + (x - \widehat{x})$$

where $\widehat{x} \in S$ and $x - \widehat{x} \in S^\perp$. Hence, $V = S + S^\perp$ and since $S \cap S^\perp = \{0\}$, we also have $V = S \odot S^\perp$. This is the projection theorem for arbitrary inner product spaces.

Theorem 13.11 *(***The projection theorem***) If S is a complete subspace of an inner product space V then*

$$V = S \odot S^\perp$$

In particular, if S is a closed subspace of a Hilbert space H then

$$H = S \odot S^\perp$$ □

Theorem 13.12 *Let S, T and T' be subspaces of an inner product space V.*
1) If $V = S \odot T$ then $T = S^\perp$.
2) If $S \odot T = S \odot T'$ then $T = T'$.
Proof. If $V = S \odot T$ then $T \subseteq S^\perp$ by definition of orthogonal direct sum. On the other hand, if $z \in S^\perp$ then $z = s + t$, for some $s \in S$ and $t \in T$. Hence,

$$0 = \langle z, s\rangle = \langle s, s\rangle + \langle t, s\rangle = \langle s, s\rangle$$

and so $s = 0$, implying that $z = t \in T$. Thus, $S^{\perp} \subseteq T$. Part 2) follows from part 1). □

Let us denote the closure of the span of a set S of vectors by cspan(S).

Theorem 13.13 *Let H be a Hilbert space.*
1) If A is a subset of H then

$$\text{cspan}(A) = A^{\perp\perp}$$

2) If S is a subspace of H then

$$\text{cl}(S) = S^{\perp\perp}$$

3) If K is a closed subspace of H then

$$K = K^{\perp\perp}$$

Proof. We leave it as an exercise to show that $[\text{cspan}(A)]^{\perp} = A^{\perp}$. Hence

$$H = \text{cspan}(A) \odot [\text{cspan}(A)]^{\perp} = \text{cspan}(A) \odot A^{\perp}$$

But since $A^{\perp}$ is closed, we also have

$$H = A^{\perp} \odot A^{\perp\perp}$$

and so by Theorem 13.12, cspan$(A) = A^{\perp\perp}$. The rest follows easily from part 1). □

In the exercises, we provide an example of a closed subspace K of an inner product space V for which $K \neq K^{\perp\perp}$. Hence, we cannot drop the requirement that H be a Hilbert space in Theorem 13.13.

Corollary 13.14 *If A is a* subset *of a Hilbert space H then* span(A) *is dense in H if and only if $A^{\perp} = \{0\}$.*
Proof. As in the previous proof,

$$H = \text{cspan}(A) \odot A^{\perp}$$

and so $A^{\perp} = \{0\}$ if and only if $H = \text{cspan}(A)$. □

Hilbert Bases

We recall the following definition from Chapter 9.

Definition *A maximal orthonormal set in a Hilbert space H is called a* **Hilbert basis** *for H.* □

Zorn's lemma can be used to show that any nontrivial Hilbert space has a Hilbert basis. Again, we should mention that the concepts of Hilbert basis and Hamel basis (a maximal linearly independent set) are quite different. We will show later in this chapter that any two Hilbert bases for a Hilbert space have the same cardinality.

Since an orthonormal set $\mathcal{O}$ is maximal if and only if $\mathcal{O}^{\perp} = \{0\}$, Corollary 13.14 gives the following characterization of Hilbert bases.

Theorem 13.15 *Let $\mathcal{O}$ be an orthonormal subset of a Hilbert space H. The following are equivalent:*
1) $\mathcal{O}$ is a Hilbert basis
2) $\mathcal{O}^{\perp} = \{0\}$
3) $\mathcal{O}$ is a **total subset** *of H, that is,* $\text{cspan}(\mathcal{O}) = H$. □

Part 3) of this theorem says that a subset of a Hilbert space is a Hilbert basis if and only if it is a total orthonormal set.

Fourier Expansions

We now want to take a closer look at best approximations. Our goal is to find an explicit expression for the best approximation to any vector x from within a closed subspace S of a Hilbert space H. We will find it convenient to consider three cases, depending on whether S has finite, countably infinite, or uncountable dimension.

The Finite-Dimensional Case

Suppose that $\mathcal{O} = \{u_1, \ldots, u_n\}$ is an orthonormal set in a Hilbert space H. Recall that the Fourier expansion of any $x \in H$, with respect to $\mathcal{O}$, is given by

$$\widehat{x} = \sum_{k=1}^{n} \langle x, u_k \rangle u_k$$

where $\langle x, u_k \rangle$ is the Fourier coefficient of x with respect to u_k. Observe that

$$\langle x - \widehat{x}, u_k \rangle = \langle x, u_k \rangle - \langle \widehat{x}, u_k \rangle = 0$$

and so $x - \widehat{x} \perp \text{span}(\mathcal{O})$. Thus, according to Theorem 13.10, the Fourier expansion $\widehat{x}$ is the best approximation to x in $\text{span}(\mathcal{O})$. Moreover, since $x - \widehat{x} \perp \widehat{x}$, we have

$$\|\widehat{x}\|^2 = \|x\|^2 - \|x - \widehat{x}\|^2 \leq \|x\|^2$$

and so

$$\|\widehat{x}\| \leq \|x\|$$

with equality if and only if $x = \widehat{x}$, which happens if and only if $x \in \text{span}(\mathcal{O})$. Let us summarize.

Theorem 13.16 *Let $\mathcal{O} = \{u_1, \ldots, u_n\}$ be a finite orthonormal set in a Hilbert space H. For any $x \in H$, the Fourier expansion $\widehat{x}$ of x is the best approximation to x in* span$(\mathcal{O})$. *We also have* **Bessel's inequality**

$$\|\widehat{x}\| \le \|x\|$$

or, equivalently

$$\sum_{k=1}^{n} |\langle x, u_k\rangle|^2 \le \|x\|^2 \tag{13.3}$$

with equality if and only if $x \in$ span$(\mathcal{O})$. □

The Countably Infinite-Dimensional Case

In the countably infinite case, we will be dealing with infinite sums and so questions of convergence will arise. Thus, we begin with the following.

Theorem 13.17 *Let $\mathcal{O} = \{u_1, u_2, \ldots\}$ be a countably infinite orthonormal set in a Hilbert space H. The series*

$$\sum_{k=1}^{\infty} r_k u_k \tag{13.4}$$

converges in H if and only if the series

$$\sum_{k=1}^{\infty} |r_k|^2 \tag{13.5}$$

converges in $\mathbb{R}$. If these series converge then they converge unconditionally (that is, any series formed by rearranging the order of the terms also converges). Finally, if the series (13.4) converges then

$$\left\| \sum_{k=1}^{\infty} r_k u_k \right\|^2 = \sum_{k=1}^{\infty} |r_k|^2$$

Proof. Denote the partial sums of the first series by s_n and the partial sums of the second series by p_n. Then for $m \le n$

$$\|s_n - s_m\|^2 = \left\| \sum_{k=m+1}^{n} r_k u_k \right\|^2 = \sum_{k=m+1}^{n} |r_k|^2 = |p_n - p_m|$$

Hence (s_n) is a Cauchy sequence in H if and only if (p_n) is a Cauchy sequence in $\mathbb{R}$. Since both H and $\mathbb{R}$ are complete, (s_n) converges if and only if (p_n) converges.

If the series (13.5) converges then it converges absolutely and hence unconditionally. (A real series converges unconditionally if and only if it converges absolutely.) But if (13.5) converges unconditionally then so does (13.4). The last part of the theorem follows from the continuity of the norm. □

Now let $\mathcal{O} = \{u_1, u_2, \dots\}$ be a countably infinite orthonormal set in H. The **Fourier expansion** of a vector $x \in H$ is defined to be the sum

$$\widehat{x} = \sum_{k=1}^{\infty} \langle x, u_k \rangle u_k \tag{13.6}$$

To see that this sum converges, observe that, for any $n > 0$, (13.3) gives

$$\sum_{k=1}^{n} |\langle x, u_k \rangle|^2 \le \|x\|^2$$

and so

$$\sum_{k=1}^{\infty} |\langle x, u_k \rangle|^2 \le \|x\|^2$$

which shows that the series on the left converges. Hence, according to Theorem 13.17, the Fourier expansion (13.6) converges unconditionally.

Moreover, since the inner product is continuous,

$$\langle x - \widehat{x}, u_k \rangle = \langle x, u_k \rangle - \langle \widehat{x}, u_k \rangle = 0$$

and so $x - \widehat{x} \in [\text{span}(\mathcal{O})]^{\perp} = [\text{cspan}(\mathcal{O})]^{\perp}$. Hence, $\widehat{x}$ is the best approximation to x in cspan$(\mathcal{O})$. Finally, since $x - \widehat{x} \perp \widehat{x}$, we again have

$$\|\widehat{x}\|^2 = \|x\|^2 - \|x - \widehat{x}\|^2 \le \|x\|^2$$

and so

$$\|\widehat{x}\| \le \|x\|$$

with equality if and only if $x = \widehat{x}$, which happens if and only if $x \in$ cspan$(\mathcal{O})$. Thus, the following analog of Theorem 13.16 holds.

Theorem 13.18 Let $\mathcal{O} = \{u_1, u_2, \dots\}$ be a countably infinite orthonormal set in a Hilbert space H. For any $x \in H$, the Fourier expansion

$$\widehat{x} = \sum_{k=1}^{\infty} \langle x, u_k \rangle u_k$$

of x converges unconditionally and is the best approximation to x in cspan$(\mathcal{O})$. We also have **Bessel's inequality**

$$\|\widehat{x}\| \le \|x\|$$

or, equivalently

$$\sum_{k=1}^{\infty}|\langle x,u_k\rangle|^2 \le \|x\|^2$$

with equality if and only if $x \in \text{cspan}(\mathcal{O})$. □

The Arbitrary Case

To discuss the case of an arbitrary orthonormal set $\mathcal{O} = \{u_k \mid k \in K\}$, let us first define and discuss the concept of the sum of an arbitrary number of terms. (This is a bit of a digression, since we could proceed without all of the coming details – but they are interesting.)

Definition Let $\mathcal{K} = \{x_k \mid k \in K\}$ be an arbitrary family of vectors in an inner product space V. The sum

$$\sum_{k\in K} x_k$$

is said to **converge** to a vector $x \in V$ and we write

$$x = \sum_{k\in K} x_k \tag{13.7}$$

if for any $\epsilon > 0$, there exists a finite set $S \subseteq K$ for which

$$T \supset S,\ T \text{ finite} \Rightarrow \left\| \sum_{k\in T} x_k - x \right\| \le \epsilon$$ □

For those readers familiar with the language of convergence of nets, the set $\mathcal{P}_0(K)$ of all finite subsets of K is a *directed set* under inclusion (for every $A, B \in \mathcal{P}_0(K)$ there is a $C \in \mathcal{P}_0(K)$ containing A and B) and the function

$$S \to \sum_{k\in S} x_k$$

is a net in H. Convergence of (13.7) is convergence of this net. In any case, we will refer to the preceding definition as the **net definition** of convergence.

It is not hard to verify the following basic properties of net convergence for arbitrary sums.

Theorem 13.19 *Let $\mathcal{K} = \{x_k \mid k \in K\}$ be an arbitrary family of vectors in an inner product space V. If*

$$\sum_{k\in K} x_k = x \textit{ and } \sum_{k\in K} y_k = y$$

then

1) (**Linearity**)

$$\sum_{k\in K}(rx_k+sy_k)=rx+sy$$

for any $r,s\in F$

2) (**Continuity**)

$$\sum_{k\in K}\langle x_k,y\rangle=\langle x,y\rangle \text{ and }\sum_{k\in K}\langle y,x_k\rangle=\langle y,x\rangle$$ □

The next result gives a useful "Cauchy type" description of convergence.

Theorem 13.20 *Let* $\mathcal{K}=\{x_k\mid k\in K\}$ *be an arbitrary family of vectors in an inner product space* V*.*

1) If the sum

$$\sum_{k\in K}x_k$$

converges then for any $\epsilon>0$*, there exists a finite set* $I\subseteq K$ *such that*

$$J\cap I=\emptyset,\ J\textit{ finite}\Rightarrow\left\|\sum_{k\in J}x_k\right\|\le\epsilon$$

2) If V *is a Hilbert space then the converse of 1) also holds.*

Proof. For part 1), given $\epsilon>0$, let $S\subseteq K$, S finite, be such that

$$T\supset S,\ T\text{ finite}\Rightarrow\left\|\sum_{k\in T}x_k-x\right\|\le\frac{\epsilon}{2}$$

If $J\cap S=\emptyset$, J finite then

$$\begin{aligned}\left\|\sum_J x_k\right\|&=\left\|(\sum_J x_k+\sum_S x_k-x)-(\sum_S x_k-x)\right\|\\&\le\left\|\sum_{J\cup S}x_k-x\right\|+\left\|\sum_S x_k-x\right\|\le\frac{\epsilon}{2}+\frac{\epsilon}{2}=\epsilon\end{aligned}$$

As for part 2), for each $n>0$, let $I_n\subseteq K$ be a finite set for which

$$J\cap I_n=\emptyset,\ J\text{ finite}\Rightarrow\left\|\sum_{j\in J}x_j\right\|\le\frac{1}{n}$$

and let

$$y_n=\sum_{k\in I_n}x_k$$

Then (y_n) is a Cauchy sequence, since

$$\begin{aligned}\|y_n - y_m\| &= \left\|\sum_{I_n} x_k - \sum_{I_m} x_k\right\| = \left\|\sum_{I_n - I_m} x_k - \sum_{I_m - I_n} x_k\right\| \\ &\le \left\|\sum_{I_n - I_m} x_k\right\| + \left\|\sum_{I_m - I_n} x_k\right\| \le \frac{1}{m} + \frac{1}{n} \to 0\end{aligned}$$

Since V is assumed complete, we have $(y_n) \to y$.

Now, given $\epsilon > 0$, there exists an N such that

$$n \ge N \Rightarrow \|y_n - y\| = \left\|\sum_{I_n} x_k - y\right\| \le \frac{\epsilon}{2}$$

Setting $n = \max\{N, 2/\epsilon\}$ gives for $T \supset I_n,\ T$ finite

$$\begin{aligned}\left\|\sum_{T} x_k - y\right\| &= \left\|\sum_{I_n} x_k - y + \sum_{T - I_n} x_k\right\| \\ &\le \left\|\sum_{I_n} x_k - y\right\| + \left\|\sum_{T - I_n} x_k\right\| \le \frac{\epsilon}{2} + \frac{1}{n} \le \epsilon\end{aligned}$$

and so $\sum_{k\in K} x_k$ converges to y. □

The following theorem tells us that convergence of an arbitrary sum implies that only countably many terms can be nonzero so, in some sense, there is no such thing as a nontrivial *uncountable* sum.

Theorem 13.21 *Let $\mathcal{K} = \{x_k \mid k \in K\}$ be an arbitrary family of vectors in an inner product space V. If the sum*

$$\sum_{k\in K} x_k$$

converges then at most a countable number of terms x_k can be nonzero.
Proof. According to Theorem 13.20, for each $n > 0$, we can let $I_n \subseteq K$, I_n finite, be such that

$$J \cap I_n = \emptyset,\ J \text{ finite} \Rightarrow \left\|\sum_{j\in J} x_j\right\| \le \frac{1}{n}$$

Let $I = \bigcup_n I_n$. Then I is countable and

$$k \notin I \Rightarrow \{k\} \cap I_n = \emptyset \text{ for all } n \Rightarrow \|x_k\| \le \frac{1}{n} \text{ for all } n \Rightarrow x_k = 0 \qquad □$$

Here is the analog of Theorem 13.17.

Theorem 13.22 *Let $\mathcal{O} = \{u_k \mid k \in K\}$ be an arbitrary orthonormal family of vectors in a Hilbert space H. The two series*

$$\sum_{k \in K} r_k u_k \text{ and } \sum_{k \in K} |r_k|^2$$

converge or diverge together. If these series converge then

$$\left\| \sum_{k \in K} r_k u_k \right\|^2 = \sum_{k \in K} |r_k|^2$$

Proof. The first series converges if and only if for any $\epsilon > 0$, there exists a finite set $I \subseteq K$ such that

$$J \cap I = \emptyset,\ J \text{ finite} \Rightarrow \left\| \sum_{k \in J} r_k u_k \right\|^2 \le \epsilon^2$$

or, equivalently

$$J \cap I = \emptyset,\ J \text{ finite} \Rightarrow \sum_{k \in J} |r_k|^2 \le \epsilon^2$$

and this is precisely what it means for the second series to converge. We leave proof of the remaining statement to the reader. □

The following is a useful characterization of arbitrary sums of nonnegative real terms.

Theorem 13.23 *Let $\{r_k \mid k \in K\}$ be a collection of nonnegative real numbers. Then*

$$\sum_{k \in K} r_k = \sup_{\substack{J \text{ finite} \\ J \subseteq K}} \sum_{k \in J} r_k \tag{13.8}$$

provided that either of the preceding expressions is finite.
Proof. Suppose that

$$\sup_{\substack{J \text{ finite} \\ J \subseteq K}} \sum_{k \in J} r_k = R < \infty$$

Then, for any $\epsilon > 0$, there exists a finite set $S \subseteq K$ such that

$$R \ge \sum_{k \in S} r_k \ge R - \epsilon$$

Hence, if $T \subseteq K$ is a finite set for which $T \supset S$ then since $r_k \geq 0$,

$$R \geq \sum_{k \in T} r_k \geq \sum_{k \in S} r_k \geq R - \epsilon$$

and so

$$\left\| R - \sum_{k \in T} r_k \right\| \leq \epsilon$$

which shows that $\sum r_k$ converges to R. Finally, if the sum on the left of (13.8) converges then the supremum on the right is finite and so (13.8) holds. □

The reader may have noticed that we have two definitions of convergence for countably infinite series: the net version and the traditional version involving the limit of partial sums. Let us write

$$\sum_{k \in \mathbb{N}^+} x_k \text{ and } \sum_{k=1}^{\infty} x_k$$

for the net version and the partial sum version, respectively. Here is the relationship between these two definitions.

Theorem 13.24 *Let H be a Hilbert space. If $x_k \in H$ then the following are equivalent:*

1) $\sum_{k \in \mathbb{N}^+} x_k$ *converges (net version) to* x

2) $\sum_{k=1}^{\infty} x_k$ *converges unconditionally to* x

Proof. Assume that 1) holds. Suppose that π is any permutation of $\mathbb{N}^+$. Given any $\epsilon > 0$, there is a finite set $S \subseteq \mathbb{N}^+$ for which

$$T \supset S,\ T \text{ finite} \Rightarrow \left\| \sum_{k \in T} x_k - x \right\| \leq \epsilon$$

Let us denote the set of integers $\{1, \ldots, n\}$ by I_n and choose a positive integer n such that $\pi(I_n) \supset S$. Then for $m \geq n$ we have

$$\pi(I_m) \supset \pi(I_n) \supset S \Rightarrow \left\| \sum_{k=1}^{m} x_{\pi(k)} - x \right\| = \left\| \sum_{k \in \pi(I_m)} x_k - x \right\| \leq \epsilon$$

and so 2) holds.

Next, assume that 2) holds, but that the series in 1) does not converge. Then there exists an $\epsilon > 0$ such that, for any finite subset $I \subseteq \mathbb{N}^+$, there exists a finite subset J with $J \cap I = \emptyset$ for which

$$\left\| \sum_{k \in J} x_k \right\| > \epsilon$$

From this, we deduce the existence of a countably infinite sequence J_n of mutually disjoint finite subsets of $\mathbb{N}^+$ with the property that

$$\max(J_n) = M_n < m_{n+1} = \min(J_{n+1})$$

and

$$\left\| \sum_{k \in J_n} x_k \right\| > \epsilon$$

Now, we choose any permutation $\pi\colon \mathbb{N}^+ \to \mathbb{N}^+$ with the following properties
1) $\pi([m_n, M_n]) \subseteq [m_n, M_n]$
2) if $J_n = \{j_{n,1}, \ldots, j_{n,u_n}\}$ then

$$\pi(m_n) = j_{n,1},\ \pi(m_n + 1) = j_{n,2}, \ldots, \pi(m_n + u_n - 1) = j_{n,u_n}$$

The intention in property 2) is that, for each n, π takes a set of consecutive integers to the integers in J_n.

For any such permutation π, we have

$$\left\| \sum_{k=m_n}^{m_n+u_n-1} x_{\pi(k)} \right\| = \left\| \sum_{k \in J_n} x_k \right\| > \epsilon$$

which shows that the sequence of partial sums of the series

$$\sum_{k=1}^{\infty} x_{\pi(k)}$$

is not Cauchy and so this series does not converge. This contradicts 2) and shows that 2) implies at least that 1) converges. But if 1) converges to $y \in H$ then since 1) implies 2) and since unconditional limits are unique, we have $y = x$. Hence, 2) implies 1). □

Now we can return to the discussion of Fourier expansions. Let $\mathcal{O} = \{u_k \mid k \in K\}$ be an arbitrary orthonormal set in a Hilbert space H. Given any $x \in H$, we may apply Theorem 13.16 to all finite subsets of $\mathcal{O}$, to deduce that

$$\sup_{\substack{J \text{ finite} \\ J\subseteq K}} \sum_{k\in J} |\langle x, u_k\rangle|^2 \le \|x\|^2$$

and so Theorem 13.23 tells us that the sum

$$\sum_{k\in K} |\langle x, u_k\rangle|^2$$

converges. Hence, according to Theorem 13.22, the **Fourier expansion**

$$\widehat{x} = \sum_{k\in K} \langle x, u_k\rangle u_k$$

of x also converges and

$$\|\widehat{x}\|^2 = \sum_{k\in K} |\langle x, u_k\rangle|^2$$

Note that, according to Theorem 13.21, $\widehat{x}$ is a countably infinite sum of terms of the form $\langle x, u_k\rangle u_k$ and so is in cspan$(\mathcal{O})$.

The continuity of infinite sums with respect to the inner product (Theorem 13.19) implies that

$$\langle x - \widehat{x}, u_k\rangle = \langle x, u_k\rangle - \langle \widehat{x}, u_k\rangle = 0$$

and so $x - \widehat{x} \in [\text{span}(\mathcal{O})]^\perp = [\text{cspan}(\mathcal{O})]^\perp$. Hence, Theorem 3.10 tells us that $\widehat{x}$ is the best approximation to x in cspan$(\mathcal{O})$. Finally, since $x - \widehat{x} \perp \widehat{x}$, we again have

$$\|\widehat{x}\|^2 = \|x\|^2 - \|x - \widehat{x}\|^2 \le \|x\|^2$$

and so

$$\|\widehat{x}\| \le \|x\|$$

with equality if and only if $x = \widehat{x}$, which happens if and only if $x \in$ cspan$(\mathcal{O})$. Thus, we arrive at the most general form of a key theorem about Hilbert spaces.

Theorem 13.25 *Let $\mathcal{O} = \{u_k \mid k \in K\}$ be an orthonormal family of vectors in a Hilbert space H. For any $x \in H$, the Fourier expansion*

$$\widehat{x} = \sum_{k\in K} \langle x, u_k\rangle u_k$$

of x converges in H and is the unique best approximation to x in cspan$(\mathcal{O})$. *Moreover, we have* **Bessel's inequality**

$$\|\widehat{x}\| \le \|x\|$$

or, equivalently

$$\sum_{k\in K}|\langle x,u_k\rangle|^2\le\|x\|^2$$

with equality if and only if $x\in\operatorname{cspan}(\mathcal{O})$. □

A Characterization of Hilbert Bases

Recall from Theorem 13.15 that an orthonormal set $\mathcal{O}=\{u_k\mid k\in K\}$ in a Hilbert space H is a Hilbert basis if and only if

$$\operatorname{cspan}(\mathcal{O})=H$$

Theorem 13.25 then leads to the following characterization of Hilbert bases.

Theorem 13.26 *Let* $\mathcal{O}=\{u_k\mid k\in K\}$ *be an orthonormal family in a Hilbert space* H*. The following are equivalent:*

1) $\mathcal{O}$ *is a Hilbert basis (a maximal orthonormal set)*
2) $\mathcal{O}^\perp=\{0\}$
3) $\mathcal{O}$ *is total (that is,* $\operatorname{cspan}(\mathcal{O})=H$*)*
4) $x=\widehat{x}$ *for all* $x\in H$
5) *Equality holds in Bessel's inequality for all* $x\in H$*, that is,*

$$\|x\|=\|\widehat{x}\|$$

for all $x\in H$.

6) **Parseval's identity**

$$\langle x,y\rangle=\langle\widehat{x},\widehat{y}\rangle$$

holds for all $x,y\in H$*, that is,*

$$\langle x,y\rangle=\sum_{k\in K}\langle x,u_k\rangle\overline{\langle y,u_k\rangle}$$

Proof. Parts 1), 2) and 3) are equivalent by Theorem 13.15. Part 4) implies part 3), since $\widehat{x}\in\operatorname{cspan}(\mathcal{O})$ and 3) implies 4) since the unique best approximation of any $x\in\operatorname{cspan}(\mathcal{O})$ is itself and so $x=\widehat{x}$. Parts 3) and 5) are equivalent by Theorem 13.25. Parseval's identity follows from part 4) using Theorem 13.19. Finally, Parseval's identity for $y=x$ implies that equality holds in Bessel's inequality. □

Hilbert Dimension

We now wish to show that all Hilbert bases for a Hilbert space H have the same cardinality and so we can define the Hilbert dimension of H to be that cardinality.

Theorem 13.27 *All Hilbert bases for a Hilbert space H have the same cardinality. This cardinality is called the* **Hilbert dimension** *of H, which we denote by* $\operatorname{hdim}(H)$.

Proof. If H has a finite Hilbert basis then that set is also a Hamel basis and so all finite Hilbert bases have size $\dim(H)$. Suppose next that $\mathcal{B} = \{b_k \mid k \in K\}$ and $\mathcal{C} = \{c_j \mid j \in J\}$ are infinite Hilbert bases for H. Then for each b_k, we have

$$b_k = \sum_{j \in J_k} \langle b_k, c_j \rangle c_j$$

where J_k is the countable set $\{j \mid \langle b_k, c_j \rangle \neq 0\}$. Moreover, since no c_j can be orthogonal to every b_k, we have $\bigcup_K J_k = J$. Thus, since each J_k is countable, we have

$$|J| = \left| \bigcup_{k \in K} J_k \right| \leq \aleph_0 |K| = |K|$$

By symmetry, we also have $|K| \leq |J|$ and so the Schröder-Bernstein theorem implies that $|J| = |K|$. □

Theorem 13.28 *Two Hilbert spaces are isometrically isomorphic if and only if they have the same Hilbert dimension.*

Proof. Suppose that $\operatorname{hdim}(H_1) = \operatorname{hdim}(H_2)$. Let $\mathcal{O}_1 = \{u_k \mid k \in K\}$ be a Hilbert basis for H_1 and $\mathcal{O}_2 = \{v_k \mid k \in K\}$ a Hilbert basis for H_2. We may define a map $\tau \colon H_1 \to H_2$ as follows

$$\tau\Big(\sum_{k \in K} r_k u_k\Big) = \sum_{k \in K} r_k v_k$$

We leave it as an exercise to verify that τ is a bijective isometry. The converse is also left as an exercise. □

A Characterization of Hilbert Spaces

We have seen that any vector space V is isomorphic to a vector space $(F^B)_0$ of all functions from B to F that have finite support. There is a corresponding result for Hilbert spaces. Let K be any nonempty set and let

$$\ell^2(K) = \Big\{ f \colon K \to \mathbb{C} \Big| \sum_{k \in K} |f(k)|^2 < \infty \Big\}$$

The functions in $\ell^2(K)$ are referred to as **square summable functions**. (We can also define a real version of this set by replacing $\mathbb{C}$ by $\mathbb{R}$.) We define an inner product on $\ell^2(K)$ by

$$\langle f, g \rangle = \sum_{k \in K} f(k) \overline{g(k)}$$

The proof that $\ell^2(K)$ is a Hilbert space is quite similar to the proof that

$\ell^2 = \ell^2(\mathbb{N})$ is a Hilbert space and the details are left to the reader. If we define $\delta_k \in \ell^2(K)$ by

$$\delta_k(j) = \delta_{k,j} = \begin{cases} 1 & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

then the collection

$$\mathcal{O} = \{\delta_k \mid k \in K\}$$

is a Hilbert basis for $\ell^2(K)$, of cardinality $|K|$. To see this, observe that

$$\langle \delta_i, \delta_j \rangle = \sum_{k \in K} \delta_i(k) \overline{\delta_j(k)} = \delta_{i,j}$$

and so $\mathcal{O}$ is orthonormal. Moreover, if $f \in \ell^2(K)$ then $f(k) \neq 0$ for only a countable number of $k \in K$, say $\{k_1, k_2, \dots\}$. If we define f' by

$$f' = \sum_{i=1}^{\infty} f(k_i)\delta_{k_i}$$

then $f' \in \text{cspan}(\mathcal{O})$ and $f'(j) = f(j)$ for all $j \in K$, which implies that $f = f'$. This shows that $\ell^2(K) = \text{cspan}(\mathcal{O})$ and so $\mathcal{O}$ is a total orthonormal set, that is, a Hilbert basis for $\ell^2(K)$.

Now let H be a Hilbert space, with Hilbert basis $\mathcal{B} = \{u_k \mid k \in K\}$. We define a map $\phi: H \to \ell^2(K)$ as follows. Since $\mathcal{B}$ is a Hilbert basis, any $x \in H$ has the form

$$x = \sum_{k \in K} \langle x, u_k \rangle u_k$$

Since the series on the right converges, Theorem 13.22 implies that the series

$$\sum_{k \in K} |\langle x, u_k \rangle|^2$$

converges. Hence, another application of Theorem 13.22 implies that the following series converges

$$\phi(x) = \sum_{k \in K} \langle x, u_k \rangle \delta_k$$

It follows from Theorem 13.19 that ϕ is linear and it is not hard to see that it is also bijective. Notice that $\phi(u_k) = \delta_k$ and so ϕ takes the Hilbert basis $\mathcal{B}$ for H to the Hilbert basis $\mathcal{O}$ for $\ell^2(K)$.

Notice also that

$$\|\phi(x)\|^2 = \langle \phi(x), \phi(x) \rangle = \sum_{k \in K} |\langle x, u_k \rangle|^2 = \left\| \sum_{k \in K} \langle x, u_k \rangle u_k \right\|^2 = \|x\|^2$$

and so ϕ is an isometric isomorphism. We have proved the following theorem.

Theorem 13.29 *If H is a Hilbert space of Hilbert dimension κ and if K is any set of cardinality κ then H is isometrically isomorphic to $\ell^2(K)$.* □

The Riesz Representation Theorem

We conclude our discussion of Hilbert spaces by discussing the Riesz representation theorem. As it happens, not all linear functionals on a Hilbert space have the form "take the inner product with...," as in the finite-dimensional case. To see this, observe that if $y \in H$ then the function

$$f_y(x) = \langle x, y \rangle$$

is certainly a linear functional on H. However, it has a special property. In particular, the Cauchy-Schwarz inequality gives, for all $x \in H$

$$|f_y(x)| = |\langle x, y \rangle| \le \|x\| \|y\|$$

or, for all $x \neq 0$,

$$\frac{|f_y(x)|}{\|x\|} \le \|y\|$$

Noticing that equality holds if $x = y$, we have

$$\sup_{x \neq 0} \frac{|f_y(x)|}{\|x\|} = \|y\|$$

This prompts us to make the following definition, which we do for linear transformations between Hilbert spaces (this covers the case of linear functionals).

Definition *Let $\tau\colon H_1 \to H_2$ be a linear transformation from H_1 to H_2. Then τ is said to be* **bounded** *if*

$$\sup_{x \neq 0} \frac{\|\tau(x)\|}{\|x\|} < \infty$$

If the supremum on the left is finite, we denote it by $\|\tau\|$ and call it the **norm** *of τ.* □

Of course, if $f\colon H \to F$ is a bounded linear functional on H then

$$\|f\| = \sup_{x\neq 0}\frac{|f(x)|}{\|x\|}$$

The set of all bounded linear functionals on a Hilbert space H is called the **continuous dual space**, or **conjugate space**, of H and denoted by H^*. Note that this differs from the algebraic dual of H, which is the set of all linear functionals on H. In the finite-dimensional case, however, since all linear functionals are bounded (exercise), the two concepts agree. (Unfortunately, there is no universal agreement on the notation for the algebraic dual versus the continuous dual. Since we will discuss only the continuous dual in this section, no confusion should arise.)

The following theorem gives some simple reformulations of the definition of norm.

Theorem 13.30 *Let $\tau\colon H_1 \to H_2$ be a bounded linear transformation.*
1) $\|\tau\| = \sup_{\|x\|=1} \|\tau(x)\|$
2) $\|\tau\| = \sup_{\|x\|\leq 1} \|\tau(x)\|$
3) $\|\tau\| = \inf\{c \in \mathbb{R} \mid \|\tau(x)\| \leq c\|x\| \textit{ for all } x \in H\}$ □

The following theorem explains the importance of bounded linear transformations.

Theorem 13.31 *Let $\tau\colon H_1 \to H_2$ be a linear transformation. The following are equivalent:*
1) *τ is bounded*
2) *τ is continuous at any point $x_0 \in H$*
3) *τ is continuous.*
Proof. Suppose that τ is bounded. Then

$$\|\tau(x) - \tau(x_0)\| = \|\tau(x - x_0)\| \leq \|\tau\|\|x - x_0\| \to 0$$

as $x \to x_0$. Hence, τ is continuous at x_0. Thus, 1) implies 2). If 2) holds then for any $y \in H$, we have

$$\|\tau(x) - \tau(y)\| = \|\tau(x - y + x_0) - \tau(x_0)\| \to 0$$

as $x \to y$, since τ is continuous at x_0 and $x - y + x_0 \to x_0$ as $y \to x$. Hence, τ is continuous at any $y \in H$ and 3) holds. Finally, suppose that 3) holds. Thus, τ is continuous at 0 and so there exists a $\delta > 0$ such that

$$\|x\| \leq \delta \Rightarrow \|\tau(x)\| \leq 1$$

In particular,

$$\|x\| = \delta \Rightarrow \frac{\|\tau(x)\|}{\|x\|} \le \frac{1}{\delta}$$

and so

$$\|x\| = 1 \Rightarrow \|\delta x\| = \delta \Rightarrow \frac{\|\tau(\delta x)\|}{\|\delta x\|} \le \frac{1}{\delta} \Rightarrow \frac{\|\tau(x)\|}{\|x\|} \le \frac{1}{\delta}$$

Thus, τ is bounded. □

Now we can state and prove the Riesz representation theorem.

Theorem 13.32 *(***The Riesz representation theorem***) Let H be a Hilbert space. For any bounded linear functional f on H, there is a unique $z_0 \in H$ such that*

$$f(x) = \langle x, z_0 \rangle$$

for all $x \in H$. Moreover, $\|z_0\| = \|f\|$.
Proof. If $f = 0$, we may take $z_0 = 0$, so let us assume that $f \ne 0$. Hence, $K = \ker(f) \ne H$ and since f is continuous, K is closed. Thus

$$H = K \odot K^{\perp}$$

Now, the first isomorphism theorem, applied to the linear functional $f\colon H \to F$, implies that $H/K \approx F$ (as vector spaces). In addition, Theorem 3.5 implies that $H/K \approx K^{\perp}$ and so $K^{\perp} \approx F$. In particular, $\dim(K^{\perp}) = 1$.

For any $z \in K^{\perp}$, we have

$$x \in K \Rightarrow f(x) = 0 = \langle x, z \rangle$$

Since $\dim(K^{\perp}) = 1$, all we need do is find a $0 \ne z \in K^{\perp}$ for which

$$f(z) = \langle z, z \rangle$$

for then $f(rz) = rf(z) = r\langle z, z \rangle = \langle rz, z \rangle$ for all $r \in F$, showing that $f(x) = \langle x, z \rangle$ for $x \in K$ as well.

But if $0 \ne z \in K^{\perp}$ then

$$z_0 = \frac{\overline{f(z)}}{\langle z, z \rangle}\, z$$

has this property, as can be easily checked. The fact that $\|z_0\| = \|f\|$ has already been established. □

Exercises

1. Prove that the sup metric on the metric space $C[a,b]$ of continuous functions on $[a,b]$ does not come from an inner product. Hint: let $f(t)=1$ and $g(t)=(t-\mathrm{a})/(b-\mathrm{a})$ and consider the parallelogram law.
2. Prove that any Cauchy sequence that has a convergent subsequence must itself converge.
3. Let V be an inner product space and let A and B be subsets of V. Show that
   a) $A\subseteq B\Rightarrow B^{\perp}\subseteq A^{\perp}$
   b) $A^{\perp}$ is a closed subspace of V
   c) $[\text{cspan}(A)]^{\perp}=A^{\perp}$
4. Let V be an inner product space and $S\subseteq V$. Under what conditions is $S^{\perp\perp\perp}=S^{\perp}$?
5. Prove that a subspace S of a Hilbert space H is closed if and only if $S=S^{\perp\perp}$.
6. Let V be the subspace of ℓ^2 consisting of all sequences of real numbers, with the property that each sequence has only a finite number of nonzero terms. Thus, V is an inner product space. Let K be the subspace of V consisting of all sequences $x=(x_n)$ in V with the property that $\Sigma x_n/n=0$. Show that K is closed, but that $K^{\perp\perp}\neq K$. Hint: For the latter, show that $K^{\perp}=\{0\}$ by considering the sequences $u=(1,\ldots,-n,\ldots)$, where the term $-n$ is in the nth coordinate position.
7. Let $\mathcal{O}=\{u_1,u_2,\ldots\}$ be an orthonormal set in H. If $x=\Sigma r_k u_k$ converges, show that
$$\|x\|^2=\sum_{k=1}^{\infty}|r_k|^2$$
8. Prove that if an infinite series
$$\sum_{k=1}^{\infty}x_k$$
converges absolutely in a Hilbert space H then it also converges in the sense of the "net" definition given in this section.
9. Let $\{r_k\mid k\in K\}$ be a collection of nonnegative real numbers. If the sum on the left below converges, show that
$$\sum_{k\in K}r_k=\sup_{\substack{J\text{ finite}\\ J\subseteq K}}\sum_{k\in J}r_k$$
10. Find a countably infinite sum of real numbers that converges in the sense of partial sums, but not in the sense of nets.
11. Prove that if a Hilbert space H has infinite Hilbert dimension then no Hilbert basis for H is a Hamel basis.
12. Prove that $\ell^2(K)$ is a Hilbert space for any nonempty set K.

13. Prove that any linear transformation between finite-dimensional Hilbert spaces is bounded.
14. Prove that if $f \in H^*$ then $\ker(f)$ is a closed subspace of H.
15. Prove that a Hilbert space is separable if and only if $\mathrm{hdim}(H) \le \aleph_0$.
16. Can a Hilbert space have countably infinite Hamel dimension?
17. What is the Hamel dimension of $\ell^2(\mathbb{N})$?
18. Let τ and σ be bounded linear operators on H. Verify the following:
    a) $\|r\tau\| = |r|\|\tau\|$
    b) $\|\tau + \sigma\| \le \|\tau\| + \|\sigma\|$
    c) $\|\tau\sigma\| \le \|\tau\|\|\sigma\|$
19. Use the Riesz representation theorem to show that $H^* \approx H$ for any Hilbert space H.

Chapter 14
Tensor Products

In the preceding chapters, we have seen several ways to construct new vector spaces from old ones. Two of the most important such constructions are the direct sum $U \oplus V$ and the vector space $\mathcal{L}(U,V)$ of all linear transformations from U to V. In this chapter, we consider another very important construction, known as the *tensor product*.

Universality

We begin by describing a general type of *universality* that will help motivate the definition of tensor product. Our description is strongly related to the formal notion of a *universal arrow (or universal element)* in category theory, but we will be somewhat less formal to avoid the need to formally define categorical concepts. Accordingly, the terminology that we shall introduce is not standard (but does not contradict any standard terminology).

Referring to Figure 14.1, consider a set A and two functions f and g, with domain A.

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[dr, "g"'] & S \arrow[d, dashed, "\tau"] \\
& X
\end{tikzcd}$$

Figure 14.1

Suppose that this diagram *commutes*, that is, that there exists a *unique* function $\tau\colon S \to X$ for which

$$g = \tau \circ f$$

What does this say about the relationship between the functions f and g?

Let us think of the "information" contained in a function $h: A \to B$ to be the way in which h *distinguishes* elements of A using *labels* from B. The relationship above implies that

$$g(a) \neq g(b) \Rightarrow f(a) \neq f(b)$$

This can be phrased by saying that whatever ability g has to distinguish elements of A is also possessed by f. Put another way, except for labeling differences, any information contained in g is also contained in f. This is sometimes expressed by saying that g can be **factored through** f.

If τ happens to be injective, then the *only* difference between f and g is the values of the labels. That is, the two functions have equal ability to distinguish elements of A. However, in general, τ is not required to be injective and so f may contain more information than g.

Suppose now that for all sets X in some family $\mathcal{S}$ of sets that includes S and for all functions $g: A \to X$ in some family $\mathcal{F}$ of functions that includes f, the diagram in Figure 14.1 commutes. This says that the information contained in *every* function in $\mathcal{F}$ is also contained in f. In other words, f captures and preserves the information in every member of $\mathcal{F}$. In this sense, $f: A \to S$ is *universal* among all functions $g: A \to X$ in $\mathcal{F}$.

Moreover, since f is a member of the family $\mathcal{F}$, we can also say that f contains the information in $\mathcal{F}$ *but no more*. In other words, the information in the universal function f is precisely the same as the information in the entire family $\mathcal{F}$. In this way, a single function $f: A \to S$ (more precisely, a single pair (S, f)) can capture a concept, as described by a family of functions, such as the concepts of basis, quotient space, direct sum and bilinearity (as we will see)!

Let us make a formal definition.

Definition *Referring to Figure 14.2, let A be a set and let $\mathcal{S}$ be a family of sets. Let $\mathcal{F}$ be a family of functions from A to members of $\mathcal{S}$. Let $\mathcal{H}$ be a family of functions on members of $\mathcal{S}$. We assume that $\mathcal{H}$ has the following structure:*

1) *$\mathcal{H}$ contains the identity function for each member of $\mathcal{S}$*
2) *$\mathcal{H}$ is closed under composition of functions*
3) *Composition of functions in $\mathcal{H}$ is associative.*

We also assume that for any $\tau \in \mathcal{H}$ and $f \in \mathcal{F}$, the composition $\tau \circ f$ is defined and is a member of $\mathcal{F}$.

$$\begin{tikzcd}
 & & S_1 \arrow[dd, "\tau_2"] \\
A \arrow[urr, "f_1"'] \arrow[r, "f_2"'] \arrow[drr, "f_3"'] & S_2 \arrow[ur, "\tau_1"'] \arrow[dr, "\tau_3"] & \\
 & & S_3
\end{tikzcd}$$

Figure 14.2

Let us refer to $\mathcal{H}$ as the **measuring set** *and its members as* **measuring functions**.

A pair $(S, f\colon A \to S)$, where $S \in \mathcal{S}$ and $f \in \mathcal{F}$ has the **universal property for** *$\mathcal{F}$* **as *measured by*** *$\mathcal{H}$, or is a* **universal pair for** *$(\mathcal{F}, \mathcal{H})$, if for any $X \in \mathcal{S}$ and any $g\colon A \to X$ in $\mathcal{F}$, there is a unique $\tau\colon S \to X$ in $\mathcal{H}$ for which the diagram in Figure 14.1 commutes, that is,*

$$g = \tau \circ f$$

Another way to express this is to say that any $g \in \mathcal{F}$ can be **factored through** *f, or that any $g \in \mathcal{F}$ can be* **lifted** *to a function $\tau \in \mathcal{H}$ on S.* □

Universal pairs are essentially unique, as the following describes.

Theorem 14.1 *Let $(S, f\colon A \to S)$ and $(T, g\colon A \to T)$ be universal pairs for $(\mathcal{F}, \mathcal{H})$. Then there is a* bijective *measuring function $\mu \in \mathcal{H}$ for which $\mu(S) = T$.*

Proof. With reference to Figure 14.3, there are unique measuring functions $\tau\colon S \to T$ and $\sigma\colon T \to S$ for which

$$g = \tau \circ f$$
$$f = \sigma \circ g$$

Hence,

$$g = (\tau \circ \sigma) \circ g$$
$$f = (\sigma \circ \tau) \circ f$$

However, referring to the third diagram in Figure 14.3, both $\sigma \circ \tau\colon S \to S$ and the identity map $\iota\colon S \to S$ are members of $\mathcal{H}$ that make the diagram commute. Hence, the uniqueness requirement implies that $\sigma \circ \tau = \iota$. Similarly $\tau \circ \sigma = \iota$ and so τ and σ are inverses of one another, making $\mu = \sigma$ the desired bijection. □

$$\begin{tikzcd} A \arrow[r, "f"] \arrow[dr, "g"'] & S \arrow[d, dashed, "\tau"] \\ & T \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "g"] \arrow[dr, "f"'] & T \arrow[d, dashed, "\sigma"] \\ & S \end{tikzcd} \qquad \begin{tikzcd} A \arrow[r, "f"] \arrow[dr, "f"'] & S \arrow[d, dashed, "\sigma\tau=\iota"] \\ & S \end{tikzcd}$$

Figure 14.3

Now let us look at some examples of the universal property. Let Vect(F) denote the family of all vector spaces over the base field F. (We use the term *family* informally to represent what in set theory is formally referred to as a *class. A class* is a "collection" that is too large to be considered a set. For example, Vect(F) is a class.)

Example 14.1 (*Bases: the universal property for set functions from a set A into a vector space, as measured by linearity*) Let A be a nonempty set. Let $\mathcal{S} = \text{Vect}(F)$ and let $\mathcal{F}$ be the family of set functions with domain A. The measuring set $\mathcal{H}$ is the family of linear transformations.

If V_A is the vector space with basis A, that is, the set of all formal linear combinations of members of A with coefficients in F, then the pair $(V_A, j\colon A \to V_A)$ where j is the injection map $j(a) = a$, is universal for $(\mathcal{F}, \mathcal{H})$. To see this, note that the condition that $g \in \mathcal{F}$ can be factored through j

$$g = \tau \circ j$$

is equivalent to the statement that $\tau(a) = g(a)$ for each basis vector $a \in A$. But this uniquely defines a linear transformation τ. Note also that Theorem 14.1 implies that if $(W, k\colon A \to W)$ is also universal for $(\mathcal{F}, \mathcal{H})$, then there is a bijective measuring function from V_A to W, that is, W and V_A are isomorphic. □

Example 14.2 (*Quotient spaces and canonical projections: the universal property for linear transformations from V whose kernel contains a particular subspace K of V, as measured by linearity*) Let V be a vector space with subspace K. Let $\mathcal{S} = \text{Vect}(F)$. Let $\mathcal{F}$ be the family of linear transformations with domain V whose kernel contains K. The measuring set $\mathcal{H}$ is the family of linear transformations. Then Theorem 3.4 says precisely that the pair $(V/K, \pi\colon V \to V/K)$, where π is the canonical projection map, has the universal property for $\mathcal{F}$ as measured by $\mathcal{H}$. □

Example 14.3 (*Direct sums: the universal property for pairs (f, g) of linear maps with the same range, where f has domain U and g has domain V, as measured by linearity*) Let U and V be vector spaces and consider the ordered pair (U,V). Let $\mathcal{S} = \text{Vect}(F)$. A member of $\mathcal{F}$ is an ordered pair $(f{:}\,U \to W, g{:}\,V \to W)$ of linear transformations, written

$$(f,g){:}\,(U,V) \to W$$

for which

$$(f,g)(u,v) = f(u) + g(v)$$

The measure set $\mathcal{H}$ is the set of all linear transformations. For $\tau \in \mathcal{H}$ and $(f,g) \in \mathcal{F}$, we set

$$\tau \circ (f,g) = (\tau \circ f, \tau \circ g)$$

We claim that the pair $(U \boxplus V, (j_1,j_2){:}\,(U,V) \to U \boxplus V)$, where

$$\begin{aligned} j_1(u) &= (u,0) \\ j_2(v) &= (0,v) \end{aligned}$$

are the canonical injections, has the universal property for $(\mathcal{F},\mathcal{H})$.

To see this, observe that for any function $(f,g){:}\,(U,V) \to W$, that is, for any pair of linear transformations $f{:}\,U \to W$ and $g{:}\,V \to W$, the condition

$$(f,g) = \tau \circ (j_1,j_2)$$

is equivalent to

$$(f,g) = (\tau \circ j_1, \tau \circ j_2)$$

or

$$f(u) + g(v) = \tau(u,0) + \tau(0,v) = \tau(u,v)$$

But the condition $\tau(u,v) = f(u) + g(v)$ does indeed define a unique linear transformation $\tau{:}\,U \boxplus V \to W$. (For those familiar with category theory, we have essentially defined the coproduct of vector spaces, which is equivalent to the product, or direct sum.)□

It should be clear from these examples that the notion of universal property is, well, universal. In fact, it happens that the most useful definition of tensor product is through its universal property.

Bilinear Maps

The universality that defines tensor products rests on the notion of a bilinear map.

Definition *Let U, V and W be vector spaces over F. Let $U \times V$ be the cartesian product of U and V* as sets. *A set function*

$$f: U \times V \to W$$

is **bilinear** *if it is linear in both variables separately, that is, if*

$$f(ru + su', v) = rf(u,v) + sf(u',v)$$

and

$$f(u, rv + sv') = rf(u,v) + sf(u,v')$$

The set of all bilinear functions from $U \times V$ to W is denoted by $\hom_F(U,V;W)$. *A bilinear function $f: U \times V \to F$, with values in the base field F, is called a* **bilinear form** *on $U \times V$.* □

It is important to emphasize that, in the definition of bilinear function, $U \times V$ is the *cartesian product of sets*, not the direct product of vector spaces. In other words, we do not consider any algebraic structure on $U \times V$ when defining bilinear functions, so equations like

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

and

$$r(x,y) = (rx, ry)$$

are false.

In fact, if V is a vector space, we have two classes of functions from $V \times V$ to W, the linear maps $\mathcal{L}(V \times V, W)$ where $V \times V$ is the direct product of vector spaces, and the bilinear maps $\hom(V,V;W)$, where $V \times V$ is just the cartesian product of sets. We leave it as an exercise to show that these two classes have only the zero map in common. In other words, the only map that is both linear and bilinear is the zero map.

We made a thorough study of bilinear forms on a finite-dimensional vector space V in Chapter 11 (although this material is not assumed here). However, bilinearity is far more important and far reaching than its application to metric vector spaces, as the following examples show. Indeed, both multiplication and evaluation are bilinear.

Example 14.4 (Multiplication is bilinear) If A is an algebra, the product map $\mu: A \times A \to A$ defined by

$$\mu(a,b) = ab$$

is bilinear. Put another way, multiplication is linear in each position. □

Example 14.5 (Evaluation is bilinear) If V and W are vector spaces, then the *evaluation map* $\phi: \mathcal{L}(V,W) \times V \to W$ defined by

$$\phi(f,v)=f(v)$$

is bilinear. In particular, the evaluation map $\phi: V^* \times V \to F$ defined by $\phi(f,v)=f(v)$ is a bilinear form on $V^* \times V$. □

Example 14.6 If V and W are vector spaces, and $f \in V^*$ and $g \in W^*$ then the product map $\phi: V \times W \to F$ defined by

$$\phi(v,w)=f(v)g(w)$$

is bilinear. Dually, if $v \in V$ and $w \in W$ then the map $\lambda: V^* \times W^* \to F$ defined by

$$\lambda(f,g)=f(v)g(w)$$

is bilinear. □

It is precisely the tensor product that will allow us to generalize the previous example. In particular, if $\tau \in \mathcal{L}(U,W)$ and $\sigma \in \mathcal{L}(V,W)$ then we would like to consider a "product" map $\phi: U \times V \to W$ defined by

$$\phi(u,v)=f(u)\ ?\ g(v)$$

The tensor product $\otimes$ is just the thing to replace the question mark, because it has the desired bilinearity property, as we will see. In fact, the tensor product is bilinear and nothing else, so it is *exactly* what we need!

Tensor Products

Let U and V be vector spaces. Our guide for the definition of the tensor product $U \otimes V$ will be the desire to have a universal property for bilinear functions, as measured by linearity. Put another way, we want $U \otimes V$ to embody the notion of bilinearity but nothing more, that is, we want it to be *universal for bilinearity*.

Referring to Figure 14.4, we seek to define a vector space T and a bilinear map $t: U \times V \to T$ so that any bilinear map f with domain $U \times V$ can be factored through t. Intuitively speaking, t is the most "general" or "universal" bilinear map with domain $U \times V$.

$$\begin{tikzcd}
U \times V \arrow[r, "t\ \text{bilinear}"] \arrow[dr, "f\ \text{bilinear}"'] & T \arrow[d, dashed, "\tau\ \text{linear}"] \\
& W
\end{tikzcd}$$

Figure 14.4

Definition *Let $U \times V$ be the cartesian product of two vector spaces over F. Let $\mathcal{S} = \mathrm{Vect}(F)$. Let*

$$\mathcal{F} = \{\hom_F(U, V; W\} \mid W \in \mathcal{S}\}$$

be the family of all bilinear maps from $U \times V$ to a vector space W. The measuring set $\mathcal{H}$ is the family of all linear transformations.

A pair $(T, t{:}\, U \times V \to T)$ is **universal for bilinearity** *if it is universal for $(\mathcal{F}, \mathcal{H})$, that is, if for every bilinear map $g{:}\, U \times V \to W$, there is a* unique *linear transformation $\tau{:}\, U \otimes V \to W$ for which*

$$g = \tau \circ t$$ □

Let us now turn to the question of the existence of a universal pair for bilinearity.

Existence I: Intuitive but Not Coordinate Free

The universal property for bilinearity captures the essence of bilinearity *and nothing more* (as is the intent for all universal properties). To understand better how this can be done, let $\mathcal{B} = \{e_i \mid i \in I\}$ be a basis for U and let $\mathcal{C} = \{f_j \mid j \in J\}$ be a basis for V. Then a bilinear map t on $U \times V$ is uniquely determined by assigning arbitrary values to the "basis" pairs (e_i, f_j). How can we do this *and nothing more*?

The answer is that we should define t on the pairs (e_i, f_j) in such a way that the images $t(e_i, f_j)$ *do not interact* and then extend by bilinearity.

In particular, for each ordered pair (e_i, f_j), we invent a new formal symbol, say $e_i \otimes f_j$ and define T to be the vector space with basis

$$\mathcal{D} = \{e_i \otimes f_j \mid e_i \in \mathcal{B}, f_j \in \mathcal{C}\}$$

Then define the map t by setting $t(e_i, f_j) = e_i \otimes f_j$ and extending by bilinearity. This uniquely defines a bilinear map t that is as "universal" as possible among bilinear maps.

Indeed, if $g{:}\, U \times V \to W$ is bilinear, the condition $g = \tau \circ t$ is equivalent to

$$\tau(e_i \otimes f_j) = g(e_i, f_j)$$

which uniquely defines a linear map $\tau{:}\, T \to W$. Hence, (T, t) has the universal property for bilinearity.

A typical element of T is a finite linear combination

$$\sum_{i,j=1}^{n} \alpha_{i,j}(e_{k_i} \otimes f_{k_j})$$

and if $u = \sum \alpha_i e_i$ and $v = \sum \beta_j f_j$ then

$$u \otimes v = t(u,v) = t\Big(\sum \alpha_i e_i, \sum \beta_j f_j\Big) = \sum \alpha_i \beta_j (e_i \otimes f_j)$$

Note that, as is customary, we have used the notation $u \otimes v$ for the image of *any* pair (u,v) under t. Strictly speaking, this is an abuse of the notation $\otimes$ as we have defined it. While it may seem innocent, it does contribute to the reputation that tensor products have for being a bit difficult to fathom.

Confusion may arise because while the elements $u_i \otimes v_j$ form a basis for T (by definition), the larger set of elements of the form $u \otimes v$ span T, but are definitely not linearly independent. This raises various questions, such as when a sum of the form $\sum u_i \otimes v_j$ is equal to 0, or when we can define a map τ on T by specifying the values $\tau(u \otimes v)$ arbitrarily. The first question seems more obviously challenging when we phrase it by asking when a sum of the form $\sum t(u_i, v_j)$ is 0, since there is no algebraic structure on the cartesian product $U \times V$, and so there is nothing "obvious" that we can do with this sum. The second question is not difficult to answer when we keep in mind that the set $\{u \otimes v\}$ is not linearly independent.

The notation $\otimes$ is used in yet another way: T is generally denoted by $U \otimes V$ and called the **tensor product** of U and V. The elements of $U \otimes V$ are called **tensors** and a tensor of the form $u \otimes v$ is said to be **decomposable**. For example, in $\mathbb{R}^2 \otimes \mathbb{R}^2$, the tensor $(1,1) \otimes (1,2)$ is decomposable but the tensor $(1,1) \otimes (1,2) + (1,2) \otimes (2,3)$ is not.

It is also worth emphasizing that the tensor product $\otimes$ is not a product in the sense of a binary operation on a set, as is the case in rings and fields, for example. In fact, even when $V = U$, the tensor product $u \otimes u$ is not in U, but rather in $U \otimes U$. It is wise to remember that the decomposable tensor $u \otimes v$ is nothing more than the image of the ordered pair (u,v) under the bilinear map t, as are the basis elements $e_i \otimes f_j$.

Existence II: Coordinate Free

The previous definition of tensor product is about as intuitive as possible, but has the disadvantage of not being coordinate free. The following customary approach to defining the tensor product does not require the choice of a basis.

Let $F_{U \times V}$ be the vector space over F with basis $U \times V$. Let S be the subspace of $F_{U \times V}$ generated by all vectors of the form

$$r(u,w)+s(v,w)-(ru+sv,w) \tag{14.1}$$

and

$$r(u,v)+s(u,w)-(u,rv+sw) \tag{14.2}$$

where $r,s\in F$ and u,v and w are in the appropriate spaces. Note that these vectors are 0 if we replace the ordered pairs by tensors according to our previous definition.

The quotient space

$$U\otimes V=\frac{F_{U\times V}}{S}$$

is also called the **tensor product** of U and V. The elements of $U\otimes V$ have the form

$$\left(\sum r_i(u_i,v_i)\right)+S=\sum r_i[(u_i,v_i)+S]$$

However, since $r(u,v)-(ru,v)\in S$ and $r(u,v)-(u,rv)\in S$, we can absorb the scalar in either coordinate, that is,

$$r[(u,v)+S]=(ru,v)+S=(u,rv)+S$$

and so the elements of $U\otimes V$ can be written simply as

$$\sum[(u_i,v_i)+S]$$

It is customary to denote the coset $(u,v)+S$ by $u\otimes v$ and so any element of $U\otimes V$ has the form

$$\sum u_i\otimes v_i$$

as before. Finally, the map $t\colon U\times V\to U\otimes V$ is defined by

$$t(u,v)=u\otimes v$$

The proof that the pair $(U\otimes V,t\colon U\times V\to U\otimes V)$ is universal for bilinearity is a bit more tedious when $U\otimes V$ is defined as a quotient space.

Theorem 14.2 *Let U and V be vector spaces. The pair*

$$(U\otimes V,t\colon U\times V\to U\otimes V)$$

has the universal property for bilinearity, as measured by linearity.

Proof. Consider the diagram in Figure 14.5. Here $F_{U\times V}$ is the vector space with basis $U\times V$.

$$\begin{tikzcd}
U \times V \arrow[r, "j"] \arrow[rr, bend left, "t"] \arrow[dr, "f"'] & F_{U \times V} \arrow[r, "\pi"] \arrow[d, dashed, "\sigma"] & U \otimes V \arrow[dl, dashed, "\tau"] \\
& W &
\end{tikzcd}$$

Figure 14.5

Since $t(u,v) = u \otimes v = \pi \circ j(u,v)$, we have

$$t = \pi \circ j$$

The universal property of vector spaces described in Example 14.1 implies that there is a unique linear transformation $\sigma\colon F_{U \times V} \to W$ for which

$$\sigma \circ j = f$$

Note that σ sends any of the vectors (14.1) and (14.2) that generate S to the zero vector and so $S \subseteq \ker(\sigma)$. For example,

$$\begin{aligned}
\sigma[r(u,w) + s(v,w) - (ru + sv, w)]& \\
&= \sigma[rj(u,w) + sj(v,w) - j(ru + sv, w)] \\
&= r\sigma j(u,w) + s\sigma j(v,w) - \sigma j(ru + sv, w) \\
&= rf(u,w) + sf(v,w) - f(ru + sv, w) \\
&= 0
\end{aligned}$$

and similarly for the second coordinate. Hence, we may apply Theorem 3.4 (the universal property described in Example 14.2), to deduce the existence of a unique linear transformation $\tau\colon U \otimes V \to W$ for which

$$\tau \circ \pi = \sigma$$

Hence,

$$\tau \circ t = \tau \circ \pi \circ j = \sigma \circ j = f$$

As to uniqueness, if $\tau' \circ t = f$ then $\sigma' = \tau' \circ \pi$ satisfies

$$\sigma' \circ j = \tau' \circ \pi \circ j = \tau' \circ t = f$$

The uniqueness of σ then implies that $\sigma' = \sigma$, which in turn implies that

$$\tau' \circ \pi = \sigma' = \sigma = \tau \circ \pi$$

and the uniqueness of τ implies that $\tau' = \tau$. □

We now have two definitions of tensor product that are equivalent, since under the second definition the tensors

$$(e_i, f_j) + S = e_i \otimes f_j$$

form a basis for $F_{U \times V}/S$, as we will prove a bit later. Accordingly, we no longer need to make a distinction between the two definitions.

Bilinearity on $U \times V$ *Equals Linearity on* $U \otimes V$

The universal property for bilinearity says that to each *bilinear* function $f: U \times V \to W$, there corresponds a unique *linear* function $\tau: U \otimes V \to W$. This establishes a correspondence

$$\phi: \hom(U, V; W) \to \mathcal{L}(U \otimes V, W)$$

given by $\phi(f) = \tau$. In other words, $\phi(f): U \otimes V \to W$ is the unique *linear* map for which

$$\phi(f)(u \otimes v) = f(u, v)$$

Observe that ϕ is linear, since if $f, g \in \hom(U, V; W)$ then

$$[r\phi(f) + s\phi(g)](u \otimes v) = rf(u, v) + sg(u, v) = (rf + sg)(u, v)$$

and so the uniqueness part of the universal property implies that

$$r\phi(f) + s\phi(g) = \phi(rf + sg)$$

Also, ϕ is surjective, since if $\tau: U \otimes V \to W$ is any linear map then $f = \tau \circ t: U \times V \to W$ is bilinear and by the uniqueness part of the universal property, we have $\phi(f) = \tau$. Finally, ϕ is injective, for if $\phi(f) = 0$ then $f = \phi(f) \circ t = 0$. We have established the following result.

Theorem 14.3 *Let U, V and W be vector spaces over F. Then the map $\phi: \hom(U, V; W) \to \mathcal{L}(U \otimes V, W)$, where $\phi(f)$ is the unique linear map satisfying $f = \phi(f) \circ t$, is an isomorphism. Thus,*

$$\hom(U, V; W) \approx \mathcal{L}(U \otimes V, W) \qquad \square$$

Armed with the definition and the universal property, we can now discuss some of the basic properties of tensor products.

When Is a Tensor Product Zero?

Let us consider the question of when a tensor $\sum u_i \otimes v_j$ is zero. The universal property proves to be very helpful in deciding this question.

First, note that the bilinearity of the tensor product gives

$$0 \otimes v = (0 + 0) \otimes v = 0 \otimes v + 0 \otimes v$$

and so $0 \otimes v = 0$. Similarly, $u \otimes 0 = 0$.

Now suppose that

$$\sum_i u_i \otimes v_i = 0$$

where we may assume that none of the vectors u_i and v_i are 0. According to the universal property of the tensor product, for any bilinear function $f\colon U \times V \to W$, there is a unique linear transformation $\tau\colon U \otimes V \to W$ for which $\tau \circ t = f$. Hence

$$0 = \tau\Big(\sum_i u_i \otimes v_i\Big) = \sum_i (\tau \circ t)(u_i, v_i) = \sum_i f(u_i, v_i)$$

The key point is that this holds for *any* bilinear function $f\colon U \times V \to W$. One possibility for f is to take two linear functionals $\alpha \in U^*$ and $\beta \in V^*$and multiply them

$$f(u, v) = \alpha(u)\beta(v)$$

which is easily seen to be bilinear and gives

$$\sum_i \alpha(u_i)\beta(v_i) = 0$$

If, for example, the vectors u_i are linearly independent, then we can consider the dual vectors u_i^*, for which $u_i^*(u_j) = \delta_{i,j}$. Setting $\alpha = u_k^*$ gives

$$0 = \sum_i u_k^*(u_i)\beta(v_i) = \beta(v_k)$$

for all linear functionals $\beta \in V^*$. This implies that $v_k = 0$. We have proved the following useful result.

Theorem 14.4 *If $u_1, \ldots, u_n$ are linearly independent vectors in U and $v_1, \ldots, v_n$ are arbitrary vectors in V then*

$$\sum u_i \otimes v_i = 0 \Rightarrow v_i = 0 \textit{ for all } i$$

In particular, $u \otimes v = 0$ if and only if $u = 0$ or $v = 0$. □

The next result says that we can get a basis for the tensor product $U \otimes V$ simply by "tensoring" any bases for each coordinate space. As promised, this shows that the two definitions of tensor product are essentially equivalent.

Theorem 14.5 *Let $\mathcal{B} = \{u_i \mid i \in I\}$ be a basis for U and let $\mathcal{C} = \{v_j \mid j \in J\}$ be a basis for V. Then the set*

$$\mathcal{D} = \{u_i \otimes v_j \mid i \in I,\ j \in J\}$$

is a basis for $U \otimes V$.

Proof. To see that the $\mathcal{D}$ is linearly independent, suppose that

$$\sum_{i,j} r_{i,j}(u_i \otimes v_j) = 0$$

This can be written

$$\sum_i u_i \otimes \Big(\sum_j r_{i,j}v_j\Big) = 0$$

and so, by Theorem 14.4, we must have

$$\sum_j r_{i,j}v_j = 0$$

for all i and hence $r_{i,j} = 0$ for all i and j. To see that $\mathcal{D}$ spans $U \otimes V$, let $u \otimes v \in U \otimes V$. Then since $u = \sum r_i u_i$ and $v = \sum s_j v_j$, we have

$$\begin{aligned} u \otimes v &= \sum_i r_i u_i \otimes \sum_j s_j v_j \\ &= \sum_i r_r \Big(u_i \otimes \sum_j s_j v_j\Big) \\ &= \sum_i r_i \Big(\sum_j s_j (u_i \otimes v_j)\Big) \\ &= \sum_{i,j} r_i s_j (u_i \otimes v_j) \end{aligned}$$

Hence, any sum of elements of the form $u \otimes v$ is a linear combination of the vectors $u_i \otimes v_j$, as desired. □

Corollary 14.6 *For finite-dimensional vector spaces,*

$$\dim(U \otimes V) = \dim(U) \cdot \dim(V) \qquad \square$$

Coordinate Matrices and Rank

If $\mathcal{B} = \{u_i \mid i \in I\}$ is a basis for U and $\mathcal{C} = \{v_j \mid j \in J\}$ is a basis for V, then any vector $z \in U \otimes V$ has a unique expression as a sum

$$z = \sum_{i \in I}\sum_{j \in J} r_{i,j}(u_i \otimes v_j)$$

where only a finite number of the coefficients $r_{i,j}$ are nonzero. In fact, for a fixed $z \in U \otimes V$, we may reindex the bases so that

$$z = \sum_{i=1}^{a}\sum_{j=1}^{b} r_{i,j}(u_i \otimes v_j)$$

where none of the rows or columns of the matrix $R = (r_{i,j})$ consists only of 0's. The matrix $R = (r_{i,j})$ is called a **coordinate matrix** of z with respect to the bases $\mathcal{B}$ and $\mathcal{C}$.

Note that a coordinate matrix R is determined only up to the order of its rows and columns. We could remove this ambiguity by considering ordered bases, but this is not necessary for our discussion and adds a complication since the bases may be infinite.

Suppose that $\mathcal{W} = \{w_i \mid i \in I\}$ and $\mathcal{X} = \{x_j \mid j \in J\}$ are also bases for U and V, respectively and that

$$z = \sum_{i=1}^{c}\sum_{j=1}^{d} s_{i,j}(w_i \otimes x_j)$$

where $S = (s_{i,j})$ is a coordinate matrix of z with respect to these bases. We claim that the coordinate matrices R and S have the same rank, which can then be defined as the **rank** of the tensor $z \in U \otimes V$.

Each $w_1, \ldots, w_c$ is a finite linear combination of basis vectors in $\mathcal{B}$, perhaps involving some of $u_1, \ldots, u_a$ and perhaps involving other vectors in $\mathcal{B}$. We can further reindex $\mathcal{B}$ so that each w_k is a linear combination of the vectors $u_1, \ldots, u_n$, where $a \le n$ and set

$$U_n = \text{span}(u_1, \ldots, u_n)$$

Next, extend $\{w_1, \ldots, w_c\}$ to a basis $\mathcal{W}' = \{w_1, \ldots, w_c, w_{c+1}, \ldots, w_n\}$ for U_n. (Since we no longer need the rest of the basis $\mathcal{W}$, we have commandeered the symbols $w_{c+1}, \ldots, w_n$, for simplicity.) Hence

$$w_i = \sum_{h=1}^{n} a_{i,h}u_h \text{ for } i = 1, \ldots, n$$

where $A = (a_{i,h})$ is invertible of size $n \times n$.

Now repeat this process on the second coordinate. Reindex the basis $\mathcal{C}$ so that the subspace $V_m = \text{span}(v_1, \ldots, v_m)$ contains $x_1, \ldots, x_d$ and extend to a basis $\mathcal{X}' = \{x_1, \ldots, x_d, x_{d+1}, \ldots, x_m\}$ for V_m. Then

$$x_j = \sum_{k=1}^{m} b_{j,k}v_k \text{ for } j = 1, \ldots, m$$

where $B = (b_{j,k})$ is invertible of size $m \times m$.

Next, write

$$z = \sum_{i=1}^{n}\sum_{j=1}^{m} r_{i,j}(u_i \otimes v_j)$$

by setting $r_{i,j} = 0$ for $i > a$ or $j > b$. Thus, the $n \times m$ matrix $R_1 = (r_{i,j})$ comes from R by adding $n - a$ rows of 0's to the bottom and then $m - b$ columns of 0's. In particular, R_1 and R have the same rank.

The expression for z in terms of the basis vectors $w_1, \ldots, w_c$ and $x_1, \ldots, x_d$ can also be extended using 0 coefficients to

$$z = \sum_{i=1}^{n}\sum_{j=1}^{m} s_{i,j}(w_i \otimes x_j)$$

where the $n \times m$ matrix $S_1 = (s_{i,j})$ has the same rank as S.

Now at last, we can compute

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{m} s_{i,j}(w_i \otimes x_j) &= \sum_{i=1}^{n}\sum_{j=1}^{m} s_{i,j}\left(\sum_{h=1}^{n} a_{i,h}u_h \otimes \sum_{k=1}^{m} b_{j,k}v_k\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{h=1}^{n}\sum_{k=1}^{m} a_{i,h}s_{i,j}b_{j,k}(u_h \otimes v_k)\\
&= \sum_{h=1}^{n}\sum_{k=1}^{m}\sum_{j=1}^{m}\sum_{i=1}^{n} (a^t_{h,i}s_{i,j})b_{j,k}(u_h \otimes v_k)\\
&= \sum_{h=1}^{n}\sum_{k=1}^{m}\sum_{j=1}^{m} (A^tS_1)_{h,j}b_{j,k}(u_h \otimes v_k)\\
&= \sum_{h=1}^{n}\sum_{k=1}^{m} (A^tS_1B)_{h,k}(u_h \otimes v_k)
\end{aligned}$$

and so

$$\sum_{i=1}^{n}\sum_{j=1}^{m} r_{i,j}(u_i \otimes v_j) = \sum_{h=1}^{n}\sum_{k=1}^{m} (A^tS_1B)_{h,k}(u_h \otimes v_k)$$

It follows that $R_1 = A^tS_1B$. Since A and B are invertible, we deduce that

$$\text{rk}(R) = \text{rk}(R_1) = \text{rk}(S_1) = \text{rk}(S)$$

In block terms

$$R_1 = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad S_1 = \begin{bmatrix} S & 0 \\ 0 & 0 \end{bmatrix}$$

and

$$A^t = \begin{bmatrix} A^t_{a,c} & * \\ * & * \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{d,b} & * \\ * & * \end{bmatrix}$$

Then $R_1 = A^t S_1 B$ implies that, for the original coordinate matrices,

$$R = A^t_{a,c} S B_{d,b}$$

where $\text{rk}(A^t_{a,c}) \geq \text{rk}(R)$ and $\text{rk}(B_{d,b}) \geq \text{rk}(R)$.

We shall soon have use for the following special case. If

$$z = \sum_{i=1}^{r} u_i \otimes v_i = \sum_{i=1}^{r} w_i \otimes x_i \tag{14.3}$$

then, in the preceding argument, $a = b = c = d = r$ and $R = S = I_r$ and

$$R_1 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad S_1 = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}$$

Hence, the equation $R = A^t_{a,c} S B_{d,b}$ becomes

$$I_r = A^t_{r,r} B_{r,r}$$

and we further have

$$w_i = \sum_{h=1}^{r} a_{i,h} u_h \text{ for } i = 1, \ldots, r$$

where $A_{r,r} = (a_{i,h})$ and

$$x_j = \sum_{k=1}^{r} b_{j,k} v_k \text{ for } j = 1, \ldots, r$$

where $B_{r,r} = (b_{j,k})$.

The Rank of a Decomposable Tensor

Recall that a tensor of the form $u \otimes v$ is said to be decomposable. If $\{u_i \mid i \in I\}$ is a basis for U and $\{v_j \mid j \in J\}$ is a basis for V then any decomposable vector has the form

$$u \otimes v = \sum_{i,j} r_i s_j (u_i \otimes v_j)$$

Hence, the rank of a decomposable vector is 1. This implies that the set of decomposable vectors is quite "small" in $U \otimes V$, as long as neither vector space has dimension 1.

Characterizing Vectors in a Tensor Product

There are several very useful representations of the tensors in $U \otimes V$.

Theorem 14.7 *Let $\{u_i \mid i \in I\}$ be a basis for U and let $\{v_j \mid j \in J\}$ be a basis for V. By a "unique" sum, we mean unique up to order and presence of zero terms. Then*

1) *Every $z \in U \otimes V$ has a unique expression as a finite sum of the form*
$$\sum_{i,j} r_{i,j} u_i \otimes v_j$$
where $r_{i,j} \in F$.
2) *Every $z \in U \otimes V$ has a unique expression as a finite sum of the form*
$$\sum_i u_i \otimes y_i$$
where $y_i \in V$.
3) *Every $z \in U \otimes V$ has a unique expression as a finite sum of the form*
$$\sum_i x_i \otimes v_i$$
where $x_i \in U$.
4) *Every nonzero $z \in U \otimes V$ has an expression of the form*
$$\sum_{i=1}^{n} x_i \otimes y_i$$
where $\{x_i\} \subseteq U$ and $\{y_i\} \subseteq V$ are linearly independent sets. As to uniqueness, n is the rank of z and so it is unique. Also, we have
$$\sum_{i=1}^{r} x_i \otimes y_i = \sum_{i=1}^{r} w_i \otimes z_i$$
where $\{w_i\} \subseteq U$ and $\{z_i\} \subseteq V$ are linearly independent sets, if and only if there exist $r \times r$ matrices $A = (a_{i,j})$ and $B = (b_{i,j})$ for which $A^t B = I$ and
$$w_i = \sum_{j=1}^{r} a_{i,j} x_j \quad \textit{and} \quad z_i = \sum_{j=1}^{r} b_{i,j} y_j$$
for $k = 1, \ldots, r$.

Proof. Part 1) merely expresses the fact that $\{u_i \otimes v_j\}$ is a basis for $U \otimes V$. From part 1), we write
$$\sum_{i,j} r_{i,j} u_i \otimes v_j = \sum_i \left[u_i \otimes \sum_j r_{i,j} v_j \right] = \sum_i u_i \otimes y_i$$

which is part 2). Uniqueness follows from Theorem 14.4. Part 3) is proved similarly. As to part 4), we start with the expression from part 2)

$$\sum_{i=1}^{n} u_i \otimes y_i$$

where we may assume that none of the y_i's are 0. If the set $\{y_i\}$ is linearly independent, we are done. If not, then we may suppose (after reindexing if necessary) that

$$y_n = \sum_{i=1}^{n-1} r_i y_i$$

Then

$$\begin{aligned}\sum_{i=1}^{n} u_i \otimes y_i &= \sum_{i=1}^{n-1} u_i \otimes y_i + \left(u_n \otimes \sum_{i=1}^{n-1} r_i y_i\right)\\ &= \sum_{i=1}^{n-1} u_i \otimes y_i + \sum_{i=1}^{n-1} (r_i u_n \otimes y_i)\\ &= \sum_{i=1}^{n-1} (u_i + r_i u_n) \otimes y_i\end{aligned}$$

But the vectors $\{u_i + r_i u_n \mid 1 \le i \le n-1\}$ are linearly independent. This reduction can be repeated until the second coordinates are linearly independent. Moreover, the identity matrix I_n is a coordinate matrix for z and so $n = \text{rk}(I_n) = \text{rk}(z)$. As to uniqueness, one direction was proved earlier; see (14.3) and the other direction is left to the reader. □

Defining Linear Transformations on a Tensor Product

One of the simplest and most useful ways to define a linear transformation σ on the tensor product $U \otimes V$ is through the universal property, for this property says precisely that a bilinear function f on $U \times V$ gives rise to a unique (and well-defined) linear transformation on $U \otimes V$. The proof of the following theorem illustrates this well. It says that a linear functional on the tensor product is nothing more or less than a tensor product of linear functionals.

Theorem 14.8 *Let U and V be finite-dimensional vector spaces. Then*

$$U^* \otimes V^* \approx (U \otimes V)^*$$

via the isomorphism $\tau: U^* \otimes V^* \to (U \otimes V)^*$ given by

$$\tau(f \otimes g)(u \otimes v) = f(u)g(v)$$

Proof. Informally, for fixed f and g, the function $(u, v) \to f(u)g(v)$ is bilinear in u and v and so there is a unique linear map $\phi_{f,g}$ taking $u \otimes v$ to $f(u)g(v)$.

The function $(f,g) \to \phi_{f,g}$ is bilinear in f and g since, *as functions*, $\phi_{af+bg,h} = a\phi_{f,h} + b\phi_{g,h}$ and so there is a unique linear map τ taking $f \otimes g$ to $\phi_{f,g}$.

A bit more formally, for fixed f and g, the map $F_{f,g}: U \times V \to F$ defined by

$$F_{f,g}(u,v) = f(u)g(v)$$

is bilinear and so the universal property of tensor products implies that there exists a unique linear functional $\phi_{f,g}$ on $U \otimes V$ for which

$$\phi_{f,g}(u \otimes v) = F_{f,g}(u,v) = f(u)g(v)$$

Next, the map $G: U^* \times V^* \to (U \otimes V)^*$ defined by

$$G(f,g) = \phi_{f,g}$$

is bilinear since, for example,

$$\begin{aligned} G(rf+sg,h)(u \otimes v) &= \phi_{rf+sg,h}(u \otimes v) \\ &= (rf+sg)(u) \cdot h(v) \\ &= rf(u)h(v) + sg(u)h(v) \\ &= (r\phi_{f,h} + s\phi_{g,h})(u,v) \\ &= (rG(f,h) + sG(g,h))(u \otimes v) \end{aligned}$$

and so

$$G(rf+sg,h) = rG(f,h) + sG(g,h)$$

which shows that G is linear in its first coordinate. Hence, the universal property implies that there exists a unique linear map

$$\tau: U^* \otimes V^* \to (U \otimes V)^*$$

for which

$$\tau(f \otimes g) = G(f,g) = \phi_{f,g}$$

that is,

$$\tau(f \otimes g)(u \otimes v) = \phi_{f,g}(u \otimes v) = f(u)g(v)$$

Finally, we must show that τ is bijective. Let $\{b_i\}$ be a basis for U, with dual basis $\{\beta_i\}$ and let $\{c_i\}$ be a basis for V, with dual basis $\{\gamma_i\}$. Then

$$\tau(\beta_i \otimes \gamma_j)(b_u \otimes c_v) = \beta_i(b_u)\gamma_j(c_v) = \delta_{i,u}\delta_{j,v} = \delta_{(i,j),(u,v)}$$

and so $\{\tau(\beta_i \otimes \gamma_j)\} \subseteq (U \otimes V)^*$ is the dual basis to the basis $\{b_u \otimes c_v\}$ for $U \otimes V$. Thus, τ takes the basis $\{\beta_i \otimes \gamma_j\}$ for $U^* \otimes V^*$ to the basis $\{\tau(\beta_i \otimes \gamma_j)\}$ and is therefore bijective. □

Combining the isomorphisms of Theorem 14.3 and Theorem 14.8, we have, for finite-dimensional vector spaces U and V,

$$U^* \otimes V^* \approx (U \otimes V)^* \approx \hom(U, V; F)$$

The Tensor Product of Linear Transformations

We wish to generalize Theorem 14.8 to arbitrary linear transformations. Let $\tau \in \mathcal{L}(U, U')$ and $\sigma \in \mathcal{L}(V, V')$. While the product $\tau(u)\sigma(v)$ does not make sense, the *tensor* product $\tau(u) \otimes \sigma(v)$ does and is bilinear in u and v

$$f(u, v) = \tau(u) \otimes \sigma(v)$$

The same informal argument that we used in the proof of Theorem 14.8 will work here. Namely, the expression $\tau(u) \otimes \sigma(v) \in U' \otimes V'$ is bilinear in u and v and so there is a unique linear map, say $(\tau \odot \sigma): U \otimes V \to U' \otimes V'$ for which

$$(\tau \odot \sigma)(u \otimes v) = \tau(u) \otimes \sigma(v)$$

Since $\tau \odot \sigma \in \mathcal{L}(U \otimes V, U' \otimes V')$, we have a function

$$\phi: \mathcal{L}(U, U') \times \mathcal{L}(V, V') \to \mathcal{L}(U \otimes V, U' \otimes V')$$

defined by

$$\phi(\tau, \sigma) = \tau \odot \sigma$$

But ϕ is bilinear, since

$$\begin{aligned} ((a\tau + b\mu) \odot \sigma)(u, v) &= (a\tau + b\mu)(u) \otimes \sigma(v) \\ &= (a\tau(u) + b\mu(u)) \otimes \sigma(v) \\ &= a[\tau(u) \otimes \sigma(v)] + b[\mu(u) \otimes \sigma(v)] \\ &= a(\tau \odot \sigma)(u, v) + b(\mu \odot \sigma)(u, v) \\ &= (a(\tau \odot \sigma) + b(\mu \odot \sigma))(u, v) \end{aligned}$$

and similarly for the second coordinate. Hence, there is a unique linear transformation

$$\theta: \mathcal{L}(U, U') \otimes \mathcal{L}(V, V') \to \mathcal{L}(U \otimes V, U' \otimes V')$$

satisfying

$$\theta(\tau \otimes \sigma) = \tau \odot \sigma$$

that is,

$$[\theta(\tau \otimes \sigma)](u \otimes v) = \tau(u) \otimes \sigma(v)$$

To see that θ is injective, if $\theta(\tau \otimes \sigma) = 0$ then $\tau(u) \otimes \sigma(v) = 0$ for all $u \in U$ and $v \in V$. If $\sigma = 0$ then $\theta \otimes \sigma = 0$. If $\sigma \neq 0$, then there is a $v \in V$ for which $\sigma(v) \neq 0$. But $\tau(u) \otimes \sigma(v) = 0$ implies that one of the factors is 0 and so

$\tau(u)=0$ for all $u\in U$, that is, $\tau=0$. Hence, $\tau\otimes\sigma=0$. In either case, we see that θ is injective.

Thus, θ is an embedding (injective linear transformation) and if all vector spaces are finite-dimensional, then

$$\dim(\mathcal{L}(U,U')\otimes\mathcal{L}(V,V'))=\dim(\mathcal{L}(U\otimes V,U'\otimes V'))$$

and so θ is also surjective and hence an isomorphism.

The embedding of $\mathcal{L}(U,U')\otimes\mathcal{L}(V,V')$ into $\mathcal{L}(U\otimes V,U'\otimes V')$ means that each $\tau\otimes\sigma$ can be thought of as the linear transformation $\tau\odot\sigma$ from $U\otimes V$ to $U'\otimes V'$, defined by

$$(\tau\odot\sigma)(u\otimes v)=\tau(u)\otimes\sigma(v)$$

In fact, the notation $\tau\otimes\sigma$ is often used to denote both the tensor product of vectors (linear transformations) and the linear map $\tau\odot\sigma$, and we will do this as well. In summary, we can say that the tensor product $\tau\otimes\sigma$ of linear transformations is a linear transformation on tensor products.

Theorem 14.9 *The linear transformation*

$$\theta\colon\mathcal{L}(U,U')\otimes\mathcal{L}(V,V')\to\mathcal{L}(U\otimes V,U'\otimes V')$$

defined by $\theta(\tau\otimes\sigma)=\tau\odot\sigma$ *where*

$$(\tau\odot\sigma)(u\otimes v)=\tau(u)\otimes\sigma(v)$$

is an embedding (injective linear transformation), and is an isomorphism if all vector spaces are finite-dimensional. Thus, the tensor product $\tau\otimes\sigma$ *of linear transformations is (via this embedding) a linear transformation on tensor products.* □

There are several special cases of this result that are of importance.

Corollary 14.10 *Let us use the symbol* $X\overset{\sim}{\hookrightarrow}Y$ *to denote the fact that there is an embedding of* X *into* Y *that is an isomorphism if* X *and* Y *are finite-dimensional.*

1) *Taking* $U'=F$ *gives*

$$U^*\otimes\mathcal{L}(V,V')\overset{\sim}{\hookrightarrow}\mathcal{L}(U\otimes V,V')$$

where

$$(f\otimes\sigma)(u\otimes v)=f(u)\sigma(v)$$

for $f\in U^*$.

2) *Taking* $U' = F$ *and* $V' = F$ *gives*

$$U^* \otimes V^* \overset{\approx}{\hookrightarrow} (U \otimes V)^*$$

where

$$(f \otimes g)(u \otimes v) = f(u)g(v)$$

3) *Taking* $V = F$ *and noting that* $\mathcal{L}(F, V') \approx V'$ *and* $U \otimes F \approx U$ *gives (letting* $W = V'$*)*

$$\mathcal{L}(U, U') \otimes W \overset{\approx}{\hookrightarrow} \mathcal{L}(U, U' \otimes W)$$

where

$$(\tau \otimes w)(u) = \tau(u) \otimes w$$

4) *Taking* $U' = F$ *and* $V = F$ *gives (letting* $W = V'$*)*

$$U^* \otimes W \overset{\approx}{\hookrightarrow} \mathcal{L}(U, W)$$

where

$$(f \otimes w)(u) = f(u)w$$ □

Change of Base Field

The tensor product gives us a convenient way to extend the base field of a vector space. (We have already discussed the complexification of a real vector space.) For convenience, let us refer to a vector space over a field F as an F-**space** and write V_F. Actually, there are several approaches to "upgrading" the base field of a vector space. For instance, suppose that K is an extension field of F, that is, $F \subseteq K$. If $\{b_i\}$ is a basis for V_F then every $x \in V_F$ has the form

$$x = \sum r_i b_i$$

where $r_i \in F$. We can define an K-space V_K simply by taking all formal linear combinations of the form

$$x = \sum \alpha_i b_i$$

where $\alpha_i \in K$. Note that the dimension of V_K as a K-space is the same as the dimension of V_F as an F-space. Also, V_K is an F-space (just restrict the scalars to F) and as such, the inclusion map $j\colon V_F \to V_K$ sending $x \in V_F$ to $j(x) = x \in V_K$, is an F-monomorphism.

The approach described in the previous paragraph uses an arbitrarily chosen basis for V_F and is therefore not coordinate free. However, we can give a coordinate–free approach using tensor products as follows. Since K is a vector space over F, we can consider the tensor product

$$W_F = K \otimes_F V_F$$

It is customary to include the subscript F on $\otimes_F$ to denote the fact that the tensor product is taken with respect to the base field F. (All relevant maps are F-bilinear and F-linear.) However, since V_F is not a K-space, the only tensor product that makes sense in $K \otimes V_F$ is the F-tensor product and so we will drop the subscript F.

The vector space W_F is an F-space by definition of tensor product, but we may make it into a K-space as follows. For $\alpha \in K$, the temptation is to "absorb" the scalar α into the first coordinate

$$\alpha(\beta \otimes v) = (\alpha\beta) \otimes v$$

But we must be certain that this is well-defined, that is, that

$$\beta \otimes v = \gamma \otimes w \Rightarrow (\alpha\beta) \otimes v = (\alpha\gamma) \otimes w$$

This becomes easy if we turn to the universal property for bilinearity. In particular, consider the map $f_\alpha: (K \times V_F) \to (K \otimes V_F)$ defined by

$$f_\alpha(\beta, v) = (\alpha\beta) \otimes v$$

This map is obviously well-defined and since it is also bilinear, the universal property of tensor products implies that there is a unique (and well-defined!) F-linear map $\tau_\alpha: (K \otimes V_F) \to (K \otimes V_F)$ for which

$$\tau_\alpha(\beta \otimes v) = (\alpha\beta) \otimes v$$

Note also that since τ_α is F-linear, it is additive and so

$$\tau_\alpha[(\beta \otimes v) + (\gamma \otimes w)] = \tau_\alpha(\beta \otimes v) + \tau_\alpha(\gamma \otimes w)$$

that is,

$$\alpha[(\beta \otimes v) + (\gamma \otimes w)] = \alpha(\beta \otimes v) + \alpha(\gamma \otimes w)$$

which is one of the properties required of a scalar multiplication. Since the other defining properties of scalar multiplication are satisfied, the set $K \otimes V_F$ is indeed a K-space under this operation (and addition), which we denote by W_K.

To be absolutely clear, we have three distinct vector spaces: the F-spaces V_F and $W_F = K \otimes V_F$ and the K-space $W_K = K \otimes V_F$, where the tensor product in both cases is with respect to F. The spaces W_F and W_K are identical as sets and as abelian groups. It is only the "permission to multiply by" that is different. Even though in W_F we can multiply only by scalars from F, we still get the same *set* of vectors. Accordingly, we can recover W_F from W_K simply by restricting scalar multiplication to scalars from F.

It follows that we can speak of "F-linear" maps τ from V_F into W_K, with the expected meaning, that is,

$$\tau(ru+sv)=r\tau(u)+s\tau(v)$$

for all scalars $r,s\in F$ (not in K).

If the dimension of K as a vector space over F is d then

$$\dim(W_F)=\dim(K\otimes V_F)=\dim(K)\cdot\dim(V_F)=d\cdot\dim(V_F)$$

As to the dimension of W_K, it is not hard to see that if $\{b_i\}$ is a basis for V_F then $\{1\otimes b_i\}$ is a basis for W_K. Hence

$$\dim(W_K)=\dim(V_F)$$

even when V_F is infinite-dimensional.

The map $\mu\colon V_F\to W_F$ defined by $\mu(v)=1\otimes v$ is easily seen to be injective and F-linear and so W_F contains an isomorphic copy of V_F. We can also think of μ as mapping V_F into W_K, in which case μ is called the **K-extension map** of V_F. This map has a universal property of its own, as described in the next theorem.

Theorem 14.11 *The K-extension map $\mu\colon V_F\to W_K$ has the universal property for the family of all F-linear maps with domain V_F and range a K-space, as measured by K-linear maps. In particular, for any F-linear map $f\colon V_F\to Y$, where Y is a K-space, there exists a unique K-linear map $\tau\colon W_K\to Y$ for which the diagram in Figure 14.6 commutes, that is,*

$$\tau\circ\mu=f$$

Proof. If such a map $\tau\colon K\otimes V_F\to W$ is to exist then it must satisfy

$$\tau(\beta\otimes v)=\beta\tau(1\otimes v)=\beta\tau\mu(v)=\beta f(v) \tag{14.4}$$

This shows that if τ exists, it is uniquely determined by f. As usual, when searching for a linear map τ on a tensor product such as $W_K=K\otimes V_F$, we look for a bilinear map. Let $g\colon(K\times V_F)\to Y$ be defined by

$$g(\beta,v)=\beta f(v)$$

Since this is bilinear, there exists a unique F-linear map τ for which (14.4) holds. It is easy to see that τ is also K-linear, since if $\alpha\in K$ then

$$\tau[\alpha(\beta\otimes v)]=\tau(\alpha\beta\otimes v)=\alpha\beta f(v)=\alpha\tau(\beta\otimes v)$$ □

$$\begin{tikzcd}
V_F \arrow[r] \arrow[dr, "f"'] & W_K \arrow[d, "\tau"] \\
 & Y
\end{tikzcd}$$

Figure 14.6

Theorem 14.11 is the key to describing how to extend an F-linear map to a K-linear map. Figure 14.7 shows an F-linear map $\tau: V \to W$ between F-spaces V and W. It also shows the K-extensions for both spaces, where $K \otimes V$ and $K \otimes W$ are K-spaces.

$$\begin{tikzcd}
V \arrow[r, "\tau"] \arrow[d, "\mu_V"'] & W \arrow[d, "\mu_W"] \\
K\otimes V \arrow[r, "\overline{\tau}"] & K\otimes W
\end{tikzcd}$$

Figure 14.7

If there is a unique K-linear map $\overline{\tau}$ that makes the diagram in Figure 14.7 commute, then this would be the obvious choice for the extension of the F-linear map u to a K-linear map.

Consider the F-linear map $\sigma = (\mu_W \circ \tau): V \to K \otimes W$ into the K-space $K \otimes W$. Theorem 14.11 implies that there is a unique K-linear map $\overline{\tau}: K \otimes V \to K \otimes W$ for which

$$\overline{\tau} \circ \mu_V = \sigma$$

that is,

$$\overline{\tau} \circ \mu_V = \mu_W \circ \tau$$

Now, $\overline{\tau}$ satisfies

$$\begin{aligned}
\overline{\tau}(\beta \otimes v) &= \beta\overline{\tau}(1 \otimes v) \\
&= \beta(\overline{\tau} \circ \mu_V)(v) \\
&= \beta(\mu_W \circ \tau)(v) \\
&= \beta(1 \otimes \tau(v)) \\
&= \beta \otimes \tau(v) \\
&= (\iota_K \otimes \tau)(\beta \otimes v)
\end{aligned}$$

and so $\overline{\tau} = \iota_K \otimes \tau$.

Theorem 14.12 *Let V and W be F-spaces, with K-extension maps μ_V and μ_W, respectively. (See Figure 14.7.) Then for any F-linear map $\tau\colon V \to W$, the map $\overline{\tau} = \iota_K \otimes \tau$ is the unique K-linear map that makes the diagram in Figure 14.7 commute, that is, for which*

$$\mu \circ \tau = \overline{\tau} \circ \nu \qquad \square$$

Multilinear Maps and Iterated Tensor Products

The tensor product operation can easily be extended to more than two vector spaces. We begin with the extension of the concept of bilinearity.

Definition *If $V_1, \ldots, V_n$ and W are vector spaces over F, a function $f\colon V_1 \times \cdots \times V_n \to W$ is said to be* **multilinear** *if it is linear in each variable separately, that is, if*

$$\begin{aligned} &f(u_1, \ldots, u_{k-1}, rv + sv', u_{k+1}, \ldots, u_n) = \\ &\quad rf(u_1, \ldots, u_{k-1}, v, u_{k+1}, \ldots, u_n) + sf(u_1, \ldots, u_{k-1}, v', u_{k+1}, \ldots, u_n) \end{aligned}$$

for all $k = 1, \ldots, n$. A multilinear function of n variables is also referred to as an **n-linear function**. *The set of all n-linear functions as defined above will be denoted by* $\hom(V_1, \ldots, V_n; W)$. *A multilinear function from $V_1 \times \cdots \times V_n$ to the base field F is called a* **multilinear form** *or* **n-form**. □

Example 14.7

1) If A is an algebra then the product map $\mu\colon A \times \cdots \times A \to A$ defined by $\mu(a_1, \ldots, a_n) = a_1 \cdots a_n$ is n-linear.
2) The determinant function $\det\colon \mathcal{M}_n \to F$ is an n-linear form on the columns of the matrices in $\mathcal{M}_n$. □

We can extend the quotient space definition of the tensor product to n-linear functions as follows.

Let $\mathcal{B}_i = \{e_{i,j} \mid j \in J_i\}$ be a basis for V_i for $i = 1, \ldots, n$. For each ordered n-tuple $(e_{1,i_1}, \ldots, e_{n,i_n})$, we invent a new formal symbol $e_{1,i_1} \otimes \cdots \otimes e_{n,i_n}$ and define T to be the vector space with basis

$$\mathcal{D} = \{e_{1,i_1} \otimes \cdots \otimes e_{n,i_n} \mid e_{k,i_k} \in \mathcal{B}_k\}$$

Then define the map t by setting $t(e_{1,i_1}, \ldots, e_{n,i_n}) = e_{1,i_1} \otimes \cdots \otimes e_{n,i_n}$ and extending by multilinearity. This uniquely defines a multilinear map t that is as "universal" as possible among multilinear maps.

Indeed, if $g\colon V_1 \times \cdots \times V_n \to W$ is multilinear, the condition $g = \tau \circ t$ is equivalent to

$$\tau(e_{1,i_1} \otimes \cdots \otimes e_{n,i_n}) = g(e_{1,i_1}, \ldots, e_{n,i_n})$$

which uniquely defines a linear map $\tau: T \to W$. Hence, (T, t) has the universal property for bilinearity.

Alternatively, we may take the coordinate-free quotient space approach as follows.

Definition *Let $V_1, \ldots, V_n$ be vector spaces over F and let T be the subspace of the free vector space $\mathcal{F}$ on $V_1 \times \cdots \times V_n$, generated by all vectors of the form*

$$r(v_1, \ldots, v_{k-1}, u, v_{k+1}, \ldots, v_n) + s(v_1, \ldots, v_{k-1}, u', v_{k+1}, \ldots, v_n) \\ - (v_1, \ldots, v_{k-1}, ru + su', v_{k+1}, \ldots, v_n)$$

for all $r, s \in F$, $u, u' \in U$ and $v_1, \ldots, v_n \in V$. The quotient space $\mathcal{F}/T$ is called the **tensor product** *of $V_1, \ldots, V_n$ and denoted by $V_1 \otimes \cdots \otimes V_n$.* □

As before, we denote the coset $(v_1, \ldots, v_n) + T$ by $v_1 \otimes \cdots \otimes v_n$ and so any element of $V_1 \otimes \cdots \otimes V_n$ is a sum of decomposable tensors, that is,

$$\sum v_{i_1} \otimes \cdots \otimes v_{i_n}$$

where the vector space operations are linear in each variable.

Let us formally state the universal property for multilinear functions.

Theorem 14.13 *(***The universal property for multilinear functions as measured by linearity***) Let $V_1, \ldots, V_n$ be vector spaces over the field F. The pair $(V_1 \otimes \cdots \otimes V_n, t)$, where*

$$t: V_1 \times \cdots \times V_n \to V_1 \otimes \cdots \otimes V_n$$

is the multilinear map defined by

$$t(v_1, \ldots, v_n) = v_1 \otimes \cdots \otimes v_n$$

has the following property. If $f: V_1 \times \cdots \times V_n \to W$ is any multilinear function from $V_1 \times \cdots \times V_n$ to a vector space W over F then there is a unique *linear transformation $\tau: V_1 \otimes \cdots \otimes V_n \to W$ that makes the diagram in Figure 14.8 commute, that is, for which*

$$\tau \circ t = f$$

Moreover, $V_1 \otimes \cdots \otimes V_n$ is unique in the sense that if a pair (X, s) also has this property then X is isomorphic to $V_1 \otimes \cdots \otimes V_n$. □

$$\begin{tikzcd}
V_1 \times \cdots \times V_n \arrow[r, "t"] \arrow[dr, "f"'] & V_1 \otimes \cdots \otimes V_n \arrow[d, dashed, "\tau"] \\
 & W
\end{tikzcd}$$

Figure 14.8

Here are some of the basic properties of multiple tensor products. Proof is left to the reader.

Theorem 14.14 The tensor product has the following properties. Note that all vector spaces are over the same field F.

1) (**Associativity**) *There exists an isomorphism*

$$\tau: (V_1 \otimes \cdots \otimes V_n) \otimes (W_1 \otimes \cdots \otimes W_m) \to V_1 \otimes \cdots \otimes V_n \otimes W_1 \otimes \cdots \otimes W_m$$

for which

$$\tau[(v_1 \otimes \cdots \otimes v_n) \otimes (w_1 \otimes \cdots \otimes w_m)] = v_1 \otimes \cdots \otimes v_n \otimes w_1 \otimes \cdots \otimes w_m$$

In particular,

$$(U \otimes V) \otimes W \approx U \otimes (V \otimes W) \approx U \otimes V \otimes W$$

2) (**Commutativity**) *Let π be any permutation of the indices $\{1, \ldots, n\}$. Then there is an isomorphism*

$$\sigma: V_1 \otimes \cdots \otimes V_n \to V_{\pi(1)} \otimes \cdots \otimes V_{\pi(n)}$$

for which

$$\sigma(v_1 \otimes \cdots \otimes v_n) = v_{\pi(1)} \otimes \cdots \otimes v_{\pi(n)}$$

3) *There is an isomorphism $\rho_1: F \otimes V \to V$ for which*

$$\rho_1(r \otimes v) = rv$$

and similarly, there is an isomorphism $\rho_2: V \otimes F \to V$ for which

$$\rho_2(v \otimes r) = rv$$

Hence, $F \otimes V \approx V \approx V \otimes F$. □

The analog of Theorem 14.3 is the following.

Theorem 14.15 *Let $V_1, \ldots, V_n$ and W be vector spaces over F. Then the map $\phi: \hom(V_1, \ldots, V_n; W) \to \mathcal{L}(V_1 \otimes \cdots \otimes V_n, W)$, defined by the fact that $\phi(f)$ is the unique linear map for which $f = \phi(f) \circ t$, is an isomorphism. Thus,*

$$\hom(V_1,\ldots,V_n;W)\approx\mathcal{L}(V_1\otimes\cdots\otimes V_n,W)$$

Moreover, if all vector spaces are finite-dimensional then

$$\dim[\hom(V_1,\ldots,V_n;W)]=\dim(W)\cdot\prod_{i=1}^{n}\dim(V_i)\qquad\square$$

Theorem 14.9 and its corollary can also be extended.

Theorem 14.16 *The linear transformation*

$$\theta\colon\mathcal{L}(U_1,U_1')\otimes\cdots\otimes\mathcal{L}(U_n,U_n')\to\mathcal{L}(U_1\otimes\cdots\otimes U_n,U_1'\otimes\cdots\otimes U_n')$$

defined by

$$\theta(\tau_1\otimes\cdots\otimes\tau_n)(u_1\otimes\cdots\otimes u_n)=\tau_1(u_1)\otimes\cdots\otimes\tau_n(u_n)$$

is an embedding, and is an isomorphism if all vector spaces are finite-dimensional. Thus, the tensor product $\tau_1\otimes\cdots\otimes\tau_n$ of linear transformations is (via this embedding) a linear transformation on tensor products. Two important special cases of this are

$$U_1^*\otimes\cdots\otimes U_n^*\overset{\sim}{\hookrightarrow}(U_1\otimes\cdots\otimes U_n)^*$$

where

$$(f_1\otimes\cdots\otimes f_n)(u_1\otimes\cdots\otimes u_n)=f_1(u_1)\cdots f_n(u_n)$$

and

$$U_1^*\otimes\cdots\otimes U_n^*\otimes V\overset{\sim}{\hookrightarrow}\mathcal{L}(U_1\otimes\cdots\otimes U_n,V)$$

where

$$(f_1\otimes\cdots\otimes f_n\otimes v)(u_1\otimes\cdots\otimes u_n)=f_1(u_1)\cdots f_n(u_n)v\qquad\square$$

Tensor Spaces

Let V be a finite-dimensional vector space. For nonnegative integers p and q, the tensor product

$$T_q^p(V)=\underbrace{V\otimes\cdots\otimes V}_{p\text{ factors}}\otimes\underbrace{V^*\otimes\cdots\otimes V^*}_{q\text{ factors}}=V^{\otimes p}\otimes(V^*)^{\otimes q}$$

is called the space of **tensors of type (p,q)**, where p is the **contravariant type** and q is the **covariant type**. If $p=q=0$ then $T_q^p(V)=F$, the base field. Here we use the notation $V^{\otimes n}$ for the n-fold tensor product of V with itself. We will also write $V^{\times n}$ for the n-fold cartesian product of V with itself.

Since all vector spaces are finite-dimensional, V and V^{**} are isomorphic and so

$$T_q^p(V) = V^{\otimes p} \otimes (V^*)^{\otimes q} \approx ((V^*)^{\otimes p} \otimes V^{\otimes q})^* \approx \hom_F((V^*)^{\times p} \times V^{\times q}, F)$$

This is the space of all multilinear functionals on

$$\underbrace{V^* \times \cdots \times V^*}_{p \text{ factors}} \times \underbrace{V \times \cdots \times V}_{q \text{ factors}}$$

In fact, tensors of type (p, q) are often defined as multilinear functionals in this way.

Note that

$$\dim(T_q^p(V)) = [\dim(V)]^{p+q}$$

Also, the associativity and commutativity of tensor products allows us to write

$$T_q^p(V) \otimes T_s^r(V) = T_{q+s}^{p+r}(V)$$

at least up to isomorphism.

Tensors of type $(p, 0)$ are called **contravariant tensors**

$$T^p(V) = T_0^p(V) = \underbrace{V \otimes \cdots \otimes V}_{p \text{ factors}}$$

and tensors of type $(0, q)$ are called **covariant tensors**

$$T_q(V) = T_q^0(V) = \underbrace{V^* \otimes \cdots \otimes V^*}_{q \text{ factors}}$$

Tensors with both contravariant and covariant indices are called **mixed tensors**.

In general, a tensor can be interpreted in a variety of ways as a multilinear map on a cartesian product, or a linear map on a tensor product. (The interpretation we mentioned above that is sometimes used as the definition is only one possibility.) We simply need to decide how many of the contravariant factors and how many of the covariant factors should be "active participants" and how many should be "passive participants."

More specifically, consider a tensor of type (p, q), written

$$v_1 \otimes \cdots \otimes v_m \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_n \otimes \cdots \otimes f_q \in T_q^p(V)$$

where $m \leq p$ and $n \leq q$. Here we are choosing the first m vectors and the first n linear functionals as active participants. This determines the number of arguments of the map. In fact, we define a map from the cartesian product

$$\underbrace{V^* \times \cdots \times V^*}_{m \text{ factors}} \times \underbrace{V \times \cdots \times V}_{n \text{ factors}}$$

to the tensor product

$$\underbrace{V \otimes \cdots \otimes V}_{p-m \text{ factors}} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{q-n \text{ factors}}$$

of the remaining factors by

$$\begin{aligned}(v_1 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q)(h_1, \ldots, h_m, x_1, \ldots, x_n)\\ = h_1(v_1)\cdots h_m(v_m) f_1(x_1)\cdots f_n(x_n) v_{m+1} \otimes \cdots \otimes v_p \otimes f_{n+1} \otimes \cdots \otimes f_q\end{aligned}$$

In words, the first group $v_1 \otimes \cdots \otimes v_m$ of (active) vectors interacts with the first set $h_1, \ldots, h_m$ of arguments to produce the scalar $h_1(v_1)\cdots h_m(v_m)$. The first group $f_1 \otimes \cdots \otimes f_n$ of (active) functionals interacts with the second group $x_1, \ldots, x_n$ of arguments to produce the scalar $f_1(x_1)\cdots f_n(x_n)$. The remaining (passive) vectors $v_{m+1} \otimes \cdots \otimes v_p$ and functionals $f_{n+1} \otimes \cdots \otimes f_q$ are just "copied" to the image vector.

It is easy to see that this map is multilinear and so there is a unique linear map from the tensor product

$$\underbrace{V^* \otimes \cdots \otimes V^*}_{m \text{ factors}} \otimes \underbrace{V \otimes \cdots \otimes V}_{n \text{ factors}}$$

to the tensor product

$$\underbrace{V \otimes \cdots \otimes V}_{p-m \text{ factors}} \otimes \underbrace{V^* \otimes \cdots \otimes V^*}_{q-n \text{ factors}}$$

defined by

$$\begin{aligned}(v_1 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q)(h_1 \otimes \cdots \otimes h_m \otimes x_1 \otimes \cdots \otimes x_n)\\ = h_1(v_1)\cdots h_m(v_m) f_1(x_1)\cdots f_n(x_n) v_{m+1} \otimes \cdots \otimes v_p \otimes f_{n+1} \otimes \cdots \otimes f_q\end{aligned}$$

(What justifies the notation $v_1 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q$ for this map?)

Let us look at some special cases. For a contravariant tensor of type $(p, 0)$

$$v_1 \otimes \cdots \otimes v_p \in T_0^p(V)$$

we get a linear map

$$\underbrace{V^* \otimes \cdots \otimes V^*}_{m \text{ factors}} \to \underbrace{V \otimes \cdots \otimes V}_{p-m \text{ factors}}$$

(where $m \le p$) defined by

$$(v_1 \otimes \cdots \otimes v_p)(h_1 \otimes \cdots \otimes h_m) = h_1(v_1)\cdots h_m(v_m) v_{m+1} \otimes \cdots \otimes v_p$$

For a covariant tensor of type $(0,q)$

$$f_1 \otimes \cdots \otimes f_q \in T_q^0(V)$$

we get a linear map from

$$\underbrace{V \otimes \cdots \otimes V}_{n \text{ factors}} \to \underbrace{V^* \otimes \cdots \otimes V^*}_{q-n \text{ factors}}$$

(where $n \le q$) defined by

$$(f_1 \otimes \cdots \otimes f_q)(x_1 \otimes \cdots \otimes x_n) = f_1(x_1) \cdots f_n(x_n) f_{n+1} \otimes \cdots \otimes f_q$$

The special case $n = q$ gives a linear functional on $V^{\otimes q}$, that is, each element of $(V^*)^{\otimes q}$ is a distinct member of $(V^{\otimes q})^*$, whence the embedding

$$V_1^* \otimes \cdots \otimes V_q^* \overset{\sim}{\hookrightarrow} (V_1 \otimes \cdots \otimes V_q)^*$$

that we described earlier.

Let us consider some small values of p and q. For a mixed tensor $v \otimes f$ of type $(1,1)$ here are the possibilities. When $m = 0$ and $n = 1$ we get the linear map $(v \otimes f): V \to V$ defined by

$$(v \otimes f)(w) = f(w)v$$

When $m = 1$ and $n = 0$ we get the linear map $(v \otimes f): V^* \to V^*$ defined by

$$(v \otimes f)(h) = h(v)f$$

Finally, when $m = n = 1$ we get a multilinear form $(v \otimes f): V^* \times V \to F$ defined by

$$(v \otimes f)(h, w) = h(v)f(w)$$

Consider also a tensor $f \otimes g$ of type $(0,2)$. When $n = q = 2$ we get a multilinear functional $f \otimes g: (V \times V) \to F$ defined by

$$(f \otimes g)(v, w) = f(v)g(w)$$

This is just a bilinear form on V. When $n = 1$ we get a multilinear map $(f \otimes g): V \to V^*$ defined by

$$(f \otimes g)(v) = f(v)g$$

Contraction

Covariant and contravariant factors can be "combined" in the following way. Consider the map

$$h: V^{\times p} \times (V^*)^{\times q} \to T_{q-1}^{p-1}(V)$$

defined by

$$h(v_1, \ldots, v_p, f_1, \ldots, f_q) = f_1(v_1)(v_2 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q)$$

This is easily seen to be multilinear and so there is a unique linear map

$$\theta: T_q^p(V) \to T_{q-1}^{p-1}(V)$$

defined by

$$\theta(v_1 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q) = f_1(v_1)(v_2 \otimes \cdots \otimes v_p \otimes f_1 \otimes \cdots \otimes f_q)$$

This is called the **contraction** in the contravariant index 1 and covariant index 1. Of course, contraction in other indices (one contravariant and one covariant) can be defined similarly.

Example 14.8 Consider the tensor space $T_1^1(V)$, which is isomorphic to $\mathcal{L}(V)$ via the fact that

$$(v \otimes f)(w) = f(w)v$$

For $p = q = 1$, the contraction takes the form

$$\theta(v \otimes f) = f(v)$$

Now, for $v \neq 0$, the operator $v \otimes f$ has kernel equal to $\ker(f)$, which has codimension 1 and so there is a nonzero vector $u \in V$ for which $V = \langle u \rangle \oplus \ker(f)$.

Now, if $f(v) \neq 0$ then $V = \langle v \rangle \oplus \ker(f)$ and

$$(v \otimes f)(v) = f(v)v$$

and so v is an eigenvector for the nonzero eigenvalue $f(v)$. Hence, $V = \mathcal{E}_{f(v)} \oplus \mathcal{E}_0$ and so the trace of $v \otimes f$ is precisely $f(v)$. Since the trace is linear, we deduce that the trace of any linear operator on V is the contraction of the corresponding vector in $T_1^1(V)$. □

The Tensor Algebra of V

Consider the contravariant tensor spaces

$$T^p(V) = T_0^p(V) = V^{\otimes p}$$

For $p = 0$ we take $T^0(V) = F$. The external direct sum

$$T(V) = \bigoplus_{p=0}^{\infty} T^p(V)$$

of these tensor spaces is a vector space with the property that

$$T^p(V) \otimes T^q(V) = T^{p+q}(V)$$

This is an example of a *graded algebra*, where $T^p(V)$ are the elements of *grade* p. The graded algebra $T(V)$ is called the **tensor algebra** over V. (We will formally define graded structures a bit later in the chapter.)

Since

$$T_q(V) = \underbrace{V^* \otimes \cdots \otimes V^*}_{q \text{ factors}} = T^q(V^*)$$

there is no need to look separately at $T_q(V)$.

Special Multilinear Maps

The following definitions describe some special types of multilinear maps.

Definition

1) *A multilinear map* $f\colon V^n \to W$ *is* **symmetric** *if interchanging any two coordinate positions changes nothing, that is, if*

$$f(v_1,\ldots,v_i,\ldots,v_j,\ldots,v_n) = f(v_1,\ldots,v_j,\ldots,v_i,\ldots,v_n)$$

for any $i \neq j$.

2) *A multilinear map* $f\colon V^n \to W$ *is* **antisymmetric** *or* **skew-symmetric** *if interchanging any two coordinate positions introduces a factor of* -1*, that is, if*

$$f(v_1,\ldots,v_i,\ldots,v_j,\ldots,v_n) = -f(v_1,\ldots,v_j,\ldots,v_i,\ldots,v_n)$$

for $i \neq j$.

3) *A multilinear map* $f\colon V^n \to W$ *is* **alternate** *or* **alternating** *if*

$$f(v_1,\ldots,v_n) = 0$$

whenever any two of the vectors v_i are equal. □

As in the case of bilinear forms, we have some relationships between these concepts. In particular, if $\operatorname{char}(F) = 2$ then

alternate $\Rightarrow$ symmetric $\Leftrightarrow$ skew-symmetric

and if $\operatorname{char}(F) \neq 2$ then

alternate $\Leftrightarrow$ skew-symmetric

A few remarks about permutations, with which the reader may very well be familiar, are in order. A **permutation** of the set $N = \{1,\ldots,n\}$ is a bijective function $\pi\colon N \to N$. We denote the group (under composition) of all such permutations by S_n. This is the **symmetric group** on n symbols. A **cycle** of length k is a permutation of the form $(i_1, i_2, \ldots, i_k)$, which sends i_1 to i_2, i_2 to $i_3, \ldots, i_{k-1}$ to i_k and i_k to i_1. All other elements of N are left fixed. Every permutation is the product (composition) of disjoint cycles.

A **transposition** is a cycle (i, j) of length 2. Every cycle (and therefore every permutation) is the product of transpositions. In general, a permutation can be expressed as a product of transpositions in many ways. However, no matter how one represents a given permutation as such a product, the number of transpositions is either always even or always odd. Therefore, we can define the **parity** of a permutation $\pi \in S_n$ to be the parity of the number of transpositions in any decomposition of π as a product of transpositions. The **sign** of a permutation is defined by

$$\mathrm{sg}(\pi) = (-1)^{\mathrm{parity}(\pi)}$$

If $\mathrm{sg}(\pi) = 1$ then π is an **even permutation** and if $\mathrm{sg}(\pi) = -1$ then π is an **odd permutation**. The sign of π is often written $(-1)^\pi$.

With these facts in mind, it is apparent that f is symmetric if and only if

$$f(v_1, \ldots, v_n) = f(v_{\pi(1)}, \ldots, v_{\pi(n)})$$

for all permutations $\pi \in S_n$ and that f is alternating if and only if

$$f(v_1, \ldots, v_n) = (-1)^\pi f(v_{\pi(1)}, \ldots, v_{\pi(n)})$$

for all permutations $\pi \in S_n$.

Graded Algebras

We need to pause for a few definitions that are useful when discussing tensor algebra. An algebra A over F is said to be a **graded algebra** if as a vector space over F, A can be written in the form

$$A = \bigoplus_{i=0}^{\infty} A_i$$

for subspaces A_i of A, and where multiplication behaves nicely, that is,

$$A_i A_j \subseteq A_{i+j}$$

The elements of A_i are said to be **homogeneous of degree** i. If $a \in A$ is written

$$a = a_{i_1} + \cdots + a_{i_n}$$

for $a_{i_k} \in A_{i_k}$, $i_k \neq i_j$, then a_{i_k} is called the **homogeneous component** of a of degree i.

The ring of polynomials $F[x]$ provides a prime example of a graded algebra, since

$$F[x] = \bigoplus_{i=0}^{\infty} F_i[x]$$

where $F_i[x]$ is the subspace of $F[x]$ consisting of all scalar multiples of x^i.

More generally, the ring $F[x_1, \ldots, x_n]$ of polynomials in several variables is a graded algebra, since it is the direct sum of the subspaces of homogeneous polynomials of degree i. (A polynomial is **homogeneous of degree** i if each term has degree i. For example, $p = x_1 x_2^2 + x_1 x_2 x_3$ is homogeneous of degree 3.)

Graded Ideals

A **graded ideal** I in a graded algebra $A = \bigotimes A_i$ is an ideal I for which, as a subspace of A,

$$I = \bigoplus_{i=0}^{\infty} (I \cap A_i)$$

(Note that $I \cap A_i$ is not, in general, an ideal.) For example, the ideal I of $F[x]$ consisting of all polynomials with zero constant term is graded. However, the ideal

$$J = \langle 1 + x \rangle = \{p(x)(1 + x) \mid p(x) \in F[x]\}$$

generated by $1 + x$ is not graded, since $F_i[x]$ contains only monomials and so $J \cap F_i[x] = \{0\}$.

Theorem 14.17 *Let A be a graded algebra. An ideal I of A is graded if and only if it is generated by homogeneous elements of A.*
Proof. *If I is graded then it is generated by the elements of the direct summands $I \cap A_i$, which are homogeneous. Conversely, suppose that $I = \langle a_k \mid k \in K \rangle$ where each a_k is homogeneous. Any $u \in I$ has the form*

$$u = \sum_i u_i a_i v_i$$

where $u_i, v_i \in A$. Since A is graded, each u_i and v_i is a sum of homogeneous terms and we can expand $u_i a_i v_i$ into a sum of homogeneous terms of the form $e_i a_i f_i$ where e_i and f_i (and a_i) are homogeneous. Hence, if

$$\deg(e_i a_i f_i) = \deg(e_i) \cdot \deg(a_i) \cdot \deg(f_i) = k$$

then $e_i a_i f_i \in A_k \cap I$ and so

$$I = \bigoplus_{i=0}^{\infty} (I \cap A_i)$$

is graded. □

If I is a graded ideal in A, then the quotient ring A/I is also graded, since it is easy to show that

$$\frac{A}{I} = \bigoplus_{i=0}^{\infty} \frac{A_i + I}{I}$$

Moreover, for $x, y \in I$ and $a_i \in A_i$,

$$[(a_j + x) + I][(a_k + y) + I] = (a_j + I)(a_k + I) = a_j a_k + I \in \frac{A_{jk} + I}{I}$$

The Symmetric Tensor Algebra

We wish to study tensors in $T^p(V)$ for $p \geq 1$ that enjoy a symmetry property. Let S_p be the symmetric group on $\{1, \ldots, p\}$. For each $\sigma \in S_p$, the multilinear map $f_\sigma \colon V^{\times p} \to T^p(V)$ defined by

$$f_\sigma(v_1, \ldots, v_p) = v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(p)}$$

determines (by universality) a unique linear operator λ_σ on $T^p(V)$ for which

$$\lambda_\sigma(v_1 \otimes \cdots \otimes v_p) = v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(p)}$$

Let $\mathcal{B} = \{e_1, \ldots, e_n\}$ be a basis for V. Since the set

$$\mathcal{B} = \{e_{i_1} \otimes \cdots \otimes e_{i_p} \mid e_{i_j} \in \mathcal{B}\}$$

is a basis for $T^p(V)$ and λ_σ is a bijection of $\mathcal{B}$, it follows that λ_σ is an isomorphism of $T^p(V)$. A tensor $t \in T^p(V)$ is **symmetric** if $\lambda_\sigma(t) = t$ for all permutations $\sigma \in S_p$.

A word of caution is in order with respect to the definition of λ_σ. The permutation λ_σ permutes the *coordinate positions* in a decomposable tensor, not the indices. Suppose, for example, that $p = 2$ and $\sigma = (12)$. If $\{e_1, e_2\}$ is a basis for V then

$$\lambda_{(12)}(e_1 \otimes e_1) = e_1 \otimes e_1$$

because λ_σ permutes the positions, not the indices. Thus, the following is not true

$$\lambda_{(12)}(e_1 \otimes e_2) \overset{\text{no}}{=} e_{\sigma(1)} \otimes e_{\sigma(1)} = e_2 \otimes e_2$$

The set of all symmetric tensors

$$ST^p(V) = \{t \in T^p(V) \mid \lambda_\sigma(t) = t \text{ for all } \sigma \in S_p\}$$

is a subspace of $T^p(V)$.

To study $ST^p(V)$ in more detail, let $e_1, \ldots, e_n$ be a basis for $T^p(V)$. Any tensor $v \in T^p(V)$ has the form

$$v = \sum_{i_1, \ldots, i_p = 1}^{n} \alpha_{i_1, \ldots, i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}$$

where $\alpha_{i_1, \ldots, i_p} \neq 0$. It will help if we group the terms in such a sum according to the multiset of indices. Specifically, for each nonempty subset S of indices $\{1, \ldots, n\}$ and each multiset $M = \{i_1, \ldots, i_p\}$ of size p with underlying set S, let G_M consist of all possible decomposable tensors

$$e_{k_1} \otimes \cdots \otimes e_{k_p}$$

where $(k_1, \ldots, k_p)$ is a permutation of $\{i_1, \ldots, i_p\}$. For example, if $M = \{2, 2, 3\}$ then

$$G_M = \{e_2 \otimes e_2 \otimes e_3, e_2 \otimes e_3 \otimes e_2, e_3 \otimes e_2 \otimes e_2\}$$

Now, ignoring coefficients, the terms in the expression for v can be organized into the groups G_M. Let us denote the set of terms of v (without coefficients) that lie in G_M by $G_M(v)$. For example, if

$$v = 2e_2 \otimes e_2 \otimes e_3 + 3e_2 \otimes e_3 \otimes e_2 + e_3 \otimes e_3 \otimes e_1$$

then

$$G_{\{2,2,3\}}(v) = \{e_2 \otimes e_2 \otimes e_3, e_2 \otimes e_3 \otimes e_2\}$$

and

$$G_{\{1,3,3\}}(v) = \{e_3 \otimes e_3 \otimes e_1\}$$

Further, let $S_M(v)$ denote the sum of the terms in v that belong to $G_M(v)$ (including the coefficients). For example,

$$S_{\{2,2,3\}}(v) = 2e_2 \otimes e_2 \otimes e_3 + 3e_2 \otimes e_3 \otimes e_2$$

Thus,

$$v = \sum_{\text{multisets } M} S_M(v)$$

Note that each permutation λ_σ is a permutation of the elements of G_M, for each multiset M. It follows that v is a symmetric tensor if and only if the following conditions are satisfied:

1) (**All or none**) For each multiset M of size p with underlying set $S \subseteq \{1, \ldots, n\}$, we have

$$G_M(v) = \emptyset \quad \text{or} \quad G_M(v) = G_M$$

2) If $G_M(v) = G_M$, then the coefficients in the sum $S_M(v)$ are the same and so

$$S_M(v) = \alpha_M(v) \cdot \sum_{t \in G_M} t$$

where $\alpha_M(v)$ is the common coefficient.

Hence, for a symmetric tensor v, we have

$$v = \sum_{\text{multisets } M} \left(\alpha_M(v) \sum_{t \in G_M} t \right)$$

Now, symmetric tensors act as though the tensor product was commutative. Of course, it is not, but we can deal with this as follows.

Let $\tau: T^p(V) \to F_p[e_1, \ldots, e_n]$ be the function from $T^p(V)$ to the vector space $F_p[e_1, \ldots, e_n]$ of all homogeneous polynomials of degree p in the formal variables $e_1, \ldots, e_n$, defined by

$$\tau\left(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\right) = \sum \alpha_{i_1,\ldots,i_p} e_{i_1} \cdots e_{i_p}$$

In this context, the product in $F_p[e_1, \ldots, e_n]$ is often denoted by the symbol $\vee$, so we have

$$\tau\left(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\right) = \sum \alpha_{i_1,\ldots,i_p} (e_{i_1} \vee \cdots \vee e_{i_p})$$

It is clear that τ is well-defined, linear and surjective. We want to use τ to explore the properties of symmetric tensors, first as a subspace of $T^p(V)$ and then as a quotient space. (The subspace approach is a bit simpler, but works only when $\operatorname{char}(F) = 0$.)

The Case $\operatorname{char}(F) = 0$

Note that τ takes every member of a group G_M to the same monomial, whose indices are precisely M. Hence, if $v \in ST^p(V)$ is symmetric then

$$\begin{aligned} \tau(v) &= \sum_{\text{multisets } M} \left(\alpha_M(v) \sum_{t \in G_M} \tau(t) \right) \\ &= \sum_{i_1 \le \cdots \le i_p} (\alpha_M(v)|G_M|)(e_{i_1} \vee \cdots \vee e_{i_p}) \end{aligned}$$

(Here we identify each multiset M with the nondecreasing sequence $i_1 \leq \cdots \leq i_p$ of its members.)

As to the kernel of τ, if $\tau(v) = 0$ for $v \in ST^p(V)$ then $\alpha_M(v)|G_M| = 0$ for all multisets M and so, if char$(F) = 0$, we may conclude that $\alpha_M(v) = 0$ for all multisets M, that is, $v = 0$. Hence, if char$(F) = 0$, the restricted map $\tau|_{ST^p(V)}$ is injective and so it is an isomorphism. We have proved the following.

Theorem 14.18 *Let V be a finite-dimensional vector space over a field F with* char$(F) = 0$*. Then the vector space $ST^p(V)$ of symmetric tensors of degree p is isomorphic to the vector space of homogeneous polynomials $F_p[e_1, \ldots, e_n]$, via the isomorphism*

$$\tau\Big(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\Big) = \sum \alpha_{i_1,\ldots,i_p} (e_{i_1} \vee \cdots \vee e_{i_p}) \qquad \square$$

The vector space $ST^p(V)$ of symmetric tensors of degree p is often called the **symmetric tensor space** of degree p for V. However, this term is also used for an isomorphic vector space that we will study next.

The direct sum

$$ST(V) = \bigoplus_{p=0}^{\infty} ST^p(V)$$

is sometimes called the **symmetric tensor algebra** of V, although this term is also used for a slightly different (but isomorphic) algebra that we will define momentarily.

We can use the vector space isomorphisms described in the previous theorem to move the product from the algebra of polynomials $F[e_1, \ldots, e_n]$ to the symmetric tensor space $ST(V)$. In other words, if char$(F) = 0$ then $ST(V)$ is a graded algebra isomorphic to the algebra of polynomials $F[e_1, \ldots, e_n]$.

The Arbitrary Case

We can define the symmetric tensor space in a different, although perhaps slightly more complex, manner that holds regardless of the characteristic of the base field. This is important, since many important fields (such as finite fields) have nonzero characteristic.

Consider again the kernel of the map τ, but this time as defined on all of $T^p(V)$, not just $ST^p(V)$. The map τ sends elements of different groups $G_M(v)$ to different monomials in $F_p[e_1, \ldots, e_n]$, and so $v \in$ ker(τ) if and only if

$$\tau(S_M(v)) = 0$$

Hence, the *sum* of the coefficients of the elements in $G_M(v)$ must be 0.

Conversely, if the *sum* of the coefficients of the elements in $G_M(v)$ is 0 for all multisets M, then $v \in \ker(\tau)$.

Suppose that $M = \{i_1, \ldots, i_p\}$ is a multiset for which

$$t = e_{i_1} \otimes \cdots \otimes e_{i_p} \in G_M(v)$$

Then each decomposable tensor in $G_M(v)$ is a permutation of t and so $S_M(v)$ may be written in the form

$$S_M(v) = \beta e_{i_1} \otimes \cdots \otimes e_{i_p} + \sum_i \alpha_i \lambda_{\sigma_i}(e_{i_1} \otimes \cdots \otimes e_{i_p})$$

where the sum is over a subset of the symmetric group S_p, corresponding to the terms that appear in $S_M(v)$ and where

$$\beta + \sum_i \alpha_i = 0$$

Substituting for β in the expression for $S_M(v)$ gives

$$S_M(v) = \sum_i \alpha_i [\lambda_{\sigma_i}(e_{i_1} \otimes \cdots \otimes e_{i_p}) - (e_{i_1} \otimes \cdots \otimes e_{i_p})]$$

It follows that v is in the subspace I_p of $T^p(V)$ generated by tensors of the form $\lambda_\sigma(t) - t$, that is

$$I_p = \langle \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p \rangle$$

and so $\ker(\tau) \subseteq I_p$. Conversely,

$$\tau(\lambda_\sigma(e_{k_1} \otimes \cdots \otimes e_{k_p}) - (e_{k_1} \otimes \cdots \otimes e_{k_p})) = 0$$

and so $I_p \subseteq \ker(\tau)$.

Theorem 14.19 *Let V be a finite-dimensional vector space over a field F. For $p \geq 1$, the surjective linear map $\tau\colon T^p(V) \to F_p[e_1, \ldots, e_n]$ defined by*

$$\tau\Big(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\Big) = \sum \alpha_{i_1,\ldots,i_p} e_{i_1} \vee \cdots \vee e_{i_p}$$

has kernel

$$I_p = \langle \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p \rangle$$

and so

$$\frac{T^p(V)}{I_p} \approx F_p[e_1, \ldots, e_n]$$

The vector space $T^p(V)/I$ is also referred to as the **symmetric tensor space** *of degree p of V. The ideal of $T(V)$ defined by*

$$I = \langle \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p, p \geq 1 \rangle$$

being generated by homogeneous elements, is graded, so that

$$I = \bigoplus_{p=0}^{\infty} I_p$$

where $I_0 = \{0\}$. *The graded algebra*

$$\frac{T(V)}{I} = \bigoplus_{i=0}^{\infty} \frac{T^p(V) + I}{I}$$

is also called the **symmetric tensor algebra** *for* V *and is isomorphic to* $F[e_1, \ldots, e_n]$. □

Before proceeding to the universal property, we note that the dimension of the symmetric tensor space $ST^p(V)$ is equal to the number of monomials of degree p in the variables $e_1, \ldots, e_n$ and this is

$$\dim(ST^p(V_n)) = \binom{n+p-1}{p}$$

The Universal Property for Symmetric p-Linear Maps

The vector space $F_p[x_1, \ldots, x_n]$ of homogeneous polynomials, and therefore also the isomorphic spaces of symmetric tensors $ST^p(V)$ and $T^p(V)/I_p$, have the universal property for *symmetric* p-linear maps.

Theorem 14.20 *(***The universal property for symmetric multilinear maps, as measured by linearity***) Let* V *be a finite-dimensional vector space. Then the pair* $(F_p[x_1, \ldots, x_n], t: V^{\times p} \to F_p[x_1, \ldots, x_n])$*, where*

$$t(v_1, \ldots, v_p) = v_1 \vee \cdots \vee v_p$$

has the universal property for symmetric p*-linear maps with domain* $V^{\times p}$*, as measured by linearity. That is, for any symmetric* p*-linear map* $f: V^{\times p} \to U$ *where* U *is a vector space, there is a unique linear map* $\tau: F_p[x_1, \ldots, x_n] \to U$ *for which*

$$\tau(v_1 \vee \cdots \vee v_p) = f(v_1, \ldots, v_p)$$

for any vectors $v_i \in V$.

Proof. The universal property requires that

$$\tau(e_{i_1} \vee \cdots \vee e_{i_p}) = f(e_{i_1}, \ldots, e_{i_p})$$

and this does indeed uniquely define a linear transformation τ, provided that it is well-defined. However,

$$e_{i_1} \vee \cdots \vee e_{i_p} = e_{j_1} \vee \cdots \vee e_{j_p}$$

if and only if the multisets $\{e_{i_1}, \ldots, e_{i_p}\}$ and $\{e_{j_1}, \ldots, e_{j_p}\}$ are the same, which implies that $f(e_{i_1}, \ldots, e_{i_p}) = f(e_{j_1}, \ldots, e_{j_p})$, since f is symmetric. □

The Symmetrization Map

When $\operatorname{char}(F) = 0$, we can define a linear map $S\colon T^p(V) \to ST^p(V)$, called the **symmetrization map**, by

$$S(t) = \frac{1}{p!}\sum_{\sigma \in S_p} \lambda_\sigma(t)$$

(Since $\operatorname{char}(F) = 0$ we have $p! \neq 0$.)

Since $\lambda_\tau \lambda_\sigma = \lambda_{\tau\sigma}$, we have

$$\lambda_\tau(S(t)) = \frac{1}{p!}\sum_{\sigma \in S_p} \lambda_\tau\lambda_\sigma(t) = \frac{1}{p!}\sum_{\sigma \in S_p} \lambda_{\tau\sigma}(t) = \frac{1}{p!}\sum_{\sigma \in S_p} \lambda_\sigma(t) = S(t)$$

and so $S(t)$ is, in fact, symmetric. The reason for the factor $1/p!$ is that if v is a symmetric tensor, then $\lambda_\sigma(v) = v$ and so

$$S(v) = \frac{1}{p!}\sum_{\sigma \in S_p} \lambda_\sigma(v) = \frac{1}{p!}\sum_{\sigma \in S_p} v = v$$

that is, the symmetrization map fixes all symmetric tensors.

It follows that for any tensor $t \in T^p(V)$

$$S^2(t) = S(S(t)) = S(t)$$

Thus, S is idempotent and is therefore the projection map of $T^p(V)$ onto $\operatorname{im}(S) = ST^p(V)$

The Antisymmetric Tensor Algebra: The Exterior Product Space

Let us repeat our discussion for antisymmetric tensors. Before beginning officially, we want to introduce a very useful and very simple concept.

Definition *Let $E = \{e_i \mid i \in I\}$ be a set, which we refer to as an* **alphabet**. *A* **word**, *or* **string** *over E of finite length $p > 0$ is a sequence $w = x_1 \cdots x_p$ where $x_i \in E$. There is one word of length 0 over E, denoted by ϵ and called the* **empty word**. *Let $\mathcal{W}_p(E)$ be the set of all words over E of length p and let $\mathcal{W}$(E) be the set of all words over E (of finite length).*

Concatenation of words is done by placing one word after another: If $v = y_1 \cdots y_q$ then $wv = x_1 \cdots x_p y_1 \cdots y_q$. Also, $\epsilon w = w = w\epsilon$.

If the alphabet E is an ordered set, we say that a word w over E is in **ascending order** *if each $e_i \in E$ appears at most once in w and if the order of the letters in w is that given by the order of E. (For example, e_3e_4 is in ascending order but e_4e_3 and e_3e_3 are not.) The empty word is in ascending order by definition. Let $\mathcal{A}_p(E)$ be the set of all words in ascending order over E of length p and let $\mathcal{A}(E)$ be the set of all words in ascending order over E.* □

We will assume throughout that $\operatorname{char}(F) \neq 2$. For each $\sigma \in S_p$, the multilinear map $f_\sigma: V^{\times p} \to T^p(V)$ defined by

$$f_\sigma(v_1, \ldots, v_p) = (-1)^\sigma v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(p)}$$

where $(-1)^\sigma$ is the sign of σ, determines (by universality) a unique linear operator λ_σ on $T^p(V)$ for which

$$\lambda_\sigma(v_1 \otimes \cdots \otimes v_p) = (-1)^\sigma v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(p)}$$

Note that if $v_i = v_j$ for some $i \neq j$, then for $\sigma = (i, j)$, we have

$$\begin{aligned} \lambda_{(ij)}&(v_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes v_j \otimes \cdots \otimes v_p) \\ &= -v_1 \otimes \cdots \otimes v_j \otimes \cdots \otimes v_i \otimes \cdots \otimes v_p \\ &= -v_1 \otimes \cdots \otimes v_i \otimes \cdots \otimes v_j \otimes \cdots \otimes v_p \end{aligned}$$

and since $\operatorname{char}(F) \neq 2$, we conclude that $\lambda_{(ij)}(v_1 \otimes \cdots \otimes v_p) = 0$.

Let $\mathcal{B} = \{e_1, \ldots, e_n\}$ be a basis for V. Since the set

$$\mathcal{B} = \{e_{i_1} \otimes \cdots \otimes e_{i_p} \mid e_{i_j} \in \mathcal{B}\}$$

is a basis for $T^p(V)$ and λ_σ is a bijection of $\mathcal{B}$, it follows that λ_σ is an isomorphism of $T^p(V)$. A tensor $t \in T^p(V)$ is **antisymmetric** if $\lambda_\sigma(t) = (-1)^\sigma t$ for all permutations $\sigma \in S_p$.

The set of all antisymmetric tensors

$$AS^p(V) = \{t \in T^p(V) \mid \lambda_\sigma(t) = (-1)^\sigma t \text{ for all } \sigma \in S_p\}$$

is a subspace of $T^p(V)$.

Let $e_1, \ldots, e_n$ be a basis for $T^p(V)$. Any tensor $v \in T^p(V)$ has the form

$$v = \sum_{i_1, \ldots, i_p = 1}^{n} \alpha_{i_1, \ldots, i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}$$

where $\alpha_{i_1, \ldots, i_p} \neq 0$. As before, we define the groups $G_M(v)$ and the sums

$S_M(v)$. Each permutation λ_σ sends an element $t \in G_M$ to another element of G_m, multiplied by $(-1)^\sigma$. It follows that v is an antisymmetric tensor if and only if the following hold:

1) If M is a multiset of size p with underlying set $S \subseteq \{1,\ldots,n\}$ and if at least one element of M has multiplicity greater than 1, that is, if M is not a set, then $G_M(v) = \emptyset$.
2) For each *subset* M of size p of $\{1,\ldots,n\}$, we have
$$G_M(v) = \emptyset \quad \text{or} \quad G_M(v) = G_M$$
3) If $G_M(v) = G_M$, then since M is a set, there is a unique member $u = e_{i_1} \otimes \cdots \otimes e_{i_p}$ of G_M for which $i_1 < \cdots < i_p$. If $\sigma_{u,t}$ denotes the permutation in S_p for which $\lambda_{\sigma_{u,t}}$ takes u to t, then
$$S_M(v) = \alpha_M(v) \cdot \sum_{t \in G_M} \mathrm{sg}(\sigma_{u,t})t$$
where $\alpha_M(v)$ is the absolute value of the coefficient of u.

Hence, for an antisymmetric tensor v, we have
$$v = \sum_M \left(\alpha_M(v) \sum_{t \in G_M} \mathrm{sg}(\sigma_{u,t})t \right)$$

Next, we need the counterpart of the polynomials $F_p[e_1,\ldots,e_n]$ in which multiplication acts anticommutatively, that is, $e_ie_j = -e_je_i$. To this end, we define a map $\phi\colon \mathcal{W}_p(E) \to \mathcal{A}_p(E) \cup \{\epsilon\}$ as follows.

For $t = e_{i_1}\cdots e_{i_p}$, let $\phi(t) = \epsilon$ if t has any repeated variables. Otherwise, there is exactly one permutation of positions that will reorder t in ascending order. If the resulting word in ascending order is u, denote this permutation by $\sigma_{t,u}$. Set
$$\phi(t) = \mathrm{sg}(\sigma_{t,u})\sigma_{t,u}(t) = \mathrm{sg}(\sigma_{t,u})u$$

We can now define our anticommutative "polynomials."

Definition *Let $E = (e_1,\ldots,e_n)$ be a sequence of independent variables. Let $F_p^-[e_1,\ldots,e_n]$ be the vector space over F generated by all words over E in ascending order. For $p = 0$, this is $F\epsilon$, which we identify with F. Define a multiplication on the direct sum*
$$F^- = F^-[e_1,\ldots,e_n] = \bigoplus_{p=0}^{\infty} F_p^-$$
as follows. For monomials $f = x_1\cdots x_p \in F_p^-$ and $g = y_1\cdots y_q \in F_q^-$ set

$$fg = \phi(x_1 \cdots x_p y_1 \cdots y_q)$$

and extend by distributivity to F^-. The resulting multiplication makes $F^-[e_1, \ldots, e_n]$ into a (noncommutative) algebra over F. □

It is customary to use the notation $\wedge$ for the product in $F^-[e_1, \ldots, e_n]$. This product is called the **wedge product** or **exterior product**. We will not name the algebra F^-, but we will name an isomorphic algebra to be defined soon.

Now we can define a function $\tau: T^p(V) \to F_p^-[e_1, \ldots, e_n]$ by

$$\tau\left(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\right) = \sum \alpha_{i_1,\ldots,i_p} \phi(e_{i_1} \wedge \cdots \wedge e_{i_p})$$

It is clear that τ is well-defined, linear and surjective.

The Case char$(F) = 0$

Just as with the symmetric tensors, if char$(F) \neq 0$, we can benefit by restricting τ to the space $AT^p(V)$. Let

$$s = e_{j_1} \otimes \cdots \otimes e_{j_p} \quad \text{and} \quad t = e_{k_1} \otimes \cdots \otimes e_{k_p}$$

belong to the same group G_M and suppose that $u = e_{i_1} \otimes \cdots \otimes e_{i_p} \in G_M$ is in ascending order, that is, $i_1 < \cdots < i_p$.

If $v \in AT^p(V)$ is antisymmetric, then

$$\begin{aligned}
\tau(v) &= \sum_M \left(\alpha_M(v) \sum_{t \in G_M} \text{sg}(\sigma_{u,t}) \tau(t) \right) \\
&= \sum_M \left(\alpha_M(v) \sum_{t \in G_M} \text{sg}(\sigma_{u,t}) \text{sg}(\sigma_{t,u}) u \right) \\
&= \sum_M \left(\alpha_M(v) \sum_{t \in G_M} u \right) \\
&= \sum_M \alpha_M(v) |G_M| u
\end{aligned}$$

Now, if $\tau(v) = 0$ for $v \in AT^p(V)$ then $\alpha_M(v)|G_M| = 0$ for all sets M and so, if char$(F) = 0$, we may conclude that $\alpha_M(v) = 0$ for all sets M, that is, $v = 0$. Hence, if char$(F) = 0$, the restricted map $\tau|_{AT^p(V)}$ is injective and so it is an isomorphism. We have proved the following.

Theorem 14.21 *Let V be a finite-dimensional vector space over a field F with* char$(F) = 0$. *Then the vector space $AT^p(V)$ of antisymmetric tensors of degree*

p is isomorphic to the vector space $F_p^-[e_1, \ldots, e_n]$, via the isomorphism

$$\tau\Big(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\Big) = \sum \alpha_{i_1,\ldots,i_p} e_{i_1} \wedge \cdots \wedge e_{i_p} \qquad \square$$

The vector space $AT^p(V)$ of antisymmetric tensors of degree p is called the **antisymmetric tensor space** of degree p for V or the **exterior product space** of degree p over V.

The direct sum

$$AT(V) = \bigoplus_{p=0}^{\infty} AT^p(V)$$

is called the **antisymmetric tensor algebra** of V or the **exterior algebra** of V.

We can use the vector space isomorphisms described in the previous theorem to move the product from the algebra $F^-[e_1, \ldots, e_n]$ to the antisymmetric tensor space $AT(V)$. In other words, if char$(F) = 0$ then $AT(V)$ is a graded algebra isomorphic to the algebra $F^-[e_1, \ldots, e_n]$.

The Arbitrary Case

We can define the antisymmetric tensor space in a different manner that holds regardless of the characteristic of the base field.

Consider the kernel of the map τ, as defined on all of $T^p(V)$. Suppose that $v \in$ ker(τ). Since τ sends elements of different groups $G_M(v)$ to different monomials in $F_p[e_1, \ldots, e_n]$, it follows that τ must send each sum $S_M(v)$ to 0

$$\tau(S_M(v)) = 0$$

Hence, the *sum* of the coefficients of the elements in $G_M(v)$ must be 0. Conversely, if the *sum* of the coefficients of the elements in $G_M(v)$ is 0 for all multisets M, then $v \in$ ker(τ).

Suppose that $M = \{i_1, \ldots, i_p\}$ is a set for which

$$u = e_{i_1} \otimes \cdots \otimes e_{i_p} \in G_M(v)$$

Then

$$S_M(v) = \beta e_{i_1} \otimes \cdots \otimes e_{i_p} + \sum_{\sigma_i} \alpha_i (-1)^{\sigma_i} \lambda_{\sigma_i}(e_{i_1} \otimes \cdots \otimes e_{i_p})$$

where the sum is over a subset of the symmetric group S_p, corresponding to the terms that appear in $S_M(v)$ and where

$$\beta + \sum_{\sigma_i} \alpha_i (-1)^{\sigma_i} = 0$$

Substituting for β in the expression for $S_M(v)$ gives

$$S_M(v) = \sum_{\sigma_i} \alpha_i [(-1)^{\sigma_i} \lambda_{\sigma_i}(e_{i_1} \otimes \cdots \otimes e_{i_p}) - (e_{i_1} \otimes \cdots \otimes e_{i_p})]$$

It follows that v is in the subspace I_p of $T^p(V)$ generated by tensors of the form $(-1)^\sigma \lambda_\sigma(t) - t$, that is

$$I_p = \langle (-1)^\sigma \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p \rangle$$

and so $\ker(\tau) \subseteq I_p$. Conversely,

$$\tau(\lambda_\sigma(e_{k_1} \otimes \cdots \otimes e_{k_p}) - (e_{k_1} \otimes \cdots \otimes e_{k_p})) = 0$$

and so $I_p \subseteq \ker(\tau)$.

Theorem 14.22 *Let V be a finite-dimensional vector space over a field F. For $p \geq 1$, the surjective linear map $\tau\colon T^p(V) \to F_p^-[e_1, \ldots, e_n]$ defined by*

$$\tau\Big(\sum \alpha_{i_1,\ldots,i_p} e_{i_1} \otimes \cdots \otimes e_{i_p}\Big) = \sum \alpha_{i_1,\ldots,i_p} e_{i_1} \wedge \cdots \wedge e_{i_p}$$

has kernel

$$I_p = \langle (-1)^\sigma \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p \rangle$$

and so

$$\frac{T^p(V)}{I_p} \approx F_p^-[e_1, \ldots, e_n]$$

The vector space $T^p(V)/I$ is also referred to as the **antisymmetric tensor space** *of degree p of V or the* **exterior algebra** *of degree p of V. The ideal of $T(V)$ defined by*

$$I = \langle (-1)^\sigma \lambda_\sigma(t) - t \mid t \in T^p(V), \sigma \in S_p, p \geq 1 \rangle$$

being generated by homogeneous elements, is graded, so that

$$I = \bigoplus_{p=1}^{\infty} I_p = \bigoplus_{p=0}^{\infty} I_p$$

where $I_0 = \{0\}$. The graded algebra

$$AT(V) = \frac{T(V)}{I} = \bigoplus_{i=0}^{\infty} \frac{T^p(V) + I}{I}$$

is also called the **antisymmetric tensor space** *of V or the* **exterior algebra** *of V and is isomorphic to $F^{-}[e_1,\ldots,e_n]$.* □

The isomorphic exterior spaces $AT^p(V)$ and $T^p(V)/I_p$ are usually denoted by $\bigwedge^p V$ and the isomorphic exterior algebras $AT(V)$ and $T(V)/I$ are usually denoted by $\bigwedge V$.

Before proceding to the universal property, we note that the dimension of the exterior tensor space $\bigwedge^p(V)$ is equal to the number of words of length p in ascending order over the alphabet $E = \{e_1,\ldots,e_n\}$ and this is

$$\dim(\bigwedge^p(V_n)) = \binom{n}{p}$$

The Universal Property for Antisymmetric p-Linear Maps

The vector space $F_p^{-}[x_1,\ldots,x_n]$ and therefore also the isomorphic spaces of antisymmetric tensors $\bigwedge^p(V)$ and $T^p(V)/I_p$, have the universal property for *symmetric* p-linear maps.

Theorem 14.23 *(***The universal property for antisymmetric multilinear maps, as measured by linearity***) Let V be a finite-dimensional vector space over a field F. Then the pair*

$$(F_p^{-}[x_1,\ldots,x_n], t\colon V^{\times p} \to F_p^{-}[x_1,\ldots,x_n])$$

where

$$t(v_1,\ldots,v_p) = v_1 \wedge \cdots \wedge v_p$$

has the universal property for antisymmetric p-linear maps with domain $V^{\times p}$, as measured by linearity. That is, for any antisymmetric p-linear map $f\colon V^{\times p} \to U$ where U is a vector space, there is a unique linear map $\tau\colon F_p^{-}[x_1,\ldots,x_n] \to U$ for which

$$\tau(v_1 \wedge \cdots \wedge v_p) = f(v_1,\ldots,v_p)$$

for any vectors $v_i \in V$.

Proof. Since f is antisymmetric, it is completely determined by the fact that it is alternate and by its values on ascending words $e_{i_1} \wedge \cdots \wedge e_{i_p}$, where $i_1 < \cdots < i_p$. Accordingly, we can define τ by

$$\tau(e_{i_1} \wedge \cdots \wedge e_{i_p}) = f(e_{i_1},\ldots,e_{i_p})$$

and this does indeed uniquely define a well-defined linear transformation τ. □

The Determinant

The universal property for antisymmetric multilinear maps has the following corollary.

Corollary 14.24 *Let V be a vector space of dimension n over a field F. Let $E=(e_1,\dots,e_n)$ be an ordered basis for V. Then there is a unique antisymmetric n-linear form $d\colon V^{\times n}\to F$ for which*

$$d(e_1,\dots,e_n)=1$$

Proof. According to the universal property for antisymmetric n-linear forms, for every such form $f\colon V^{\times n}\to F$, there is a unique linear map $\tau_f\colon \bigwedge^n V\to F$ for which

$$\tau_f(e_1\wedge\cdots\wedge e_n)=f(e_1,\dots,e_n)=1$$

But the dimension of $\bigwedge^n V$ is $\binom{n}{n}=1$ and $\{e_1\wedge\cdots\wedge e_n\}$ is a basis for $\bigwedge^n(V)$. Hence, there is only one linear map $\sigma\colon \bigwedge^n V\to F$ with $\sigma(e_1\wedge\cdots\wedge e_n)=1$. It follows that if f and g are two such forms, then

$$f(e_1,\dots,e_n)=\sigma(e_1\wedge\cdots\wedge e_n)=g(e_1,\dots,e_n)$$

and the antisymmetry of f and g imply that f and g agree on every permutation of $(e_1,\dots,e_n)$. Since f and g are multilinear, we must have $f=g$. □

We now wish to construct the unique antisymmetric form d guaranteed by the previous result. For any $v\in V$, write $[v]_{E,i}$ for the ith coordinate of the coordinate matrix $[v]_E$. Thus,

$$v=\sum_i [v]_{E,i}e_i$$

For clarity, and since we will not change the basis, let us write $[v]_i$ for $[v]_{E,i}$.

Consider the map $d\colon V^{\times n}\to F$ defined by

$$d(v_1,\dots,v_n)=\sum_{\sigma\in S_n}(-1)^\sigma[v_1]_{\sigma(1)}\cdots[v_n]_{\sigma(n)}$$

Then d is multilinear since

$$\begin{aligned}
d(av_1+bu_1,\ldots,v_n) &= \sum_{\sigma\in S_n}(-1)^\sigma[av_1+bu_1]_{\sigma(1)}\cdots[v_n]_{\sigma(n)}\\
&= \sum_{\sigma\in S_n}(-1)^\sigma(a[v_1]_{\sigma(1)}+b[u_1]_{\sigma(1)})\cdots[v_n]_{\sigma(n)}\\
&= a\sum_{\sigma\in S_n}(-1)^\sigma[v_1]_{\sigma(1)}\cdots[v_n]_{\sigma(n)}\\
&\qquad + b\sum_{\sigma\in S_n}(-1)^\sigma[u_1]_{\sigma(1)}\cdots[v_n]_{\sigma(n)}\\
&= ad(v_1,\ldots,v_n)+bd(u_1,v_2,\ldots,v_n)
\end{aligned}$$

and similarily for any coordinate position.

The map d is alternating, and therefore antisymmetric since $\mathrm{char}(F)\neq 2$. To see this, suppose for instance that $v_1=v_2$. For any permutation $\sigma\in S_n$, let $\sigma(1)=a$ and $\sigma(2)=b$. Then the permutation $\sigma'=(ab)\sigma$ satisfies

1) For $x\neq 1$ and $x\neq 2$, $\sigma'(x)=\sigma(x)$
2) $\sigma'(1)=(ab)\sigma(1)=(ab)(a)=b=\sigma(2)$
3) $\sigma'(2)=(ab)\sigma(2)=(ab)(b)=a=\sigma(1)$.

Hence, $\sigma'\neq\sigma$ and it is easy to check that $(\sigma')'=\sigma$. It follows that if the sets $\{\sigma,\sigma'\}$ and $\{\rho,\rho'\}$ intersect, then they are identical. In other words, the distinct sets $\{\sigma,\sigma'\}$ form a partition of S_n.

Hence,

$$\begin{aligned}
d(v_1,v_1,\ldots,v_n) &= \sum_{\sigma\in S_n}(-1)^\sigma[v_1]_{\sigma(1)}[v_1]_{\sigma(2)}\cdots[v_n]_{\sigma(n)}\\
&= \sum_{\text{pairs }\{\sigma,\sigma'\}}\Big[(-1)^\sigma[v_1]_{\sigma(1)}[v_1]_{\sigma(2)}\cdots[v_n]_{\sigma(n)}\\
&\qquad\qquad + (-1)^{\sigma'}[v_1]_{\sigma'(1)}[v_1]_{\sigma'(2)}\cdots[v_n]_{\sigma'(n)}\Big]
\end{aligned}$$

But

$$[v_1]_{\sigma(1)}[v_1]_{\sigma(2)}=[v_1]_{\sigma'(1)}[v_1]_{\sigma'(2)}$$

and since $(-1)^\sigma=-(-1)^{\sigma'}$, the sum of the two terms involving the pair $\{\sigma,\sigma'\}$ is 0. Hence, $d(v_1,v_1,\ldots,v_n)=0$. A similar argument holds for any coordinate pair.

Finally, we have

$$\begin{aligned} d(e_1,\ldots,e_n) &= \sum_{\sigma\in S_n}(-1)^\sigma [e_1]_{\sigma(1)}\cdots[e_n]_{\sigma(n)} \\ &= \sum_{\sigma\in S_n}(-1)^\sigma \delta_{1,\sigma(1)}\cdots\delta_{n,\sigma(n)} \\ &= 1 \end{aligned}$$

Thus, the map d is indeed the unique antisymmetric n-linear form on $V^{\times n}$ for which $d(e_1,\ldots,e_n)=1$.

Given the ordered basis $E=(e_1,\ldots,e_n)$, we can view V as the space F^n of coordinate vectors and view $V^{\times n}$ as the space $M_n(F)$ of $n\times n$ matrices, via the isomorphism

$$(v_1,\ldots,v_n)\mapsto \begin{bmatrix} [v_1]_1 & \cdots & [v_n]_1 \\ \vdots & & \vdots \\ [v_1]_n & \cdots & [v_n]_n \end{bmatrix}$$

where all coordinate matrices are with respect to E.

With this viewpoint, d becomes an antisymmetric n-form on the columns of a matrix $A=(a_{i,j})$ given by

$$d(A)=\sum_{\sigma\in S_n}(-1)^\sigma a_{1,\sigma(1)}\cdots a_{n,\sigma(n)}$$

This is called the **determinant** of the matrix A.

Properties of the Determinant

Let us explore some of the properties of the determinant function.

Theorem 14.25 *If* $A\in M_n(F)$ *then* $d(A)=d(A^t)$.
Proof. We know that

$$d(A)=\sum_{\sigma\in S_n}(-1)^\sigma a_{1,\sigma(1)}\cdots a_{n,\sigma(n)}$$

which can be written in the form

$$d(A)=\sum_{\sigma\in S_n}(-1)^\sigma a_{\sigma^{-1}(\sigma(1)),\sigma(1)}\cdots a_{\sigma^{-1}(\sigma(n)),\sigma(n)}$$

But we can reorder the factors in each term so that the second indices are in ascending order, giving

$$\begin{aligned} d(A) &= \sum_{\sigma\in S_n} (-1)^{\sigma} a_{\sigma^{-1}(1),1}\cdots a_{\sigma^{-1}(n),n} \\ &= \sum_{\sigma^{-1}\in S_n} (-1)^{\sigma^{-1}} a_{\sigma^{-1}(1),1}\cdots a_{\sigma^{-1}(n),n} \\ &= \sum_{\sigma\in S_n} (-1)^{\sigma} a_{\sigma(1),1}\cdots a_{\sigma(n),n} \\ &= d(A^t) \end{aligned}$$

as desired. □

Theorem 14.26 *If* $A, B \in M_n(F)$ *then* $d(AB) = d(A)d(B)$.
Proof. Consider the map $f_A\colon M_n(F) \to F$ defined by

$$f_A(X) = d(AX)$$

We can consider f_A as a function on the columns of X and write

$$f_A\colon (X^{(1)}, \ldots, X^{(n)}) \mapsto (AX^{(1)}, \ldots, AX^{(n)}) \mapsto d(AX)$$

Now, this map is multilinear since multiplication by A is distributive and the determinant is multilinear. For example, let $y \in F^n$ and let X' come from X by replacing the first column by y. Then

$$\begin{aligned} f_A(aX^{(1)} + by, \ldots, X^{(n)}) &\mapsto (aAX^{(1)} + bAy, \ldots, AX^{(n)}) \\ &\mapsto ad(AX) + bd(AX') \\ &= af_A(X) + bf_A(X') \end{aligned}$$

The map f_A is also alternating since d is alternating and interchanging two coordinates in $(X^{(1)}, \ldots, X^{(n)})$ is equivalent to interchanging the corresponding columns of AX.

Thus, f_A is an antisymmetric n-linear form and so must be a scalar multiple of the determinant function, say $f_A(X) = \gamma d(X)$. Then

$$d(AX) = f_A(X) = \gamma d(X)$$

Setting $X = I_n$ gives $d(A) = \gamma$ and so

$$d(AX) = d(A)d(X)$$

as desired. □

If $P \in M_n(F)$ is invertible, then $PP^{-1} = I_n$ and so

$$d(P)d(P^{-1}) = 1$$

which shows that $d(P) \neq 0$ and $d(P^{-1}) = 1/d(P)$.

But any matrix $A \in M_n(F)$ is equivalent to a diagonal matrix

$$A = PDQ$$

where P and Q are invertible and D is diagonal with 1's and 0's on the main diagonal. Hence,

$$d(A) = d(P)d(D)d(Q)$$

and so if $d(A) \neq 0$ then $d(D) \neq 0$. But this can happen only if $D = I_n$, whence A is invertible. We have proved the following.

Theorem 14.27 A matrix $A \in M_n(F)$ is invertible if and only if $d(A) \neq 0$. □

Exercises

1. Show that if $\tau\colon W \to X$ is a linear map and $b\colon U \times V \to W$ is bilinear then $\tau \circ b\colon U \times V \to X$ is bilinear.
2. Show that the only map that is both linear and n-linear (for $n \geq 2$) is the zero map.
3. Find an example of a bilinear map $\tau\colon V \times V \to W$ whose image $\operatorname{im}(\tau) = \{\tau(u, v) \mid u, v \in V\}$ is not a subspace of W.
4. Prove that the universal property of tensor products defines the tensor product up to isomorphism only. That is, if a pair $(X, s\colon U \times V \to X)$ has the universal property then X is isomorphic to $U \otimes V$.
5. Prove that the following property of a pair $(W, g\colon U \times V \to W)$ with g bilinear characterizes the tensor product $(U \otimes V, t\colon U \times V \to U \otimes V)$ up to isomorphism, and thus could have been used as the definition of tensor product: For a pair $(W, g\colon U \times V \to W)$ with g bilinear if $\{u_i\}$ is a basis for U and $\{v_i\}$ is a basis for V then $\{g(u_i, v_j)\}$ is a basis for W.
6. Prove that $U \otimes V \approx V \otimes U$.
7. Let X and Y be nonempty sets. Use the universal property of tensor products to prove that $\mathcal{F}_{X \times Y} \approx \mathcal{F}_X \otimes \mathcal{F}_Y$.
8. Let $u, u' \in U$ and $v, v' \in V$. Assuming that $u \otimes v \neq 0$, show that $u \otimes v = u' \otimes v'$ if and only if $u' = ru$ and $v' = r^{-1}v$, for $r \neq 0$.
9. Let $\mathcal{B} = \{b_i\}$ be a basis for U and $\mathcal{C} = \{c_i\}$ be a basis for V. Show that any function $f\colon U \times V \to W$ can be extended to a linear function $\overline{f}\colon U \otimes V \to W$. Deduce that the function f can be extended in a unique way to a bilinear map $\widehat{f}\colon U \times V \to W$. Show that all bilinear maps are obtained in this way.
10. Let S_1, S_2 be subspaces of U. Show that

$$(S_1 \otimes V) \cap (S_2 \otimes V) \approx (S_1 \cap S_2) \otimes V$$

11. Let $S \subseteq U$ and $T \subseteq V$ be subspaces of vector spaces U and V, respectively. Show that

$$(S \otimes V) \cap (U \otimes T) \approx S \otimes T$$

12. Let $S_1, S_2 \subseteq U$ and $T_1, T_2 \subseteq V$ be subspaces of U and V, respectively. Show that
$$(S_1 \otimes T_1) \cap (S_2 \otimes T_2) \approx (S_1 \cap S_2) \otimes (T_1 \otimes T_2)$$
13. Find an example of two vector spaces U and V and a nonzero vector $x \in U \otimes V$ that has at least two distinct (not including order of the terms) representations of the form
$$x = \sum_{i=1}^{n} u_i \otimes v_i$$
where the u_i's are linearly independent and so are the v_i's.
14. Let ι_X denote the identity operator on a vector space X. Prove that $\iota_V \odot \iota_W = \iota_{V \otimes W}$.
15. Suppose that $\tau_1: U \to V$, $\tau_2: V \to W$ and $\sigma_1: U' \to V_K$, $\sigma_2: V_K \to W'$. Prove that
$$(\tau_2 \circ \tau_1) \odot (\sigma_2 \circ \sigma_1) = (\tau_2 \odot \sigma_2) \circ (\tau_1 \odot \sigma_1)$$
16. Connect the two approaches to extending the base field of an F-space V to K (at least in the finite-dimensional case) by showing that $F^n \otimes_F K \approx (K)^n$.
17. Prove that in a tensor product $U \otimes U$ for which $\dim(U) \geq 2$ not all vectors have the form $u \otimes v$ for some $u, v \in U$. *Hint*: Suppose that $u, v \in U$ are linearly independent and consider $u \otimes v + v \otimes u$.
18. Prove that for the block matrix
$$M = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}_{\text{block}}$$
we have $d(M) = d(A)d(C)$.
19. Let $A, B \in M_n(F)$. Prove that if either A or B is invertible, then the matrices $A + \alpha B$ are invertible except for a finite number of α's.

The Tensor Product of Matrices

20. Let $A = (a_{i,j})$ be the matrix of a linear operator $\tau \in \mathcal{L}(V)$ with respect to the ordered basis $\mathcal{A} = (u_1, \ldots, u_n)$. Let $B = (b_{i,j})$ be the matrix of a linear operator $\sigma \in \mathcal{L}(V)$ with respect to the ordered basis $\mathcal{B} = (v_1, \ldots, v_m)$. Consider the ordered basis $\mathcal{C} = (u_i \otimes v_j)$ ordered by lexicographic order, that is $u_i \otimes v_j < u_\ell \otimes v_k$ if $i < \ell$ or $i = \ell$ and $j < k$. Show that the matrix of $\tau \otimes \sigma$ with respect to $\mathcal{C}$ is
$$A \otimes B = \begin{pmatrix} a_{1,1}B & a_{1,2}B & \cdots & a_{1,n}B \\ a_{2,1}B & a_{2,2}B & \cdots & a_{2,n}B \\ \vdots & \vdots & & \vdots \\ a_{n,1}B & a_{n,2}B & \cdots & a_{n,n}B \end{pmatrix}_{\text{block}}$$

This matrix is called the **tensor product**, **Kronecker product** or **direct product** of the matrix A with the matrix B.

21. Show that the tensor product is not, in general, commutative.
22. Show that the tensor product $A \otimes B$ is bilinear in both A and B.
23. Show that $A \otimes B = 0$ if and only if $A = 0$ or $B = 0$.
24. Show that
    a) $(A \otimes B)^t = A^t \otimes B^t$
    b) $(A \otimes B)^* = A^* \otimes B^*$ (when $F = \mathbb{C}$)
25. Show that if $u, v \in F^n$ then (as row vectors) $u^t v = u^t \otimes v$.
26. Suppose that $A_{m,n}, B_{p,q}, C_{n,k}$ and $D_{q,r}$ are matrices of the given sizes. Prove that

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$$

    Discuss the case $k = r = 1$.
27. Prove that if A and B are nonsingular, then so is $A \otimes B$ and

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

28. Prove that tr$(A \otimes B) =$ tr$(A) \cdot$ tr(B)
29. Suppose that F is algebraically closed. Prove that if A has eigenvalues $\lambda_1, \ldots, \lambda_n$ and B has eigenvalues $\mu_1, \ldots, \mu_m$ both lists including multiplicity then $A \otimes B$ has eigenvalues $\{\lambda_i \mu_j \mid i \le n, j \le m\}$, again counting multiplicity.
30. Prove that det$(A_{n,n} \otimes B_{m,m}) = (det(A_{n,n}))^m(det(B_{m,m}))^n$.

Chapter 15
Positive Solutions to Linear Systems: Convexity and Separation

Given a matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$ consider the homogeneous system of linear equations

$$Ax = 0$$

It is of obvious interest to determine conditions that guarantee the existence of *positive* solutions to this system, in a manner made precise by the following definition.

Definition *Let* $v = (a_1, \ldots, a_n) \in \mathbb{R}^n$*. Then*

1) v *is* **nonnegative***, written* $v \geq 0$ *if*

$$a_i \geq 0 \textit{ for all } i = 1, \ldots, n$$

(Note that the term **positive** *is also used in the literature for this property.) The set of all nonnegative vectors in* $\mathbb{R}^n$ *is the* **nonnegative orthant** *in* $\mathbb{R}^n$*.*

2) v *is* **strictly positive***, written* $v > 0$ *if* v *is nonnegative but not* 0*, that is, if*

$$a_i \geq 0 \textit{ for all } i = 1, \ldots, n \textit{ and } a_j > 0 \textit{ for at least one } j = 1, \ldots, n$$

The set $\mathbb{R}^n_+$ *of all strictly positive vectors in* $\mathbb{R}^n$ *is the* **strictly positive orthant** *in* $\mathbb{R}^n$*.*

3) v *is* **strongly positive***, written* $v \gg 0$ *if*

$$a_i > 0 \textit{ for all } i = 1, \ldots, n$$

The set $\mathbb{R}^n_{++}$ *of all strongly positive vectors in* $\mathbb{R}^n$ *is the* **strongly positive orthant** *in* $\mathbb{R}^n$*.*□

We are interested in conditions under which the system $Ax = 0$ has strictly positive or strongly positive solutions. Since the strictly and strongly positive orthants in $\mathbb{R}^n$ are not subspaces of $\mathbb{R}^n$, it is difficult to use strictly linear methods in studying this issue: we must also use geometric methods, in particular, methods of convexity.

Let us pause briefly to consider an important application of strictly positive solutions to the system $Ax = 0$. If $X = (x_1, \ldots, x_n)$ is a strictly positive solution then so is the vector

$$\Pi = \frac{1}{\Sigma x_i} X = \frac{1}{\Sigma x_i}(x_1, \ldots, x_n) = (\pi_1, \ldots, \pi_n)$$

which is a *probability distribution*. Note that if we replace "strictly" with "strongly" then the probability distribution has the property that each probability is positive.

Now, the product $A\Pi$ is the expected value of the columns of A with respect to the probability distribution Π. Hence, $Ax = 0$ has a strictly (strongly) positive solution if and only if there is a strictly (strongly) positive probability distribution for which the columns of A have expected value 0. If these columns represent the payoffs from a game of chance then the game is fair when the expected value of the columns is 0. Thus, $Ax = 0$ has a strictly (strongly) positive solution if and only if the "game" A, where in the strongly positive case, all outcomes are possible, is fair.

As another (related) example, in discrete option pricing models of mathematical finance, the absence of arbitrage opportunities in the model is equivalent to the fact that a certain vector describing the gains in a portfolio does not intersect the strictly positive orthant in $\mathbb{R}^n$. As we will see in this chapter, this is equivalent to the existence of a strongly positive solution to a homogeneous system of equations. This solution, when normalized to a probability distribution, is called a *martingale measure*.

Of course, the equation $Ax = 0$ has a strictly positive solution if and only if $\ker(A)$ contains a strictly positive vector, that is, if and only if

$$\ker(A) = \text{RowSpace}(A)^{\perp}$$

meets the strictly positive orthant in $\mathbb{R}^n$. Thus, we wish to characterize the subspaces S of $\mathbb{R}^n$ for which $S^{\perp}$ meets the strictly positive orthant in $\mathbb{R}^n$, in symbols

$$S^{\perp} \cap \mathbb{R}^n_+ \neq \emptyset$$

for these are precisely the row spaces of the matrices A for which $Ax = 0$ has a strictly positive solution. A similar statement holds for strongly positive solutions.

Looking at the real plane $\mathbb{R}^2$, we can divine the answer with a picture. A one-dimensional subspace S of $\mathbb{R}^2$ has the property that its orthogonal complement

$S^\perp$ meets the strictly positive orthant (quadrant) in $\mathbb{R}^2$ if and only if S is the x-axis, the y-axis or a line with negative slope. For the case of the strongly positive orthant, S must have negative slope. Our task is to generalize this to $\mathbb{R}^n$.

This will lead us to the following results, which are quite intuitive in $\mathbb{R}^2$ and $\mathbb{R}^3$

$$S^\perp \cap \mathbb{R}^n_{++} \neq \emptyset \text{ if and only if } S \cap \mathbb{R}^n_+ = \emptyset \tag{15.1}$$

and

$$S^\perp \cap \mathbb{R}^n_{+} \neq \emptyset \text{ if and only if } S \cap \mathbb{R}^n_{++} = \emptyset \tag{15.2}$$

Let us apply this to the matrix equation $Ax = 0$. If $S = \text{RowSpace}(A)$ then $S^\perp = \ker(A)$ and so we have

$$\ker(A) \cap \mathbb{R}^n_{++} \neq \emptyset \text{ if and only if } \text{RowSpace}(A) \cap \mathbb{R}^n_+ = \emptyset$$

and

$$\ker(A) \cap \mathbb{R}^n_{+} \neq \emptyset \text{ if and only if } \text{RowSpace}(A) \cap \mathbb{R}^n_{++} = \emptyset$$

Now,

$$\text{RowSpace}(A) \cap \mathbb{R}^n_+ = \{vA \mid vA > 0\}$$

and

$$\text{RowSpace}(A) \cap \mathbb{R}^n_{++} = \{vA \mid vA \gg 0\}$$

and so these statements become

$$Ax = 0 \text{ has a strongly positive solution if and only if } \{vA \mid vA > 0\} = \emptyset$$

and

$$Ax = 0 \text{ has a strictly positive solution if and only if } \{vA \mid vA \gg 0\} = \emptyset$$

We can rephrase these results in the form of a *theorem of the alternative*, that is, a theorem that says that exactly one of two conditions holds.

Theorem 15.1 *Let* $A \in \mathcal{M}_{m,n}(\mathbb{R})$.

1) *Exactly one of the following holds:*
   a) $Au = 0$ *for some strongly positive* $u \in \mathbb{R}^n$
   b) $vA > 0$ *for some* $v \in \mathbb{R}^m$
2) *Exactly one of the following holds:*
   a) $Au = 0$ *for some strictly positive* $u \in \mathbb{R}^n$
   b) $vA \gg 0$ *for some* $v \in \mathbb{R}^m$. □

Before proving statements (15.1) and (15.2), we require some background.

Convex, Closed and Compact Sets

We shall need the following concepts.

Definition

1) *Let $x_1, \ldots, x_k \in \mathbb{R}^n$. Any linear combination of the form*

$$t_1x_1 + \cdots + t_kx_k$$

where $t_1 + \cdots + t_k = 1, 0 \leq t_i \leq 1$ is called a **convex combination** *of the vectors $x_1, \ldots, x_k$.*

2) *A subset $X \subseteq \mathbb{R}^n$ is* **convex** *if whenever $x, y \in X$ then the entire line segment between x and y also lies in X, in symbols*

$$\{sx + ty \mid s + t = 1, 0 \leq s, t \leq 1\} \subseteq X$$

3) *A subset $X \subseteq \mathbb{R}^n$ is* **closed** *if whenever (x_n) is a convergent sequence of elements of X, then the limit is also in X. Simply put, a subset is closed if it is closed under the taking of limits.*
4) *A subset $X \subseteq \mathbb{R}^n$ is* **compact** *if it is both closed and bounded.*
5) *A subset $X \subseteq \mathbb{R}^n$ is a* **cone** *if $x \in X$ implies that $ax \in X$ for all $a \geq 0$.* □

We will also have need of the following facts from analysis.

1) A continuous function that is defined on a compact set X in $\mathbb{R}^n$ takes on its maximum and minimum values at some points within the set X.
2) A subset X of $\mathbb{R}^n$ is compact if and only if every sequence in X has a subsequence that converges in X.

Theorem 15.2 *Let X and Y be subsets of $\mathbb{R}^n$. Define*

$$X + Y = \{a + b \mid a \in X, \, b \in Y\}$$

1) *If X and Y are convex then so is $X + Y$*
2) *If X is compact and Y is closed then $X + Y$ is closed.*

Proof. For 1) let $x_0 + y_0$ and $x_1 + y_1$ be in $X + Y$. The line segment between these two points is

$$\begin{aligned} t(x_0 + y_0) \; + \; & (1-t)(x_1 + y_1) \\ & = tx_0 + (1-t)x_1 + ty_0 + (1-t)y_1 \\ & \in X + Y \end{aligned}$$

where $0 \leq t \leq 1$ and so $X + Y$ is convex.

For part 2) let $x_n + y_n$ be a convergent sequence in $X + Y$. Suppose that $x_n + y_n \to z$. We must show that $z \in X + Y$. Since x_n is a sequence in the compact set X, it has a convergent subsequence x_{n_k} whose limit x lies in X. Since $a_{n_k} + b_{n_k} \to z$ and $a_{n_k} \to x$ we can conclude that $b_{n_k} \to z - x$. Since Y is closed, we must have $z - x \in Y$ and so $z = x + (z - x) \in X + Y$, as desired. □

Convex Hulls

We will have use for the notion of convex hull.

Definition *The* **convex hull** *of a set $S = \{x_1, \ldots, x_k\}$ of vectors in $\mathbb{R}^n$ is the smallest convex set in $\mathbb{R}^n$ that contains the vectors $x_1, \ldots, x_k$. We denote the convex hull of S by $\mathcal{C}(S)$.* □

Here is a characterization of convex hulls.

Theorem 15.3 *Let $S = \{x_1, \ldots, x_k\}$ be a set of vectors in $\mathbb{R}^n$. Then the convex hull $\mathcal{C}(S)$ is the set Δ of all convex combinations of vectors in S, that is,*

$$\mathcal{C}(S) = \Delta \coloneqq \{t_1x_1 + \cdots + t_kx_k \mid 0 \le t_i \le 1, \Sigma t_i = 1\}$$

Proof. First, we show that Δ is convex. Let

$$\begin{aligned} X &= t_1x_1 + \cdots + t_kx_k \\ Y &= s_1x_1 + \cdots + s_kx_k \end{aligned}$$

be convex combinations of S and let $a + b = 1, 0 \le a, b \le 1$. Then

$$\begin{aligned} aX + bY &= a(t_1x_1 + \cdots + t_kx_k) + b(s_1x_1 + \cdots + s_kx_k) \\ &= (at_1 + bs_1)x_1 + \cdots + (at_k + bs_k)x_k \end{aligned}$$

But this is also a convex combination of the vectors in S because

$$0 \le at_i + bs_i \le (a + b)\max(s_i, t_i) = \max(s_i, t_i) \le 1$$

and

$$\sum_{i=1}^{k}(at_i + bs_i) = a\sum_{i=1}^{k}t_i + b\sum_{i=1}^{k}s_i = a + b = 1$$

Thus,

$$X, Y \in \Delta \Rightarrow aX + bY \in \Delta$$

which says that Δ is convex. Since $S \subseteq \Delta$, we have $\mathcal{C}(S) \subseteq \Delta$. Clearly, if D is a convex set that contains S then D also contains Δ. Hence $\Delta \subseteq \mathcal{C}(S)$. □

Theorem 15.4 *The convex hull $\mathcal{C}(S)$ of a* finite *set $S = \{x_1, \ldots, x_k\}$ of vectors in $\mathbb{R}^n$ is a compact set.*

Proof. Let

$$D = \left\{ (t_1, \ldots, t_k) \mid 0 \le t_i \le 1 \text{ and } \sum_i t_i = 1 \right\}$$

and define a function $f\colon D \to \mathbb{R}^n$ as follows. If $t = (t_1, \ldots, t_k)$ then

$$f(t) = t_1 x_1 + \cdots + t_k x_k$$

To see that f is continuous, let $s = (s_1, \ldots, s_k)$ and let $M = \max(\|x_i\|)$. For $\epsilon > 0$, if $\|s - t\| < \epsilon / kM$ then

$$|s_i - t_i| \le \|s - t\| < \frac{\epsilon}{nM}$$

and so

$$\begin{aligned} \|f(s) - f(t)\| &= \|(s_1 - t_1)x_1 + \cdots + (s_k - t_k)x_k\| \\ &\le |s_1 - t_1|\|x_1\| + \cdots + |s_k - t_k|\|x_k\| \\ &\le kM\|s - t\| \\ &= \epsilon \end{aligned}$$

Finally, since f maps the compact set D onto $\mathcal{C}(S)$, we deduce that $\mathcal{C}(S)$ is compact. □

Linear and Affine Hyperplanes

We next discuss hyperplanes in $\mathbb{R}^n$. A **linear hyperplane** in $\mathbb{R}^n$ is an $(n-1)$-dimensional subspace of $\mathbb{R}^n$. As such, it is the solution set of a linear equation of the form

$$a_1 x_1 + \cdots + a_n x_n = 0$$

or

$$\langle N, x \rangle = 0$$

where $N = (a_1, \ldots, a_n)$ is nonzero and $x = (x_1, \ldots, x_n)$. Geometrically speaking, this is the set of all vectors in $\mathbb{R}^n$ that are perpendicular (normal) to the vector N.

An **(affine) hyperplane** is a linear hyperplane that has been translated by a vector. Thus, it is the solution set to an equation of the form

$$a_1(x_1 - b_1) + \cdots + a_n(x_n - b_n) = 0$$

or

$$a_1 x_1 + \cdots + a_n x_n = a_1 b_1 + \cdots a_n b_n$$

or finally

$$\langle N, x\rangle = \langle N, B\rangle$$

where $B = (b_1, \ldots, b_n)$.

Let us write $\mathcal{H}(N, b)$, where $N \in \mathbb{R}^n$ and $b \in \mathbb{R}$, to denote the hyperplane

$$\mathcal{H}(N, b) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle = b\}$$

Note that the hyperplane

$$\mathcal{H}(N, \|N\|^2) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle = \|N\|^2\}$$

contains the point N, which is the point of $\mathcal{H}(N, b)$ closest to the origin, since Cauchy's inequality gives

$$\|N\|^2 = \langle N, x\rangle \leq \|N\|\|x\|$$

and so $\|N\| \leq \|x\|$ for all $x \in \mathcal{H}(N, \|N\|^2)$. Moreover, any hyperplane has the form $\mathcal{H}(N, \|N\|^2)$ for any appropriate vector N.

A hyperplane defines two (nondisjoint) **closed half-spaces**

$$\mathcal{H}_+(N, b) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle \geq b\}$$
$$\mathcal{H}_-(N, b) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle \leq b\}$$

and two (disjoint) **open half-spaces**

$$\mathcal{H}_+^\circ(N, b) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle > b\}$$
$$\mathcal{H}_-^\circ(N, b) = \{x \in \mathbb{R}^n \mid \langle N, x\rangle < b\}$$

It is not hard to show that

$$\mathcal{H}_+(N, b) \cap \mathcal{H}_-(N, b) = \mathcal{H}(N, b)$$

and that $\mathcal{H}_+^\circ(N, b)$, $\mathcal{H}_-^\circ(N, b)$ and $\mathcal{H}(N, b)$ are pairwise disjoint and

$$\mathcal{H}_+^\circ(N, b) \cup \mathcal{H}_-^\circ(N, b) \cup \mathcal{H}(N, b) = \mathbb{R}^n$$

Definition *The subsets X and Y of $\mathbb{R}^n$ are* **strictly separated** *by a hyperplane $\mathcal{H}(N, b)$ if X lies in one open half-space determined by $\mathcal{H}(N, b)$ and Y lies in the other. Thus, one of the following holds:*

1) $\langle N, x\rangle < b < \langle N, y\rangle$ *for all* $x \in X, y \in Y$
2) $\langle N, y\rangle < b < \langle N, x\rangle$ *for all* $x \in X, y \in Y$. □

Note that 1) holds for N and b if and only if 2) holds for $-N$ and $-b$, and so we need only consider one of the conditions to demonstrate that two sets X and Y are *not* stricctly separated. In particular, suppose that 1) fails for all N and b. Then the condition

$$\langle -N, y\rangle < -b < \langle -N, x\rangle$$

also fails and so 1) and 2) both fail for all N and b and X and Y are not strictly separated.

The following type of separation is stronger than strict separation.

Definition *The subsets X and Y of $\mathbb{R}^n$ are* **strongly separated** *by a hyperplane $\mathcal{H}(N, b)$ if there is an $e > 0$ for which one of the following holds:*

1) $\langle N, x\rangle < b - e < b + e < \langle N, y\rangle$ for all $x \in X, y \in Y$
2) $\langle N, y\rangle < b - e < e + b < \langle N, x\rangle$ for all $x \in X, y \in Y$□

Note that, as before, we need only consider one of the conditions to show that two sets are *not* strongly separated.

Separation

Now that we have the preliminaries out of the way, we can get down to some theorems. The first is a well known *separation theorem* that is the basis for many other separation theorems. It says that if a closed convex set C in $\mathbb{R}^n$ does not contain a vector b, then C can be strongly separated from b by a hyperplane.

Theorem 15.5 *Let C be a closed convex subset of $\mathbb{R}^n$.*

1) C contains a unique *vector N of minimum norm, that is, there is a unique vector $N \in C$ for which*

$$\|N\| < \|x\|$$

for all $x \in C, x \neq N$.

2) If C does not contain the origin then C lies in the closed half-space

$$\langle N, x\rangle \geq \|N\|^2 > 0$$

where $N \neq 0$ is the vector in C of minimum norm. Hence, 0 and C are strongly separated by the hyperplane $\mathcal{H}(N, \|N\|^2/2)$.

3) If $b \notin C$ then b and C are strongly separated.

Proof. For part 1), we first show that C contains a vector N of minimum norm. Recall that the Euclidean norm (distance) is a continuous function. Although C need not be compact, if we choose a real number s such that the closed ball

$$B_s(0) = \{z \in \mathbb{R}^n \mid \|z\| \leq s\}$$

intersects C, then that intersection $C' = C \cap B_s(0)$ is both closed and bounded and so is compact. The distance function therefore achieves its minimum on C', say at the point $N \in C' \subseteq C$. It is clear that if for some $v \in C$ we have $\|v\| < \|N\|$ then $v \in B_{\|N\|}(0) \subseteq C'$, which is a contradiction to the minimality of N. Hence, N is a vector of minimum norm in C. Let us write $\|N\| = a$.

Suppose now that $x \neq N$ is another vector in C with $\|x\| = a$. Since C is convex, the line segment from N to x must be contained in C. In particular, the vector $z = (1/2)(x + N)$ is in C. Since x cannot be a scalar multiple of N, the Cauchy-Schwarz inequality is strict

$$\langle N, x \rangle < \|N\|\|x\| = a^2$$

Hence

$$\begin{aligned} \|z\|^2 &= \frac{1}{4}\|x + N\|^2 \\ &= \frac{1}{4}(\|N\|^2 + 2\langle N, x \rangle + \|x\|^2) \\ &= \frac{1}{2}(a^2 + \langle N, x \rangle) \\ &< a^2 \end{aligned}$$

But this contradicts the minimality of a and so $x = N$. Thus, C has a unique vector of minimum norm.

For part 2), if there is an $x \in C$ for which

$$\langle N, x \rangle < \|N\|^2 = a^2$$

then again setting $z = (1/2)(x + N) \in C$ we find that z has norm less than a, which is not possible. Hence,

$$\langle N, x \rangle \geq \|N\|^2$$

for all $x \in C$.

For part 3), if $b \notin C$ is not the origin, then 0 is not in the closed convex set

$$C - \{b\} = \{c - b \mid c \in C\}$$

Hence, by part 2), there is a nonzero $N \in C$ for which

$$\langle N, x \rangle \geq \|N\|^2$$

for all $x \in C - \{b\}$. But as x ranges over $C - \{b\}$, the vector $x - b$ ranges over C so we have

$$\langle N, x - b \rangle \geq \|N\|^2$$

for all $x \in C$. This can be written

$$\langle N, x \rangle \geq \|N\|^2 + \langle N, b \rangle$$

for all $x \in C$. Hence

$$\langle N, b \rangle < \|N\|^2 + \langle N, b \rangle \le \langle N, x \rangle$$

from which it follows that b and C are strongly separated. □

The next result brings us closer to our goal by replacing the origin with a subspace S disjoint from C. However, we must strengthen the requirements on C a bit.

Theorem 15.6 *Let C be a compact convex subset of $\mathbb{R}^n$ and let S be a subspace of $\mathbb{R}^n$ such that $C \cap S = \emptyset$. Then there exists a nonzero $N \in S^\perp$ such that*

$$\langle N, x \rangle \ge \|N\|^2$$

for all $x \in C$. Hence, the hyperplane $\mathcal{H}(N, \|N\|^2/2)$ strongly separates S and C.

Proof. Consider the set $S + C$, which is closed since S is closed and C is compact. It is also convex since S and C are convex. Furthermore, $0 \notin S + C$ because if $0 = s + c$ then $c = -s$ would be in $C \cap S = \emptyset$.

According to Theorem 15.5, the set $S + C$ can be strongly separated from the origin. Hence, there is a nonzero $N \in \mathbb{R}^n$ such that

$$\langle N, x \rangle \ge \|N\|^2$$

for all $x = s + c \in S + C$, that is,

$$\langle N, s \rangle + \langle N, c \rangle = \langle N, s + c \rangle \ge \|N\|^2$$

for all $s \in S$ and $c \in C$. Now, if $\langle N, s \rangle$ is nonzero for any $s \in S$, we can replace s by an appropriate scalar multiple of s to make the left side negative, which is impossible. Hence, we must have $\langle N, s \rangle = 0$ for all $s \in S$. Thus, $N \in S^\perp$ and

$$\langle N, c \rangle \ge \|N\|^2$$

for all $c \in C$, as desired. □

We can now prove (15.1) and (15.2).

Theorem 15.7 *Let S be a subspace of $\mathbb{R}^n$. Then*
1) $S \cap \mathbb{R}^n_+ = \emptyset$ if and only if $S^\perp \cap \mathbb{R}^n_{++} \neq \emptyset$
2) $S \cap \mathbb{R}^n_{++} = \emptyset$ if and only if $S^\perp \cap \mathbb{R}^n_+ \neq \emptyset$

Proof. For part 1), it is clear that there cannot exist vectors $u \in \mathbb{R}^n_{++}$ and $v \in \mathbb{R}^n_+$ that are orthogonal. Hence, $S \cap \mathbb{R}^n_+$ and $S^\perp \cap \mathbb{R}^n_{++}$ cannot both be nonempty, so if $S^\perp \cap \mathbb{R}^n_{++} \neq \emptyset$ then $S \cap \mathbb{R}^n_+ = \emptyset$. The converse is more interesting.

Suppose that $S \cap \mathbb{R}_+^n = \emptyset$. A good candidate for an element of $S^\perp \cap \mathbb{R}_{++}^n$ would be a normal to a hyperplane that separates S from a subset of $\mathbb{R}_+^n$. Note that our theorems do not allow us to separate S from $\mathbb{R}_+^n$, because it is not compact. So consider instead the convex hull Δ of the standard basis vectors $\epsilon_1, \dots, \epsilon_n$ in $\mathbb{R}_+^n$

$$\Delta = \{t_1\epsilon_1 + \cdots + t_n\epsilon_n \mid 0 \le t_i \le 1, \Sigma t_i = 1\}$$

It is clear that Δ is convex and $\Delta \subseteq \mathbb{R}_+^n$ and so $\Delta \cap S = \emptyset$. Also, Δ is closed and bounded and therefore compact. Hence, by Theorem 15.6, there is a nonzero vector $N = (a_1, \dots, a_n) \in S^\perp$ such that

$$\langle N, \delta \rangle \ge \|N\|^2$$

for all $\delta \in \Delta$. Taking $\delta = \epsilon_i$ gives

$$a_i = \langle N, \epsilon_i \rangle \ge \|N\|^2 > 0$$

and so $N \in S^\perp \cap \mathbb{R}_{++}^n$, which is therefore nonempty.

To prove part 2), again we note that there cannot exist vectors $u \in \mathbb{R}_{++}^n$ and $v \in \mathbb{R}_+^n$ that are orthogonal. Hence, $S \cap \mathbb{R}_{++}^n$ and $S^\perp \cap \mathbb{R}_+^n$ cannot both be nonempty, so if $S^\perp \cap \mathbb{R}_+^n \ne \emptyset$ then $S \cap \mathbb{R}_{++}^n = \emptyset$.

To prove that

$$S \cap \mathbb{R}_{++}^n = \emptyset \Rightarrow S^\perp \cap \mathbb{R}_+^n \ne \emptyset$$

note first that a subspace contains a strictly positive vector N if and only if it contains a strictly positive vector whose coordinates sum to 1.

Let $\mathcal{B} = \{B_1, \dots, B_k\}$ be a basis for S and consider the matrix

$$M = (m_{i,j}) = (B_1 \mid B_2 \mid \cdots \mid B_k)$$

whose columns are the basis vectors in $\mathcal{B}$. Let the rows of M be denoted by $R_1, \dots, R_n$. Note that $R_i \in \mathbb{R}^k$, where $k = \dim(S)$.

Now, $S^\perp$ contains a strictly positive vector $N = (a_1, \dots, a_n)$ if and only if

$$a_1R_1 + \cdots + a_nR_n = 0$$

for coefficients $a_i \ge 0$ satisfying $\Sigma a_i = 1$, that is, if and only if 0 is contained in the convex hull C of the vectors $R_1, \dots, R_n$ in $\mathbb{R}^k$. Hence,

$$S^\perp \cap \mathbb{R}_+^n \ne \emptyset \text{ if and only if } 0 \in C$$

Thus, we wish to prove that

$$S \cap \mathbb{R}_{++}^n = \emptyset \Rightarrow 0 \in C$$

or, equivalently,

$$0 \neq C \Rightarrow S \cap \mathbb{R}^n_{++} \neq \emptyset$$

Now we have something to separate. Since C is closed and convex, it follows from Theorem 15.5 that there is a nonzero vector $B = (b_1, \dots, b_k) \in \mathbb{R}^k$ for which

$$\langle B, x \rangle \geq \|B\|^2 > 0$$

for all $x \in C$. Consider the vector

$$v = b_1 B_1 + \cdots + b_k B_k \in S$$

The ith coordinate of v is

$$b_1 m_{i,1} + \cdots + b_k m_{i,k} = \langle B, R_i \rangle \geq \|B\|^2 > 0$$

and so v is strongly positive. Hence, $v \in S \cap \mathbb{R}^n_{++}$ and so this set is nonempty. This completes the proof. □

Nonhomogeneous Systems

We now turn our attention to nonhomogeneous systems

$$Ax = b$$

The following lemma is required.

Lemma 15.8 *Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$. Then the set*

$$C = \{Ay \mid y \in \mathbb{R}^n, y \geq 0\}$$

is a closed, convex cone.
Proof. We leave it as an exercise to prove that C is a convex cone and omit the proof that C is closed. □

Theorem 15.9 *(***Farkas's lemma***) Let $A \in \mathcal{M}_{m,n}(\mathbb{R})$ and let $b \in \mathbb{R}^m$ be nonzero. Then exactly one of the following holds:*
1) There is a strictly positive solution $u \in \mathbb{R}^n$ to the system $Ax = b$.
2) There is a vector $v \in \mathbb{R}^m$ for which $vA \leq 0$ and $\langle v, b \rangle > 0$.
Proof. Suppose first that 1) holds. If 2) also holds, then

$$(vA)u = v(Au) = \langle v, b \rangle > 0$$

However, $vA \leq 0$ and $u > 0$ imply that $(vA)u \leq 0$. This contradiction implies that 2) cannot hold.

Assume now that 1) fails to hold. By Lemma 15.8, the set

$$C = \{Ay \mid y \in \mathbb{R}^n, y \geq 0\} \subseteq \mathbb{R}^m$$

is closed and convex. The fact that 1) fails to hold is equivalent to $b \notin C$.

Hence, there is a hyperplane that strongly separates b and C. All we require is that b and C be strictly separated, that is, for some $\alpha \in \mathbb{R}$ and $v \in \mathbb{R}^m$

$$\langle v, x\rangle < \alpha < \langle v, b\rangle \text{ for all } x \in C$$

Since $0 \in C$ it follows that $\alpha > 0$ and so $\langle v, b\rangle > 0$. Also, the first inequality is equivalent to $\langle v, Ay\rangle < \alpha$, that is,

$$\langle A^t v, y\rangle < \alpha$$

for all $y \in \mathbb{R}^n, y \geq 0$. We claim that this implies that $A^t v$ cannot have any positive coordinates and thus $vA \leq 0$. For if the ith coordinate $(A^t v)_i$ is positive, then taking $y = \lambda e_i$ for $\lambda > 0$ we get

$$\lambda (A^t v)_i < \alpha$$

which does not hold for large λ. Thus, 2) holds. □

In the exercises, we ask the reader to show that the previous result cannot be improved by replacing $vA \leq 0$ in statement 2) with $vA \ll 0$.

Exercises

1. If A is an $m \times n$ matrix prove that the set $\{Ax \mid x \in \mathbb{R}^n, x > 0\}$ is a convex cone in $\mathbb{R}^m$.
2. If A and B are strictly separated subsets of $\mathbb{R}^n$ and if A is finite, prove that A and B are strongly separated as well.
3. Let V be a vector space over a field F with $\operatorname{char}(F) \neq 2$. Show that a subset X of V is closed under the taking of convex combinations of any two of its points if and only if X is closed under the taking of arbitrary convex combinations, that is, for all $n \geq 1$

$$x_1, \ldots, x_n \in X,\ \sum_{i=1}^{n} r_i = 1, 0 \leq r_i \leq 1 \Rightarrow \sum_{i=1}^{n} r_i x_i \in X$$

4. Explain why an $(n-1)$-dimensional subspace of $\mathbb{R}^n$ is the solution set of a linear equation of the form $a_1 x_1 + \cdots + a_n x_n = 0$.
5. Show that

$$\mathcal{H}_+(N, b) \cap \mathcal{H}_-(N, b) = \mathcal{H}(N, b)$$

and that $\mathcal{H}_+^\circ(N, b)$, $\mathcal{H}_-^\circ(N, b)$ and $\mathcal{H}(N, b)$ are pairwise disjoint and

$$\mathcal{H}_+^\circ(N, b) \cup\ \mathcal{H}_-^\circ(N, b) \cup \mathcal{H}(N, b) = \mathbb{R}^n$$

6. A function $T\colon \mathbb{R}^n \to \mathbb{R}^m$ is **affine** if it has the form $T(v) = \tau(v) + b$ for $b \in \mathbb{R}^m$, where $\tau \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)$. Prove that if $C \subseteq \mathbb{R}^n$ is convex then so is $T(C) \subseteq \mathbb{R}^m$.

7. Find a cone in $\mathbb{R}^2$ that is not convex. Prove that a subset X of $\mathbb{R}^n$ is a convex cone if and only if $x, y \in X$ implies that $\lambda x + \mu y \in X$ for all $\lambda, \mu \geq 0$.
8. Prove that the convex hull of a set $\{x_1, \ldots, x_n\}$ in $\mathbb{R}^n$ is bounded, without using the fact that it is compact.
9. Suppose that a vector $x \in \mathbb{R}^n$ has two distinct representations as convex combinations of the vectors $v_1, \ldots, v_n$. Prove that the vectors $v_2 - v_1, \ldots, v_n - v_1$ are linearly dependent.
10. Suppose that C is a nonempty convex subset of $\mathbb{R}^n$ and that $\mathcal{H}(N, b)$ is a hyperplane disjoint from C. Prove that C lies in one of the open half-spaces determined by $\mathcal{H}(N, b)$.
11. Prove that the conclusion of Theorem 15.6 may fail if we assume only that C is closed and convex.
12. Find two nonempty convex subsets of $\mathbb{R}^2$ that are strictly separated but not strongly separated.
13. Prove that X and Y are strongly separated by $\mathcal{H}(N, b)$ if and only if

$$\langle N, x' \rangle > b \text{ for all } x' \in X_\epsilon \text{ and } \langle N, y' \rangle < b \text{ for all } y' \in Y_\epsilon$$

where $X_\epsilon = X + \epsilon B(0,1)$ and $Y_\epsilon = Y + \epsilon B(0,1)$ and where $B(0,1)$ is the closed unit ball.
14. Show that Farkas's lemma cannot be improved by replacing $vA \leq 0$ in statement 2) with $vA \ll 0$. *Hint*: A nice counterexample exists for $m = 2, n = 3$.

Chapter 16
Affine Geometry

In this chapter, we will study the geometry of a finite-dimensional vector space V, along with its structure-preserving maps. *Throughout this chapter, all vector spaces are assumed to be finite-dimensional.*

Affine Geometry

The cosets of a quotient space have a special geometric name.

Definition *Let S be a subspace of a vector space V. The coset*

$$v+S=\{v+s\mid s\in S\}$$

is called a **flat** *in V with* **base** *S. We also refer to $v+S$ as a* **translate** *of S. The set $\mathcal{A}(V)$ of all flats in V is called the* **affine geometry** *of V. The* **dimension** $\dim(\mathcal{A}(V))$ *of $\mathcal{A}(V)$ is defined to be* $\dim(V)$. □

Here are some simple yet useful observations about flats.

1) A flat $X=x+S$ is a subspace if and only if $x=0$, that is, if and only if $X=S$.
2) A subset X is a flat if and only if for any $x\in X$ the translate $S=-x+X$ is a subspace.
3) If $X=x+S$ is a flat and $0\in z+X$ then $z+X=S$.

Definition *Two flats $x=x+S$ and $Y=y+T$ are said to be* **parallel** *if $S\subseteq T$ or $T\subseteq S$. This is denoted by $X\parallel Y$.* □

We will denote subspaces of V by the letters $S,T,\ldots$ and flats in V by $X,Y,\ldots$.

Here are some of the basic intersection properties of flats.

Theorem 16.1 Let S and T be subspaces of V and let $X = x + S$ and $Y = y + T$ be flats in V.

1) *The following are equivalent:*
   a) $x + S = y + S$
   b) $x \in y + S$
   c) $x - y \in S$
2) *The following are equivalent:*
   a) $w + X \subseteq Y$ *for some* $w \in V$
   b) $v + S \subseteq T$ *for some* $v \in V$
   c) $S \subseteq T$
3) *The following are equivalent:*
   a) $w + X = Y$ *for some* $w \in V$
   b) $v + S = T$ *for some* $v \in V$
   c) $S = T$
4) $X \cap Y \neq \emptyset,\ S \subseteq T \Leftrightarrow X \subseteq Y$
5) $X \cap Y \neq \emptyset,\ S = T \Leftrightarrow X = Y$
6) *If* $X \parallel Y$ *then* $X \subseteq Y$, $Y \subseteq X$ *or* $X \cap Y = \emptyset$
7) $X \parallel Y$ *if and only if some translation of one of these flats is contained in the other.*

Proof. We leave proof of part 1) for the reader. To prove 2), if 2a) holds then $-y + w + x + S \subseteq T$ and so 2b) holds. Conversely, if 2b) holds then

$$y + v - x + (x + S) \subseteq y + T = Y$$

and so 2a) holds. Now, if 2b) holds then $v = v + 0 \in T$ and so $S \subseteq -v + T \subseteq T$, which is 2c). Finally, if $S \subseteq T$ then just take $v = 0$ to get 2b). Part 3) is proved in a similar manner.

For part 4), we know that $w + X \subseteq Y$ for some $w \in V$. However, if $z \in X \cap Y$ then $w + z \in Y$ and so $w \in -z + Y = Y$, whence $X \subseteq Y$. Part 5) follows similarly. We leave proof of 6) and 7) to the reader. □

Part 1) of the previous theorem says that, in general, a flat can be represented in many ways as a translate of the base S. If $X = x + S$, then x is called a **flat representative**, or **coset representative** of X. Any element of a flat is a flat representative.

On the other hand, if $x + S = y + T$ then $x - y + S = T$ and the previous theorem tells us that $S = T$. Thus, the base of a flat is uniquely determined by the flat and we can make the following definition.

Definition *The* **dimension** *of a flat* $x + S$ *is* $\dim(S)$. *A flat of dimension* k *is called a* k**-flat**. *A* 0*-flat is a* **point**, *a* 1*-flat is a* **line** *and a 2-flat is a* **plane**. *A flat of dimension* $\dim(\mathcal{A}(V)) - 1$ *is called a* **hyperplane**. □

Affine Combinations

If $r_i \in F$ and $r_1 + \cdots + r_n = 1$ then the linear combination

$$r_1x_1 + \cdots + r_nx_n$$

is referred to as an **affine combination** of the vectors $x_1, \ldots, x_n$.

Our immediate goal is to show that, while the subspaces of V are precisely the subsets of V that are closed under the taking of linear combinations, the flats of V are precisely the subsets of V that are closed under the taking of affine combinations.

First, we need the following.

Theorem 16.2 *If* char$(F) \neq 2$*, then the following are equivalent for a subset X of V.*

1) *X is closed under the taking of affine combinations of any two of its points, that is,*
$$x, y \in X \Rightarrow rx + (1-r)y \in X$$
2) *X is closed under the taking of arbitrary affine combinations, that is,*
$$x_1, \ldots, x_n \in X,\ r_1 + \cdots + r_n = 1 \Rightarrow r_1x_1 + \cdots + r_nx_n \in X$$

Proof. It is clear that 2) implies 1). For the converse, we proceed by induction on $n \geq 2$. Part 1) is the case $n = 2$. Assume the result true for $n-1$ and consider the affine combination

$$z = r_1x_1 + \cdots + r_nx_n$$

If one of r_1 or r_2 is different from 1, say $r_1 \neq 1$ then we may write

$$z = r_1x_1 + (1-r_1)\Big(\frac{r_2}{1-r_1}x_2 + \cdots + \frac{r_n}{1-r_1}x_n\Big)$$

and since the sum of the coefficients inside the large parentheses is 1, the induction hypothesis implies that this sum is in X. Then 1) shows that $z \in X$. On the other hand, if $r_1 = r_2 = 1$ then since char$(F) \neq 2$, we may write

$$z = 2\left[\frac{1}{2}x_1 + \frac{1}{2}x_2\right] + r_3x_3 + \cdots + r_nx_n$$

and since 1) implies that $(1/2)x_1 + (1/2)x_2 \in X$, we may again deduce from the induction hypothesis that $z \in X$. In any case, $z \in X$ and so 2) holds. □

Note that the requirement char$(F) \neq 2$ is necessary, for if $F = \mathbb{Z}_2$ then the subset

$$X = \{(0,0),\ (1,0),\ (0,1)\}$$

satisfies condition 1) but not condition 2). We can now characterize flats.

Theorem 16.3

1) *A subset X of V is a flat in V if and only if it is closed under the taking of affine combinations, that is, if and only if*

$$x_1,\dots,x_n \in X,\ r_1+\cdots+r_n = 1 \Rightarrow r_1x_1+\cdots+r_nx_n \in X$$

2) *If* $\operatorname{char}(F) \neq 2$*, a subset X of V is a flat if and only if x contains the line through any two of its points, that is, if and only if*

$$x,y \in X \Rightarrow rx+(1-r)y \in X$$

Proof. Suppose that $X = x+S$ is a flat and $x_1,\dots,x_n \in X$. Then $x_i = x+s_i$, for $s_i \in S$ and so if $\Sigma r_i = 1$, we have

$$\sum_i r_ix_i = \sum_i r_i(x+s_i) = x + \sum_i r_is_i \in x+S$$

and so X is closed under affine combinations. Conversely, suppose that X is closed under the taking of affine combinations. It is sufficient (and necessary) to show that for a given $x_0 \in X$, the set $S = -x_0 + X$ is a subspace of V. But if

$$-x_0+x_1, -x_0+x_2 \in S$$

then for any $r_1, r_2 \in F$

$$\begin{aligned} r_1(-x_0+x_1)+r_2(-x_0+x_2) &= -(r_1+r_2)x_0 + r_1x_1 + r_2x_2 \\ &= -x_0 + [(1-r_1-r_2)x_0 + r_1x_1 + r_2x_2] \\ &\in -x_0 + X \end{aligned}$$

Hence, S is a subspace of V. Part 2) follows from part 1) and Theorem 16.2. □

Affine Hulls

The following definition gives the analog of the subspace spanned by a collection of vectors.

Definition *Let C be a nonempty set of vectors in V. The* **affine hull** *of C, denoted by* $\operatorname{hull}(C)$*, is the smallest flat containing C.* □

Theorem 16.4 The affine hull of a nonempty subset C of V is the **affine span** of C, that is, the set of all affine combinations of vectors in C

$$\operatorname{hull}(C) = \left\{ \sum_{i=1}^n r_ix_i \,\middle|\, n \geq 1,\ x_1,\dots,x_n \in C,\ \sum_{i=1}^n r_i = 1 \right\}$$

Proof. According to Theorem 16.3, any flat containing C must contain all affine combinations of vectors in C. It remains to show that the set X of all affine combinations of C is a flat, or equivalently, that for any $y \in X$, the set

$$S = X - y$$

is a subspace of V. To this end, let

$$y = \sum_{i=1}^{n} r_{0,i} x_i, \quad y_1 = \sum_{i=1}^{n} r_{1,i} x_i \quad \text{and} \quad y_2 = \sum_{i=1}^{n} r_{2,i} x_i$$

where $x_i \in C$ and $\Sigma r_{0,i} = \Sigma r_{1,i} = \Sigma r_{2,i} = 1$. Hence, any linear combination of $y_1 - y$ and $y_2 - y$ has the form

$$\begin{aligned} z &= s(y_1 - y) + t(y_2 - y) \\ &= s\sum_{i=1}^{n} r_{1,i} x_i + t\sum_{i=1}^{n} r_{2,i} x_i - (s+t)y \\ &= \sum_{i=1}^{n} (sr_{1,i} + tr_{2,i}) x_i - (s+t-1)y - y \\ &= \sum_{i=1}^{n} \Big(sr_{1,i} + tr_{2,i} - (s+t-1) r_{0,i} \Big) x_i - y \end{aligned}$$

But, since $\Sigma r_{0,i} = \Sigma r_{1,i} = \Sigma r_{2,i} = 1$, the sum of the coefficients of x_i is equal to 1 and so the sum is an affine sum, which shows that $z \in S$. Hence, S is a subspace of V. □

The affine hull of a finite set of vectors is denoted by hull$\{x_1, \ldots, x_n\}$. We leave it as an exercise to show that for any i

$$\begin{aligned} &\text{hull}\{x_1, \ldots, x_n\} \\ &\quad = x_i + \langle x_1 - x_i, \ldots, x_{i-1} - x_i, x_{i+1} - x_i, \ldots, x_n - x_i \rangle \end{aligned} \tag{16.1}$$

where $\langle\rangle$ denotes the subspace spanned by the vectors within the brackets. It follows that

$$\dim(\text{hull}\{x_1, \ldots, x_n\}) \le n - 1$$

The affine hull of a pair of distinct points is the line through those points, denoted by

$$\overline{xy} = \{rx + (1-r)y \mid r \in F\} = y + \langle x - y \rangle$$

The Lattice of Flats

Since flats are subsets of V, they are partially ordered by set inclusion.

Theorem 16.5 *The intersection of a nonempty collection* $\mathcal{C} = \{x_i + S_i \mid i \in K\}$ *of flats in* V *is either empty or is a flat. If the intersection is nonempty, then*

$$\bigcap_{i\in K}(x_i + S_i) = x + \bigcap_{i\in K} S_i$$

for any vector x *in the intersection. In other words, the base of the intersection is the intersection of the bases.*
Proof. If

$$x \in \bigcap_{i\in K}(x_i + S_i)$$

then $x_i + S_i = x + S_i$ for all $i \in K$ and so

$$\bigcap_{i\in K}(x_i + S_i) = \bigcap_{i\in K}(x + S_i) = x + \bigcap_{i\in K} S_i$$ □

Definition *The* **join** *of a nonempty collection* $\mathcal{C} = \{x_i + S_i \mid i \in K\}$ *of flats in* V *is the smallest flat containing all flats in* $\mathcal{C}$*. We denote the join of the collection* $\mathcal{C}$ *of flats by* $\bigvee\mathcal{C}$*, or by*

$$\bigvee_{i\in K}(x_i + S_i)$$

The join of two flats is written $(x + S) \vee (y + T)$*.* □

Theorem 16.6 *Let* $\mathcal{C} = \{x_i + S_i \mid i \in K\}$ *be a nonempty collection of flats in the vector space* V*.*
1) $\bigvee\mathcal{C}$ *is the intersection of all flats that contain all flats in* $\mathcal{C}$*.*
2) $\bigvee\mathcal{C}$ *is* hull$(\bigcup\mathcal{C})$*, where* $\bigcup\mathcal{C}$ *is the union of all flats in* $\mathcal{C}$*.* □

Theorem 16.7 *Let* $X = x + S$ *and* $Y = y + T$ *be flats in* V*. Then*
1)

$$X \vee Y = x + (\langle x - y\rangle + S + T)$$

2) If $X \cap Y \neq \emptyset$ *then*

$$X \vee Y = x + (S + T)$$

Proof. For part 1), let

$$X \vee Y = (x + S) \vee (y + T) = z + U$$

for some $z \in V$ and subspace U of V. Of course, $S + T \subseteq U$. But since $x, y \in z + U$, we also have $x - y \in U$. (Note that $x - y$ is not necessarily in $S + T$.)

Let $W = \langle x - y\rangle + S + T$. Then $W \subseteq U$ and so $x + W \subseteq x + U = X \vee Y$. On the other hand

$$X = x + S \subseteq x + W$$

and

$$Y = y + T = x - (x - y) + T \subseteq x + W$$

and so $X \vee Y \subseteq x + W$. Thus, $X \vee Y = x + W$, as desired.

For part 2), if $X \cap Y \neq \emptyset$ then we may take the flat representatives for X and Y to be any element $z \in X \cap Y$, in which case part 1) gives

$$X \vee Y = z + (\langle z - z \rangle + S + T) = z + S + T$$

and since $x \in X \vee Y$ we also have $X \vee Y = x + S + T$. □

We can now describe the dimension of the join of two flats.

Theorem 16.8 *Let $X = x + S$ and $Y = y + T$ be flats in V.*
1) If $X \cap Y \neq \emptyset$ then

$$\dim(X \vee Y) = \dim(S + T) = \dim(X) + \dim(Y) - \dim(X \cap Y)$$

2) If $X \cap Y = \emptyset$ then

$$\dim(X \vee Y) = \dim(S + T) + 1$$

Proof. According to Theorem 16.7, if $X \cap Y \neq \emptyset$ then

$$X \vee Y = x + S + T$$

and so by definition of the dimension of a flat

$$\dim(X \vee Y) = \dim(S + T)$$

On the other hand, if $X \cap Y = \emptyset$ then

$$X \vee Y = x + \langle x - y \rangle + S + T$$

and since $\dim(\langle x - y \rangle) = 1$, we get

$$\dim(X \vee Y) = \dim(S + T) + 1$$

Finally, we have

$$\dim(S + T) = \dim(S) + \dim(T) - \dim(S \cap T)$$

Therefore, if $z \in (x + S) \cap (y + T)$ then $x + S = z + S$ and $y + T = z + T$ and so

$$\begin{aligned} \dim(S \cap T) &= \dim(z + [S \cap T]) \\ &= \dim([z + S] \cap [z + T]) \\ &= \dim(X \cap Y) \end{aligned}$$

□

Affine Independence

We now discuss the affine counterpart of linear independence.

Theorem 16.9 *Let $X = \{x_1, \ldots, x_n\}$ be a nonempty set of vectors in V. The following are equivalent:*

1) *$H = \operatorname{hull}\{x_1, \ldots, x_n\}$ has dimension $n - 1$.*
2) *The set*
$$\{x_1 - x_i, \ldots, x_{i-1} - x_1, x_{i+1} - x_i, \ldots, x_n - x_i\}$$
is linearly independent for all $i = 1, \ldots, n$.
3) *$x_i \notin \operatorname{hull}\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ for all $i = 1, \ldots, n$.*
4) *If $\Sigma r_j x_j$ and $\Sigma s_j x_j$ are affine combinations then*
$$\sum_j r_j x_j = \sum_j s_j x_j \Rightarrow r_j = s_j \textit{ for all } j$$

A set $X = \{x_1, \ldots, x_n\}$ of vectors satisfying any (any hence all) of these conditions is said to be **affinely independent**.

Proof. The fact that 1) and 2) are equivalent follows directly from (16.1). If 3) does not hold, we have

$$\operatorname{hull}\{x_1, \ldots, x_n\} = \operatorname{hull}\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$$

where by (16.1), the latter has dimension at most $n - 2$. Hence, 1) cannot hold and so 1) implies 3).

Next we show that 3) implies 4). Suppose that 3) holds and that $\Sigma r_j x_j = \Sigma s_j x_j$. Setting $t_j = r_j - s_j$ gives

$$\sum_j t_j x_j = 0 \text{ and } \sum_j t_j = 0$$

But if any of the t_j's are nonzero, say $t_1 \neq 0$ then dividing by t_1 gives

$$x_1 + \sum_{j>1} (t_j/t_1) x_j = 0$$

or

$$x_1 = -\sum_{j>1} (t_j/t_1) x_j$$

where

$$-\sum_{j>1} (t_j/t_1) = 1$$

Hence, $x_1 \in \text{hull}\{x_2, \ldots, x_n\}$. This contradiction implies that $t_j = 0$ for all j, that is, $r_j = s_j$ for all j. Thus, 3) implies 4).

Finally, we show that 4) implies 2). For concreteness, let us show that 4) implies that $\{x_2 - x_1, \ldots, x_n - x_1\}$ is linearly independent. Indeed, if $\alpha_2, \ldots, \alpha_n \in F$ and $\Sigma\alpha_j = \alpha$ then

$$\sum_{j\geq 2}\alpha_j(x_j - x_1) = 0 \Rightarrow \sum_{j\geq 2}\alpha_j x_j = \alpha x_1 \Rightarrow (1-\alpha)x_1 + \sum_{j\geq 2}\alpha_j x_j = x_1$$

But the latter is an equality between two affine combinations and so corresponding coefficients must be equal, which implies that $\alpha_j = 0$ for all $j = 2, \ldots, n$. This shows that 4) implies 2). □

Affinely independent sets enjoy some of the basic properties of linearly independent sets. For example, a nonempty subset of an affinely independent set is affinely independent. Also, any nonempty set X contains an affinely independent set.

Since the affine hull $H = \text{hull}(X)$ of an affinely independent set X is not the affine hull of any proper subset of X, we deduce that X is a minimal affine spanning set of its affine hull.

Note that if $H = \text{hull}(X)$ where $X = \{x_1, \ldots, x_n\}$ is affinely independent then for any i, the set

$$\{x_1 - x_i, \ldots, x_{i-1} - x_1, x_{i+1} - x_i, \ldots, x_n - x_i\}$$

is a basis for the base subspace of H. Conversely, if $H = x + S \neq S$ and $B = \{b_1, \ldots, b_n\}$ is a basis for S then

$$B' = \{x, x + b_1, \ldots, x + b_n\}$$

is affinely independent and since $\text{hull}(B')$ has dimension n and is contained in $x + S$ we must have $B' = H$. This provides a way to go between "affine bases" B' of a flat and linear bases B of the base subspace of the flat.

Theorem 16.10 *If X is a flat of dimension n then there exist $n+1$ vectors $x_1, \ldots, x_{n+1}$ for which every vector $x \in X$ has a* unique *expression as an affine combination*

$$x = r_1 x_1 + \cdots + r_{n+1}x_{n+1}$$

The coefficients r_i are called the **barycentric coordinates** *of x with respect to the vectors $x_1, \ldots, x_{n+1}$.* □

Affine Transformations

Now let us discuss some properties of maps that preserve affine structure.

Definition *A function $f\colon V \to V$ that preserves affine combinations, that is, for which*

$$\sum_i r_i = 1 \Rightarrow f\left(\sum_i r_i x_i\right) = \sum_i r_i f(x_i)$$

is called an **affine transformation** *(or* **affine map**, *or* **affinity**). □

We should mention that some authors require that f be bijective in order to be an affine map. The following theorem is the analog of Theorem 16.2.

Theorem 16.11 *If* $\text{char}(F) \neq 2$ *then a function $f\colon V \to V$ is an affine transformation if and only if it preserves affine combinations of any two of its points, that is, if and only if*

$$f(rx + (1-r)y) = rf(x) + (1-r)f(y) \qquad \square$$

Thus, if $\text{char}(F) \neq 2$ then a map f is an affine transformation if and only if it sends the line through x and y to the line through $f(x)$ and $f(y)$. It is clear that linear transformations are affine transformations. So are the following maps.

Definition *Let $v \in V$. The affine map $T_v\colon V \to V$ defined by*

$$T_v(x) = x + v$$

for all $x \in V$, is called **translation** *by v.* □

It is not hard to see that any map of the form "linear operator followed by translation," that is, $T_v \circ \tau$, where $\tau \in \mathcal{L}(V)$, is affine. Conversely, any affine map must have this form.

Theorem 16.12 *A function $f\colon V \to V$ is an affine transformation if and only if it is a linear operator followed by a translation,*

$$f = T_v \circ \tau$$

where $v \in V$ and $\tau \in \mathcal{L}(V)$.

Proof. We leave proof that $T_v \circ \tau$ is an affine transformation to the reader. Conversely, suppose that f is an affine map. If we expect f to have the form $T_v \circ \tau$ then $f(0)$ will equal $T_v \circ \tau(0) = T_v(0) = v$. So let $v = f(0)$. We must show that $\tau = T_{-f(0)} \circ f$ is a linear operator on V. However, for any $x \in V$

$$\tau(x) = f(x) - f(0)$$

and so

$$\begin{aligned}\tau(ru+sv) &= f(ru+sv)-f(0)\\ &= f(ru+sv+(1-r-s)0)-f(0)\\ &= rf(u)+sf(v)+(1-r-s)f(0)-f(0)\\ &= r\tau(u)+s\tau(v)\end{aligned}$$

Thus, τ is linear. □

Corollary 16.13

1) *The composition of two affine transformations is an affine transformation.*
2) *An affine transformation $f = T_v \circ \tau$ is bijective if and only if τ is bijective.*
3) *The set* aff(V) *of all bijective affine transformations on V is a group under composition of maps, called the* **affine group** *of V.* □

Let us make a few group-theoretic remarks about aff(V). The set trans(V) of all translations of V is a subgroup of aff(V). We can define a function ϕ: aff$(V) \to \mathcal{L}(V)$ by

$$\phi(T_v \circ \tau) = \tau$$

It is not hard to see that ϕ is a well-defined group homomorphism from aff(V) onto $\mathcal{L}(V)$, with kernel trans(V). Hence, trans(V) is a normal subgroup of aff(V) and

$$\frac{\text{aff}(V)}{\text{trans}(V)} \approx \mathcal{L}(V)$$

Projective Geometry

If $\dim(V) = 2$, the join (affine hull) of any two distinct points in V is a line. On the other hand, it is not the case that the intersection of any two lines is a point, since the lines may be parallel. Thus, there is a certain asymmetry between the concepts of points and lines in V. This asymmetry can be removed by constructing the *projective plane*. Our plan here is to very briefly describe one possible construction of projective geometries of all dimensions.

By way of motivation, let us consider Figure 16.1.

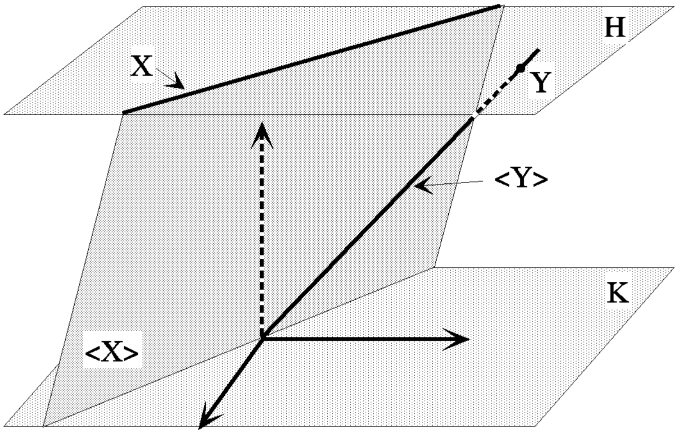

Figure 16.1

Note that H is a hyperplane in a 3-dimensional vector space V and that $0 \notin H$. Now, the set $\mathcal{A}(H)$ of all flats of V that lie in H is an affine geometry of dimension 2. (According to our definition of affine geometry, H must be a vector space in order to define $\mathcal{A}(H)$. However, we hereby extend the definition of affine geometry to include the collection of all flats contained in a flat of V.)

Figure 16.1 shows a one-dimensional flat X and its linear span $\langle X\rangle$, as well as a zero-dimensional flat Y and its span $\langle Y\rangle$. Note that, for any flat X in H, we have

$$\dim(\langle X\rangle) = \dim(X) + 1$$

Note also that if L_1 and L_2 are any two distinct lines in H, the corresponding planes $\langle L_1$ and $\langle L_2\rangle$ have the property that their intersection is a line through the origin, *even if the lines are parallel*. We are now ready to define projective geometries.

Let V be a vector space of any dimension and let H be a hyperplane H in V not containing the origin. To each flat X in H, we associate the subspace $\langle X\rangle$ of V generated by X. Thus, the linear span function from $P\colon \mathcal{A}(H) \to \mathcal{S}(V)$ maps affine subspaces X of H to subspaces $\langle X\rangle$ of V. The span function is not surjective: Its image is the set of all subspaces that are *not* contained in the base subspace K of the flat H.

The linear span function is one-to-one and its inverse is intersection with H

$$P^{-1}(U) = U \cap H$$

for any subspace U not contained in K.

The affine geometry $\mathcal{A}(H)$ is, as we have remarked, somewhat incomplete. In the case $\dim(H) = 2$ every pair of points determines a line but not every pair of lines determines a point.

Now, since the linear span function P is injective, we can identify $\mathcal{A}(H)$ with its image $P(\mathcal{A}(H))$, which is the set of all subspaces of V not contained in the base subspace K. This view of $\mathcal{A}(H)$ allows us to "complete" $\mathcal{A}(H)$ by including the base subspace K. In the three-dimensional case of Figure 16.1, the base plane, in effect, adds a projective line at infinity. With this inclusion, every pair of lines intersects, parallel lines intersecting at a point on the line at infinity. This two-dimensional projective geometry is called the **projective plane**.

Definition *Let V be a vector space. The set $\mathcal{S}(V)$ of all subspaces of V is called the* **projective geometry** *of V. The* **projective dimension** $\operatorname{pdim}(S)$ *of $S \in \mathcal{S}(V)$ is defined as*

$$\operatorname{pdim}(S) = \dim(S) - 1$$

The **projective dimension** *of $\mathcal{P}(V)$ is defined to be* $\operatorname{pdim}(V) = \dim(V) - 1$. *A subspace of projective dimension 0 is called a* **projective point** *and a subspace of projective dimension 1 is called a* **projective line**. □

Thus, referring to Figure 16.1, a projective point is a line through the origin and, provided that it is not contained in the base plane K, it meets H in an affine point. Similarly, a projective line is a plane through the origin and, provided that it is not K, it will meet H in an affine line. In short,

$$\begin{aligned}\text{span}(\text{affine point}) &= \text{line through the origin} = \text{projective point}\\ \text{span}(\text{affine line}) &= \text{plane through the origin} = \text{projective line}\end{aligned}$$

The linear span function has the following properties.

Theorem 16.14 *The linear span function $P\colon \mathcal{A}(H) \to \mathcal{S}(V)$ from the affine geometry $\mathcal{A}(H)$ to the projective geometry $\mathcal{S}(V)$ defined by $P(X) = \langle X \rangle$ satisfies the following properties:*

1) *The linear span function is injective, with inverse given by*

$$P^{-1}(U) = U \cap H$$

for all subspaces U not contained in the base subspace K of H.

2) *The image of the span function is the set of all subspaces of V that are not contained in the base subspace K of H.*
3) *$X \subseteq Y$ if and only if $\langle X \rangle \subseteq \langle Y \rangle$*
4) *If X_i are flats in H with nonempty intersection then*

$$\text{span}\Big(\bigcap_{i\in K} X_i\Big) = \bigcap_{i\in K} \text{span}(X_i)$$

5) *For any collection of flats in* H,

$$\operatorname{span}\left(\bigvee_{i\in K} X_i\right) = \bigoplus_{i\in K}\operatorname{span}(X_i)$$

6) *The linear span function preserves dimension, in the sense that*

$$\operatorname{pdim}(\operatorname{span}(X)) = \dim(X)$$

7) $X \parallel Y$ *if and only if one of* $\langle X\rangle \cap K$ *and* $\langle Y\rangle \cap K$ *is contained in the other.*

Proof. To prove part 1), let $x+S$ be a flat in H. Then $x\in H$ and so $H = x+K$, which implies that $S\subseteq K$. Note also that $\langle x+S\rangle = \langle x\rangle + S$ and

$$z\in\langle x+S\rangle\cap H = (\langle x\rangle + S)\cap(x+K) \Rightarrow z = rx+s = x+k$$

for some $s\in S$, $k\in K$ and $r\in F$. This implies that $(1-r)x\in K$, which implies that either $x\in K$ or $r=1$. But $x\in H$ implies $x\notin K$ and so $r=1$, which implies that $z = x+s\in x+S$. In other words,

$$\langle x+S\rangle\cap H\subseteq x+S$$

Since the reverse inclusion is clear, we have

$$\langle x+S\rangle\cap H = x+S$$

This establishes 1).

To prove 2), let U be a subspace of V that is not contained in K. We wish to show that U is in the image of the linear span function. Note first that since $U\nsubseteq K$ and $\dim(K) = \dim(V)-1$, we have $U+K=V$ and so

$$\dim(U\cap K) = \dim(U)+\dim(K)-\dim(U+K) = \dim(U)-1$$

Now, let $0\neq x\in U-K$. Then

$$\begin{aligned} x\notin K &\Rightarrow \langle x\rangle + K = V\\ &\Rightarrow rx+k\in H \text{ for some } 0\neq r\in F,\ k\in K\\ &\Rightarrow rx\in H\end{aligned}$$

Thus, $rx\in U\cap H$ for some $0\neq r\in F$. Hence, the flat $rx+(U\cap K)$ lies in H and

$$\dim(rx+(U\cap K)) = \dim(U\cap K) = \dim(U)-1$$

which implies that $\operatorname{span}(rx+(U\cap K)) = \langle rx\rangle + (U\cap K)$ lies in U and has the same dimension as U. In other words,

$$\operatorname{span}(rx+(U\cap K)) = \langle rx\rangle + (U\cap K) = U$$

We leave proof of the remaining parts of the theorem as exercises. □

Exercises

1. Show that if $x_1,\ldots,x_n\in V$ then the set $S=\{\Sigma r_i x_i \mid \Sigma r_i=0\}$ is a subspace of V.
2. Prove that hull$\{x_1,\ldots,x_n\}=x_1+\langle x_2-x_1,\ldots,x_n-x_1\rangle$.
3. Prove that the set $X=\{(0,0),\ (1,0),\ (0,1)\}$ in $(\mathbb{Z}_2)^2$ is closed under the formation of lines, but not affine hulls.
4. Prove that a flat contains the origin 0 if and only if it is a subspace.
5. Prove that a flat X is a subspace if and only if for some $x\in X$ we have $rx\in X$ for some $1\neq r\in F$.
6. Show that the join of a collection $\mathcal{C}=\{x_i+S_i \mid i\in K\}$ of flats in V is the intersection of all flats that contain all flats in $\mathcal{C}$.
7. Is the collection of all flats in V a lattice under set inclusion? If not, how can you "fix" this?
8. Suppose that $X=x+S$ and $Y=y+T$. Prove that if $\dim(X)=\dim(Y)$ and $X\parallel Y$ then $S=T$.
9. Suppose that $X=x+S$ and $Y=y+T$ are disjoint hyperplanes in V. Show that $S=T$.
10. (The parallel postulate) Let X be a flat in V and $v\notin X$. Show that there is exactly one flat containing v, parallel to X and having the same dimension as X.
11. a) Find an example to show that the join $X\vee Y$ of two flats may not be the set of all lines connecting all points in the union of these flats.
    b) Show that if X and Y are flats with $X\cap Y\neq\emptyset$ then $X\vee Y$ is the union of all lines $\overline{xy}$ where $x\in X$ and $y\in Y$.
12. Show that if $X\parallel Y$ and $X\cap Y=\emptyset$ then

$$\dim(X\vee Y)=\max\{\dim(X),\dim(Y)\}+1$$

13. Let $\dim(V)=2$. Prove the following:
    a) The join of any two distinct points is a line.
    b) The intersection of any two nonparallel lines is a point.
14. Let $\dim(V)=3$. Prove the following:
    a) The join of any two distinct points is a line.
    b) The intersection of any two nonparallel planes is a line.
    c) The join of any two lines whose intersection is a point is a plane.
    d) The intersection of two coplanar nonparallel lines is a point.
    e) The join of any two distinct parallel lines is a plane.
    f) The join of a line and a point not on that line is a plane.
    g) The intersection of a plane and a line not on that plane is a point.
15. Prove that $f\colon V\to V$ is a surjective affine transformation if and only if $f=\tau\circ T_w$ for some $w\in V$ and $\tau\in\mathcal{L}(V)$.
16. Verify the group-theoretic remarks about the group homomorphism ϕ: aff$(V)\to\mathcal{L}(V)$ and the subgroup trans(V) of aff(V).

Chapter 17
Operator Factorizations: QR and Singular Value

The QR Decomposition

Let V be a finite-dimensional inner product space over F, where $F = \mathbb{R}$ or $F = \mathbb{C}$. Let us recall a definition.

Definition *A linear operator τ on V is* **upper triangular** *with respect to an ordered basis $\mathcal{B} = (v_1, \ldots, v_n)$ if the matrix $[\tau]_{\mathcal{B}}$ is upper triangular, that is, if for all $i = 1, \ldots, n$*

$$\tau(v_i) \in \langle v_1, \ldots, v_i \rangle$$

The operator τ is **upper triangularizable** *if there is an ordered basis with respect to which τ is upper triangular.* □

Given any orthonormal basis $\mathcal{B}$ for V, it is possible to find a unitary operator ν for which $\nu\tau$ is upper triangular with respect to $\mathcal{B}$. In matrix terms, this is equivalent to the fact that any matrix A can be factored into a product $A = QR$ where Q is unitary (orthogonal) and R is upper triangular. This is the well known **QR factorization** of a matrix. Before proving this fact, let us repeat one more definition.

Definition *For a nonzero $v \in V$, the unique operator H_v for which*

$$H_v v = -v, \ (H_v)|_{\langle v \rangle^\perp} = \iota$$

is called a **reflection** *or a* **Householder transformation**. □

According to Theorem 10.11, if $\|v\| = \|w\| \neq 0$, then H_{v-w} is the unique reflection sending v to w, that is, $H_{v-w}(v) = w$.

Theorem 17.1 *(***QR-Factorization of an operator***) Let τ be a linear operator on a finite-dimensional real or complex vector space V. Then for any ordered orthonormal basis $\mathcal{B} = (u_1, \ldots, u_n)$ for V, there is a unitary (orthogonal) operator ν and an operator ρ that is upper triangular with respect to $\mathcal{B}$, that is,*

$$\rho(u_i) \in \langle u_1, \ldots, u_i \rangle$$

for all $i = 1, \ldots, n$, for which

$$\tau = \nu \circ \rho$$

Moreover, if τ is invertible, then ρ can be chosen with positive eigenvalues, in which case both ρ and ν are unique.

Proof. Let $b_i = \|\tau u_i\|$. If $x_1 = \tau u_1 - b_1 u_1$ then

$$(H_{x_1} \circ \tau)(u_1) = (H_{\tau u_1 - b_1 u_1} \circ \tau)(u_1) = b_1 u_1 \in \langle u_1 \rangle$$

where, if τ is invertible then b_1 is positive.

Assume for the purposes of induction that, for a given $1 \leq k < n$, we have found reflections $H_{x_1}, \ldots, H_{x_k}$ for which, setting $H^{(k)} = H_{x_k} \cdots H_{x_1}$

$$(H^{(k)} \circ \tau) u_i \in \langle u_1, \ldots, u_i \rangle$$

for all $i \leq k$. Assume also that if τ is invertible, the coefficient of u_i in $(H^{(k)} \circ \tau) u_i$ is positive.

We seek a reflection $H_{x_{k+1}}$ for which

$$(H_{x_{k+1}} \circ H^{(k)} \circ \tau) u_i \in \langle u_1, \ldots, u_i \rangle$$

for $i \leq k+1$ and for which, if τ is invertible, the coefficient of u_i in $(H_{x_{k+1}} \circ H^{(k)} \circ \tau) u_i$ is positive.

Note that if $x_{k+1} \in \langle u_{k+1}, \ldots, u_n \rangle$ then $H_{x_{k+1}}$ is the identity on $\langle u_1, \ldots, u_k \rangle$ and so, at least for $i \leq k$ we have

$$(H_{x_{k+1}} \circ H^{(k)} \circ \tau) u_i \in H_{x_{k+1}}(\langle u_1, \ldots, u_i \rangle) = \langle u_1, \ldots, u_i \rangle$$

as desired. But we also want to choose x_{k+1} so that

$$(H_{x_{k+1}} \circ H^{(k)} \circ \tau) u_{k+1} \in \langle u_1, \ldots, u_{k+1} \rangle$$

Let us write

$$(H^{(k)} \circ \tau) u_{k+1} = v + w$$

where $v \in \langle u_1, \ldots, u_k \rangle$ and $w \in \langle u_{k+1}, \ldots, u_n \rangle$. We can accomplish our goal by reflecting w onto the subspace $\langle u_{k+1} \rangle$. In particular, let $x_{k+1} = w - \|w\| u_{k+1}$.

Since $x_{k+1} \in \langle u_{k+1}, \ldots, u_n \rangle$, the operator $H_{x_{k+1}}$ is the identity on $\langle u_1, \ldots, u_k \rangle$ and so as noted earlier

$$(H_{x_{k+1}} \circ H^{(k)} \circ \tau)u_i \in \langle u_1, \ldots, u_i \rangle, \text{ for } i \leq k$$

Also

$$\begin{aligned}(H_{x_{k+1}} \circ H^{(k)} \circ \tau)u_{k+1} &= H_{w-\|w\|u_{k+1}}(v+w) \\ &= v + \|w\|u_{k+1} \in \langle u_1, \ldots, u_{k+1} \rangle\end{aligned}$$

and if τ is invertible, then $w \neq 0$ and so $\|w\| > 0$. Thus, we have found $H^{(k+1)}$ and by induction,

$$\rho = H_{x_n} \cdots H_{x_1} \tau$$

is upper triangular with respect to $\mathcal{B}$, which proves the first part of the theorem. It remains only to prove the uniqueness statement.

Suppose that τ is invertible and that $\tau = \nu_1 \rho_1 = \nu_2 \rho_2$ and that the coefficients of u_i in $\rho_1 u_i$ and $\rho_2 u_i$ are positive. Then $\mu = \nu_2^{-1} \nu_1 = \rho_2 \rho_1^{-1}$ is both unitary and upper triangular with respect to $\mathcal{B}$ and the coefficient of u_i in μu_i is positive. We leave it to the reader to show that μ must be the identity and so $\nu_1 = \nu_2$ and $\rho_1 = \rho_2$. □

Here is the matrix version of the preceding theorem.

Theorem 17.2 (**The QR factorization**) Any real or complex matrix A can be written in the form $A = QR$ where Q is unitary (orthogonal) and R is upper triangular. Moreover, if A is nonsingular then the diagonal entries of R may be taken to be positive, in which case the factorization is unique.

Proof. According to Theorem 17.1, there is a unitary (orthogonal) operator U for which $[U\tau_A]_{\mathcal{E}} = R$ is upper triangular. Hence

$$A = [\tau_A]_{\mathcal{E}} = [U^*]_{\mathcal{E}} R = QR$$

where Q is a unitary (orthogonal) matrix. □

The QR decomposition has important applications. For example, a system of linear equations $Ax = u$ can be written in the form

$$QRx = u$$

and since $Q^{-1} = Q^*$, we have

$$Rx = Q^* u$$

This is an upper triangular system, which is easily solved by *back substitution* that is, starting from the bottom and working up.

Singular Values

Let U and V be finite-dimensional inner product spaces over $\mathbb{C}$ or $\mathbb{R}$. The spectral theorem can be of considerable help in understanding the relationship between a linear transformation $\tau \in \mathcal{L}(U,V)$ and its adjoint $\tau^* \in \mathcal{L}(V,U)$. This relationship is shown in Figure 17.1. (We assume that U and V are finite-dimensional.)

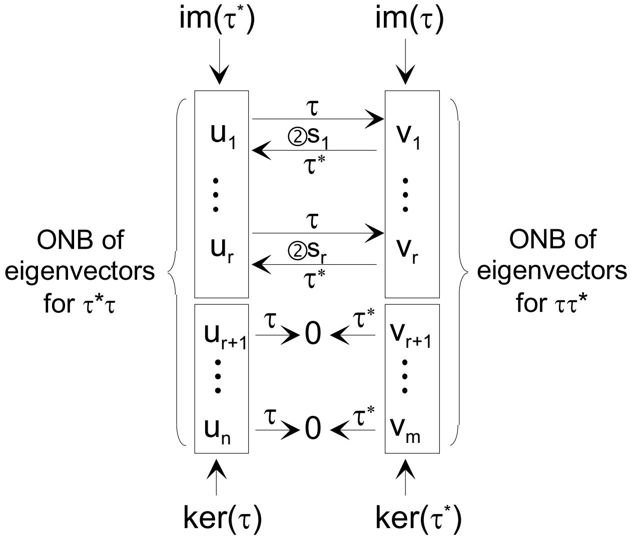

Figure 17.1

We begin with a simple observation: If $\tau \in \mathcal{L}(U,V)$ then $\tau^*\tau \in \mathcal{L}(U)$ is a positive Hermitian operator. Hence, if $r = \text{rk}(\tau) = \text{rk}(\tau^*\tau)$ then U has an ordered orthonormal basis $\mathcal{B} = (u_1, \ldots, u_r, u_{r+1}, \ldots, u_n)$ of eigenvectors for $\tau^*\tau$, where the corresponding (not necessarily unique) eigenvalues satisfy

$$\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_n$$

The numbers $s_i = +\sqrt{\lambda_i}$, for $i = 1, \ldots, r$ are called the **singular values** of τ and for $i = 1, \ldots, n$ we have

$$\tau^*\tau u_i = s_i^2 u_i$$

where $s_i = 0$ for $i > r$.

It is not hard to show that $(u_{r+1}, \ldots, u_n)$ is an ordered orthonormal basis for $\ker(\tau)$ and so $(u_1, \ldots, u_r)$ is an ordered orthonormal basis for $\ker(\tau)^\perp = \text{im}(\tau^*)$. For if $i > r$ then

$$\langle \tau u_i, \tau u_i \rangle = \langle \tau^*\tau u_i, u_i \rangle = 0$$

and so $\tau(u_i) = 0$, that is, $u_i \in \ker(\tau)$. On the other hand, if $x = \Sigma a_i u_i$ is in $\ker(\tau)$ then

$$0 = \langle \tau(x), \tau(x) \rangle = \Big\langle \sum_i a_i \tau(u_i), \sum_j a_j \tau(u_j) \Big\rangle = \sum_i |a_i^2| s_i^2$$

and so $a_i = 0$ for $i \le r$ and so $\ker(\tau) \subseteq \langle u_{r+1}, \ldots, u_n \rangle$.

We can achieve some "symmetry" here between τ and τ^* by setting $v_i = (1/s_i)\tau u_i$ for each $i \le r$, giving

$$\tau u_i = \begin{cases} s_i v_i & i \le r \\ 0 & i > r \end{cases}$$

and

$$\tau^* v_i = \begin{cases} s_i u_i & i \le r \\ 0 & i > r \end{cases}$$

The vectors $v_1, \ldots, v_r$ are orthonormal, since if $i, j \le r$ then

$$\langle v_i, v_j \rangle = \frac{1}{s_i s_j} \langle \tau u_i, \tau u_j \rangle = \frac{1}{s_i s_j} \langle \tau^* \tau u_i, u_j \rangle = \frac{s_i}{s_j} \langle u_i, u_j \rangle = \delta_{i,j}$$

Hence, $(v_1, \ldots, v_r)$ is an orthonormal basis for $\operatorname{im}(\tau) = \ker(\tau^*)^\perp$, which can be extended to an orthonormal basis $\mathcal{C} = (v_1, \ldots, v_m)$ for V, the extension $(v_{r+1}, \ldots, v_m)$ being an orthonormal basis for $\ker(\tau^*)$. The vectors u_i are called the **right singular vectors** for τ and the vectors v_i are called the **left singular vectors** for τ.

Moreover, since

$$\tau \tau^* v_i = s_i \tau u_i = s_i^2 v_i$$

the vectors $v_1, \ldots, v_r$ are eigenvectors for $\tau\tau^*$ with the same eigenvalues $\lambda_i = s_i^2$ as for $\tau^*\tau$. This completes the picture in Figure 17.1.

Theorem 17.3 *Let U and V be finite-dimensional inner product spaces over $\mathbb{C}$ or $\mathbb{R}$ and let $\tau \in \mathcal{L}(U, V)$ have rank r. Then there is an ordered orthonormal basis $\mathcal{B} = (u_1, \ldots, u_r, u_{r+1}, \ldots, u_n)$ of U and an ordered orthonormal basis $\mathcal{C} = (v_1, \ldots, v_r, v_{r+1}, \ldots, v_m)$ of V with the following properties:*

1) *$\mathcal{B}_r = (u_1, \ldots, u_r)$ is an orthonormal basis for* $\ker(\tau)^\perp = \operatorname{im}(\tau^*)$
2) *$(u_{r+1}, \ldots, u_n)$ is an orthonormal basis for* $\ker(\tau)$
3) *$\mathcal{C}_r = (v_1, \ldots, v_r)$ is an orthonormal basis for* $\ker(\tau^*)^\perp = \operatorname{im}(\tau)$
4) *$(v_{r+1}, \ldots, v_m)$ is an orthonormal basis for* $\ker(\tau^*)$
5) *The operators τ and τ^* behave "symmetrically" on $\mathcal{B}_r$ and $\mathcal{C}_r$, specifically, for $i \le r$,*

$$\tau(u_i) = s_i v_i$$
$$\tau^*(v_i) = s_i u_i$$

where $s_i > 0$ *are called the* **singular values** *of* τ.
The vectors u_i *are called the* **right singular vectors** *for* τ *and the vectors* v_i *are called the* **left singular vectors** *for* τ. □

The matrix version of the previous discussion leads to the well known **singular value decomposition** of a matrix. The matrix of τ under the ordered orthonormal bases $\mathcal{B} = (u_1, \ldots, u_n)$ and $\mathcal{C} = (v_1, \ldots, v_m)$ is

$$[\tau]_{\mathcal{B},\mathcal{C}} = \Sigma = \text{diag}(s_1, s_2, \ldots, s_r, 0, \ldots, 0)$$

Given any matrix $A \in \mathcal{M}_{m,n}$ of rank r, let $\tau = \tau_A$ be multiplication by A. Then $A = [\tau_A]_{\mathcal{E}_n,\mathcal{E}_m}$ where $\mathcal{E}_n$ and $\mathcal{E}_m$ are the standard bases for U and V, respectively. By changing orthonormal bases to $\mathcal{B}$ and $\mathcal{C}$ we get

$$A = [\tau_A]_{\mathcal{E}_n,\mathcal{E}_m} = M_{\mathcal{C},\mathcal{E}_m}[\tau_A]_{\mathcal{B},\mathcal{C}} M_{\mathcal{E}_n,\mathcal{B}} = P\Sigma Q^*$$

where $P = M_{\mathcal{C},\mathcal{E}_m}$ is unitary (orthogonal for $F = \mathbb{R}$) with ith column equal to $[v_i]_{\mathcal{E}_m}$ and $Q = M_{\mathcal{B},\mathcal{E}_n}$ is unitary (orthogonal for $F = \mathbb{R}$) with ith column equal to $[u_i]_{\mathcal{E}_n}$.

As to uniqueness, if $A = P\Sigma Q^*$ is a singular value decomposition then

$$A^*A = (P\Sigma Q^*)^* P\Sigma Q^* = Q\Sigma^*\Sigma Q^*$$

and since $\Sigma^*\Sigma = \text{diag}(s_1^2, s_2^2, \ldots, s_r^2, 0, \ldots, 0)$, it follows that s_i^2 is an eigenvalue of A^*A. Hence, since $s_i > 0$, we deduce that the singular values are uniquely determined by A.

We state without proof the following uniqueness facts. For a proof, the reader may wish to consult reference [HJ1]. If $n \le m$ and if the eigenvalues λ_i are distinct then P is uniquely determined up to multiplication on the right by a diagonal matrix of the form $D = \text{diag}(z_1, \ldots, z_m)$ with $|z_i| = 1$. If $n < m$ then Q is never uniquely determined. If $m = n = r$ then for any given P there is a unique Q. Thus, we see that, in general, the singular value decomposition is not unique.

The Moore–Penrose Generalized Inverse

Singular values lead to a generalization of the inverse of an operator that applies to all linear transformations. The setup is the same as in Figure 17.1. Referring to that figure, we are prompted to define a linear transformation $\tau^+ \colon V \to U$ by

$$\tau^+ v_i = \begin{cases} \frac{1}{s_i} u_i & \text{for } i \le r \\ 0 & \text{for } i > r \end{cases}$$

for then

$$\begin{aligned}(\tau^+\tau)|_{\langle u_1,\ldots,u_r\rangle} &= \iota \\ (\tau^+\tau)|_{\langle u_{r+1},\ldots,u_n\rangle} &= 0\end{aligned}$$

and

$$\begin{aligned}(\tau\tau^+)|_{\langle v_1,\ldots,v_r\rangle} &= \iota \\ (\tau\tau^+)|_{\langle v_{r+1},\ldots,v_m\rangle} &= 0\end{aligned}$$

Hence, if $n = m = r$ then $\tau^+ = \tau^{-1}$. The transformation τ^+ is called the **Moore–Penrose generalized inverse** or **Moore–Penrose pseudoinverse** of τ. We abbreviate this as **MP inverse**.

Note that the composition $\tau^+\tau$ is the identity on the largest possible subspace of U upon which any composition of the form $\sigma\tau$ could be the identity, namely, the orthogonal complement of the kernel of τ. A similar statement holds for the composition $\tau\tau^+$. Hence, τ^+ is as "close" to an inverse for τ as is possible.

We have said that if τ is invertible then $\tau^+ = \tau^{-1}$. More is true: If τ is injective then $\tau^+\tau = \iota$ and so τ^+ is a left inverse for τ. Also, if τ is surjective then τ^+ is a right inverse for τ. Hence the MP inverse τ^+ generalizes the one-sided inverses as well.

Here is a characterization of the MP inverse.

Theorem 17.4 *Let $\tau \in \mathcal{L}(U,V)$. The MP inverse τ^+ of τ is completely characterized by the following four properties:*

1) $\tau\tau^+\tau = \tau$

2) $\tau^+\tau\tau^+ = \tau^+$

3) $\tau\tau^+$ *is Hermitian*

4) $\tau^+\tau$ *is Hermitian*

Proof. We leave it to the reader to show that τ^+ does indeed satisfy conditions 1)–4) and prove only the uniqueness. Suppose that ρ and σ satisfy 1)–4) when substituted for τ^+. Then

$$\begin{aligned}\rho &= \rho\tau\rho \\ &= (\rho\tau)^*\rho \\ &= \tau^*\rho^*\rho \\ &= (\tau\sigma\tau)^*\rho^*\rho \\ &= \tau^*\sigma^*\tau^*\rho^*\rho \\ &= (\sigma\tau)^*\tau^*\rho^*\rho \\ &= \sigma\tau\tau^*\rho^*\rho \\ &= \sigma\tau\rho\tau\rho \\ &= \sigma\tau\rho\end{aligned}$$

and

$$\begin{aligned}\sigma &= \sigma\tau\sigma \\ &= \sigma(\tau\sigma)^* \\ &= \sigma\sigma^*\tau^* \\ &= \sigma\sigma^*(\tau\rho\tau)^* \\ &= \sigma\sigma^*\tau^*\rho^*\tau^* \\ &= \sigma\sigma^*\tau^*(\tau\rho)^* \\ &= \sigma\sigma^*\tau^*\tau\rho \\ &= \sigma\tau\sigma\tau\rho \\ &= \sigma\tau\rho\end{aligned}$$

which shows that $\rho = \sigma$. □

The MP inverse can also be defined for matrices. In particular, if $A \in M_{m,n}(F)$ then the matrix operator τ_A has an MP inverse τ_A^+. Since this is a linear transformation from F^n to F^m, it is just multiplication by a matrix $\tau_A^+ = \tau_B$. This matrix B is the **MP inverse** for A and is denoted by A^+.

Since $\tau_A^+ = \tau_{A^+}$ and $\tau_{AB} = \tau_A\tau_B$, the matrix version of Theorem 17.4 implies that A^+ is completely characterized by the four conditions

1) $AA^+A = A$
2) $A^+AA^+ = A^+$
3) AA^+ is Hermitian
4) A^+A is Hermitian

Moreover, if

$$A = U_1\Sigma U_2^*$$

is the singular value decomposition of the matrix A then

$$A^+ = U_2\Sigma' U_1^*$$

where Σ' is obtained from Σ by replacing all nonzero entries by their multiplicative inverses. This follows from the characterization above and also from the fact that, for $i \le r$

$$U_2\Sigma' U_1^* v_i = U_2\Sigma'\epsilon_i = s_i^{-1}U_2\epsilon_i = s_i^{-1}u_i$$

and for $i > r$

$$U_2\Sigma' U_1^* v_i = U_2\Sigma'\epsilon_i = 0$$

Least Squares Approximation

Let us now discuss the most important use of the MP inverse. Consider the system of linear equations

$$Ax = v$$

where $A \in M_{m,n}(F)$. (As usual, $F = \mathbb{C}$ or $F = \mathbb{R}$.) Of course, this system has a solution if and only if $v \in \mathrm{im}(\tau_A)$. If the system has no solution, then it is of considerable practical importance to be able to solve the system

$$Ax = \widehat{v}$$

where $\widehat{v}$ is the unique vector in $\mathrm{im}(\tau_A)$ that is closest to v, as measured by the unitary (or Euclidean) distance. This problem is called the **linear least squares problem**. Any solution to the system $Ax = \widehat{v}$ is called a **least squares solution** to the system $Ax = v$. Put another way, a least squares solution to $Ax = v$ is a vector x for which $\|Ax - v\|$ is minimized.

Suppose that w and z are least squares solutions to $Ax = v$. Then

$$Aw = \widehat{v} = Az$$

and so $w - z \in \ker(A)$. (We will write A for τ_A.) Thus, if w is a particular least squares solution, then the set of all least squares solutions is $w + \ker(A)$. Among all solutions, the most interesting is the solution of minimum norm. Note that if there is a least squares solution w that lies in $\ker(A)^{\perp}$, then for any $z \in \ker(A)$, we have

$$\|w + z\|^2 = \|w\|^2 + \|z\|^2 \geq \|w\|^2$$

and so w will be the unique least squares solution of minimum norm.

Before proceeding, we remind the reader of our discussion related to the projection theorem (Theorem 9.12) to the effect that if S is a subspace of a finite-dimensional inner product space V, then the best approximation to a vector $v \in V$ from within S is the unique vector $\widehat{v} \in S$ for which $v - \widehat{v} \perp S$.

Now we can see how the MP inverse comes into play.

Theorem 17.5 *Let $A \in M_{m,n}(F)$. Among the least squares solutions to the system*

$$Ax = \widehat{v}$$

there is a unique solution of minimum norm, given by $A^{+}v$, where A^{+} is the MP inverse of A.

Proof. A vector w is a least squares solution if and only if $Aw = \widehat{v}$. Using the characterization of the best approximation $\widehat{v}$, we see that w is a solution to $Aw = \widehat{v}$ if and only if

$$Aw - v \perp \operatorname{im}(A)$$

Since $\operatorname{im}(A)^\perp = \ker(A^*)$ this is equivalent to

$$A^*(Aw - v) = 0$$

or

$$A^*Aw = A^*v$$

This system of equations is called the **normal equations** for $Ax = v$. Its solutions are precisely the least squares solutions to the system $Ax = v$.

To see that $w = A^+v$ is a least squares solution, recall that, in the notation of Figure 17.1

$$AA^+v_i = \begin{cases} v_i & i \le r \\ 0 & i > r \end{cases}$$

and so

$$A^*A(A^+v_i) = \begin{cases} A^*v_i & i \le r \\ 0 & i > r \end{cases} = A^*v_i$$

and since $\mathcal{C} = (v_1, \ldots, v_m)$ is a basis for V we conclude that A^+v satisfies the normal equations.

Finally, since $A^+v \in \ker(A)^\perp$, we deduce by the preceding remarks that A^+v is the unique least squares solution of minimum norm. □

Exercises

QR-Factorization

1. Suppose that μ is unitary and upper triangular with respect to an orthonormal basis $\mathcal{B}$ and that the coefficient of u_i in μu_i is positive. Show that μ must be the identity.
2. Assume that τ is a nonsingular operator on a finite-dimensional inner product space. Use the Gram–Schmidt process to obtain the QR-factorization of τ.
3. Prove that for reflections, $H_u = H_v$ if and only if u is a scalar multiple of v.
4. For any nonzero $v \in F^n$, show that the reflection H_v is given by
$$H_v = I_n - \frac{2vv^*}{\|v\|^2}$$
5. Use the QR factorization to show that any matrix that is similar to an upper triangular matrix is also similar to an upper triangular matrix via a unitary (orthogonal) matrix.

6. Let $S, \tau \in \mathcal{L}(V)$ and suppose that $S\tau = \tau S$. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues for S and $\mu_1, \ldots, \mu_n$ be the eigenvalues for τ. Show that the eigenvalues of $S + \tau$ are

$$\lambda_1 + \mu_{i_1}, \ldots, \lambda_n + \mu_{i_n}$$

where $(i_1, \ldots, i_n)$ is a permutation of $(1, \ldots, n)$. Hence, for commuting operators,

$$\sigma(S + \tau) \subseteq \sigma(S) + \sigma(\tau)$$

7. Let $S, \tau \in \mathcal{L}(V)$ be commuting operators. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues for S and $\mu_1, \ldots, \mu_n$ be the eigenvalues for τ. Using the previous exercise, show that if all of the sums $\lambda_i + \mu_j$ are nonzero, then $S + \tau$ is invertible.
8. Let J be the matrix

$$J = \begin{bmatrix} 0 & \cdots & 0 & 1 \\ \vdots & \iddots & 1 & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \cdots & 0 \end{bmatrix}$$

that has 1's on the diagonal that moves up from left to right and 0's elsewhere. Find J^{-1} and J^*. Compare JA with A. Compare AJ with A. Compare JAJ^* with A. Show that any upper triangular matrix is unitarily equivalent to a lower triangular matrix.
9. If $\tau \in \mathcal{L}(V)$ and $\mathcal{B} = (u_1, \ldots, u_n)$ is a basis for which

$$\tau u_i \in \langle u_1, \ldots, u_i \rangle$$

then find a basis $\mathcal{C} = (x_1, \ldots, x_n)$ for which

$$\tau x_i \in \langle x_i, \ldots, x_n \rangle$$

10. (**Cholsky decomposition**) We have seen that a linear operator τ is positive if and only if it has the form $\tau = \sigma^*\sigma$ for some operator σ. Using the QR-factorization of σ, prove the following result, known as the **Cholsky decomposition**. A linear operator $\tau \in \mathcal{L}(V)$ is positive if and only if it has the form $\tau = \rho^*\rho$ where ρ is upper triangularizable. Moreover, if τ is invertible then ρ can be chosen with positive eigenvalues, in which case the factorization is unique.

Singular Values

11. Let $\tau \in \mathcal{L}(U)$. Show that the singular values of τ^* are the same as those of τ.
12. Find the singular values and the singular value decomposition of the matrix

$$A = \begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix}$$

Find A^+.

13. Find the singular values and the singular value decomposition of the matrix

$$A = \begin{bmatrix} 1 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}$$

Find A^+. *Hint*: is it better to work with A^*A or AA^*?

14. Let $X = (x_1\ x_2\ \cdots\ x_m)^t$ be a column matrix over $\mathbb{C}$. Find a singular value decomposition of X.

15. Let $A \in M_{m,n}(F)$ and let $B \in M_{m+n,m+n}(F)$ be the square matrix

$$B = \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix}_{\text{block}}$$

Show that, counting multiplicity, the nonzero eigenvalues of B are precisely the singular values of A together with their negatives. *Hint*: Let $A = U_1 \Sigma U_2^*$ be a singular–value decomposition of A and try factoring B into a product USU^* where U is unitary. Do not read the following second hint unless you get stuck. *Second Hint*: verify the block factorization

$$B = \begin{bmatrix} 0 & U_1 \\ U_2 & 0 \end{bmatrix} \begin{bmatrix} 0 & \Sigma^* \\ \Sigma & 0 \end{bmatrix} \begin{bmatrix} 0 & U_2^* \\ U_1^* & 0 \end{bmatrix}$$

What are the eigenvalues of the middle factor on the right? (Try $\epsilon_1 + \epsilon_{n+1}$ and $\epsilon_1 - \epsilon_{n+1}$.)

16. Use the results of the previous exercise to show that a matrix $A \in M_{m,n}(F)$, its adjoint A^*, its transpose A^t and its conjugate $\overline{A}$ all have the same singular values. Show also that if U and U' are unitary then A and UAU' have the same singular values.

17. Let $A \in M_n(F)$ be nonsingular. Show that the following procedure produces a singular-value decomposition $A = U_1 \Sigma U_2^*$ of A.

   a) Write $A = UDU^*$ where $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and the λ_i's are positive and the columns of U form an orthonormal basis of eigenvectors for A. (We never said that this was a practical procedure.)

   b) Let $\Sigma = \text{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})$ where the square roots are nonnegative. Also let $U_1 = U$ and $\text{U}_2 = A^* U \Sigma^{-1}$.

18. If $A = (a_{i,j})$ is an $n \times m$ matrix then the **Frobenius norm** of A is

$$\|A\|_F = \left(\sum_{i,j} a_{i,j}^2 \right)^{1/2}$$

Show that $\|A\|_F^2 = \sum s_i^2$ is the sum of the squares of the singular values of A.

Chapter 18
The Umbral Calculus

In this chapter, we give a brief introduction to an area called the *umbral calculus*. This is a linear-algebraic theory used to study certain types of polynomial functions that play an important role in applied mathematics. We give only a brief introduction to the subject, emphasizing the algebraic aspects rather than the applications. For more on the umbral calculus, may we suggest *The Umbral Calculus*, by Roman [1984]?

One bit of notation: The **lower factorial numbers** are defined by

$$(n)_k = n(n-1)\cdots(n-k+1)$$

Formal Power Series

We begin with a few remarks concerning formal power series. Let $\mathcal{F}$ denote the algebra of formal power series in the variable t, with complex coefficients. Thus, $\mathcal{F}$ is the set of all formal sums of the form

$$f(t) = \sum_{k=0}^{\infty} a_k t^k \tag{18.1}$$

where $a_k \in \mathbb{C}$ (the complex numbers). Addition and multiplication are purely formal

$$\sum_{k=0}^{\infty} a_k t^k + \sum_{k=0}^{\infty} b_k t^k = \sum_{k=0}^{\infty} (a_k + b_k) t^k$$

and

$$\left(\sum_{k=0}^{\infty} a_k t^k\right)\left(\sum_{k=0}^{\infty} b_k t^k\right) = \sum_{k=0}^{\infty}\left(\sum_{j=0}^{k} a_j b_{k-j}\right) t^k$$

The **order** $o(f)$ of f is the smallest exponent of t that appears with a nonzero coefficient. The order of the zero series is defined to be $+\infty$. Note that a series

f has a multiplicative inverse, denoted by f^{-1}, if and only if $o(f) = 0$. We leave it to the reader to show that

$$o(fg) = o(f) + o(g)$$

and

$$o(f + g) \geq \min\{o(f), o(g)\}$$

If f_k is a sequence in $\mathcal{F}$ with $o(f_k) \to \infty$ as $k \to 0$ then for any series

$$g(t) = \sum_{k=0}^{\infty} b_k t^k$$

we may substitute f_k for t^k to get the series

$$h(t) = \sum_{k=0}^{\infty} b_k f_k(t)$$

which is well-defined since the coefficient of each power of t is a finite sum. In particular, if $o(f) \geq 1$ then $o(f^k) \to \infty$ and so the **composition**

$$(g \circ f)(t) = g(f(t)) = \sum_{k=0}^{\infty} b_k f^k(t)$$

is well-defined. It is easy to see that $o(g \circ f) = o(g)o(f)$.

If $o(f) = 1$ then f has a compositional inverse, denoted by $\overline{f}$ and satisfying

$$(f \circ \overline{f})(t) = (\overline{f} \circ f)(t) = t$$

A series f with $o(f) = 1$ is called a **delta series**.

The sequence of powers f^k of a delta series f forms a **pseudobasis** for $\mathcal{F}$, in the sense that for any $g \in \mathcal{F}$, there exists a unique sequence of constants a_k for which

$$g(t) = \sum_{k=0}^{\infty} a_k f^{\,k}(t)$$

Finally, we note that the formal derivative of the series (18.1) is given by

$$\partial_t f(t) = f'(t) = \sum_{k=1}^{\infty} k a_k t^{k-1}$$

The operator ∂_t is a derivation, that is,

$$\partial_t(fg) = \partial_t(f)g + f\partial_t(g)$$

The Umbral Algebra

Let $\mathcal{P} = \mathbb{C}[x]$ denote the algebra of polynomials in a single variable x over the complex field. One of the starting points of the umbral calculus is the fact that any formal power series in $\mathcal{F}$ can play three different roles: as a formal power series, as a linear functional on $\mathcal{P}$ and as a linear operator on $\mathcal{P}$. Let us first explore the connection between formal power series and linear functionals.

Let $\mathcal{P}^*$ denote the vector space of all linear functionals on $\mathcal{P}$. Note that $\mathcal{P}^*$ is the algebraic dual space of $\mathcal{P}$, as defined in Chapter 2. It will be convenient to denote the action of $L \in \mathcal{P}^*$ on $p(x) \in \mathcal{P}$ by

$$\langle L \mid p(x) \rangle$$

(This is the "bra-ket" notation of Paul Dirac.) The vector space operations on $\mathcal{P}^*$ then take the form

$$\langle L + M \mid p(x) \rangle = \langle L \mid p(x) \rangle + \langle M \mid p(x) \rangle$$

and

$$\langle rL \mid p(x) \rangle = r\langle L \mid p(x) \rangle, \; r \in \mathbb{C}$$

Note also that since any linear functional on $\mathcal{P}$ is uniquely determined by its values on a basis for $\mathcal{P}$, the functional $L \in \mathcal{P}^*$ is uniquely determined by the values $\langle L \mid x^n \rangle$ for $n \geq 0$.

Now, any formal series in $\mathcal{F}$ can be written in the form

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

and we can use this to define a linear functional $f(t)$ by setting

$$\langle f(t) \mid x^n \rangle = a_n$$

for $n \geq 0$. In other words, the linear functional $f(t)$ is defined by

$$f(t) = \sum_{k=0}^{\infty} \frac{\langle f(t) \mid x^k \rangle}{k!} t^k$$

where the expression $f(t)$ on the left is just a formal power series. Note in particular that

$$\langle t^k \mid x^n \rangle = n!\delta_{n,k}$$

where $\delta_{n,k}$ is the Kronecker delta function. This implies that

$$\langle t^k \mid p(x) \rangle = p^{(k)}(0)$$

and so t^k is the functional "kth derivative at 0." Also, t^0 is evaluation at 0.

As it happens, any linear functional L on $\mathcal{P}$ has the form $f(t)$. To see this, we simply note that if

$$f_L(t)=\sum_{k=0}^{\infty}\frac{\langle L\mid x^k\rangle}{k!}t^k$$

then

$$\langle f_L(t)\mid x^n\rangle=\langle L\mid x^n\rangle$$

for all $n\ge 0$ and so as linear functionals, $L=f_L(t)$.

Thus, we can define a map $\phi\colon\mathcal{P}^*\to\mathcal{F}$ by $\phi(L)=f_L(t)$.

Theorem 18.1 *The map $\phi\colon\mathcal{P}^*\to\mathcal{F}$ defined by $\phi(L)=f_L(t)$ is a vector space isomorphism from $\mathcal{P}^*$ onto $\mathcal{F}$.*

Proof. To see that ϕ is injective, note that

$$f_L(t)=f_M(t)\Rightarrow\langle L\mid x^n\rangle=\langle M\mid x^n\rangle\text{ for all }n\ge 0\Rightarrow L=M$$

Moreover, the map ϕ is surjective, since for any $f\in\mathcal{F}$, the linear functional $L=f(t)$ has the property that $\phi(L)=f_L(t)=f(t)$. Finally,

$$\begin{aligned}\phi(rL+sM)&=\sum_{k=0}^{\infty}\frac{\langle rL+sM\mid x^k\rangle}{k!}t^k\\&=r\sum_{k=0}^{\infty}\frac{\langle L\mid x^k\rangle}{k!}t^k+s\sum_{k=0}^{\infty}\frac{\langle M\mid x^k\rangle}{k!}t^k\\&=r\phi(L)+s\phi(M)\end{aligned}$$

□

From now on, we shall identify the vector space $\mathcal{P}^*$ with the vector space $\mathcal{F}$, using the isomorphism $\phi\colon\mathcal{P}^*\to\mathcal{F}$. Thus, we think of linear functionals on $\mathcal{P}$ simply as formal power series. The advantage of this approach is that $\mathcal{F}$ is more than just a vector space—it is an algebra. Hence, we have automatically defined a multiplication of linear functionals, namely, the product of formal power series. The algebra $\mathcal{F}$, when thought of as both the algebra of formal power series and the algebra of linear functionals on $\mathcal{P}$, is called the **umbral algebra**.

Let us consider an example.

Example 18.1 For $a\in\mathbb{C}$, the **evaluation functional** $\epsilon_a\in\mathcal{P}^*$ is defined by

$$\langle\epsilon_a\mid p(x)\rangle=p(a)$$

In particular, $\langle \epsilon_a \mid x^n \rangle = a^n$ and so the formal power series representation for this functional is

$$f_{\epsilon_a}(t) = \sum_{k=0}^{\infty} \frac{\langle \epsilon_a \mid x^k \rangle}{k!} t^k = \sum_{k=0}^{\infty} \frac{a^k}{k!} t^k = e^{at}$$

which is the exponential series. If e^{bt} is evaluation at b then

$$e^{at} e^{bt} = e^{(a+b)t}$$

and so the product of evaluation at a and evaluation at b is evaluation at $a + b$. □

When we are thinking of a delta series $f \in \mathcal{F}$ as a linear functional, we refer to it as a **delta functional**. Similarly, an invertible series $f \in \mathcal{F}$ is referred to as an **invertible functional**. Here are some simple consequences of the development so far.

Theorem 18.2

1) *For any $f \in \mathcal{F}$,*

$$f(t) = \sum_{k=0}^{\infty} \frac{\langle f(t) \mid x^k \rangle}{k!} t^k$$

2) *For any $p \in \mathcal{P}$,*

$$p(x) = \sum_{k \geq 0} \frac{\langle t^k \mid p(x) \rangle}{k!} x^k$$

3) *For any $f, g \in \mathcal{F}$,*

$$\langle f(t)g(t) \mid x^n \rangle = \sum_{k=0}^{n} \binom{n}{k} \langle f(t) \mid x^k \rangle \langle g(t) \mid x^{n-k} \rangle$$

4) $o(f(t)) > \deg p(x) \Rightarrow \langle f(t) \mid p(x) \rangle = 0$

5) *If $o(f_k) = k$ for all $k \geq 0$ then*

$$\left\langle \sum_{k=0}^{\infty} a_k f_k(t) \middle| p(x) \right\rangle = \sum_{k \geq 0} a_k \langle f_k(t) \mid p(x) \rangle$$

where the sum on the right is a finite one.

6) *If $o(f_k) = k$ for all $k \geq 0$ then*

$$\langle f_k(t) \mid p(x) \rangle = \langle f_k(t) \mid q(x) \rangle \text{ for all } k \geq 0 \Rightarrow p(x) = q(x)$$

7) *If $\deg p_k(x) = k$ for all $k \geq 0$ then*

$$\langle f(t) \mid p_k(x) \rangle = \langle g(t) \mid p_k(x) \rangle \text{ for all } k \geq 0 \Rightarrow f(t) = g(t)$$

Proof. We prove only part 3). Let

$$f(t)=\sum_{k=0}^{\infty}\frac{a_k}{k!}\,t^k \text{ and } g(t)=\sum_{j=0}^{\infty}\frac{b_j}{j!}\,t^j$$

Then

$$f(t)g(t)=\sum_{m=0}^{\infty}\Big(\frac{1}{m!}\sum_{k=0}^{m}\binom{m}{k}a_kb_{m-k}\Big)t^m$$

and applying both sides of this (as linear functionals) to x^n gives

$$\langle f(t)g(t)\mid x^n\rangle=\sum_{k=0}^{n}\binom{n}{k}a_kb_{n-k}$$

The result now follows from the fact that part 1) implies $a_k=\langle f(t)\mid x^k\rangle$ and $b_{n-k}=\langle g(t)\mid x^{n-k}\rangle$. □

We can now present our first "umbral" result.

Theorem 18.3 For any $f(t)\in\mathcal{F}$ and $p(x)\in\mathcal{P}$,

$$\langle f(t)\mid xp(x)\rangle=\langle\partial_t f(t)\mid p(x)\rangle$$

Proof. By linearity, we need only establish this for $p(x)=x^n$. But, if

$$f(t)=\sum_{k=0}^{\infty}\frac{a_k}{k!}\,t^k$$

then

$$\begin{aligned}\langle\partial_t f(t)\mid x^n\rangle&=\Big\langle\sum_{k=1}^{\infty}\frac{a_k}{(k-1)!}\,t^{k-1}\Big|x^n\Big\rangle\\&=\sum_{k=1}^{\infty}\frac{a_k}{(k-1)!}\,\delta_{k-1,n}\\&=a_{n+1}\\&=\langle f(t)\mid x^{n+1}\rangle\end{aligned}$$

□

Let us consider a few examples of important linear functionals and their power series representations.

Example 18.2

1) We have already encountered the **evaluation functional** e^{at}, satisfying

$$\langle e^{at}\mid p(x)\rangle=p(a)$$

2) The **forward difference functional** is the delta functional $e^{at}-1$, satisfying

$$\langle e^{at}-1 \mid p(x)\rangle = p(a)-p(0)$$

3) The **Abel functional** is the delta functional $t\mathrm{e}^{at}$, satisfying

$$\langle t\mathrm{e}^{at} \mid p(x)\rangle = p'(a)$$

4) The invertible functional $(1-t)^{-1}$ satisfies

$$\langle (1-t)^{-1} \mid p(x)\rangle = \int_0^\infty p(u)e^{-u}\,du$$

as can be seen by setting $p(x)=x^n$ and expanding the expression $(1-t)^{-1}$.

5) To determine the linear functional f satisfying

$$\langle f(t) \mid p(x)\rangle = \int_0^a p(u)\,du$$

we observe that

$$f(t)=\sum_{k=0}^\infty \frac{\langle f(t) \mid x^k\rangle}{k!}\,t^k = \sum_{k=0}^\infty \frac{a^{k+1}}{(k+1)!}\,t^k = \frac{eat^{at}-1}{t}$$

The inverse $t/(e^{at}-1)$ of this functional is associated with the Bernoulli polynomials, which play a very important role in mathematics and its applications. In fact, the numbers

$$B_n = \left\langle \frac{t}{e^{at}-1} \middle| x^n \right\rangle$$

are known as the **Bernoulli numbers**. □

Formal Power Series as Linear Operators

We now turn to the connection between formal power series and linear operators on $\mathcal{P}$. Let us denote the kth derivative operator on $\mathcal{P}$ by t^k. Thus,

$$t^k p(x) = p^{(k)}(x)$$

We can then extend this to formal series in t

$$f(t)=\sum_{k=0}^\infty \frac{a_k}{k!}\,t^k \tag{18.2}$$

by defining the linear operator $f(t)\colon\mathcal{P}\to\mathcal{P}$ by

$$f(t)p(x)=\sum_{k=0}^{\infty}\frac{a_k}{k!}\,[t^k p(x)]=\sum_{k\geq 0}\frac{a_k}{k!}p^{(k)}(x)$$

the latter sum being a finite one. Note in particular that

$$f(t)x^n=\sum_{k=0}^{n}\binom{n}{k}a_k x^{n-k} \tag{18.3}$$

With this definition, we see that each formal power series $f\in\mathcal{F}$ plays three roles in the umbral calculus, namely, as a formal power series, as a linear functional and as a linear operator. The two notations $\langle f(t)\mid p(x)\rangle$ and $f(t)p(x)$ will make it clear whether we are thinking of f as a functional or as an operator.

It is important to note that $f=g$ in $\mathcal{F}$ if and only if $f=g$ as linear functionals, which holds if and only if $f=g$ as linear operators. It is also worth noting that

$$[f(t)g(t)]p(x)=f(t)[g(t)p(x)]$$

and so we may write $f(t)g(t)p(x)$ without ambiguity. In addition,

$$f(t)g(t)p(x)=g(t)f(t)p(x)$$

for all $f,g\in\mathcal{F}$ and $p\in\mathcal{P}$.

When we are thinking of a delta series f as an operator, we call it a **delta operator**. The following theorem describes the key relationship between linear functionals and linear operators of the form $f(t)$.

Theorem 18.4 *If $f,g\in\mathcal{F}$ then*

$$\langle f(t)g(t)\mid p(x)\rangle=\langle f(t)\mid g(t)p(x)\rangle$$

for all polynomials $p(x)\in\mathcal{P}$.
Proof. If f has the form (18.2) then by (18.3),

$$\langle t^0\mid f(t)x^n\rangle=\left\langle t^0\middle|\sum_{k=0}^{n}\binom{n}{k}a_k x^{n-k}\right\rangle=a_n=\langle f(t)\mid x^n\rangle \tag{18.4}$$

By linearity, this holds for x^n replaced by any polynomial $p(x)$. Hence, applying this to the product fg gives

$$\begin{aligned}\langle f(t)g(t)\mid p(x)\rangle&=\langle t^0\mid f(t)g(t)p(x)\rangle\\&=\langle t^0\mid f(t)[g(t)p(x)]\rangle=\langle f(t)\mid g(t)p(x)\rangle\end{aligned}$$ □

Equation (18.4) shows that applying the linear functional $f(t)$ is equivalent to applying the operator $f(t)$ and then following by evaluation at $x=0$.

Here are the operator versions of the functionals in Example 18.2.

Example 18.3

1) The operator e^{at} satisfies

$$e^{at}x^n = \sum_{k=0}^{\infty} \frac{a^k}{k!} t^k x^n = \sum_{k=0}^{n} \binom{n}{k} a^k x^{n-k} = (x+a)^n$$

and so

$$e^{at}p(x) = p(x+a)$$

for all $p \in \mathcal{P}$. Thus e^{at} is a **translation operator**.

2) The **forward difference operator** is the delta operator $e^{at} - 1$, where

$$(e^{at} - 1)p(x) = p(x+a) - p(a)$$

3) The **Abel operator** is the delta operator te^{at}, where

$$t\mathrm{e}^{at}p(x) = p'(x+a)$$

4) The invertible operator $(1-t)^{-1}$ satisfies

$$(1-t)^{-1}p(x) = \int_0^{\infty} p(x+u)e^{-u}\mathrm{du}$$

5) The operator $(e^{at} - 1)/t$ is easily seen to satisfy

$$\frac{e^{at} - 1}{t}\, p(x) = \int_x^{x+a} p(u)\, du \qquad \square$$

We have seen that all linear functionals on $\mathcal{P}$ have the form $f(t)$, for $f \in \mathcal{F}$. However, not all linear operators on $\mathcal{P}$ have this form. To see this, observe that

$$\deg\, [f(t)p(x)] \le \deg\, p(x)$$

but the linear operator $\phi\colon \mathcal{P} \to \mathcal{P}$ defined by $\phi(p(x)) = xp(x)$ does not have this property.

Let us characterize the linear operators of the form $f(t)$. First, we need a lemma.

Lemma 18.5 *If T is a linear operator on P and $Tf(t) = f(t)T$ for some delta series $f(t)$ then* $\deg(Tp(x)) \le \deg(p(x))$.

Proof. For any $m \ge 0$

$$\deg(Tx^m) - 1 = \deg(f(t)Tx^m) = \deg(Tf(t)x^m)$$

and so

$$\deg(Tx^m) = \deg(Tf(t)x^m) + 1$$

Since $\deg(f(t)x^m) = m - 1$ we have the basis for an induction. When $m = 0$ we get $\deg(T1) = 1$. Assume that the result is true for $m - 1$. Then

$$\deg(Tx^m) = \deg(Tf(t)x^m) + 1 \le m - 1 + 1 = m \qquad \square$$

Theorem 18.6 *The following are equivalent for a linear operator $T: \mathcal{P} \to \mathcal{P}$.*

1) *T has the form $f(t)$, that is, there exists an $f \in \mathcal{F}$ for which $T = f(t)$, as linear operators.*
2) *T commutes with the derivative operator, that is, $Tt = tT$.*
3) *T commutes with any delta operator $g(t)$, that is, $Th(t) = h(t)T$.*
4) *T commutes with any translation operator, that is, $Te^{at} = e^{at}T$.*

Proof. It is clear that 1) implies 2). For the converse, let

$$g(t) = \sum_{k=0}^{\infty} \frac{\langle t^0 \mid Tx^k \rangle}{k!} t^k$$

Then

$$\langle g(t) \mid x^n \rangle = \langle t^0 \mid Tx^k \rangle$$

Now, since T commutes with t, we have

$$\begin{aligned} \langle t^n \mid Tx^k \rangle &= \langle t^0 \mid t^n Tx^k \rangle \\ &= \langle t^0 \mid Tt^n x^k \rangle \\ &= (k)_n \langle t^0 \mid Tx^{k-n} \rangle \\ &= (k)_n \langle t^0 \mid g(t)x^{k-n} \rangle \\ &= \langle t^n \mid g(t)x^k \rangle \end{aligned}$$

and since this holds for all n and k we get $T = g(t)$. We leave the rest of the proof as an exercise. □

Sheffer Sequences

We can now define the principal object of study in the umbral calculus. When referring to a sequence $s_n(x)$ in $\mathcal{P}$, we shall always assume that $\deg s_n(x) = n$ for all $n \ge 0$.

Theorem 18.7 *Let f be a delta series, let g be an invertible series and consider the geometric sequence*

$$g,\ gf, gf^2, gf^3, \ldots$$

in $\mathcal{F}$. Then there is a unique sequence $s_n(x)$ in $\mathcal{P}$ satisfying the **orthogonality conditions**

$$\langle g(t)f^{\,k}(t) \mid s_n(x)\rangle = n!\delta_{n,k} \tag{18.5}$$

for all $n, k \geq 0$.

Proof. The uniqueness follows from Theorem 18.2. For the existence, if we set

$$s_n(x) = \sum_{j=0}^{n} a_{n,j}x_j$$

and

$$g(t)f^k(t) = \sum_{i=k}^{\infty} b_{k,i}t^i$$

where $b_{k,k} \neq 0$ then (18.5) is

$$\begin{aligned} n!\delta_{n,k} &= \left\langle \sum_{i=k}^{\infty} b_{k,i}t^i \middle| \sum_{j=0}^{n} a_{n,j}x_j \right\rangle \\ &= \sum_{i=k}^{\infty}\sum_{j=0}^{n} b_{k,i}a_{n,j}\langle t^i \big| x_j\rangle \\ &= \sum_{i=k}^{n} b_{k,i}a_{n,i}i! \end{aligned}$$

Taking $k = n$ we get

$$a_{n,n} = \frac{1}{b_{n,n}}$$

For $k = n - 1$ we have

$$0 = b_{n-1,n-1}a_{n,n-1}(n-1)! + b_{n-1,n}a_{n,n}n!$$

and using the fact that $a_{n,n} = 1/b_{n,n}$ we can solve this for $a_{n,n-1}$. By successively taking $k = n, n-1, n-2, \ldots$ we can solve the resulting equations for the coefficients $a_{n,k}$ of the sequence $s_n(x)$. □

Definition *The sequence* $s_n(x)$ *in (18.5) is called the* **Sheffer sequence** *for the ordered pair* $(g(t), f(t))$. *We shorten this by saying that* $s_n(x)$ *is* **Sheffer for** $(g(t), f(t))$. □

Two special types of Sheffer sequences deserve explicit mention.

Definition *The Sheffer sequence for a pair of the form* $(1, f(t))$ *is called the* **associated sequence** *for* $f(t)$. *The Sheffer sequence for a pair of the form* $(g(t), t)$ *is called the* **Appell sequence** *for* $g(t)$. □

Note that the sequence $s_n(x)$ is Sheffer for $(g(t),f(t))$ if and only if

$$\langle g(t)f^{\,k}(t) \mid s_n(x)\rangle = n!\delta_{n,k}$$

which is equivalent to

$$\langle f^{\,k}(t) \mid g(t)s_n(x)\rangle = n!\delta_{n,k}$$

which, in turn, is equivalent to saying that the sequence $p_n(x) = g(t)s_n(x)$ is the associated sequence for $f(t)$.

Theorem 18.8 *The sequence $s_n(x)$ is Sheffer for $(g(t),f(t))$ if and only if the sequence $p_n(x) = g(t)s_n(x)$ is the associated sequence for $f(t)$.* □

Before considering examples, we wish to describe several characterizations of Sheffer sequences. First, we require a key result.

Theorem 18.9 *(***The expansion theorems***) Let $s_n(x)$ be Sheffer for $(g(t),f(t))$.*
1) For any $h \in \mathcal{F}$,

$$h(t) = \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x)\rangle}{k!} g(t)f^{\,k}(t)$$

2) For any $p \in \mathcal{P}$,

$$p(x) = \sum_{k\geq 0} \frac{\langle g(t)f^{\,k}(t) \mid p(x)\rangle}{k!} s_k(x)$$

Proof. Part 1) follows from Theorem 18.2, since

$$\begin{aligned}\Big\langle \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x)\rangle}{k!} g(t)f^{\,k}(t) \Big| s_n(x)\Big\rangle &= \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x)\rangle}{k!} n!\delta_{n,k} \\ &= \langle h(t) \mid s_n(x)\rangle\end{aligned}$$

Part 2) follows in a similar way from Theorem 18.2. □

We can now begin our characterization of Sheffer sequences, starting with the generating function. The idea of a generating function is quite simple. If $r_n(x)$ is a sequence of polynomials, we may define a formal power series of the form

$$g(t,x) = \sum_{k=0}^{\infty} \frac{r_k(x)}{k!} t^k$$

This is referred to as the (**exponential**) **generating function** for the sequence $r_n(x)$. (The term exponential refers to the presence of $k!$ in this series. When this is not present, we have an ordinary generating function.) Since the series is a formal one, knowing $g(t,x)$ is equivalent (in theory, if not always in practice) to knowing the polynomials $r_n(x)$. Moreover, a knowledge of the generating function of a sequence of polynomials can often lead to a deeper understanding of the sequence itself, that might not be otherwise easily accessible. For this reason, generating functions are studied quite extensively.

For the proofs of the following characterizations, we refer the reader to Roman [1984].

Theorem 18.10 *(***Generating function***)*

1) *The sequence $p_n(x)$ is the associated sequence for a delta series $f(t)$ if and only if*

$$e^{y\overline{f}(t)}=\sum_{k=0}^{\infty}\frac{p_k(y)}{k!}t^k$$

where $\overline{f}(t)$ is the compositional inverse of $f(t)$.

2) *The sequence $s_n(x)$ is Sheffer for $(g(t),f(t))$ if and only if*

$$\frac{1}{g(\overline{f}(t))}e^{y\overline{f}(t)}=\sum_{k=0}^{\infty}\frac{s_k(y)}{k!}t^k$$

The sum on the right is called the **generating function** *of $s_n(x)$.*

Proof. Part 1) is a special case of part 2). For part 2), the expression above is equivalent to

$$\frac{1}{g(t)}e^{yt}=\sum_{k=0}^{\infty}\frac{s_k(y)}{k!}f^k(t)$$

which is equivalent to

$$e^{yt}=\sum_{k=0}^{\infty}\frac{s_k(y)}{k!}g(t)f^k(t)$$

But if $s_n(x)$ is Sheffer for $(f(t),g(t))$ then this is just the expansion theorem for e^{yt}. Conversely, this expression implies that

$$s_n(y)=\langle e^{yt}\mid s_n(x)\rangle=\sum_{k=0}^{\infty}\frac{s_k(y)}{k!}\langle g(t)f^{\,k}(t)\mid s_n(x)\rangle$$

and so $\langle g(t)f^{\,k}(t)\mid s_n(x)\rangle=n!\delta_{n,k}$, which says that $s_n(x)$ is Sheffer for (f,g). □

We can now give a representation for Sheffer sequences.

Theorem 18.11 (Conjugate representation)

1) *A sequence $p_n(x)$ is the associated sequence for $f(t)$ if and only if*

$$p_n(x) = \sum_{k=0}^{n} \frac{1}{k!} \langle \overline{f}(t)^k \mid x^n \rangle x^k$$

2) *A sequence $s_n(x)$ is Sheffer for $(g(t), f(t))$ if and only if*

$$s_n(x) = \sum_{k=0}^{n} \frac{1}{k!} \langle g(\overline{f}(t))^{-1} \overline{f}(t)^k \mid x^n \rangle x^k$$

Proof. We need only prove part 2). We know that $s_n(x)$ is Sheffer for $(g(t), f(t))$ if and only if

$$\frac{1}{g(\overline{f}(t))} e^{y\overline{f}(t)} = \sum_{k=0}^{\infty} \frac{s_k(y)}{k!} t^k$$

But this is equivalent to

$$\left\langle \frac{1}{g(\overline{f}(t))} e^{y\overline{f}(t)} \mid x^n \right\rangle = \left\langle \sum_{k=0}^{\infty} \frac{s_k(y)}{k!} t^k \middle| x^n \right\rangle = s_n(y)$$

Expanding the exponential on the left gives

$$\sum_{k=0}^{\infty} \frac{\langle g(\overline{f}(t))^{-1} \overline{f}(t)^k \mid x^n \rangle}{k!} y^k = \left\langle \sum_{k=0}^{\infty} \frac{s_k(y)}{k!} t^k \middle| x^n \right\rangle = s_n(y)$$

Replacing y by x gives the result. □

Sheffer sequences can also be characterized by means of linear operators.

Theorem 18.12 (**Operator characterization**)

1) *A sequence $p_n(x)$ is the associated sequence for $f(t)$ if and only if*
   a) $p_n(0) = \delta_{n,0}$
   b) $f(t)p_n(x) = np_{n-1}(x)$ *for* $n \geq 0$
2) *A sequence $s_n(x)$ is Sheffer for $(g(t), f(t))$, for some invertible series $g(t)$ if and only if*

$$f(t)s_n(x) = ns_{n-1}(x)$$

for all $n \geq 0$.

Proof. For part 1), if $p_n(x)$ is associated with $f(t)$ then

$$p_n(0) = \langle e^{0t} \mid p_n(x) \rangle = \langle f(t)^0 \mid p_n(x) \rangle = 0!\delta_{n,0}$$

and

$$\begin{aligned}\langle f(t)^k \mid f(t)p_n(x)\rangle &= \langle f(t)^{k+1} \mid p_n(x)\rangle \\ &= n!\delta_{n,k+1} \\ &= n(n-1)!\delta_{n-1,k} \\ &= n\langle f(t)^k \mid p_{n-1}(x)\rangle\end{aligned}$$

and since this holds for all $k \geq 0$ we get 1b). Conversely, if 1a) and 1b) hold then

$$\begin{aligned}\langle f(t)^k \mid p_n(x)\rangle &= \langle t^0 \mid f(t)^k p_n(x)\rangle \\ &= (n)_k p_{n-k}(0) \\ &= (n)_k \delta_{n-k,0} \\ &= n!\delta_{n,k}\end{aligned}$$

and so $p_n(x)$ is the associated sequence for $f(t)$.

As for part 2), if $s_n(x)$ is Sheffer for $(g(t), f(t))$ then

$$\begin{aligned}\langle g(t)f(t)^k \mid f(t)s_n(x)\rangle &= \langle g(t)f(t)^{k+1} \mid s_n(x)\rangle \\ &= n!\delta_{n,k+1} \\ &= n(n-1)!\delta_{n-1,k} \\ &= n\langle g(t)f(t)^k \mid s_{n-1}(x)\rangle\end{aligned}$$

and so $f(t)s_n(x) = ns_{n-1}(x)$, as desired. Conversely, suppose that

$$f(t)s_n(x) = ns_{n-1}(x)$$

and let $p_n(x)$ be the associated sequence for $f(t)$. Let T be the invertible linear operator on V defined by

$$Ts_n(x) = p_n(x)$$

Then

$$Tf(t)s_n(x) = nTs_{n-1}(x) = np_{n-1}(x) = f(t)p_n(x) = f(t)Ts_n(x)$$

and so Theorem 18.5 implies that $T = g(t)$ for some invertible series $g(t)$. Then

$$\begin{aligned}\langle g(t)f(t)^k \mid s_n(x)\rangle &= \langle f(t)^k \mid g(t)s_n(x)\rangle \\ &= \langle t^0 \mid f(t)^k p_n(x)\rangle \\ &= (n)_k p_{n-k}(0) \\ &= (n)_k \delta_{n-k,0} \\ &= n!\delta_{n,k}\end{aligned}$$

and so $s_n(x)$ is Sheffer for $(g(t), f(t))$. □

We next give a formula for the action of a linear operator $h(t)$ on a Sheffer sequence.

Theorem 18.13 *Let $s_n(x)$ be a Sheffer sequence for $(g(t), f(t))$ and let $p_n(x)$ be associated with $f(t)$. Then for any $h(t)$ we have*

$$h(t)s_n(x) = \sum_{k=0}^{n} \binom{n}{k} \langle h(t) \mid s_k(x) \rangle p_{n-k}(x)$$

Proof. By the expansion theorem

$$h(t) = \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x) \rangle}{k!} g(t) f^{\,k}(t)$$

we have

$$\begin{aligned} h(t)s_n(x) &= \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x) \rangle}{k!} g(t) f^{\,k}(t) s_n(x) \\ &= \sum_{k=0}^{\infty} \frac{\langle h(t) \mid s_k(x) \rangle}{k!} (n)_k p_{n-k}(x) \end{aligned}$$

which is the desired formula. □

Theorem 18.14

1) (**The binomial identity**) *A sequence $p_n(x)$ is the associated sequence for a delta series $f(t)$ if and only if it is of* **binomial type**, *that is, if and only if it satisfies the identity*

$$p_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} p_k(y) p_{n-k}(x)$$

for all $y \in \mathbb{C}$.

2) (**The Sheffer identity**) *A sequence $s_n(x)$ is Sheffer for $(g(t), f(t))$, for some invertible $g(t)$ if and only if*

$$s_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} p_k(y) s_{n-k}(x)$$

for all $y \in \mathbb{C}$, where $p_n(x)$ is the associated sequence for $f(t)$.

Proof. To prove part 1), if $p_n(x)$ is an associated sequence then taking $h(t) = e^{yt}$ in Theorem 18.13 gives the binomial identity. Conversely, suppose that the sequence $p_n(x)$ is of binomial type. We will use the operator characterization to show that $p_n(x)$ is an associated sequence. Taking $x = y = 0$ we have for $n = 0$

$$p_0(0) = p_0(0)p_0(0)$$

and so $p_0(0) = 1$. Also,

$$p_1(0) = p_0(0)p_1(0) + p_1(0)p_0(0) = 2p_1(0)$$

and so $p_1(0) = 0$. Assuming that $p_i(0) = 0$ for $i = 1, \ldots, m-1$ we have

$$p_m(0) = p_0(0)p_m(0) + p_m(0)p_0(0) = 2p_m(0)$$

and so $p_m(0) = 0$. Thus, $p_n(0) = \delta_{n,0}$.

Next, define a linear functional $f(t)$ by

$$\langle f(t) \mid p_n(x)\rangle = \delta_{n,1}$$

Since $\langle f(t) \mid 1\rangle = \langle f(t) \mid p_0(x)\rangle = 0$ and $\langle f(t) \mid p_1(x)\rangle = 1 \neq 0$ we deduce that $f(t)$ is a delta series. Now, the binomial identity gives

$$\begin{aligned}\langle f(t) \mid e^{yt}p_n(x)\rangle &= \sum_{k=0}^{n} \binom{n}{k} p_k(y)\langle f(t) \mid p_{n-k}(x)\rangle \\ &= \sum_{k=0}^{n} \binom{n}{k} p_k(y)\delta_{n-k,1} \\ &= np_{n-1}(y)\end{aligned}$$

and so

$$\langle e^{yt} \mid f(t)p_n(x)\rangle = \langle e^{yt} \mid np_{n-1}(x)\rangle$$

and since this holds for all y, we get $f(t)p_n(x) = np_{n-1}(x)$. Thus, $p_n(x)$ is the associated sequence for $f(t)$.

For part 2), if $s_n(x)$ is a Sheffer sequence then taking $h(t) = e^{yt}$ in Theorem 18.13 gives the Sheffer identity. Conversely, suppose that the Sheffer identity holds, where $p_n(x)$ is the associated sequence for $f(t)$. It suffices to show that $g(t)s_n(x) = p_n(x)$ for some invertible $g(t)$. Define a linear operator T by

$$Ts_n(x) = p_n(x)$$

Then

$$e^{yt}Ts_n(x) = e^{yt}p_n(x) = p_n(x+y)$$

and by the Sheffer identity

$$Te^{yt}s_n(x) = \sum_{k=0}^{n} \binom{n}{k} p_k(y)Ts_{n-k}(x) = \sum_{k=0}^{n} \binom{n}{k} p_k(y)p_{n-k}(x)$$

and the two are equal by part 1). Hence, T commutes with e^{yt} and is therefore of the form $g(t)$, as desired. □

Examples of Sheffer Sequences

We can now give some examples of Sheffer sequences. While it is often a relatively straightforward matter to verify that a given sequence is Sheffer for a given pair $(g(t), f(t))$, it is quite another matter to find the Sheffer sequence for a given pair. The umbral calculus provides two formulas for this purpose, one of which is direct, but requires the usually very difficult computation of the series $(f(t)/t)^{-n}$. The other is a recurrence relation that expresses each $s_n(x)$ in terms of previous terms in the Sheffer sequence. Unfortunately, space does not permit us to discuss these formulas in detail. However, we will discuss the recurrence formula for associated sequences later in this chapter.

Example 18.4 The sequence $p_n(x) = x^n$ is the associated sequence for the delta series $f(t) = t$. The generating function for this sequence is

$$e^{yt} = \sum_{k=0}^{\infty} \frac{y^k}{k!} t^k$$

and the binomial identity is the well known binomial formula

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

Example 18.5 The **lower factorial polynomials**

$$(x)_n = x(x-1)\cdots(x-n+1)$$

form the associated sequence for the forward difference functional

$$f(t) = e^t - 1$$

discussed in Example 18.2. To see this, we simply compute, using Theorem 18.12. Since $(0)_0$ is defined to be 1, we have $(0)_n = \delta_{n,0}$. Also,

$$\begin{aligned}(e^t-1)(x)_n &= (x+1)_n - (x)_n \\ &= [(x+1)x(x-1)\cdots(x-n+2)] - [x(x-1)\cdots(x-n+1)] \\ &= x(x-1)\cdots(x-n+2)[(x+1)-(x-n+1)] \\ &= nx(x-1)\cdots(x-n+2) \\ &= n(x)_{n-1}\end{aligned}$$

The generating function for the lower factorial polynomials is

$$e^{y\log(1+t)} = \sum_{k=0}^{\infty} \frac{(y)_k}{k!} t^k$$

which can be rewritten in the more familiar form

$$(1+t)^y = \sum_{k=0}^{\infty} \binom{y}{k} t^k$$

Of course, this is a formal identity, so there is no need to make any restrictions on t. The binomial identity in this case is

$$(x+y)_n = \sum_{k=0}^{n} \binom{n}{k} (x)_k (y)_{n-k}$$

which can also be written in the form

$$\binom{x+y}{n} = \sum_{k=0}^{n} \binom{x}{k} \binom{y}{n-k}$$

This is known as the **Vandermonde convolution formula**.

Example 18.6 The **Abel polynomials**

$$A_n(x;a) = x(x-an)^{n-1}$$

form the associated sequence for the **Abel functional**

$$f(t) = t\mathrm{e}^{at}$$

also discussed in Example 18.2. We leave verification of this to the reader. The generating function for the Abel polynomials is

$$e^{y\overline{f}(t)} = \sum_{k=0}^{\infty} \frac{y(y-ak)^{k-1}}{k!} t^k$$

Taking the formal derivative of this with respect to y gives

$$\overline{f}(t)e^{y\overline{f}(t)} = \sum_{k=0}^{\infty} \frac{k(y-a)(y-ak)^{k-1}}{k!} t^k$$

which, for $y=0$, gives a formula for the compositional inverse of the series $f(t) = te^{at}$,

$$\overline{f}(t) = \sum_{k=1}^{\infty} \frac{(-a)^k k^{k-1}}{(k-1)!} t^k$$

Example 18.7 The famous **Hermite polynomials** $H_n(x)$ form the Appell sequence for the invertible functional

$$g(t) = e^{t^2/2}$$

We ask the reader to show that $s_n(x)$ is the Appell sequence for $g(t)$ if and only if $s_n(x)=g(t)^{-1}x^n$. Using this fact, we get

$$H_n(x)=e^{-t^2/2}x^n=\sum_{k\geq 0}(-\frac{1}{2})^k\frac{(n)_{2k}}{k!}\,x^{n-k}$$

The generating function for the Hermite polynomials is

$$e^{yt-t^2/2}=\sum_{k=0}^{\infty}\frac{H_k(y)}{k!}\,t^k$$

and the Sheffer identity is

$$H_n(x+y)=\sum_{k=0}^{n}\binom{n}{k}H_k(x)y^{n-k}$$

We should remark that the Hermite polynomials, as defined in the literature, often differ from our definition by a multiplicative constant. □

Example 18.8 The well known and important **Laguerre polynomials** $L_n^{(\alpha)}(x)$ of order α form the Sheffer sequence for the pair

$$g(t)=(1-t)^{-\alpha-1},\ f(t)=\frac{t}{t-1}$$

It is possible to show (although we will not do so here) that

$$L_n^{(\alpha)}(x)=\sum_{k=0}^{n}\frac{n!}{k!}\binom{\alpha+n}{n-k}(-x)^k$$

The generating function of the Laguerre polynomials is

$$\frac{1}{(1-t)^{\alpha+1}}e^{yt/(t-1)}=\sum_{k=0}^{\infty}\frac{L_k^{(\alpha)}(x)}{k!}\,t^k$$

As with the Hermite polynomials, some definitions of the Laguerre polynomials differ by a multiplicative constant. □

We presume that the few examples we have given here indicate that the umbral calculus applies to a significant range of important polynomial sequences. In Roman [1984], we discuss approximately 30 different sequences of polynomials that are (or are closely related to) Sheffer sequences.

Umbral Operators and Umbral Shifts

We have now established the basic framework of the umbral calculus. As we have seen, the umbral algebra plays three roles: as the algebra of formal power series in a single variable, as the algebra of all linear functionals on $\mathcal{P}$ and as the algebra of all linear operators on $\mathcal{P}$ that commute with the derivative operator. Moreover, since $\mathcal{F}$ is an algebra, we can consider geometric sequences

$$g,\ gf, gf^2, gf^3, \ldots$$

in $\mathcal{F}$, where $o(g) = 0$ and $o(f) = 1$. We have seen by example that the orthogonality conditions

$$\langle g(t)f^{\,k}(t) \mid s_n(x)\rangle = n!\delta_{n,k}$$

define important families of polynomial sequences.

While the machinery that we have developed so far does unify a number of topics from the classical study of polynomial sequences (for example, special cases of the expansion theorem include Taylor's expansion, the Euler-MacLaurin formula and Boole's summation formula), it does not provide much new insight into their study. Our plan now is to take a brief look at some of the deeper results in the umbral calculus, which center around the interplay between operators on $\mathcal{P}$ and their adjoints, which are operators on the umbral algebra $\mathcal{F} = \mathcal{P}^*$.

We begin by defining two important operators on $\mathcal{P}$ associated with each Sheffer sequence.

Definition *Let $s_n(x)$ be Sheffer for $(g(t), f(t))$. The linear operator $\lambda_{g,f}\colon \mathcal{P} \to \mathcal{P}$ defined by*

$$\lambda_{g,f}(x^n) = s_n(x)$$

is called the **Sheffer operator** *for the pair $(g(t), f(t))$, or for the sequence $s_n(x)$. If $p_n(x)$ is the associated sequence for $f(t)$, the Sheffer operator*

$$\lambda_f(x^n) = p_n(x)$$

is called the **umbral operator** *for $f(t)$, or for $p_n(x)$.* □

Definition *Let $s_n(x)$ be Sheffer for $(g(t), f(t))$. The linear operator $\theta_{g,f}\colon \mathcal{P} \to \mathcal{P}$ defined by*

$$\theta_{g,f}[s_n(x)] = s_{n+1}(x)$$

is called the **Sheffer shift** *for the pair $(g(t), f(t))$, or for the sequence $s_n(x)$. If $p_n(x)$ is the associated sequence for $f(t)$, the Sheffer operator*

$$\theta_f[p_n(x)] = p_{n+1}(x)$$

is called the **umbral shift** *for $f(t)$, or for $p_n(x)$.* □

It is clear that each Sheffer sequence uniquely determines a Sheffer operator and vice versa. Hence, knowing the Sheffer operator of a sequence is equivalent to knowing the sequence.

Continuous Operators on the Umbral Algebra

It is clearly desirable that a linear operator T on the umbral algebra $\mathcal{F}$ pass under infinite sums, that is, that

$$T\Big(\sum_{k=0}^{\infty} a_k f_k(t)\Big) = \sum_{k=0}^{\infty} a_k T[f_k(t)] \tag{18.6}$$

whenever the sum on the left is defined, which is precisely when $o(f_k(t)) \to \infty$ as $k \to \infty$. Not all operators on $\mathcal{F}$ have this property, which leads to the following definition.

Definition *A linear operator T on the umbral algebra $\mathcal{F}$ is* **continuous** *if it satisfies* (18.6). □

The term continuous can be justified by defining a topology on $\mathcal{F}$. However, since no additional topological concepts will be needed, we will not do so here. Note that in order for (18.6) to make sense, we must have $o(T[f_k(t)]) \to \infty$. It turns out that this condition is also sufficient.

Theorem 18.15 A linear operator T on $\mathcal{F}$ is continuous if and only if

$$o(f_k) \to \infty \Rightarrow o(T(f_k)) \to \infty \tag{18.7}$$

Proof. The necessity is clear. Suppose that (18.7) holds and that $o(f_k) \to \infty$. For any $m \geq 0$, we have

$$\Big\langle T\sum_{k=0}^{\infty} a_k f_k(t) \Big| x^n \Big\rangle = \Big\langle T\sum_{k=0}^{m} a_k f_k(t) \Big| x^n \Big\rangle + \Big\langle T\sum_{k>m} a_k f_k(t) \Big| x^n \Big\rangle \tag{18.8}$$

Since

$$o\left(\sum_{k>m} a_k f_k(t)\right) \to \infty$$

(18.7) implies that we may choose m large enough so that

$$o\Big(T\sum_{k>m} a_k f_k(t)\Big) > n$$

and

$$o(T[f_k(t)]) > n \text{ for } k > m$$

Hence, (18.8) gives

$$\begin{aligned}\left\langle T\sum_{k=0}^{\infty} a_k f_k(t)\middle| x^n\right\rangle &= \left\langle T\sum_{k=0}^{m} a_k f_k(t)\middle| x^n\right\rangle \\ &= \left\langle \sum_{k=0}^{m} a_k T[f_k(t)]\middle| x^n\right\rangle \\ &= \left\langle \sum_{k=0}^{\infty} a_k T[f_k(t)]\middle| x^n\right\rangle\end{aligned}$$

which implies the desired result. □

Operator Adjoints

If $\tau\colon \mathcal{P} \to \mathcal{P}$ is a linear operator on $\mathcal{P}$ then its (operator) adjoint $\tau^\times$ is an operator on $\mathcal{P}^* = \mathcal{F}$ defined by

$$\tau^\times[h(t)] = h(t) \circ \tau$$

In the symbolism of the umbral calculus, this is

$$\langle \tau^\times h(t) \mid p(x)\rangle = \langle h(t) \mid \tau p(x)\rangle$$

(We have reduced the number of parentheses used to aid clarity.)

Let us recall the basic properties of the adjoint from Chapter 3.

Theorem 18.16 *For $\tau, \sigma \in \mathcal{L}(\mathcal{P})$,*

1) $(\tau + \sigma)^\times = \tau^\times + \sigma^\times$
2) $(r\tau)^\times = r\tau^\times$ *for any* $r \in \mathbb{C}$
3) $(\tau\sigma)^\times = \sigma^\times \tau^\times$
4) $(\tau^{-1})^\times = (\tau^\times)^{-1}$ *for any invertible* $\tau \in \mathcal{L}(\mathcal{P})$ □

Thus, the map $\phi\colon \mathcal{L}(\mathcal{P}) \to \mathcal{L}(\mathcal{F})$ that sends $\tau\colon \mathcal{P} \to \mathcal{P}$ to its adjoint $\tau^\times\colon \mathcal{F} \to \mathcal{F}$ is a linear transformation from $\mathcal{L}(\mathcal{P})$ to $\mathcal{L}(\mathcal{F})$. Moreover, since $\tau^\times = 0$ implies that $\langle h(t) \mid \tau p(x)\rangle = 0$ for all $h(t) \in \mathcal{F}$ and $p(x) \in \mathcal{P}$, which in turn implies that $\tau = 0$, we deduce that ϕ is injective. The next theorem describes the range of ϕ.

Theorem 18.17 *A linear operator $T \in \mathcal{L}(\mathcal{F})$ is the adjoint of a linear operator $\mathcal{L} \in \mathcal{L}(\mathcal{P})$ if and only if T is continuous.*

Proof. First, suppose that $T = \tau^\times$ for some $\tau \in \mathcal{L}(\mathcal{P})$ and let $o(f_k(t)) \to \infty$. If $n \geq 0$ then for all $0 \leq i \leq n$ we have

$$\langle \tau^\times f_k(t) \mid x^i\rangle = \langle f_k(t) \mid \tau x^i\rangle$$

and so it is only necessary to take k large enough so that $o(f_k(t)) >$ deg $\tau(x^i)$ for all $0 \leq i \leq n$, whence

$$\langle \tau^{\times} f_k(t) \mid x^i \rangle = 0$$

for all $0 \le i \le n$ and so $o(\tau^{\times} f_k(t)) > n$. Thus, $o(\tau^{\times} f_k(t)) \to \infty$ and $\tau^{\times}$ is continuous.

For the converse, assume that T is continuous. If T did have the form $\tau^{\times}$ then

$$\langle Tt^k \mid x^n \rangle = \langle \tau^{\times} t^k \mid x^n \rangle = \langle t^k \mid \tau x^n \rangle$$

and since

$$\tau x^n = \sum_{k \ge 0} \frac{\langle t^k \mid \tau x^n \rangle}{k!} x^k$$

we are prompted to *define* τ by

$$\tau x^n = \sum_{k \ge 0} \frac{\langle Tt^k \mid x^n \rangle}{k!} x^k$$

This makes sense since $o(Tt^k) \to \infty$ as $k \to \infty$ and so the sum on the right is a finite sum. Then

$$\langle \tau^{\times} t^m \mid x^n \rangle = \langle t^m \mid \tau x^n \rangle = \sum_{k \ge 0} \frac{\langle Tt^k \mid x^n \rangle}{k!} \langle t^m \mid x^k \rangle = \langle Tt^m \mid x^n \rangle$$

which implies that $Tt^m = \tau^{\times} t^m$ for all $m \ge 0$. Finally, since T and $\tau^{\times}$ are both continuous, we have $T = \tau^{\times}$. □

Umbral Operators and Automorphisms of the Umbral Algebra

Figure 18.1 shows the map ϕ, which is an isomorphism from the vector space $\mathcal{L}(\mathcal{P})$ onto the space of all continuous linear operators on $\mathcal{F}$. We are interested in determining the images under this isomorphism of the set of umbral operators and the set of umbral shifts, as pictured in Figure 18.1.

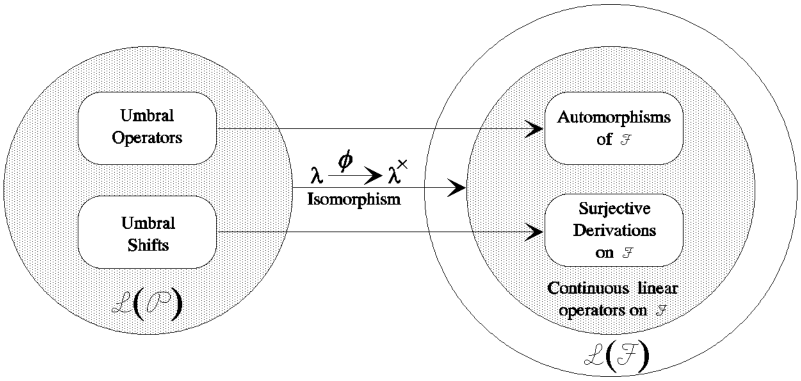
Umbral Operators
Umbral Shifts
$\mathcal{L}(\mathcal{P})$
$\lambda \xrightarrow{\phi} \lambda^\times$
Isomorphism
Automorphisms of $\mathcal{F}$
Surjective Derivations on $\mathcal{F}$
Continuous linear operators on $\mathcal{F}$
$\mathcal{L}(\mathcal{F})$

Figure 18.1

Let us begin with umbral operators. Suppose that λ_f is the umbral operator for the associated sequence $p_n(x)$, with delta series $f(t) \in \mathcal{F}$. Then

$$\langle \lambda_f^\times f(t)^k \mid x^n \rangle = \langle f(t)^k \mid \lambda_f x^n \rangle = \langle f(t)^k \mid p_n(x) \rangle = n!\delta_{n,k} = \langle t^k \mid x^n \rangle$$

for all k and n. Hence, $\lambda_f^\times f(t)^k = t^k$ and the continuity of $\lambda_f^\times$ implies that

$$\lambda_f^\times t^k = \overline{f}(t)^k$$

More generally, for any $h(t) \in \mathcal{F}$,

$$\lambda_f^\times h(t) = h(\overline{f}(t)) \tag{18.9}$$

In words, $\lambda_f^\times$ is composition by $\overline{f}(t)$.

From (18.9), we deduce that $\lambda_f^\times$ is a vector space isomorphism and that

$$\lambda_f^\times [g(t)h(t)] = g(\overline{f}(t))h(\overline{f}(t)) = \lambda_f^\times g(t)\lambda_f^\times h(t)$$

Hence, $\lambda_f^\times$ is an automorphism of the umbral algebra $\mathcal{F}$. It is a pleasant fact that this characterizes umbral operators. The first step in the proof of this is the following, whose proof is left as an exercise.

Theorem 18.18 *If T is an automorphism of the umbral algebra then T preserves order, that is, $o(Tf(t)) = o(f(t))$. In particular, T is continuous.* □

Theorem 18.19 *A linear operator λ on $\mathcal{P}$ is an umbral operator if and only if its adjoint is an automorphism of the umbral algebra $\mathcal{F}$. Moreover, if λ_f is an umbral operator then*

$$\lambda_f^\times h(t) = h(\overline{f}(t))$$

for all $h(t) \in \mathcal{F}$. *In particular,* $\lambda_f^\times f(t) = t$.

Proof. We have already shown that the adjoint of λ_f is an automorphism satisfying (18.9). For the converse, suppose that $\lambda^\times$ is an automorphism of $\mathcal{F}$. Since $\lambda^\times$ is surjective, there is a unique series $f(t)$ for which $\lambda^\times f(t) = t$. Moreover, Theorem 18.18 implies that $f(t)$ is a delta series. Thus,

$$n!\delta_{n,k} = \langle t^k \mid x^n \rangle = \langle \lambda^\times f(t)^k \mid x^n \rangle = \langle f(t)^k \mid \lambda x^n \rangle$$

which shows that λx^n is the associated sequence for $f(t)$ and hence that λ is an umbral operator. □

Theorem 18.19 allows us to fill in one of the boxes on the right side of Figure 18.1. Let us see how we might use Theorem 18.19 to advantage in the study of associated sequences.

We have seen that the isomorphism $\lambda \mapsto \lambda^\times$ maps the set $\mathcal{U}$ of umbral operators on $\mathcal{P}$ onto the set aut$(\mathcal{F})$ of automorphisms of $\mathcal{F} = \mathcal{P}^*$. But aut$(\mathcal{F})$ is a group under composition. So if

$$\lambda_f\colon x^n \to p_n(x) \text{ and } \lambda_g\colon x^n \to q_n(x)$$

are umbral operators then since

$$(\lambda_g \circ \lambda_f)^\times = \lambda_f^\times \circ \lambda_g^\times$$

is an automorphism of $\mathcal{F}$, it follows that the composition $\lambda_g \circ \lambda_f$ is an umbral operator. In fact, since

$$(\lambda_g \circ \lambda_f)^\times f(g(t)) = \lambda_f^\times \circ \lambda_g^\times f(g(t)) = \lambda_f^\times f(t) = t$$

we deduce that $\lambda_g \circ \lambda_f = \lambda_{f \circ g}$. Also, since

$$\lambda_{\overline{f}} \circ \lambda_f = \lambda_{f \circ \overline{f}} = \lambda_t = \iota$$

we have $\lambda_f^{-1} = \lambda_{\overline{f}}$.

Thus, the set $\mathcal{U}$ of umbral operators is a group under composition with

$$\lambda_g \circ \lambda_f = \lambda_{f \circ g}$$

and

$$\lambda_f^{-1} = \lambda_{\overline{f}}$$

Let us see how this plays out with respect to associated sequences. If the associated sequence for $f(t)$ is

$$p_n(x) = \sum_{k=0}^{n} p_{n,k}x^k$$

then $\lambda_f\colon x^n \to p_n(x)$ and so $\lambda_{f\circ g} = \lambda_g \circ \lambda_f$ is the umbral operator for the associated sequence

$$(\lambda_g \circ \lambda_f)x^n = \lambda_g p_n(x) = \sum_{k=0}^{n} p_{n,k}\lambda_g x^k = \sum_{k=0}^{n} p_{n,k}q_k(x)$$

This sequence, denoted by

$$p_n(\boldsymbol{q}(x)) = \sum_{k=0}^{n} p_{n,k}q_k(x) \tag{18.10}$$

is called the **umbral composition** of $p_n(x)$ with $q_n(x)$. The umbral operator $\lambda_{\overline{f}} = \lambda_f^{-1}$ is the umbral operator for the associated sequence $r_n(x) = \Sigma r_{n,k}x^k$ where

$$\lambda_f^{-1}x^n = r_n(x)$$

and so

$$x^n = \sum_{k=0}^{n} r_{n,k}p_k(x)$$

Let us summarize.

Theorem 18.20

1) *The set $\mathcal{U}$ of umbral operators on $\mathcal{P}$ is a group under composition, with*

$$\lambda_g \circ \lambda_f = \lambda_{f\circ g} \quad \textit{and} \quad \lambda_f^{-1} = \lambda_{\overline{f}}$$

2) *The set of associated sequences forms a group under umbral composition*

$$p_n(\boldsymbol{q}(x)) = \sum_{k=0}^{n} p_{n,k}q_k(x)$$

In particular, the umbral composition $p_n(\boldsymbol{q}(x))$ is the associated sequence for the composition $f \circ g$, that is

$$\lambda_{f\circ g}\colon x^n \to p_n(\boldsymbol{q}(x))$$

The identity is the sequence x^n and the inverse of $p_n(x)$ is the associated sequence for the compositional inverse $\overline{f}(t)$.

3) Let $\lambda_f \in \mathcal{U}$ and $g(t) \in \mathcal{F}$. Then as operators

$$\lambda_f g(t) = \lambda_f^{-1} g(t) \lambda_f$$

4) Let $\lambda_f \in \mathcal{U}$ and $g(t) \in \mathcal{F}$. Then

$$\lambda_f g(\overline{f}(t)) = g(t)\lambda_f$$

Proof. We prove 3) as follows. For any $h(t) \in \mathcal{F}$ and $p(x) \in P$

$$\begin{aligned} \langle h(t) \mid \lambda^{\times} g(t) p(x)\rangle &= \langle [\lambda^{\times} g(t)] h(t) \mid p(x)\rangle \\ &= \langle \lambda^{\times}[g(t)(\lambda^{-1})^{\times} h(t)] \mid p(x)\rangle \\ &= \langle g(t)(\lambda^{-1})^{\times} h(t) \mid \lambda p(x)\rangle \\ &= \langle (\lambda^{-1})^{\times} h(t) \mid g(t)\lambda p(x)\rangle \\ &= \langle h(t) \mid \lambda^{-1} g(t) \lambda p(x)\rangle \end{aligned}$$

which gives the desired result. Part 4) follows immediately from part 3) since λ_f is composition by $\overline{f}$. □

Sheffer Operators

If $s_n(x)$ is Sheffer for (g, f) then the linear operator $\lambda_{g,f}$ defined by

$$\lambda_{g,f}(x^n) = s_n(x)$$

is called a **Sheffer operator**. Sheffer operators are closely related to umbral operators, since if $p_n(x)$ is associated with $f(t)$ then

$$s_n(x) = g^{-1}(t)p_n(x) = g^{-1}(t)\lambda_f x^n$$

and so

$$\lambda_{g,f} = g^{-1}(t)\lambda_f$$

It follows that the Sheffer operators form a group with composition

$$\begin{aligned} \lambda_{g,f} \circ \lambda_{h,k} &= g^{-1}(t)\lambda_f h^{-1}(t)\lambda_k \\ &= g^{-1}(t)h^{-1}(f(t))\lambda_f\lambda_k \\ &= [g(t)h(f(t))]^{-1}\lambda_{k\circ f} \\ &= \lambda_{g\cdot(h\circ f),k\circ f} \end{aligned}$$

and inverse

$$\lambda_{g,f}^{-1} = \lambda_{g^{-1}(\overline{f}),\overline{f}}$$

From this, we deduce that the umbral composition of Sheffer sequences is a Sheffer sequence. In particular, if $s_n(x)$ is Sheffer for (g, f) and $t_n(x) = \Sigma t_{n,k} x^k$ is Sheffer for (h, k) then

$$\begin{aligned}\lambda_{g,f} \circ \lambda_{h,k}(x^n) &= \sum_{k=0}^{n} t_{n,k} \lambda_{g,f} x^k \\ &= \sum_{k=0}^{n} t_{n,k} s_k(x) \\ &= t_n(\boldsymbol{s}(x))\end{aligned}$$

is Sheffer for $(g \cdot (h \circ f), k \circ f)$.

Umbral Shifts and Derivations of the Umbral Algebra

We have seen that an operator on $\mathcal{P}$ is an umbral operator if and only if its adjoint is an automorphism of $\mathcal{F}$. Now suppose that $\theta_f \in \mathcal{L}(\mathcal{P})$ is the umbral shift for the associated sequence $p_n(x)$, associated with the delta series $f(t) \in \mathcal{F}$. Then

$$\begin{aligned}\langle \theta_f^{\times} f(t)^k \mid p_n(x) \rangle &= \langle f(t)^k \mid \theta_f p_n(x) \rangle \\ &= \langle f(t)^k \mid p_{n+1}(x) \rangle \\ &= (n+1)! \delta_{n+1,k} \\ &= k(k-1)! \delta_{n,k-1} \\ &= \langle k f(t)^{k-1} \mid p_n(x) \rangle\end{aligned}$$

and so

$$\theta_f^{\times} f(t)^k = k f(t)^{k-1} \tag{18.11}$$

This implies that

$$\theta_f^{\times} [f(t)^k f(t)^j] = \theta_f^{\times} [f(t)^k] f(t)^j + f(t)^k \theta_f^{\times} [f(t)^j] \tag{18.12}$$

and further, by continuity, that

$$\theta_f^{\times} [g(t)h(t)] = [\theta_f^{\times} g(t)] h(t) + g(t) [\theta_f^{\times} g(t)] \tag{18.13}$$

Let us pause for a definition.

Definition *Let $\mathcal{A}$ be an algebra. A linear operator ∂ on $\mathcal{A}$ is a* **derivation** *if*

$$\partial(ab) = (\partial a)b + a\partial b$$

for all $a, b \in \mathcal{A}$. □

Thus, we have shown that the adjoint of an umbral shift is a derivation of the umbral algebra $\mathcal{F}$. Moreover, the expansion theorem and (18.11) show that $\theta_f^{\times}$ is surjective. This characterizes umbral shifts. First we need a preliminary result on surjective derivations.

Theorem 18.21 *Let ∂ be a surjective derivation on the umbral algebra $\mathcal{F}$. Then $\partial c = 0$ for any* constant $c \in \mathcal{F}$ *and $o(\partial f(t)) = o(f(t)) - 1$, if $o(f(t)) \geq 1$. In particular, ∂ is continuous.*

Proof. We begin by noting that

$$\partial 1 = \partial 1^2 = \partial 1 + \partial 1 = 2\partial 1$$

and so $\partial c = c\partial 1 = 0$ for all constants $c \in \mathcal{F}$. Since ∂ is surjective, there must exist an $h(t) \in \mathcal{F}$ for which

$$\partial h(t) = 1$$

Writing $h(t) = h_0 + th_1(t)$, we have

$$1 = \partial[h_0 + th_1(t)] = (\partial t)h_1(t) + t\partial h_1(t)$$

which implies that $o(\partial t) = 0$. Finally, if $o(h(t)) = k \geq 1$ then $h(t) = t^k h_1(t)$, where $o(h_1(t)) = 0$ and so

$$o[\partial h(t)] = o[\partial t^k h_1(t)] = o[t^k \partial h(t) + kt^{k-1} h_1(t) \partial t] = k - 1 \qquad \square$$

Theorem 18.22 *A linear operator θ on $\mathcal{P}$ is an umbral shift if and only if its adjoint is a surjective derivation of the umbral algebra $\mathcal{F}$. Moreover, if θ_f is an umbral shift then $\theta_f^\times = \partial_f$ is derivation with respect to $f(t)$, that is,*

$$\theta_f^\times f(t)^k = kf(t)^{k-1}$$

for all $k \geq 0$. In particular, $\theta_f^\times f(t) = 1$.

Proof. We have already seen that $\theta_f^\times$ is derivation with respect to $f(t)$. For the converse, suppose that $\theta^\times$ is a surjective derivation. Theorem 18.21 implies that there is a delta functional $f(t)$ such that $\theta^\times f(t) = 1$. If $p_n(x)$ is the associated sequence for $f(t)$ then

$$\begin{aligned}
\langle f(t)^k \mid \theta p_n(x)\rangle &= \langle \theta^\times f(t)^k \mid p_n(x)\rangle \\
&= \langle kf(t)^{k-1}\theta^\times f(t) \mid p_n(x)\rangle \\
&= \langle kf(t)^{k-1} \mid p_n(x)\rangle \\
&= (n+1)!\delta_{n+1,k} \\
&= \langle f(t)^k \mid p_{n+1}(x)\rangle
\end{aligned}$$

Hence, $\theta p_n(x) = p_{n+1}(x)$, that is, $\theta = \theta_f$ is the umbral shift for $p_n(x)$. $\square$

We have seen that the fact that the set of all automorphisms on $\mathcal{F}$ is a group under composition shows that the set of all associated sequences is a group under umbral composition. The set of all surjective derivations on $\mathcal{F}$ does not form a group. However, we do have the chain rule for derivations!

Theorem 18.23 *(***The chain rule***) Let ∂_f and ∂_g be surjective derivations on $\mathcal{F}$. Then*

$$\partial_g = (\partial_g f(t))\partial_f$$

Proof. This follows from

$$\partial_g f(t)^k = kf(t)^{k-1}\partial_g f(t) = (\partial_g f(t))\partial_f f(t)^k$$

and so continuity implies the result. □

The chain rule leads to the following umbral result.

Theorem 18.24 *If θ_f and θ_g are umbral shifts then*

$$\theta_g = \theta_f \circ \partial_g f(t)$$

Proof. Taking adjoints in the chain rule gives

$$\theta_g = \theta_f \circ (\partial_g f(t))^\times = \theta_f \circ \partial_g f(t) \qquad \square$$

We leave it as an exercise to show that $\partial_g f(t) = [\partial_f g(t)]^{-1}$. Now, by taking $g(t) = t$ in Theorem 18.24 and observing that $\theta_t x^n = x^{n+1}$ and so θ_t is multiplication by x, we get

$$\theta_f = x\partial_f t = x[\partial_t f(t)]^{-1} = x[f'(t)]^{-1}$$

Applying this to the associated sequence $p_n(x)$ for $f(t)$ gives the following important recurrence relation for $p_n(x)$.

Theorem 18.25 *(***The recurrence formula***) Let $p_n(x)$ be the associated sequence for $f(t)$. Then*

1) $p_{n+1}(x) = x[f'(t)]^{-1}p_n(x)$

2) $p_{n+1}(x) = x\lambda_f[\overline{f}(t)]'x^n$

Proof. The first part is proved. As to the second, using Theorem 18.20 we have

$$\begin{aligned} p_{n+1}(x) &= x[f'(t)]^{-1}p_n(x) \\ &= x[f'(t)]^{-1}\lambda_f x^n \\ &= x\lambda_f[f'(\overline{f}(t))]^{-1}x^n \\ &= x\lambda_f[\overline{f}(t)]'x^n \end{aligned} \qquad \square$$

Example 18.9 The recurrence relation can be used to find the associated sequence for the forward difference functional $f(t) = e^t - 1$. Since $f'(t) = e^t$, the recurrence relation is

$$p_{n+1}(x) = xe^{-t}p_n(x) = xp_n(x-1)$$

Using the fact that $p_0(x) = 1$, we have

$$p_1(x) = x,\ p_2(x) = x(x-1),\ p_3(x) = x(x-1)(x-2)$$

and so on, leading easily to the lower factorial polynomials

$$p_n(x) = x(x-1)\cdots(x-n+1) = (x)_n \qquad \square$$

Example 18.10 Consider the delta functional

$$f(t) = \log(1+t)$$

Since $\overline{f}(t) = e^t - 1$ is the forward difference functional, Theorem 18.20 implies that the associated sequence $\phi_n(x)$ for $f(t)$ is the inverse, under umbral composition, of the lower factorial polynomials. Thus, if we write

$$\phi_n(x) = \sum_{k=0}^{n} S(n,k)x^k$$

then

$$x^n = \sum_{k=0}^{n} S(n,k)(x)_k$$

The coefficients $S(n,k)$ in this equation are known as the **Stirling numbers of the second kind** and have great combinatorial significance. In fact, $S(n,k)$ is the number of partitions of a set of size n into k blocks. The polynomials $\phi_n(x)$ are called the **exponential polynomials**.

The recurrence relation for the exponential polynomials is

$$\phi_{n+1}(x) = x(1+t)\phi_n(x) = x(\phi_n(x) + \phi_n'(x))$$

Equating coefficients of x^k on both sides of this gives the well known formula for the Stirling numbers

$$S(n+1,k) = S(n,k-1) + kS(n,k)$$

Many other properties of the Stirling numbers can be derived by umbral means. □

Now we have the analog of part 3) of Theorem 18.20.

Theorem 18.26 *Let θ_f be an umbral shift. Then*

$$\theta_f^{\times} g(t) = g(t)\theta_f - \theta_f g(t)$$

Proof. We have

$$\begin{aligned}\langle f^k(t) \mid \theta_f^\times g(t)p_n(x)\rangle &= \langle [\theta_f^\times g(t)]f^k(t) \mid p_n(x)\rangle \\ &= \langle \theta_f^\times [g(t)f^k(t)] - g(t)\theta_f^\times f^k(t) \mid p_n(x)\rangle \\ &= \langle \theta_f^\times [g(t)f^k(t)] \mid p_n(x)\rangle - \langle kg(t)f^{k-1}(t) \mid p_n(x)\rangle \\ &= \langle f^k(t) \mid g(t)\theta_f p_n(x)\rangle - \langle kf^{k-1}(t) \mid g(t)p_n(x)\rangle \\ &= \langle f^k(t) \mid g(t)\theta_f p_n(x)\rangle - \langle \theta_f^\times f^k(t) \mid g(t)p_n(x)\rangle \\ &= \langle f^k(t) \mid g(t)\theta_f p_n(x)\rangle - \langle f^k(t) \mid \theta_f g(t)p_n(x)\rangle\end{aligned}$$

from which the result follows. □

If $f(t) = t$ then $\theta_f$is multiplication by x and $\theta_f^\times$ is the derivative with respect to t and so the previous result becomes

$$g'(t) = g(t)x - xg(t)$$

as operators on $\mathcal{P}$. The right side of this is called the **Pincherle derivative** of the operator $g(t)$. (See [Pin].)

Sheffer Shifts

Recall that the linear map

$$\theta_{g,f}[s_n(x)] = s_{n+1}(x)$$

where $s_n(x)$ is Sheffer for $(g(t), f(t))$ is called a Sheffer shift. If $p_n(x)$ is associated with $f(t)$ then $g(t)s_n(x) = p_n(x)$ and so

$$g^{-1}(t)p_{n+1}(x) = \theta_{g,f}[g^{-1}(t)p_n(x)]$$

and so

$$\theta_{g,f} = g^{-1}(t)\theta_f g(t)$$

From Theorem 18.26, the recurrence formula and the chain rule, we have

$$\begin{aligned}\theta_{g,f} &= g^{-1}(t)\theta_f g(t) \\ &= g^{-1}(t)[g(t)\theta_f - \theta_f^\times g(t)] \\ &= \theta_f - g^{-1}(t)\partial_f g(t) \\ &= \theta_f - g^{-1}(t)\partial_f g(t) \\ &= \theta_f - g^{-1}(t)\partial_f t\partial_t g(t) \\ &= x[f'(t)]^{-1} - g^{-1}(t)[f'(t)]^{-1}g'(t) \\ &= \left[x - \frac{g'(t)}{g(t)}\right]\frac{1}{f'(t)}\end{aligned}$$

We have proved the following.

Theorem 18.27 *Let $\theta_{g,f}$ be a Sheffer shift. Then*

1) $\theta_{g,f} = \left[x - \frac{g'(t)}{g(t)}\right]\frac{1}{f'(t)}$

2) $s_{n+1}(x) = \left[x - \frac{g'(t)}{g(t)}\right]\frac{1}{f'(t)}s_n(x)$ □

The Transfer Formulas

We conclude with a pair of formulas for the computation of associated sequences.

Theorem 18.28 (The transfer formulas) *Let $p_n(x)$ be the associated sequence for $f(t)$. Then*

1) $p_n(x) = f'(t)\left(\frac{f(t)}{t}\right)^{-n-1}x^n$

2) $p_n(x) = x\left(\frac{f(t)}{t}\right)^{-n}x^{n-1}$

Proof. First we show that 1) and 2) are equivalent. Write $g(t) = f(t)/t$. Then

$$\begin{aligned} f'(t)g(t)^{-n-1}x^n &= [tg(t)]'g(t)^{-n-1}x^n \\ &= g(t)^{-n}x^n + tg'(t)g(t)^{-n-1}x^n \\ &= g(t)^{-n}x^n + ng'(t)g(t)^{-n-1}x^{n-1} \\ &= g(t)^{-n}x^n + [g(t)^{-n}]'x^{n-1} \\ &= g(t)^{-n}x^n - [g(t)^{-n}x - xg(t)^{-n}]x^{n-1} \\ &= xg(t)^{-n}x^{n-1} \end{aligned}$$

To prove 1), we verify the operation conditions for an associated sequence for the sequence $q_n(x) = f'(t)g(t)^{-n-1}x^n$. First, when $n \geq 1$ the fourth equality above gives

$$\begin{aligned} \langle t^0 \mid q_n(x)\rangle &= \langle t^0 \mid f'(t)g(t)^{-n-1}x^n\rangle \\ &= \langle t^0 \mid g(t)^{-n}x^n - [g(t)^{-n}]'x^{n-1}\rangle \\ &= \langle g(t)^{-n} \mid x^n\rangle - \langle [g(t)^{-n}]' \mid x^{n-1}\rangle \\ &= \langle g(t)^{-n} \mid x^n\rangle - \langle g(t)^{-n} \mid x^n\rangle \\ &= 0 \end{aligned}$$

If $n = 0$ then $\langle t^0 \mid q_n(x)\rangle = 1$ and so, in general, we have $\langle t^0 \mid q_n(x)\rangle = \delta_{n,0}$ as required.

For the second required condition,

$$\begin{aligned} f(t)q_n(x) &= f(t)f'(t)g(t)^{-n-1}x^n \\ &= tg(t)f'(t)g(t)^{-n-1}x^n \\ &= nf'(t)g(t)^{-n-1}x^{n-1} \\ &= nq_{n-1}(x) \end{aligned}$$

Thus, $q_n(x)$ is the associated sequence for $f(t)$. □

A Final Remark

Unfortunately, space does not permit a detailed discussion of examples of Sheffer sequences nor the application of the umbral calculus to various classical problems. In [Rom1], one can find a discussion of the following polynomial sequences:

The lower factorial polynomials and Stirling numbers
The exponential polynomials and Dobinski's formula
The Gould polynomials
The central factorial polynomials
The Abel polynomials
The Mittag–Leffler polynomials
The Bessel polynomials
The Bell polynomials
The Hermite polynomials
The Bernoulli polynomials and the Euler–Maclaurin expansion
The Euler polynomials
The Laguerre polynomials
The Bernoulli polynomials of the second kind
The Poisson–Charlier polynomials
The actuarial polynomials
The Meixner polynomials of the first and second kinds
The Pidduck polynomials
The Narumi polynomials
The Boole polynomials
The Peters polynomials
The squared Hermite polynomials
The Stirling polynomials
The Mahler polynomials
The Mott polynomials

and more. In [Rom1], we also find a discussion of how the umbral calculus can be used to approach the following types of problems:

The connection constants problem
Duplication formulas
The Lagrange inversion formula
Cross sequences
Steffensen sequences
Operational formulas
Inverse relations
Sheffer sequence solutions to recurrence relations
Binomial convolution

Finally, it is possible to generalize the classical umbral calculus that we have described in this chapter to provide a context for studying polynomial sequences such as those of the name Gegenbauer, Chebyshev and Jacobi. Also, there is a q-version of the umbral calculus that involves the **q-binomial coefficients** (also known as the **Gaussian coefficients**)

$$\binom{n}{k}_q = \frac{(1-q)\cdots(1-q^n)}{(1-q)\cdots(1-q^k)(1-q)\cdots(1-q^{n-k})}$$

in place of the binomial coefficients. There is also a logarithmic version of the umbral calculus, which studies the *harmonic logarithms* and sequences of *logarithmic type*. For more on these topics, please see [LR], [Rom2] and [Rom3].

Exercises

1. Prove that $o(fg) = o(f) + o(g)$, for any $f, g \in \mathcal{F}$.
2. Prove that $o(f+g) \geq \min\{o(f), o(g)\}$, for any $f, g \in \mathcal{F}$.
3. Show that any delta series has a compositional inverse.
4. Show that for any delta series f, the sequence f^k is a pseudobasis.
5. Prove that ∂_t is a derivation.
6. Show that $f \in \mathcal{F}$ is a delta functional if and only if $\langle f \mid 1 \rangle = 0$ and $\langle f \mid x \rangle \neq 0$.
7. Show that $f \in \mathcal{F}$ is invertible if and only if $\langle f \mid 1 \rangle \neq 0$.
8. Show that $\langle f(at) \mid p(x) \rangle = \langle f(t) \mid p(ax) \rangle$ for any $\mathrm{a} \in \mathbb{C}$, $f \in \mathcal{F}$ and $p \in \mathcal{P}$.
9. Show that $\langle te^{at} \mid p(x) \rangle = p'(\mathrm{a})$ for any polynomial $p(x) \in \mathcal{P}$.
10. Show that $f = g$ in $\mathcal{F}$ if and only if $f = g$ as linear functionals, which holds if and only if $f = g$ as linear operators.
11. Prove that if $s_n(x)$ is Sheffer for $(g(t), f(t))$ then $f(t)s_n(x) = ns_{n-1}(x)$. *Hint*: Apply the functionals $g(t)f^k(t)$ to both sides.
12. Verify that the Abel polynomials form the associated sequence for the Abel functional.
13. Show that a sequence $s_n(x)$ is the Appell sequence for $g(t)$ if and only if $s_n(x) = g(t)^{-1}x^n$.
14. If f is a delta series, show that the adjoint $\lambda_f^\times$ of the umbral operator λ_f is a vector space isomorphism of $\mathcal{F}$.
15. Prove that if T is an automorphism of the umbral algebra then T preserves order, that is, $o(Tf(t)) = o(f(t))$. In particular, T is continuous.
16. Show that an umbral operator maps associated sequences to associated sequences.
17. Let $p_n(x)$ and $q_n(x)$ be associated sequences. Define a linear operator α by $\alpha\colon p_n(x) \to q_n(x)$. Show that α is an umbral operator.
18. Prove that if ∂_f and ∂_g are surjective derivations on $\mathcal{F}$ then $\partial_g f(t) = [\partial_f g(t)]^{-1}$.

References

General References

[HJ1] Horn, R. and Johnson, C., *Matrix Analysis*, Cambridge University Press, 1985.

[HJ2] Horn, R. and Johnson, C., *Topics in Matrix Analysis*, Cambridge University Press, 1991.

[J1] Jacobson, N., *Basic Algebra I*, Second Edition, W.H. Freeman, 1985.

[KM] Kostrikin, A. and Manin, Y., *Linear Algebra and Geometry*, Gordon and Breach Science Publishers, 1997.

[MM1] Marcus, M., *Finite Dimensional Multilinear Algebra, Part I*, Marcel Dekker, 1971.

[MM2] Marcus, M., *Finite Dimensional Multilinear Algebra, Part II*, Marcel Dekker, 1975.

[Sh] Shapiro, H., A survey of canonical forms and invariants for unitary similarity, *Linear Algebra and Its Applications* 147:101-167 (1991).

[ST] Snapper, E. and Troyer, R., *Metric Affine Geometry*, Dover Publications, 1971.

References on the Umbral Calculus

[LR] Loeb, D. and Rota, G.-C., Formal Power Series of Logarithmic Type, *Advances in Mathematics*, Vol. 75, No. 1, (May 1989) 1–118.

[Pin] Pincherle, S. "Operatori lineari e coefficienti di fattoriali." *Alti Accad. Naz. Lincei, Rend. Cl. Fis. Mat. Nat*. (6) 18, 417–519, 1933.

[Rom1] Roman, S., *The Umbral Calculus*, Pure and Applied Mathematics vol. 111, Academic Press, 1984.

[Rom2] Roman, S., The logarithmic binomial formula, *American Mathematical Monthly* 99 (1992) 641–648.

[Rom3] Roman, S., The harmonic logarithms and the binomial formula, *Journal of Combinatorial Theory*, series A, 63 (1992) 143–163.

Index